A Taxonomy of Decision Support Systems

Steven Alter, University of Southern California

Computer systems for decision making and decision implementation vary widely in terms of *what* they do and *how* they do it. The decision support system is not a homogeneous category. Through a series of case examples, the author develops a taxonomy of seven distinct types of decision support systems. The end result is a guideline for designing and implementing systems, and a framework for further communication and research. *Ed.*

Introduction

As evidenced by the title of a recent well-attended conference,[1] *decision support system* (DSS) is a buzzword whose time has arrived. Now that most corporations have survived the growing pains of learning to develop and use data processing systems, many of the most innovative new systems clearly fall under the general heading of decision support. Unfortunately, there is relatively little organized knowledge about DSSs. Although a certain amount of conjecture has been generated concerning the *nature* of decision support systems and the significance of various system characteristics, even the conjectures are often contradictory.

This article discusses a taxonomic scheme that was one of the main findings of an exploratory study of decision support systems[2] undertaken in response to this dearth of information. The purpose of the study was to gain a better understanding of the dynamics of these systems and the key issues leading to their success or lack of success. As will be discussed, one of the conclusions in this regard was that these key issues differ across various types of DSSs.

In embarking upon this research a definitional question arose immediately: What are decision support systems? How can they be recognized and distinguished from other systems? Taking the approach of looking first and defining later, the following general distinctions emerged. Business computer applications can be stereotyped into two categories: electronic data processing (EDP) systems and decision support systems (DSS). The main difference between DSS and EDP systems is related to their basic

[1] Conference on Decision Support Systems, San Jose, California, January 24-26, 1977, sponsored by IBM San Jose Research Laboratory, Sloan School of Management, M.I.T., Wharton Business School, and ACM-SIGBDP.

[2] See Alter [2].

purposes. EDP systems are designed to automate or expedite transaction processing, record keeping, and business reporting; DSSs are designed to aid in decision making and decision implementation. While most DSSs are used to facilitate management, planning, or staff activities, EDP systems emphasize intrinsically clerical activities. Whereas the general orientation of EDP systems is toward mechanical efficiency, that of DSSs is more toward the overall effectiveness of individuals or organizations. The manner of usage is also quite different. Unlike the EDP user, who typically receives reports on a periodic basis, the DSS user often initiates each instance of system use, either directly or through a staff intermediary.

Although the DSS vs. EDP dichotomy is weakened by overlaps due to the multiple purposes and orientations of many systems, implementers and users who participated in the study had no real difficulty in identifying DSSs used in their organizations. Starting with detailed case studies of eight systems[3] a sample of fifty-six DSSs was eventually compiled. The data used in the analysis consisted of mini-case studies of each of the systems. Each mini-case was a structured story of the system in terms of interview responses to questions under the following headings:

- General background,
- System history and characteristics,
- Types of use and impact,
- Limitations and types of disuse or abuse,
- Factors in favor of or opposed to getting started, and
- Factors in favor of or opposed to successful implementation.

As the sample grew, it became increasingly clear that the *decision support system* is not a homogeneous category. Quite to the contrary, many of the systems in the sample differed vastly in what they did and how they did it. This led me to wonder why people who talk about DSSs often seemed to talk about DSSs *in general*. It appeared that this was much like talking about pets in general, without distinguishing between dogs and cats and piranha fish and turtles. I concluded that one of the main products of the research should be a taxonomy of decision support systems which differentiated the sample in a useful and understandable manner.

A first step in attempting to develop such a taxonomy was to examine the usefulness of commonly used system-labeling schemes such as:

- *Functional Area*: marketing, production, finance;
- *Decision Perspective*: operational control, management control, strategic planning;[4]
- *Problem Type*: structured vs. unstructured;[5]

[3] See Alter [1].

[4] See Gorry and Scott Morton [3].

[5] See Simon [8], Gorry and Scott Morton [3], and Mason and Mitroff [5].

- *Computer Technology*: interactive vs. batch;
- *Modeling Approach*: simulation vs. optimization.

Unfortunately, few significant conclusions seemed to emerge when the systems in the sample were grouped in terms of these schemes. For instance, financial projection systems for operational planning seem very similar in concept and structure to several systems for strategic planning. Likewise, the significance of interactive computation seemed to diminish greatly when decision makers were not hands-on users of systems. Difficulties in deciding whether one repetitive business problem was more versus less structured than another also reduced the usefulness of that distinction.

A Taxonomy of Decision Support Systems

The taxonomy that seemed most useful in categorizing the systems in the sample was based on what can be called the *degree of action implication of system outputs* (i.e., the degree to which the system's outputs could directly determine the decision). This is related to a spectrum of generic operations which can be performed by decision support systems. These generic operations extend along a single dimension ranging from extremely data oriented to extremely model oriented:

- Retrieving a single item of information,
- Providing a mechanism for ad hoc data analysis,
- Providing prespecified aggregations of data in the form of reports,
- Estimating the consequences of proposed decisions,
- Proposing decisions, and
- Making decisions.

The idea here is that a decision support system can be categorized in terms of the generic operations it performs, independent of the type of problem, functional area, decision perspective, etc.

Clustered from this viewpoint, the fifty-six systems in the sample fell into seven reasonably distinct types which can be labeled as follows:[6]

A. *File drawer systems* allow immediate access to data items.

B. *Data analysis systems* allow the manipulation of data by means of operators tailored to the task and setting or operators of a general nature.

C. *Analysis information systems* provide access to a series of data bases and small models.

D. *Accounting models* calculate the consequences of planned actions based on accounting definitions.

[6] Mason [4] describes a parallel but more abstract taxonomy suggested by Churchman.

E. *Representational models* estimate the consequences of actions based on models which are partially non-definitional.

F. *Optimization models* provide guidelines for action by generating the optimal solutions consistent with a series of constraints.

G. *Suggestion models* perform mechanical work leading to a specific suggested decision for a fairly structured task.

Figure 1 illustrates that this taxonomy can be collapsed into a simple dichotomy between data-oriented and model-oriented systems. Such a simplification loses a great deal of information, however, by grouping systems which differ in many significant ways.

Each of the seven types of DSS will be discussed briefly with references to specific examples. The last section of the article will summarize some of the differences in key issues across the various types.

A. File Drawer Systems

File drawer systems are basically mechanized versions of manual filing systems. The purpose of file drawer systems is to provide on-line access to particular data items (e.g., status information concerning entities ranging from overdue invoices and available seats on future airplane flights through inventory items, stock portfolios, lots flowing through a shop, etc.).

System A1 is a CRT-based inventory control system used in manufacturing complicated, high technology hardware on a one-of-a-kind basis.

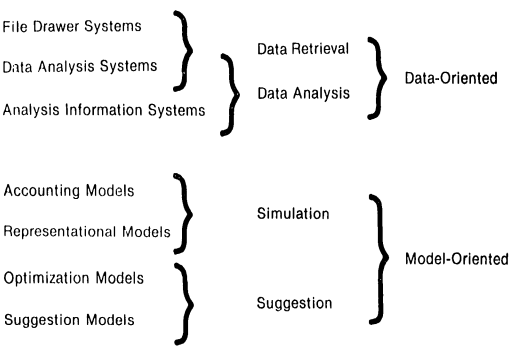


Figure 1 Data-Oriented vs. Model-Oriented Decision Support System Types

Since a single missing part can halt the progress of this complex assembly process, it is very important that the location and status of all available parts are known at all times. In addition to day-to-day use in finding and reallocating parts, the system is used by plant management in a weekly meeting. At this meeting the current needs of various projects are examined, and existing inventory is reallocated or transferred from project to project to expedite a smooth work flow.

System A2 is a CRT-based shop floor information system which tracks the flow of production lots of integrated circuits through a manufacturing process which involves over fifty steps and suffers from a serious yield problem. The input to the system consists of daily work reports submitted by operators. The system stores this information and maintains a history of each lot by step. Included are the yield, the release date, identification of the person who did the work, and so on. In addition, various aggregations of productivity data by operators and by lot are provided. The data is accessed by means of thirteen standard retrieval commands. The system is used by section chiefs to monitor work flow and to detect yield problems and production bottlenecks.

Typically, the hands-on users of such systems are nonmanagerial personnel ranging from clerks to foremen who use the system to support their day-to-day operational tasks. The concept is a very simple one: people performing ongoing operational tasks should have immediate access to the information they need, and should be able to obtain the most current version of that information. In some cases, it is proprietary commercial information to which access is sold (e.g., commodity and service trading systems which provide availability information concerning ships for charter, lumber in stock, apartments for rent, etc.).

B. Data Analysis Systems

Data analysis systems are generally used by nonmanagerial line or staff personnel in analyzing files of current or historical data.

System B1 is a CRT-based analysis system in a bank. It is used by division comptrollers to generate customized monthly variance reports which are used in budget control sessions with responsible cost center managers. By allowing these comptrollers to screen, analyze, and annotate budget variances before they are sent to cost center managers, the system expedites the budget control process and facilitates communication between the involved parties. This system also contains planning aids which help division planners generate their one-year and five-year budgets by doing simple projections, comparisons, and extensions of line items.

System B2 is a generalized financial analysis system which is basically an interpretive language for use by financial analysts. It is used in an oil company to analyze investment opportunities and to consolidate plans at the corporate level. The purpose of the system was to improve upon a disorderly series of programs which did particular calculations and consoli-

dations. It replaced all these programs with a unified system which could handle most financial analysis needs and which would produce reasonably consistent reports.

The data analysis systems in the sample fell into two categories: *tailored analysis systems* and *generalized analysis systems*. Tailored analysis systems are designed specifically to meet particular analysis requirements related to a definite job or task. The data in these systems is often historical, although current status information may be included. These systems allow analysts to manipulate the data and to produce analysis reports on an ad hoc basis. Generalized analysis systems are specialized programming languages whose purpose is to allow users to perform fairly general kinds of analysis of data bases and to program simple models. Such systems are viewed as off-the-shelf tools for use in many settings. Given a data base in an appropriate format, some of these systems provide the user with the capability to analyze the data by means of operations such as data retrieval, pictorial representation, summarization of the data, and calculations. Others are oriented more toward facilitating the creation of simple models. Unlike tailored analysis systems which address the special analysis needs of particular tasks, generalized analysis systems are designed to be readily transferable and relatively context free. The border between file drawer systems and tailored analysis systems is fuzzy. Although there exist systems whose sole purpose is the retrieval of data items and other systems whose sole purpose is the analysis of files of information, systems also exist which attempt to serve both functions.

C. Analysis Information Systems

Throughout the first twenty years of computer-based *management information systems*, one of the most common complaints was that these systems simply were not flexible enough to satisfy the changing information needs of managers. Typically, such management information systems were basically transaction processing and record keeping systems. Although these systems could be used conveniently to generate standard periodic reports, their requirements for efficiency precluded the generation of management information relevant to decisions or situations whose essential components varied over time. The purpose of *analysis information systems* is to provide management information through the use of a series of decision-oriented data bases and small models.

System C1 is a growing marketing information system in a consumer products company. Its data bases include internal sales, advertising, promotion, and pricing data, plus a number of proprietary marketing data bases which are purchased from marketing research firms. The system is used by staff personnel for many types of ad hoc reporting. In addition, it is used by a research group in developing methodologies for forecasting sales and analyzing the effectiveness of competitive actions. The data is accessed through a report generator and a statistical package.

System C2 is a sales analysis system developed at an industrial equipment company. In addition to detailed sales data, it contains internally generated and purchased information about customers and potential customers, plus forecasts from industry sector economic models. It is used for product planning through the development of growth forecasts by industry sector, and of corresponding forecasts for product growth within industry sectors. Viewed by its originators as a tool kit of data and small models for the purpose of supporting planning, an attempt is made to limit the system's use for day-to-day reporting and analysis.

The basic idea underlying these analysis information systems is to recognize the incongruities between transaction processing systems and decision-oriented information systems, and to proceed accordingly. Analysis information systems are designed to extract relevant data from EDP systems and to augment this data with external data. By maintaining this type of analysis data base, it is possible to access that data freely and without being constrained by the operational requirements of scheduling and running a large-scale corporate data center efficiently. In some cases, such systems are basically vehicles by which a staff man or staff group tries to have an impact on the ways in which decisions are made. The *modus operandi* is highly incremental: start with an existing data base and set of models, identify a new business problem, develop a solution that extends the system, and use the credit gained to expand the scope of future efforts.

D. Accounting Models

Accounting models use definitional relationships and formulas to calculate the consequences of particular actions.

System D1 is a voyage profitability estimator used by a shipping company via time sharing. This program performs a standard profit calculation which is used to decide what charter rate should be charged for a particular voyage. The formula that is used involves ship and voyage characteristics including tonnage, speed, rate of fuel consumption, port costs, and so on. Because much of this data is stored in advance, what formerly required fifteen to twenty minutes of calculations now requires three to five minutes of specification using the terminal. In addition to merely saving time, this makes it possible for charter clerks to explore tradeoffs between speed and fuel consumption.

System D2 is an on-line source and application of funds budget used for operational decision making and financial planning over a two year horizon in an insurance company. The inputs are cash flow projections from various lines of insurance and investment areas. The output is an overall cash flow by month. The system output is used at weekly meetings of an investment committee to help in allocating funds across investment areas and to minimize the amount of cash that is left idle in banks.

The accounting models in the sample were used to facilitate planning by generating estimates of income statements, balance sheets, or other

outcome measures. The inputs to these systems were estimates by business unit (product, department, etc.) of various elements of costs and revenues. Using accounting definitions and estimated line items rather than actuals, these systems performed the kinds of extensions and additions that are performed by a clerk or a computer in producing a business statement. Such systems contained little or no sense of any mechanism whereby the firm's actions are related to outcomes in the market. For instance, it was typical to use sales as a fixed input rather than as a function of price or other competitive actions. On the other hand, one of the key attributes of these models was their understandability by managers.

E. Representational Models

Representational models include all simulation models which are not primarily accounting definitions (i.e., which use at least partially non-definitional relationships in estimating the consequences of various actions, environmental conditions, or relationships). Whereas an accounting model might start with product sales and prices that were determined external to the system, a representational model might start with only price and then calculate sales based on a model representing the causal mechanism by which price determines sales. On the boundary between accounting definitions and representational models are systems such as E1 — some of whose statements are definitions, while others are cost accounting approximations to the relationships between variables.

System E1 is used by a large chemical company to simulate and cost out flows of materials among a hierarchy of mining sites, production facilities, inventory depots, and sales locations. The purpose of the model is to provide a quick and reliable method for evaluating a variety of yearly budgeting alternatives at a variety of planning levels. Based on individual models of each point in the flow pattern, the system calculates volumetric outputs and costs which are then aggregated upward. This helps rationalize the development of yearly budgets by allowing management to examine the profit impact of alternatives involving changes in volumes, prices, distribution patterns, production costs, transportation modes, inventory levels, raw materials sources, etc.

System E2 is an aggregate market response model which relates levels of advertising, promotions, and pricing to levels of sales for a particular brand. The model is used by a consumer product company to track the marketplace and the effects of competitive actions. The model was developed in a team setting by reconciling an analysis of historical information with individuals' subjective opinions concerning response parameters.

It is possible to classify simulation models in terms of the uncertainty inherent in the relationships in the models themselves. Thus, simulation models can be viewed along the following dimension: (1) accounting definitions, (2) models in which the form of the relationship is accurate while parameter values may be inaccurate, and (3) models in which the form of

the relationship may not be a good representation of the underlying process. Clear-cut accounting models are on one end of the spectrum, and representational models are on the other end. Many models fall between the two extremes.

The location of a model along the above continuum has many implications for its potential usefulness and acceptance. Accounting models are typically viewed as specialized adding machines that perform calculations a person would otherwise perform manually. Much of the effort in building such a model involves the clarification of accounting definitions and relationships that are internal to the company. On the other hand, representational models are frequently viewed as attempts to develop an understanding of the possible relationship between future actions and future outcomes. Much of the effort in building these models involves the creation of approximate relationships that attempt to roughly describe the linkages between actions and outcomes. In using an accounting model, the accuracy of the model itself should not be an issue; rather, the main questions should concern the quality of the estimated values provided as inputs. In using a representational model, one of the main issues is whether or not the model is a reasonable representation of the situation being studied. At the same time, an important part of the benefit of such a model comes from the increased understanding that is gained by trying to develop explicit relationships describing how part of the business environment works. Related to accuracy, but only partially, are the credibility and acceptance of a model. In many cases, representational models tend to have credibility problems. Because they are approximations, it is often possible to question important relationships and to wonder whether these relationships produce misleading results. At the same time, however, representational models that pass the test of credibility can be a very valuable source of understanding concerning the interaction between internal and external forces in the future.

F. Optimization Models

Optimization models are used in studying situations that can be described mathematically as complicated puzzles whose goals involve combining the pieces in a way that attains a specific objective such as maximizing profit or minimizing cost.

System F1 aids in determining the start dates of three-week, twenty-member training classes in a training school for personnel who exhibit a high attrition rate. The inputs to the model include the company's forecasted service demands, current staffing levels, the acceptable level of shortfall during peak periods, and so on. Constrained by these inputs and some complex rules concerning consecutive start dates and school administration, the system uses a smoothing algorithm to generate a set of start dates with relatively (although not necessarily *optimally*) low cost. The system is used iteratively in developing an understanding of the effect on

the current year's plan of potential modifications in policy inputs such as the maximum shortfall acceptable during peak periods.

System F2 was a linear programming model used by a consumer products company faced with short-run supply problems. For many of the raw materials the company used, both availability and supply had suddenly begun to fluctuate. One way to make the best of a bad situation was to respond to these fluctuations by adjusting product recipes in a way that met production requirements at minimum cost. It took a staff analyst two weeks to set up a small linear programming model that produced an optimal set of product recipes based on somewhat simplified assumptions concerning the flexibility of production facilities. The system was used sporadically over the course of a year as a way of providing guidelines for production adjustments.

The systems classified here as optimization models are used as analysis tools rather than as a way of generating a definitive answer that can be acted upon directly. In other words, this approach for supporting decisions can be used in situations that have enough structure to develop an optimizing model that can be used as part of the analysis. Many applications of optimization techniques such as linear programming are of this type. There are other types of applications, however, in which there exists enough structure that a model can produce a direct suggestion of action. These models will be described in the next section.

G. Suggestion Models

Suggestion models generate suggested actions based on formulas or mathematical procedures which can range from decision rules to optimization methods. The purpose of such systems is to expedite or bypass other procedures for generating the suggestion. In a sense, suggestion systems are even more structured than optimization systems, since their output is pretty much *the answer*, rather than a way of viewing tradeoffs, the importance of constraints, and so on.

System G1 performs some complicated calculations which are needed in adjusting the rates on particular group insurance policies based on the historical relationship between premiums and claims for those policies. The system was developed to eliminate part of the clerical burden associated with renewal underwriting and to help assure that rate calculations are consistent and accurate. Using the system has become part of the job of a large number of underwriters in an insurance company. Instead of calculating renewal rates by hand and in a relatively undisciplined manner, the underwriter fills out coded input sheets for the system, which calculates a renewal rate under a series of standard statistical and actuarial assumptions that may or may not apply to the policy. Upon receiving the output, the underwriter reviews the accompanying documentation and decides whether these calculations correctly represent the situation. If not, the coding sheet is modified in an appropriate manner and resubmitted.

System G2 was used to expedite the assembly of a standard piece of electronic equipment over the course of a one-year production contract. Each unit of equipment contained ten diodes, each of which had a particular resonant frequency. Due to problems in producing the diodes, this measurable frequency varied from one diode to the next. Due to peculiarities of the electronics, 200 among the millions of different combinations of diodes of particular frequencies could be used in any unit of the equipment. The weekly input to the system was the inventory on hand of each type of diode. Using linear programming, the system maximized the number of units produced with this inventory. The output of the model fed a program which generated a separate circuit diagram for each unit to be assembled. In this way, a complicated manual matching problem (analogous to little league scheduling) was automated.

The suggestion models in the sample were a potpourri of applications which had a single common theme (i.e., performing a calculation whose output was a specific recommendation for action). These applications differed greatly in impact and significance. The user of an optimal bond bidding model stated that it had increased the profits of his bank because neither he nor any other person could possibly match the model's performance in generating solutions to an intrinsically combinatoric problem of choosing bond coupon rates which satisfy a series of complicated constraints at minimal cost to the bond underwriter. The developer of a system which calculated rates for group insurance policies felt that this system had probably saved money by preventing rate errors which had occasionally gone unnoticed. The implementer of a system which forecasts production requirements by product line and type felt that this system had an important impact on production planning since only very aggregate forecasts had been available previously. On the other hand, most of the remaining suggestion systems in the sample had their primary impact through saving time and/or aggravation by allowing someone to avoid spending several hours each week doing a task manually (and somewhat less optimally).

Comparative Findings

By merely asking what type of operation a decision support system performs, it was possible to classify each of fifty-six DSSs into one of seven categories. The categories range from type A, systems whose basic purpose was to retrieve simple aggregations of raw data, through type G, systems whose basic purpose was to suggest actions based on formulas or mathematical procedures. Aside from performing different types of operations, do the various types of DSS actually differ in significant ways?

Figure 2 summarizes some of the important characteristics of the systems of each type encountered in the sample. Each entry in Figure 2 is an attempt to describe in a single qualitative phrase the commonalities or predominant values of each characteristic within the systems of each type.

FIGURE 2 CHARACTERISTICS OF PARTICULAR DECISION SUPPORT SYSTEM TYPES

CHARACTERISTICS	DECISION SUPPORT SYSTEM TYPES						
	A FILE DRAWER	B DATA ANALYSIS	C ANALYSIS INFORMATION	D ACCOUNTING	E REPRESEN-TATIONAL	F OPTIMIZATION	G SUGGESTION
TYPE OF TASK	operational	operational or analysis	analysis	planning	planning	planning	operational
HANDS-ON USER	nonmanagerial line personnel	nonmanagerial line personnel or staff analyst	staff analyst	staff analyst or manager	staff analyst	staff or nonmanagerial line personnel	nonmanagerial line personnel
DECISION MAKER	nonmanagerial line personnel	nonmanagerial line personnel manager, or planner	manager or planner	manager, planner, or line personnel	manager	manager or nonmana-gerial line personnel	nonmanagerial line personnel
KEY ROLE	hands-on user	hands-on user	intermediary	intermediary, feeder	intermediary	intermediary	hands-on user

KEY USAGE PROBLEM	user motivation and training	can people figure out what to do with the system	how effective is the intermediary	integration into planning process	understanding	understanding	user motivation and understanding
SYSTEM INITIATOR	managerial	entrepreneurial	entrepreneurial	user or managerial	entrepreneurial	mixed	mixed
KEY DESIGN AND IMPLEMENTATION PROBLEM	defining the data; procedural changes	deciding how to use system; assessing impact on decisions	focusing usage and development: control mix of projects	getting people to participate seriously in planning process	richness vs. understandability	richness vs. linearity and understanding	designing rules sensibly
KEY CHANGE ISSUE	changing information sources and procedures	unfreezing job image and way of approaching problems	using system as a vehicle for change	unfreezing procedures people are familiar with	unfreezing ways of approaching problems	unfreezing ways of approaching problems	unfreezing standard procedures; avoiding a fear reaction
KEY TECHNICAL PROBLEM	system crashes; retrieval from large data base	flexible retrieval from broad data base: generality vs. power	flexible retrieval from broad data base	checking consistency of intention, meaning of numbers	modeling technology	modeling and solution technology	task modeling

(The sample contained seven, eight, three, eleven, twelve, six, and nine mini-cases of systems in categories A through G respectively.) Without getting into an elaborate methodological discussion, there is clearly some question of whether or not these mini-cases constitute a sufficient basis for generalizations by type. On the other hand, many of the commonalities by the system type in the data were relatively striking (e g., in many instances, most or all of the occurrences of a particular problem were within one type of system or two types with a similar characteristic).

To the extent to which its summary characterizations are accurate, Figure 2 indicates that systems of various types do differ in many significant ways. Consider, for instance, the notion of the *key role* in successful system usage. Since the planning and analysis systems (C through F) were often used through intermediaries who structured and performed much of the analysis, the success of these systems was especially dependent on the ability of the intermediary to maintain effective communication with decision makers. In the systems for operational tasks (especially A and G), intermediaries were not a main issue because the hands-on user was the decision maker.

The key usage problem varied greatly across the sample. In the systems for operational tasks (A and G), user motivation and training were major issues, especially since the system development efforts were often initiated by the users' superiors. In the data analysis systems (B), a recurrent problem was that system implementers and proponents incorrectly assumed that potential users would figure out how to apply the systems; in the more successful cases, either users were trained to use the system in a relatively repetitive manner or the implementers themselves were the users. For the representational and optimization models (E and F), the key impediment to successful usage was a lack of understanding of how the model worked and what it really represented. This was a direct consequence of the fact that the users of these models were typically intermediaries rather than decision makers.

Although the implementation patterns of the systems varied greatly, it was interesting that most of the data analysis systems, analysis information systems, and representational models (B, C, and E) were initiated by internal or external entrepreneurs. These individuals often found themselves in a position of attempting to sell their innovative ideas to managers and potential users. On the other hand, the need for most of the file drawer systems and accounting models (A and D) was identified by users or their superiors. One possible inference is that these latter types of DSS are more easily visualized and appreciated by nontechnical personnel.

Key design and implementation problems varied by system type. Since the file drawer systems (A) were all used by a large number of people, and often involved procedural changes in the way data was collected and reported on a day-to-day basis, the process of defining the data and handling the procedural changes was especially important. The data analysis

systems (B) were typically viewed as a way of making it convenient to analyze specialized data bases. In addition to the previously mentioned problem of deciding how to use these systems in changing situations, it was often difficult to assess the degree to which the analysis had a significant impact on decisions. The analysis information systems (C) in the sample were entrepreneurial efforts that grew incrementally; a key issue noted by the developer in each case was that of focusing usage appropriately and controlling the mix of projects that were undertaken. The purpose of most of the accounting models (D) in the sample was to compute the combined result of planning inputs submitted by people in different parts of the company; a significant problem for these systems was to get people to participate seriously in the planning process by submitting numbers that were well thought out. The tradeoff between richness and understandability was a key issue for both representational models and optimization models (E and F); as these models became richer and more detailed, they also became more difficult to explain. For suggestion models (G), the key design issue was whether or not it was actually possible to develop a standard method or set of rules for computing a suggested decision. In half of the sample cases, the specification of the method was considered a major breakthrough.

Systems of different types brought different kinds of change. File drawer systems, accounting models, and suggestion models (A, D, and G) brought changes in organizational procedures and information handling methods. The successful use of data analysis systems (B) by line rather than staff personnel seemed to require major changes in the user's job image. The success of advanced models (primarily E and F) often required changes in the way people thought about situations and solved problems.

Finally, the main technical challenges varied in a manner quite consistent with the generic operation performed by the system. In the data-oriented systems, the main technical challenge involved attainment of an appropriate balance between flexibility and efficiency in retrieval from a data base. In model-oriented systems, developing the model itself was the main technical challenge since current modeling methods are insufficient for many types of analysis of the future.

Conclusions and Implications

This article has attempted to support the hypothesis that a particular taxonomy is appropriate and useful. Since there is no statistical methodology for supporting such a hypothesis, the article has proceeded by proposing an organizing principle (generic operations), describing a taxonomy based on that principle, describing two examples of each type of system, and comparing the types of systems in terms of key characteristics. If it has been demonstrated that the taxonomy is an appropriate classification scheme, the question that remains is whether or why it is useful. I believe that the taxonomy is useful in a number of ways:

- As a guideline for designing systems,
- As a guideline for implementing systems, and
- As a framework for communication and research.

A Guideline for Designing Systems

One of the main implications of the taxonomy itself is that there are many different ways to use computers in supporting decision making. In designing a DSS, one of the first steps is to choose the type of system that will be developed. A potential use of the taxonomy is as a guideline in this process. In other words, a system designer might attempt to sketch out a system of each type as a potential *solution* to the *system design problem,* and would then combine the most useful features of each solution into his final design. Thus, the taxonomy would provide a substantive framework which would help in generating quite different approaches for supporting a particular decision. Whether this would actually be a fruitful exercise is a researchable question that has not yet been explored. Be this as it may, the sample did contain indirect supporting evidence in the form of several cases which at least suggest that the exercise of generating alternative designs might be useful. In one of these cases, a consultant felt very strongly that a representational model was needed for advertising decisions, whereas several users were more worried about the unavailability of data. After a period of trial and error, an effective procedure was developed in which a staff specialist provided briefings based in part on his use of a representational model (E) and in part on his use of a data analysis system (B). In another case, a plan to build a very expensive detailed simulation model (E) for raw materials allocation was abandoned when a staff man in a different department demonstrated that the same analysis could be done inexpensively with a rather simple optimization model (F). In a third case, a portfolio analysis system was installed to help portfolio managers think about portfolios from many different viewpoints (e.g., risk profiles, industry breakdowns, detailed sorted listings, etc.). After initial experience with this data analysis approach (B), it became clear that many portfolio managers wanted displays of what a portfolio would look like if particular decisions were made. To handle these *what if* inquiries, an accounting model (D) was added to the system. As a result, system usage increased. In all three cases, the consideration of different types of systems led to a better overall solution.

The taxonomy also provides insight for system user groups and system development groups concerning the types of systems that are currently installed and are on the drawing boards in their organizations. If none or few of the types of DSS are being used, the taxonomy provides a reference point in asking why existing applications encompass only a limited number of approaches for supporting decisions. One possible conclusion is that for this particular organization, computer-based decision support simply does

not have high priority. Alternatively, the users may not be familiar with the different approaches that can be used, and the designers may have been reluctant to try to initiate types of systems that are new and untested in the organizational setting. The fact that most of the B, C, and E systems in the sample were initiated by internal or external entrepreneurs gives added credence to the notion that people whose main activities are not computer-related may have a very limited appreciation of how computers can aid in decision making. For these individuals, the taxonomy may be valuable as a framework for understanding the technical approaches that are or are not being suggested or used by resident systems groups.

A Guideline for Implementing Systems

The comparative findings in the previous section indicate that key implementation issues vary across the different types of DSS. These findings complement the growing body of knowledge concerning the general topic of implementation.[7] This knowledge is useful because it provides guidelines for implementers and alerts them to early warning signals that may be symptomatic of incipient implementation difficulties. The comparative findings provide an additional framework for anticipating and avoiding potential problems. For instance, while implementing a data analysis system, a designer should be especially concerned about the user's willingness and/or ability to figure out how to apply the system in novel situations. In developing an accounting model, the implementer should put special effort into assuring that the input estimates are well thought out. In developing a representational model or optimization model, the implementer should be concerned about possible misunderstandings of what the model means and how it can or cannot be used.

Although it is obviously impossible to assure implementation success by merely understanding what did or did not work in the past, implementation success rates should benefit from organized knowledge about implementation if this information can be used to anticipate and avoid potential problems. To the degree to which the various types of DSS really are quite different, it is not only desirable, but also necessary to accumulate and analyze empirical data about the various types of systems. From the viewpoint of the MIS or DSS researcher, a stronger restatement might be as follows: unless taxonomies of this sort are taken into account in the research design, contradictory or inconclusive results can be *expected* because taxonomic contingencies (rather than *noise* per se) may well swamp the effects being studied.

A Framework for Communication and Research

Finally, the taxonomy provides a framework for communication and research. In light of the continual state of confusion that has surrounded

[7] See [6], and Schultz and Slevin [7].

terms such as management information system (which usually is not used by management), interactive system (which rarely interacts with decision makers), distributed processing (which has many meanings currently), the need for understandable taxonomies in the computer applications field should be clear. The findings described here — both the taxonomy itself and the fact that DSSs of various types differ in many important ways — illustrate that the term *decision support system* can have vastly different connotations for different people. Consider, for instance, the respective viewpoints of a user of a file drawer system and of a user of a very large optimization model. Whereas the file drawer user might conclude that the essence of decision support lies in on-line access to data, the optimization user might feel that on-line access is completely beside the point since each run of his DSS might require two hours of preparation and setup. Rather, he would probably identify accurate and complete modeling as the key issue in producing a useful DSS. Although the opinions of both users might be appropriate with regard to their own systems, neither conclusion would be appropriate for all or even most DSSs. Thus, whether it is this particular taxonomy or another, a classification scheme for DSSs is needed merely to help users and implementers communicate their experience in this emerging area.

References

[1] Alter, S. "Eight Case Studies of Decision Support Systems." Mimeographed. Cambridge, MA: Center for Information Systems Research, Sloan School of Management, M.I.T., 1974.

[2] Alter, S. "A Study of Computer Aided Decision Making in Organizations." Ph.D. dissertation, M.I.T., 1975.

[3] Gorry, G. A., and Scott Morton, M. S. "A Framework for Management Information Systems." *Sloan Management Review*, Fall 1971, pp. 55-70.

[4] Mason, R. "Basic Concepts for Designing Management Information Systems." In *Information for Decision Making: Quantitative and Behavioral Dimensions*, compiled by A. Rapaport, pp. 2-16. Englewood Cliffs, NJ: Prentice-Hall, 1970.

[5] Mason, R., and Mitroff, I. "A Program for Research on Management Information Systems." *Management Science* 19 (1973): 475-487.

[6] "Special Issue on Implementation," *Management Science*, in press.

[7] Schultz, R., and Slevin, D. *Implementing Operations Research/Management Science*. New York: American Elsevier, 1975.

[8] Simon, H. *The New Science of Management Decision*. New York: Harper and Row, 1960.

International Journal of Information Management 46 (2019) 93–103

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Business intelligence and analytics for value creation: The role of absorptive capacity

Katerina Božič*, Vlado Dimovski

University of Ljubljana, Faculty of Economics, Kardeljeva ploščad 17, 1000 Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Business intelligence and analytics
Value
Insights
Absorptive capacity
Assets

ABSTRACT

Firms continuously report increased competitive value gains from the use of business intelligence and analytics (BI&A), however, little is known about how insights from BI&A are transformed to added value to date. We have conducted fourteen in-depth, semi-structured interviews with a sample of informants in CEO positions, IT managers, CIO, Heads of R&D, as well as Market Managers from nine medium or large-sized European firms. Applying the absorptive capacity's theoretical lens, we have provided evidence that absorptive capacity's capabilities are an underlying foundation in the process of transforming BI&A triggered insights into valuable knowledge. Moreover, this process is supported by technological, human, and relationship assets.

1. Introduction

The amount of data and information generated on a daily basis continuously increases, forcing firms to increasingly rely on external knowledge and information to enhance firm innovation and performance (Benner & Tushman, 2015; Ireland, Hitt, & Vaidyanath, 2002). With the quick development of computer intelligence as well as the appearance of "big data" concept, business intelligence and analytics has become an increasingly important concept for researchers and practitioners (Chen, Chiang, & Storey, 2012). Although BI&A were initially used for decision-making support activities, they have been increasingly considered for organizational learning and adjustments, improving operational efficiency, and strengthening organizational intelligence (Trieu, 2017). A survey conducted by IBM Institute for Business Value and MIT Sloan Management Review reported that firms are increasingly gaining competitive advantage from analytics (58% of the more than 4500 respondents reported competitive value gains from analytics) (Kiron & Shockley, 2011). Not surprisingly, Gartner's survey on IT Spending found BI&A to be a top priority for most of the analyzed firms and predicted that BI&A would remain one of the top foci for the leading firms (Gartner, 2013).

On the other hand, firms have acknowledged the potential of BI&A to generate insights and knowledge from both external and internal sources of knowledge (Shehzad, Khan, & Naeem, 2013; Wang, 2014; Wixom, Watson, & Werner, 2011; Yeoh & Koronios, 2010). As the complexity of the data is increasing, humans have difficulties interpreting the external information due to limited mental capacities (Jansen, Van Den Bosch, & Volberda, 2005; Sammut & Sartawi, 2012). More information is not necessarily beneficial for the organization since its information and knowledge processing capacity is limited as well (Simsek, 2009). As a result, organizations develop information filters and routines to cope with bounded rationality (March, 1978; Nelson & Winter, 1982). BI&A have found it possible to expand the human mental capacity as well as the firm's absorptive capacity by increasing the ability of individuals and firms to receive, store, analyze and transfer information with fewer errors (Brynjolfsson & Hitt, 2000; Elbashir, Collier, Sutton, Davern, & Leech, 2013; Simon, 1991). While various streams of studies have provided research on the BI&A potential, there has been little attention given to the improvement of understanding the role of BI&A in the process of knowledge generation from external data and with it, the underlying mechanisms that facilitate this process.

Despite the prominence of BI&A as a source of competitive advantage with an abundance of studies acknowledging the ability of BI&A to derive business value, anecdotal evidence has been made to capture the BI&A value creation process (Chen, Preston, & Swink, 2015; Fink, Yogev, & Even, 2017; Trieu, 2017; Vidgen, Shaw, & Grant, 2017). Prior research in the information systems (IS) research field has examined the role of BI&A for insight generation; however, predominantly from the technological aspect (Bose, 2009; Chaudhuri, Dayal, & Narasayya, 2011; Ranjan, 2009). Only a few studies have investigated the role of BI&A from an organizational aspect; such as, organizational learning, organizational capabilities, effective use, and customer relationship management (Elbashir, Collier, & Sutton, 2011,

* Corresponding author.
E-mail addresses: katerina.bozic@ef.uni-lj.si (K. Božič), vlado.dimovski@ef.uni-lj.si (V. Dimovski).

https://doi.org/10.1016/j.ijinfomgt.2018.11.020
Received 22 February 2018; Received in revised form 24 November 2018; Accepted 30 November 2018
Available online 07 December 2018
0268-4012/

2013; Forsgren & Sabherwal, 2015; Real, Roldán, & Leal, 2014; Trieu, 2017; Yeoh & Popovič, 2016). Despite the strong technological focus, valuable customer insights are usually a result of a meaningful transformation of BI&A insights into meaningful knowledge that is subsequently dispersed across business units to be acted upon (Fan, Lau, & Zhao, 2015; Shollo & Galliers, 2016).

Ergo, more recent studies (Fink et al., 2017; Shollo & Galliers, 2016) have criticized overemphasizing technology without accounting for the human 'sense-making' processes. As Sharma, Mithas, and Kankanhalli, 2014, p. 435) "insights emerge out of an active process of engagement between analysts and business managers using the data and analytic tools to uncover new knowledge." Accordingly, Shollo and Galliers (2016) have provided empirical evidence of the BI&A agency in data selection and problem articulation for the active process of knowing. Moreover, Fink et al. (2017) have presented and empirically tested a model of BI& value creation which identified BI team and infrastructure assets that were transformed through operational and strategic BI capabilities into operational and strategic value; a process moderated by exploitative and explorative learning. Although they attempted to theoretically advance the BI&A research through the lens of organizational learning, they offered a limited understanding of the underlying processes, therefore, calling for further research to strengthen the theoretical foundation of BI&A research. Moreover, as Trieu (2017) noted in his most recent, exhaustive literature review study, there is a lack of research that studies the complementary links between BI impacts and organizational BI assets to help the organization better understand the value creation process, and has suggested applying an inductive inquiry approach to explore this complex phenomenon. Extending the discourse, we seek to answer the following research question*: "How are BI &A triggered insights transformed into valuable knowledge?*

To address this research question, we have conducted qualitative research involving fourteen key informant interviews in nine European firms. Following Trieu's (2017) recent findings, we consider the existing absorptive capacity's theoretical lens as a sensing device to analyze empirical data. Although concepts such as absorptive capacity capability have been used in prior studies (e.g., Elbashir et al., 2011; Ramamurthy, Sen, & Sinha, 2008; Trieu, 2017), it has remained unclear how the underlying capabilities and resources contribute to business value creation. Using the abductive method of inquiry, we have attempted to elaborate on existing theories, focusing on the role of BI&A in the organizational knowing processes and its underlying capabilities and assets which facilitate value generation process. This includes but is unrestrained to decision-making.

Our research identified the role of the four absorptive capacity's capabilities in insight generation and exploitation. Secondly, we studied the assets needed to allow full realization of the identified absorptive capacity capabilities. Our findings suggest that absorptive capacity allows external business insights from BI&A to be successfully assimilated and transformed into valuable business knowledge. Internal human, technological, and relationship resources have appeared to be the prerequisites necessary for the insights transformation process. A better understanding of the former has contributed to previous IS and management research. Also, practitioners can benefit from a comprehensive overview of the capabilities and resources needed to turn BI&A insight into meaningful actions and decisions, allowing them to adjust their efforts accordingly. Therefore, we offer a holistic and systematic understanding of the underlying capabilities and the underpinning assets that allow knowledge extraction from BI&A insights.

The remainder of this paper is structured as follow. In the first section, we review the concept of BI&A and the absorptive capacity theory. In the second section, we present the research context and the methodology, followed by the overview of findings. The last section concludes with a discussion of the findings, implications for theory and practice, and limitations and further research suggestions.

2. Literature review

This section offers a review of the current literature revolving around the BI&A value creation process; focusing primarily on the organizational impacts that result from BI&A use. Next, we present the Absorptive Capacity Theory and discuss the resources necessary for the full realization of the absorptive capacity capability.

2.1. BI&A definition

Existing literature offers several definitions of BI&A, none of which has been well-accepted. Namely, from the first appearance of Luhn (1958) the BI&A term was most commonly used to describe systematic processes (Lonnqvist & Pirttimaki, 2006), methodologies (Ranjan, 2009), technologies (Bose, 2009; Kimball & Ross, 2011), analytical tools (Davenport & Harris, 2007; Elbashir, Collier, & Davern, 2008; Watson & Wixom, 2007), and techniques (Lim & Lee, 2010) that use computer-supported systems to collect, analyze, and disseminate information for effective business activities and better decision-making. The current, most widely used definition is Chen et al. (2012), p. 1166) encompassing definition that covers most of the existing literature perspectives and refers to BI&A as "the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions". This perspective diversely leads to different aspects of the value creation process. Thus, a literature review is required to define the current study and determine what we already know about the BI&A value creation process.

2.2. BI&A value creation process: A literature review

Recent academic and practitioner literature emphasize the ability of organizations to create value through the use of BI&A (Chen et al., 2012; Larson & Chang, 2016; Lavalle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011; McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012; Mithas, Lee, Earley, Murugesan, & Djavanshir, 2013). For instance, Lavalle et al. (2011) found top-performing organizations to have substantial experience using BI&A to create value. Similarly, Chen et al. (2012) have recognized the role of BI&A in acquiring intelligence on customer needs and opinions, leading to new business opportunities. Despite increased interest, the process of transforming the insights triggered by BI&A into valuable business knowledge remains vague. Thereby causing many authors such as Sharma et al. (2014) and Ross, Beath, and Quaadgras, (2013) to call for a deeper analysis of how organizations can create value from the use of BI&A and understand the underlying resource allocation processes.

Over the past decade, a widespread interest of researchers and academia have stimulated a remarkable body of research to determine the added value of investing in BI&A technology. Many studies are contributing to this knowledge in different ways. A general presumption from the extant literature is that BI&A use leads to improved efficiency for the decision-making process. Thus, a common premise of this view is that BI&A allow identification, capture, and production of new insights and knowledge, later used for decision-making (Acharya, Singh, Pereira, & Singh, 2018; Hou, 2012; Kowalczyk, Buxmann, & Besier, 2013). For instance, Popovič, Turk, and Jaklič, (2010) proposed a research model for deriving business value from BI&A and found BI&A maturity and absorbable capabilities to facilitate the use of quality information enabled by BI&A in business processes and decision-making. Similarly, Elbashir et al. (2008) in their survey-based study, found BI&A to deliver value through improved business processes (business partner relations, internal process efficiency, and customer intelligence benefits). Trkman, McCormack, De Oliveira, and Ladeira, (2010) found firms that support analytical capabilities with good IS to perform better in delivering decisions. Furthermore, IşıK, Jones, and Sidorova, (2013) empirically found the importance of technological capabilities and

high-quality data to support decision-making and accessibility to all users across different decision-making environments. Hannula and Pirttimaki (2003), in their survey-based study, found the most significant benefits, provided by BI&A, were the acquirement of better quality information for decision making, improvement in the ability to anticipate threats and opportunities as well as the growth of knowledge base and time savings. Despite the importance of these studies in the identification of factors influencing delivered value through improved decision-support, it remains unclear how new knowledge is obtained as a result.

Other studies research the type and measurement of value that is generated from BI&A use. Hence, Watson (2009) and Watson and Wixom (2007) found BI&A to generate a range of benefits from local impacts (such as, cost savings from data consolidation, time savings from the user), to global ones, such as, more and better information, better decisions, improvement of business processes, which are ultimately difficult to assess due to their "soft" nature. In addition, Clark, Jones, and Armstrong, (2007) have presented a theoretical model of benefits from BI&A and other decisional support technologies and have found value to be difficult to measure, since many organizational factors such as culture, the use of information, management commitment can heavily influence the BI&A perceived value and are also difficult to assess. Nonetheless, Davenport and Harris (2007), in a multiple case study research, found BI&A resources to be an insufficient source of value if not coupled with sufficient data analytical capability and a strong analytical culture. Even though they discussed the potential benefits from BI&A use in a more detailed or cursory fashion, one gets the impression that technology delivers value in some inert form that can be transferred and controlled.

Until now, researchers have examined the BI&A value creation process using a variety of theories and empirical approaches. Since most of the studies of IT value creation ground their studies on a Resource-Based Theory (RBT), Dynamic Capabilities Perspective, and Information Processing View (Bharadwaj, 2000; Melville, Kraemer, & Gurbaxani, 2004; Ryu & Lee, 2013; Santhanam & Hartono, 2003; Trkman et al., 2010; Wang & Ahmed, 2007), recent studies of BI&A value creation have used similar theoretical foundations (Cao, Duan, & Li, 2015; Cao, Duan, & Cadden, 2019; Côrte-Real, Oliveira, & Ruivo, 2017; Fan et al., 2015; Fink et al., 2017; Kowalczyk & Buxmann, 2014). In accordance to Trieu (2017) call for consideration of firm factors (such as organizational size, scope, and absorptive capacity) while investigating the relationship between BI&A assets and impact on understanding the dependence of its value on organizational resources allocation, we reviewed the theoretical foundation of the Absorptive Capacity Theory and the BI&A assets to identify conceptual ideas as a guideline in preparing interviews.

2.3. Theoretical foundation

2.3.1. Absorptive capacity

In their research on innovation, Cohen and Levinthal (1990), p. 128) conceptualized a firm's absorptive capacity as "the ability of a firm to recognize the value of new, external information, assimilate it, and apply it to commercial ends." It depends on the prior related knowledge which allows firms to better evaluate the signals for technological advances and development. Absorptive capacity, therefore, allows identification of new knowledge by an organization from outside the organization and the assimilation and integration of knowledge within existing knowledge internally (Arora & Gambardella, 1994; Cohen & Levinthal, 1990; Kogut & Zander, 1992). Thus, there is not only new knowledge generation but also competence leverage is required to exploit existing technologies (Danneels, 2002). This classifies absorptive capacity as integral to dynamic capabilities since it allows for continuous acquisition, search, and management of knowledge (Pavlou & El Sawy, 2010). An absorptive capacity enhances the interaction of the organization with the external environment through greater external knowledge assimilation as well as improving knowledge sharing and learning between organizational subunits (Rosenkopf & Nerkar, 2001).

It is important for the firm, however, to constantly invest in the development of its absorptive capacity, since the firm might become not be aware of the technological opportunities and signals in a given field (Cohen & Levinthal, 1990; Kogut & Zander, 1992). Thus, organizations with high levels of absorptive capacity are proactively exploiting technologies and market opportunities in the environment independent of their current performance by combining both internal and external knowledge sources. Organizations with a modest level of absorptive capacity are more reactive, responding to some performance criterion (Cohen & Levinthal, 1990; Lavie & Rosenkopf, 2006; Rothaermel & Alexandre, 2009). The higher the level of absorptive capacity, the higher the spillovers between internal and external knowledge sourcing (Rothaermel & Alexandre, 2009).

Roberts, Galluch, Dinger, and Grover, (2012) emphasized a few important assumptions underlying the absorptive capacity. Firstly, absorptive capacity is domain-specific. The ability to determine the value of the external knowledge depends on prior-related knowledge. Although it is important that the firm's existing knowledge overlap with external knowledge for successful acquisition, a strong overlap will limit new opportunities and insight generation (Lord & Ranft, 2000). Secondly, absorptive capacity is firm-specific. Absorptive capacity depends on the absorptive capacities of a firm's individuals; however, it is not just the sum of the individuals' capacities, but is also composed of the overlapping of individuals' knowledge and the knowledge transfer across and within subunits (Cohen & Levinthal, 1990; Roberts et al., 2012). Effective information and knowledge obtainment require both structures and processes that increase the quality and quantity of information and knowledge and can integrate it into collective action (Sheremata, 2000, p. 405). As individuals from various departments obtain and interpret knowledge in various ways, internal communication and integration are important for improving the quality of learning (Brown & Eisenhardt, 1995; Nonaka, 2007). The increased information flow can cross-functionally enhance the quality of learning. Thirdly, absorptive capacity is path-dependent. Absorptive capacity accumulation in one period will, therefore, allow more efficient absorptive capacity accumulation into the next. For effective organizational learning, there must be a balance between inward and outward-looking absorptive capacity, since excessive dominance by one of them is dysfunctional (Cohen & Levinthal, 1990; Grant, 1996). If the body of expertise becomes sufficiently specialized, it could impede the assimilation of external knowledge, resulting in the so-called Not-Invented-Here Syndrome, in which firms reject innovative ideas from the environment (Cohen & Levinthal, 1990). Lack of external knowledge openness and past experiences lacking in reward may reflect organizational myopia towards different external sources (Levinthal & March, 1993). Path-dependency allows firms to predict more accurately the potential of technological advances.

As Grant (1991) discussed, assets are the basic unit of analysis, whereas capabilities are integrated arrangements of assets. Hence, as Fink et al. (2017) and Lin and Wu (2014) argued capabilities represent the primary source of the value and are often seen as a converter of organizational resources/assets into a competitive advantage. Several conceptualizations of the construct of absorptive capacity can be found in the literature (Jansen et al., 2005; Todorova & Durisin, 2007; Zahra & George, 2002). In line with prior research, we define absorptive capacity as a second-order dynamic capability that builds, integrates, and reconfigures underlying first-order capabilities and zero-order assets to create and deploy knowledge (Gao, Yeoh, Wong, & Scheepers, 2017; Wang & Ahmed, 2007; Zahra & George, 2002). Hence, absorptive capacity is captured by four first-order capabilities that reflect dynamic processes, such as acquisition, assimilation (potential absorptive capacity), transformation, and exploitation (realized absorptive capacity) (Flatten, Engelen, Zahra, & Brettel, 2011; Lane, Koka, & Pathak, 2006). The first capability – acquisition, refers to the identification and

obtainment of information through external sources relevant to the firm's operations and is affected, as well, by the prior knowledge (Cohen & Levinthal, 1990; Zahra & George, 2002). The second on - assimilation, refers to the firm's ability to analyze, interpret and understand externally acquired information (Cohen & Levinthal, 1990; Flatten et al., 2011). The third capability - transformation, refers to developing routines that facilitate a combination of existing knowledge with new, acquired knowledge and internalization of this knowledge (Zahra & George, 2002). The last one – exploitation, refers to an application of the acquired, transformed knowledge to commercial ends (Cohen & Levinthal, 1990). The four first-order capabilities of absorptive capacity together enable firms to exploit new knowledge, enhance the firm's performance and achieve competitive advantage through new product innovation. However, absorptive capacity not merely connects the first-order capabilities but combine them creating synergistic outcomes (Lichtenthaler, 2009; Raisch & Birkinshaw, 2008; Wang & Ahmed, 2007). Nevertheless, following recent Gao et al's (2017) recommendations, we examined absorptive capacity on an organizational level of analysis in the behavioral domain of study, which refers to activities and application of the technical domain.

2.3.2. Underlying BI&A assets

As the process of knowledge extraction does not happen in isolation, different BI&A resources either facilitate or inhibit knowledge accumulation and utilization. Thus, the BI&A business value has been found to be contingent on the underlying BI&A resources/assets (Fink et al., 2017; Wieneke & Lehrer, 2016). Extant literature has already presented some potential resources/assets that could impact the value creation process. For instance, Cosic, Shanks, and Maynard, (2015) presented four categories of organizational resources and capabilities, such as governance, culture, people, and technology capabilities. Further, Shuradze and Wagner (2016) proposed three groups of assets for data analytics, such as technological infrastructure, personal expertise, and relational infrastructure. Similarly, Wieneke and Lehrer (2016) presented physical, human, and organizational resources as a basis for social-media insight exploitation. Nevertheless, Castro, Delgado-Verde, Amores-Salvadó, and Navas-López, (2013) described human, technological, and relational assets for intellectual capital creation and product innovation. Based on the reviewed literature, we identified technological, human, and relationship assets that may underpin the first- and second-order dynamic capabilities of absorptive capacity, influencing the knowledge creation process in the BI&A context. Thus, we consider assets as raw material that would affect the capabilities' development process (Ravichandran & Lertwongsatien, 2005; Wade & Hulland, 2004).

Here, technological assets refer to technical platforms, IT infrastructure, physical IT assets, data repositories, communication technologies, and IT architectures (Bharadwaj, 2000; Wade & Hulland, 2004). Technological assets, such as databases and networks are easily acquired in the market, in contrast to sophisticated IT infrastructure and communication technologies which are difficult to imitate. IT technological assets have found to enhance a firm's absorptive capacity (Roberts et al., 2012; Yeoh & Popovič, 2016). Technology allows firms to codify, process, store and recover information that has been acquired (Argote, McEvily, & Reagans, 2003). Next, it facilitates knowledge diffusion across different business units or networks for further transformation and exploitation (Lee & Choi, 2003). In summary, it enables firms to acquire, process, manage and share data and information for meaningful insights generation and, furthermore, allows fast and cost-effective integration of new technologies with existing ones (Ravichandran & Lertwongsatien, 2005).

On the other hand, human assets refer to workforce business knowledge, technical skills, work experience and relationships (Barney, 1991; Teece, 1998). Prior research has shown the importance of human assets for absorptive capacity capability. Namely, employees with strong business knowledge and technical skills are more efficient in recognizing and valuing new external knowledge, therefore, increasing the knowledge level in the firm (Lund Vinding, 2006; Mangematin & Nesta, 1999). Moreover, greater work experience increases the accumulation of firm-specific knowledge, increasing the ability to transform and exploit assimilated knowledge (Zahra & George, 2002).

Nonetheless, relationship assets encompass inter-divisional relationships, external (client) networks, management sponsorship and culture (Ross, Beath, & Goodhue, 1996; Wade & Hulland, 2004). The knowledge transfer across different business units enable intra-organizational knowledge flows and knowledge consolidation (Cohen & Levinthal, 1990), which in turn, increases both the recipient's knowledge base and organization's knowledge base (Pawlowski & Robey, 2004). Organizational culture strongly influences these processes (Verona & Ravasi, 2003).

3. Research context and the methodology

3.1. Sample and procedures

The main objective of the exploratory inquiry was to examine how BI&A triggered insights are transformed into valuable knowledge and what the underlying capabilities and assets are. We found the exploratory methodology of the research suitable since the phenomenon is new, broad and complex, so it is difficult to identify causal relationships (Corbin & Strauss, 1990; Eisenhardt, 1989; Pare, 2004). The exploratory analysis aids to extend existing theory, offering additional insights into the complex relationship between the constructs (Denzin & Lincoln, 2005; Eisenhardt & Graebner, 2007). We apply abductive scientific reasoning (Mantere & Ketokivi, 2013; Strauss & Corbin, 1998), where initial inductive insights from empirical data are engaged with existing theoretical knowledge to explain the empirical puzzle then. We assume the semi-structured interview to be the most effective method of gathering information for our research since it is flexible and accessible enough (Alvesson, 2003; Brinkmann, 2014; Holstein & Gubrium, 1995).

We followed the theoretical, purposeful sampling approach in selecting participants in the study to ensure a relevant representation of the actual state (Denzin & Lincoln, 2005). Nine European firms from different sectors: high-tech, manufacturing, telecommunicative, service-oriented, retail, financial, and energy were selected to conduct the interviews. Acknowledging the fact that larger firms are more able to invest in different IT technologies with related employee training (Chatterjee, Grewal, & Sambamurthy, 2002; Elbashir et al., 2013), we have, therefore, considered medium and large-sized firms. The expert interviewees had to fulfil l the following screening criteria: (1) having deep knowledge about the organization; (2) having more than three years of experience in BI&A initiatives, and (3) being at leading IT or management position. According to the needs of this study, we selected fourteen expert interviewees/key informants, in positions within the variety of Chief Executive Officer, Chief Information Officer, IT manager, Head of R&D, or Market Research Manager. Out of fourteen, five key informants were selected through the snowballing method. All of them possessed and actively used BI&A in their everyday work. Thus, over a two-year period (between February 2016 and October 2018), we carried fourteen interviews involving nine firms. Table 1 provides a breakdown of the informants included. The relatively small sample size of interviews was, however, sufficient to generate theoretical saturation, whereas, the new interviews provided no additional data that lead to any new emergent themes, as discussed by many authors (Boyce & Neale, 2006; Crouch & McKenzie, 2006; Urquhart & Fernandez, 2016). Moreover, increasing the sample size may question the ability of the researchers to devote sufficient attention to dataset analysis (Marshall, Cardon, Poddar, & Fontenot, 2013). All the firms, included in the research, had used BI&A for several years at that time and were appropriate candidates to illuminate the BI&A value generation process when the study was conducted.

With each interviewee, we conducted a semi-structured interview

Table 1
Informants data.

Firm	Number of informants/ Position	Country	Industry sector	Mode
A	1: CEO	Croatia	Services	On-site
B	1: Chief Information Officer	Slovenia	Software	On-site
C	2: CEO; IT manager	Austria	High-tech industry	Skype
D	1: Market Research Manager	Germany	Software, IoT	On-site
E	2: IT Manager, Head of R&D	Germany	Manufacturing	Skype
F	2: IT manager, Managing Director	Germany	Telecommunications	Skype
G	1: Chief Information Officer	Slovenia	Retail	On-site
H	2: IT manager, Managing Director	Slovenia	Financial	On-site
I	2: IT Manager, Head of R&D	Slovenia	Energy	On-site

based on the interview protocol (Appendix), with an average duration of one hour. We asked all informants participating in the study to speak as the representative voice of the collective. Firstly, we collected data about interviewees' position as well as experience and general data about the firm. Next, after presenting the goal of the research, we asked the interviewees to describe their understanding of BI&A, discussing the highlighted topic and the use of it in as much detail as possible. Since the specific purpose of the interview was to learn as much as possible about the interviewees' perceptions and concerns about BI&A, we asked a set of open-ended questions. At the end of the interview, each participant was asked for other details that might be relevant to the interview. Since some of the interviewees did not allow tape-recording, we took detailed field notes during the interviews, complementing them with detailed notes immediately after they were completed. Although we acknowledge that taping would provide richer and more accurate data, we had to consider the participants' requirements. After each interview, a systematic analysis of the notes taken was completed.

3.2. Data analysis

The data analysis procedure followed the guidelines specified for methods of naturalistic inquiry and constant comparison (Charmaz, 2006; Glasser & Strauss, 1967; Schwandt, Lincoln, & Guba, 2007). The latter allowed us to adjust iteratively theoretical categories and delineate aggregated dimensions. Each interview was systematically examined and systematized within the categories. To assure better quality and accuracy of the coding process two independent reviewers coded the same data. We started with identifying initial, first order codes that were informant-centric (Corbin & Strauss, 1990). Next, we used axial coding, seeking similarities and differences between and amongst these categories, assembling first-order codes into theoretical categories. Finally, after coding saturation regarding the refining categories that had been reached, we distill the emergent theoretical categories into aggregate theoretical dimensions. We have, however, finished these steps in a recursive analytic procedure (Locke, 2002). At the end of the coding process, we calculated the interrater reliability among the two coders, and we reached a high level of agreement (0.92), considering to be a justifiable verification of the coding procedure. The final data structured is summarized in Fig. 1 and details have been described in Section 4. To assure better quality and accuracy of the coding process, we used peer debriefing (Creswell & Miller, 2000; Schwandt et al., 2007; Spall, 1998). We have, hence, invited two external peers (departmental members), that were not included in the research to evaluate and reflect on the data collection and analysis procedures. A detailed search for disconfirming evidence was conducted until we reached a strong level of agreement.

4. Findings

This section presents our findings, drawn on the interview data. The origin of data, presented in the quotation marks below, is extracted from the field quotes and the observation field notes, verbatim.

4.1. BI&A definition and characteristics

We asked interviewees to describe how they define BI&A and what is the importance of that technology for their firm. Before offering a concrete definition, different interviewees highlighted different properties depending on the degree of BI&A use. As Interviewee 3 noted: *"In general, business intelligence and analytics means an advanced analysis to generate intelligence from business data for improving business. We apply advanced techniques, like data mining, semantic and network analysis, and machine learning to understand our customer preferences, mostly real-time. This allows us to articulate the problem, which is something that was difficult with transactional data and conventional analytics"*. Similarly, Interviewee 7 said: *"It is the advancement of BI technology and techniques that fit into new developments and gather timely information. Without real-time or close to real-time market information, we lag immediately behind the competition"*. Moreover, Interviewee 12 noted, "*It is a bunch of technologies, that allows us to create a real-time relevant knowledge, based on prior and current customer information."* Interviewee 14 elaborated "*BI&A are technologies and methodologies that help our firm predict future trends to enhance the reliability of the decision-taken. Hence, we mainly rely on predictive analytics, machine learning, and regression for better pattern recognition*". On the other hand, some interviewees did not agree with the strict distinction between traditional BI and BI&A. Thus, Interviewee 1 commented: *"It is nothing, but traditional database business intelligence suited for larger datasets that are used for knowledge discovery. Currently, some fads appear to be new, revolutionary but are only an evolution of existing technologies and techniques"*.

However, most of the interviewees highlighted the so-called "big data" challenge; regarding increased volume, variety, and velocity of data generation. Thus, Interviewee 2 noted: *"The main difference between the older BI and BI&A is in the challenge to analyze and store large, unstructured and complex datasets, presently known as big data, requiring unique technologies. Here, we try to develop new approaches to make sense of the massive amount of data we collect as well as generating previously unknown insights. All of which is, however, easier said than done"*.

Nonetheless, they emphasized the importance of the value of information as a crucial component in obtaining competitive advantage: *"We found the main difference between traditional BI and business analytics to be in the specific approaches made to realize the value in understanding the customer, competition, and market behavior. We collect an enormous amount of data from smartphones, social media, and the Internet. It is, however, not about how much you have, but what you have in your hand when deciding. It requires us to develop different abilities. Otherwise, we could quickly become obsolete."* (Interviewee 9). Thus, most of them agreed that new, valuable insights are the greatest benefit from BI&A use.

4.2. Underlying capabilities: absorptive capacity perspective

The interviewees recognized the role of BI&A processes for insight and knowledge generation. To incorporate BI&A's generated insights into the value creation processes, we draw on four distinct, but complementary absorptive capacity capabilities: acquisition, assimilation, transformation, and exploitation, as suggested by Carlsson (2003) and Zahra and George (2002). Thus, we used the absorptive capacity's theoretical lens as a sensing device to analyze empirical data. However, all the claims are grounded in empirical, field data.

The analysis of the interviews emphasized the role of strong acquisition capability for being able to identify and obtain valuable data from external sources. Namely, the overwhelming amount of data from external sources requires careful cleaning, conditioning, and integration of data sets to make them usable. This is a demanding process since data is often the origin in heterogeneous sources, coming with noise,

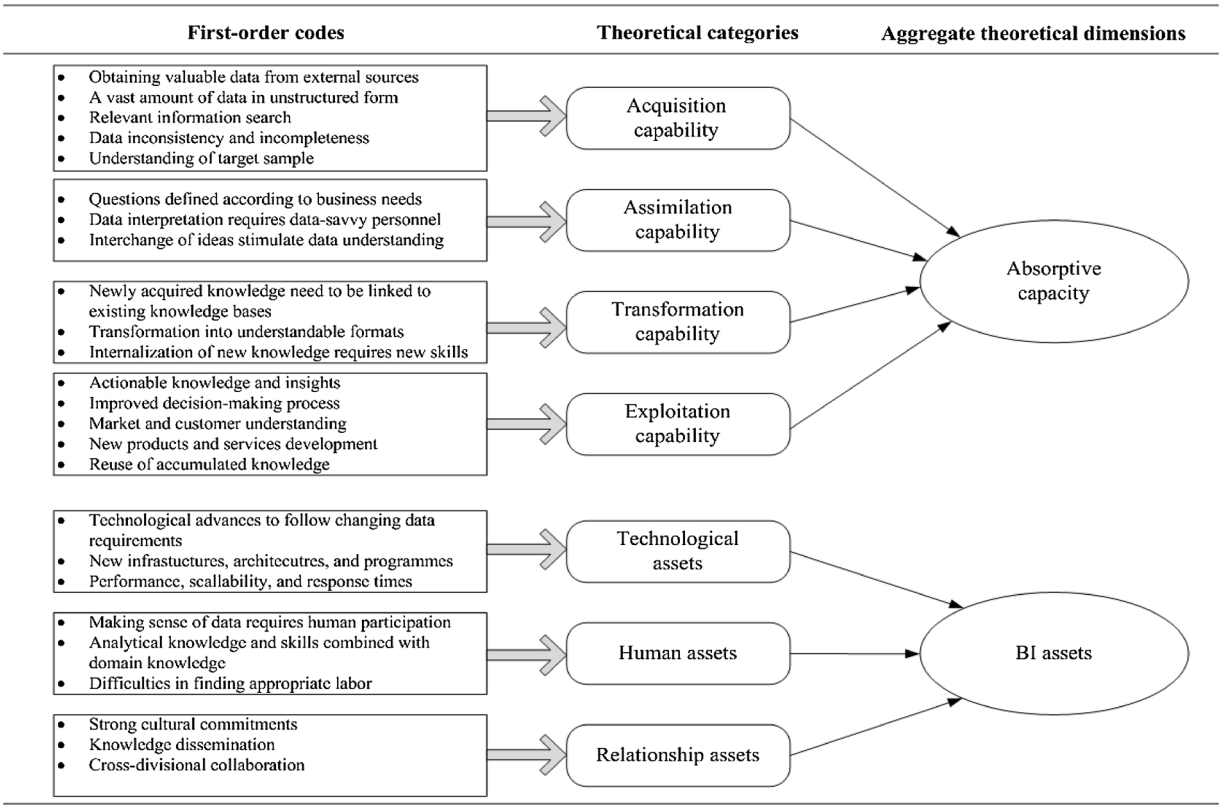


Fig. 1. Overview of the data structure.

opacity, and incompetencies (Gandomi & Haider, 2015). The increased volume and unstructuredness of data has increased storage costs, adding real-cost to firms. Thus, it is crucial to have the ability to determine the value of retaining data for future actions, as noted by Interviewee 4: *"Although data can be a source of useful insights, we have quite often a problem with data inconsistency and incompleteness. We spend hours and hours cleaning data, deciding which attributes to keep, how to represent and analyze them, which, however, often results in wasted time and money"*. Moreover, firms reported an increased collection of information from social networks as a source of data (such as Facebook, Twitter, LinkedIn), has shown, many times, to have a limited scope as well as limited quality. Instead of assuming social networks are representing the whole population, firms should assume users as a sample, requiring critical examination and understanding of the target sample. As Interviewee 6 noted: *"We made a huge mistake by collecting all datasets related to our business problem. This lead to storage problems, endless search, and inconsistent insights. Unfortunately, it cost us a fortune to figure out that we needed to target collect data"*.

Additionally, the interviewees highlighted the importance of a firm's assimilation capability in processing internal and external data. To be able to capitalize on the generated knowledge, firms must define the business needs and objectives, while defining the right questions. As mentioned by Interviewee 2: *"Our IT professionals require specific questions to be asked upfront that would clean and prepare samples out of the whole dataset. When we don't know what to ask or don't understand the sample, the abundance of noise in datasets lead to weird conclusions."* Hence, Interviewee 13 commented "*We use challenging as a technique for asking a relevant research question. The one that orders particular analysis must come with a clear estimate about what is the potential added value of the analysis, and what are the potential outcomes"*. Therefore, human intervention is crucial in making sense of data. Interpretation of externally acquired information requires data savvy decision-makers who can estimate and understand the potential insights' value. Since developing IT skills for management personnel can be a time-consuming process, the interviewees suggested a close collaboration between IT and managers. For instance, Interviewee 5 commented: *"We ask our managers to work closely with IT personnel to tackle the challenge of data interpretation, which offers us a strong basis of heterogeneous expertise in IT, marketing, and customer relationships."* Similarly, interviewee 10 noted: "*Our data scientists work together with the reporting staff to create useful recommendations for the decision-makers*." Nevertheless, Interviewee 13 remarked: "*We believe we are much stronger when complementing our knowledge and capabilities. Hence, we don't want a single person that is knowing everything. Instead, we create teams, that complement each other's knowledge. Otherwise, you will reach an adverse effect, having people that are mediocre in everything"*.

Transformation capability was found to play an important role in facilitating the internalization of newly, acquired knowledge within an existing knowledge basis. Having amazing insights does not mean one has succeeded. These insights need first to relate to an existing knowledge base and then to be disseminated within the organization to reach persons that need them for decision-making or an action-taking improvement. However, interviewees emphasized the importance of an insight's format that is clearly and easily understood: *"We ask our teams to transform results in graphics, dashboards, and interactive visualizations so that other less data-savvy personnel can access and comprehend their value in a more intuitive manner"* (Interviewee 3). Similarly, Interviewee 11 elaborated: *"We are increasingly relying on data visualization to present the mined information in a comprehensible manner*." In contrary, an employee's resistance to BI&A, as a response to information overload, appeared to impede the internalization of new knowledge. Following that, Interviewee 6 noted: *"We were forced to organize different educational courses and workshops to develop skills of information assessment and interpretation."* Nevertheless, employees became more confident in BI&A use after equipped with the appropriate skills and knowledge.

Finally, the exploitation capability has been found to allow transformed knowledge commercialization. Our interviewees reported BI&A insight use for a different purpose: to understand customer and market

behavior, optimize business processes (decision-making, supply-chain), optimize advertisement campaigns and pricing strategies, develop new products and services, manage financial risks, improve efficiency, identify faults and to provide proactive machine maintenance. However, as noted by Interviewee 8: *"It requires a lot of effort to maximize the value of generated insights. We motivate our employees to search our databases for generated knowledge, continuously, to support their everyday decisions"*. Thus, applying reuse as well as the formation of already generated knowledge can increase BI&A insights value, which in turn could enhance firm performance.

4.3. Underpinning assets

We identified technological, human and relationship assets to form the basis for absorptive capacity capabilities. In the following subsections, we elaborate on each in detail.

4.3.1. Technological assets

Most of the interviewees highlighted the importance of advancement in technology to maintain changing data requirements. Thus, a major shift from transactional to behavioral data pushes firms to upgrade their technological assets, regarding infrastructures, architectures, and programs. Interviewees reported using different BI&A assets, such as data and text mining, regression, OLAP, search engines, multivariate analysis, process, and network mining, cloud computing, parallel programming, opinion mining, sentiment analysis, visualization, social media analysis, and natural language processing. Accordingly, Interviewee 2 discussed the following: *"We were forced to go beyond traditional, relational databases to fit the requirements of new, unstructured, massive datasets. Thus, we adopted the MapReduce parallel computing tool and Hadoop database technology, which allows us to integrate new, external data with internal data."* Despite the excitement about the possibilities of advanced programs, infrastructures, and architecture offer, most of the interviewees complained about drawbacks, like poor performance, scalability and long response times. Moreover, the needed learning processes needed to achieve skillfulness are both time and finance-consuming. Hence, Interviewee 10 remarked: *"One of the greatest obstacle related with advancing usage of BI&A is the time constraint. The employees are already overloaded. Thus, successful usage of the BI&A solution requires additional financial investments"*.

Nonetheless, interviewees pointed out the role of BI&A assets in facilitating internalization of acquired knowledge. With the increasing storage costs, however, firms decreased their appetites towards valueless data retaining. Only valuable knowledge is saved in joint repositories, which make it available for future knowledge discovery. Thus, the technological advancement offered powerful visualization techniques for knowledge discovery. As one IT manager noted: *"Our managers often have difficulties in understanding complex data. For that reason, we try to present results in the form of interactive visualizations or graphs, and then, together, analyze details and potential applications"* (Interviewee 6). Similarly, Interviewee 12 commented, "*We can hardly force continuous usage of the generated information if not presented in a synthe sized way in the form of easy-understandable graphs and visualization, complemented with recommendations and specifications, where could be this information relevantly used."* Although there is a noticeable advance in the visualization approach, they still have been found by our informants to be scarce and time-demanding.

4.3.2. Human assets

Although much of the current enthusiasm refers to technological assets, human assets have begun to be emphasized as a critical milestone in succeeding with BI&A. While not neglecting important technological breakthroughs, informants have stressed the role of the human factor in making sense and use of data and insights. Ideally, these firms need multidisciplinary data scientists that own a combination of data, analytics and business knowledge which would allow them to communicate with, and understand, the broader business environment. However, most of the IT that professional firms have employed are trained in Computer Sciences, Statistics, and Mathematics, lacking overall business knowledge and struggling to interpret data for a firm's performance enhancement. According to Interviewee 4: *"It is extremely hard to find a suitable workforce that has considerable expertise in both analytics and business issues. Usually, they come with strong data and a computational focus."* Moreover, as Interviewee 13 commented "*We don't even look for data scientists that have advanced business knowledge. Although they are welcomed, we are at the first place looking for IT professionals, that can use a "common sense" and are good team players. Then, we create teams consisted of different professionals to work together on a particular project"*. Therefore, firms have reported recruiting professionals with strong technical and analytical skills to model, analyze and manipulate data; then, organizing them into teams with business managers, where IT expertise is combined with deep domain knowledge for collaborative data exploration.

Additionally, interviews reported that data analysis shortage seriously constrains the possibility of insight generation. As is exemplified by Interviewee 7: *"We can find IT professionals, however, not all of them are equipped with the needed technical, data and analytical skills necessary to exploit the technology fully. Thereby, we started a collaboration with universities to develop an educational curriculum that would address this labor issue."* Still, recruiting technically and analytically sound data scientists remains to be a large challenge.

4.3.3. Relationship assets

Although they had been using BI&A for some time, interviewees mentioned some organizational factors that notably influence successful use. Firstly, it was mostly agreed upon that strong decision-making culture could significantly impact on creating a competitive advantage with analytics: *"Our higher-level management is often reluctant to use BI&A to support their actions and decisions. Some of them still believe their experience and intuition are the most secure source of knowledge when deriving strategic decisions. Unfortunately, this impacts on lower level management, leading them to be skeptical about the advantages of utilizing this technology"* (Interviewee 4). Similarly, Interviewee 11 remarked *"We still rely to a great extent on intuition. It is difficult to convince the decision-makers that mixing both is beneficial"*. Hence, Interviewee 12 added "*I suppose I should not talk like this, regarding the fact I am a data scientist, but we believe data triggered insights just complement intuition. Prior related knowledge and experience are very important in making the correct decision."*

Unsurprisingly, many interviewees emphasized the effort it takes to build strong, cultural commitments while incorporating BI&A into day-to-day activities. As Interviewee 7 noted: *"We started using BI&A without considering the level of commitment it requires to be successful. Culture became a greater obstacle than the technology itself. It was a long process to make the technology trustworthy to our employees."* Hence, as Interviewee 13 commented *"Cultural commitments could be built only if you prove your employees that data provide added value"*.

Many firms prompted their employees to collaborate cross-divisionally, to compensate a potential lack of skills and capabilities. Moreover, reliable, information-centralized knowledge bases have been found important; since it allows further knowledge dissemination, transformation and exploitation, especially when data-mindset is a prevalent cultural pattern. This, however, requires aligning an existing, overall firm strategy with the contemplated data strategy.

5. Discussion

The increasing prevalence of BI&A research impacted scholarly attention to understanding the mechanism trough witch BI&A use creates value. We add to this line of inquiry by examining the issue of how BI&A triggered insights are transformed into valuable knowledge. The extant literature on the BI&A value creation highlights the process mostly

regarding improved decision-making that could drive business performance (Chen et al., 2015; Fan et al., 2015; Sharma et al., 2014; Wieder & Ossimitz, 2015). Here, scholars have relied on the presumption that BI&A uncover useful information that is used by decision-makers across different business levels to make better, and more informed decisions. Rather than viewing the technology as a 'passive container,' which produces knowledge that is ultimately used in a decision-making process, our analysis exhibits the absorptive capacity to underlie the process of raw data transformation into valuable knowledge for action-taking and decision-making. Instead of positioning BI&A exclusively as a decision-supporting tool, our analysis underlines that firms should develop higher-order dynamic capabilities to allow continuous acquisition, search, and management of knowledge. This is in-line with some recent works (Fink et al., 2017; Shollo & Galliers, 2016) that warn a limited understanding of the concept when overlooking the role of BI&A in organizational knowing, but extending by showing how different BI&A resources and lower-order knowledge capabilities are integrated and reconfigured by the higher-order dynamic capability of absorptive capacity for knowledge creation.

5.1. Theoretical contributions

This study offers several theoretical contributions to the understanding of the BI&A value creation process. Our study suggests that might be insufficient to focus on the improved-decision making, without considering how knowledge creation occurred in the first place. We contribute to this research vein by focusing specifically on the mechanism trough witch different knowledge creation capabilities interplay with organizational resources to create useful organizational knowledge. Thus, this article offers few implications for research on business intelligence, knowledge management, and dynamic capabilities. First, it integrates prior research on BI&A use and absorptive capacity by specifying the underlying, first-order capabilities of absorptive capacity in the context of BI&A. Beyond that, our paper emphasizes the importance of the underpinning technological, human, and relational assets, while specifying the role of absorptive capacity as a second-order dynamic capability that builds, integrates, and reconfigures the underlying capabilities of acquisition, assimilation, transformation, and exploitation and the underpinning assets. As such, this study echoes the call by Gao et al. (2017) to establish the importance of the absorptive capacity in the BI&A domain, while considering the call by Trieu (2017) for considering a firm's factors when investigating the relationship between BI&A assets and BI&A impacts. With this integrated perspective, scholars might have better awareness of the BI&A value creation process.

Although technological appropriateness of the BI&A has been widely argued to be an essential catalyst of the successful BI&A use (Chaudhuri et al., 2011; Chen et al., 2012; Watson & Wixom, 2007), the importance of human assets that underpin the knowledge creation processes has only recently started to be investigated (McAfee et al., 2012; Ransbotham, Kiron, & Prentice, 2016). Our research extends this stream of thinking by identifying human assets as crucial for successfully delivering value from BI&A use. Namely, our informants emphasized the importance of having personnel equipped with both, domain and data knowledge, so pattern identification and insight discovery are possible. Since the value of the information, contained in some data, depends mainly on the intended application and the contextualization (Popovič, Hackney, Tassabehji, & Castelli, 2016), firms must set in place strong human assets, equipped with domain-specific knowledge (Wixom, Yen, & Relich, 2013) that are able to ask relevant business question. Considering the importance of technological assets, the empirical findings have emphasized the role of human assets in making sense and use of data, since the technology itself is outpacing the ability of the firm to deploy technology effectively. In line with some recent research (Davenport & Patil, 2012) we found important for firms to allow close collaboration of IT and management personnel to cope with the shortage of skills successfully. Moreover, since learning is a cumulative process (Cohen & Levinthal, 1990), richness and relevance of prior, related knowledge will allow better knowledge assimilation. The interpretation of externally acquired information is possible when the "modern" gatekeepers are equipped with multidisciplinary knowledge and skills; which allows them to estimate and understand information for a potential benefit (Altman, Nagle, & Tushman, 2014; Staggers & Nelson, 2015).

We complement this research inquiry by showing how technological and relational assets underpin the first-order capabilities of knowledge creation, something that prior research has considered in isolation. Consistent with some prior research, our findings suggest that the inadequate, complex presentation of the BI&A triggered insights might jeopardize the use of information (McAfee et al., 2012). Hence, the technological assets should allow a presentation of newly generated knowledge in formats that are more palatable (graphics, dashboards, visualizations), so that it can be easily comprehended and accessed by less data skilled personnel. Thus, the study turns attention to the potential drawbacks of BI&A use, regarding poor, technological performance, scalability, long response times, high storage costs, labor shortage, and long learning processes, which ultimately lead to reluctance in BI&A use.

Nonetheless, our findings indicate that the commercialization of transformed knowledge through the exploitation capability requires continuous search and reuse of generated knowledge, which further allows improvement of different business processes, improvements of the development of products and services, the understanding of customer and market behavior and managing risks, etc. We have found that the relationship assets significantly influence the realized absorptive capacity capabilities (transformation, exploitation). Moreover, our findings revealed skepticism about the advantages of BI&A amongst higher management levels, leading to some hesitation to incorporate BI&A triggered insights into decision-making or action-taking processes. Therefore, companies need to invest in cultural changes to achieve a decision-making culture which blends the analytics' insight with a managers' intuition that would produce better, effective results than each could individually. Although previous research has also discussed how overturning intuition and consequential management could limit the potential value of BI&A for firms (Bronzo et al., 2013; Fallik, 2014; Ransbotham et al., 2016; Trelles, Prins, Snir, & Jansen, 2011), we extend this research vein by assessing the impact on the first-order knowledge capabilities of absorptive capacity. Therefore, this study highlights the need for aligning existing firm with the considered data strategy, while developing a strong data culture, tolerable for mistakes.

5.2. Implications for practice

In addition to theoretical contributions, this study suggests several important implications for practicing managers. First and foremost, our study emphasized the crucial role of absorptive capacity in building, integrating, and reconfiguring assets and first-order knowledge capabilities for knowledge creation from BI&A. Our findings pointed out that the value of the information in the first place comes from the intended application. Hence, firms should avoid irrelevant business questions, which are possible only if sufficient domain-specific knowledge and IT expertise are set in place. Organizations should provide systematic training and education to develop data-savvy personnel or create teams of data scientists and business professionals that could together translate the results of a complex model into simple information to digest.

Extending this discourse, we highlighted the importance necessary to align existing firm culture with the required capabilities. Our findings suggest that a firm needs strong cultural commitments and symbiotic data and strategies to eliminate organizational barriers for delivering BI&A value. A continuous dialogue between human intuition and analytic statistics will allow better decision-making, based on real-time

evidence. However, along with the openness to new ideas from data analytics, tolerance for mistakes must be present, since people cannot know which results would work out (Ransbotham et al., 2016). Failure to align the required capabilities, assets and culture could lead to defective decision-making (Erevelles, Fukawa, & Swayne, 2016; Jaklič, Grublješič, & Popovič, 2018; Matzler, Bailom, & Mooradian, 2007). Thus, all management levels must be aware of the ability of BI&A to provide more holistic and accurate market intelligence, which requires continuous organizational effort.

Nevertheless, our study suggests that consistent BI&A use in day-to-day activities, as well as decisions, are only possible if the technology is trustworthy. Firms need to upgrade existing BI&A infrastructures, architectures, and software to fit the data changing requirements. Our findings emphasized the importance of a presentation of information in forms of interactive visualizations and graphs, which further reduce the effort it takes to interpret and manage new insights. However, poor performance, scalability, long response times, and high-costs could be an important obstacle that leads to potential BI&A underuse, limiting the potential in BI&A value.

5.3. Limitations and outlook

A few limitations of this study are worth noting. First, the empirical data was collected from a sample of nine medium and large-sized firms from European countries. Although we believe that the analysis has provided insights that are valuable in context with small-sized firms, we cannot make claims that small businesses, within the often-limited market of technological knowledge, could benefit from BI&A at the same level as larger firms. Moreover, since we have selected European firms only, we could not observe how the BI&A value creation process would vary across different cultural contexts. Our focus on these selected firms, from eight high-knowledge, intensive sectors used to conduct the analysis, could also be seen as a limitation. An in-depth analysis across other, less knowledge-intensive industries may reveal additional insights. Accordingly, we encourage future studies to investigate the similarities and differences in context with this study regarding a firm's size, cultural context, and industry. Next, future research could test the theory and draw causal inferences in quantitative research design to complement the findings we have outlined here. Finally, even though we have carefully and thoroughly studied and taken notes, we are aware that notes taken do not provide a complete verbal record (Muswazi & Nhamo, 2013). Therefore, the note-taking made may have affected the accuracy to reconstruct what the interviewees have said.

5.4. Conclusion

BI&A has been often promoted to deliver competitive value gains. The findings of the present study shed light on how knowledge is created from BI&A triggered insights. Applying the absorptive capacity's theoretical lens, we explain the interplay of the absorptive capacity's underlying capabilities with the underpinning assets, providing a theoretical explanation of the process of delivering value regarding knowledge creation. Hence, we establish the importance of absorptive capacity in the BI&A domain while considering the impact of BI&A assets, providing an important basis for future research on the BI&A value creation process.

Note

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Interview questions and protocol

1. Background information

- Information about the firm
- Information about the interviewee's position, experience in the industry and the firm, and major responsibilities.

2. Brief introduction to the research project: We are trying to get a sense of how business intelligence and analytics use process results in insight generation, and hence in value creation.

Questions were as follows:

a What is your understanding of business intelligence and analytics? How would you describe it in your words?
b To ensure a common understanding of the term, we suggest the following theory-based definition: "Business intelligence and analytics (BI&A) refer to the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions" (Chen et al., 2012, p. 1166). How do you agree? Which BI&A techniques you widely use in your organization?
c How does BI&A use results in insight generation in your organization?
d How do you use BI&A generated insights?
e What are the specific technological requirements to gather and process data into valuable knowledge?
f What human skills requirements need to be met for BI&A facilitated knowledge generation?
g What organizational factors influence the value creation process?
h What are the main problems that you have witnessed or heard about?
i Thinking back over your remarks-Are there any other issues that we have not discussed and that you find worrisome? Anything else of importance you want to add?

References

Acharya, A., Singh, S. K., Pereira, V., & Singh, P. (2018). Big data, knowledge co-creation and decision making in fashion industry. *International Journal of Information Management, 42*, 90–101. https://doi.org/10.1016/j.ijinfomgt.2018.06.008.

Altman, E. J., Nagle, F., & Tushman, M. (2014). *Innovating without information constraints: Organizations, communities, and innovation when information costs approach zero.* Harvard Business School Organizational Behavior Unit Working Paper (14-043).

Alvesson, M. (2003). Beyond neopositivists, romantics, and localists: A reflexive approach to interviews in organizational research. *Academy of management review, 28*(1), 13–33.

Argote, L., McEvily, B., & Reagans, R. (2003). Managing knowledge in organizations: An integrative framework and review of emerging themes. *Management Science, 49*(4), 571–582.

Arora, A., & Gambardella, A. (1994). The changing technology of technological change: General and abstract knowledge and the division of innovative labour. *Research Policy, 23*(5), 523–532.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99–120.

Benner, M. J., & Tushman, M. L. (2015). Reflections on the 2013 decade award—"Exploitation, exploration, and process management: The productivity dilemma revisited" ten years later. *The Academy of Management Review, 40*(4), 497–514.

Bharadwaj, A. S. (2000). A resource-based perspective on information technology capability and firm performance: An empirical investigation. *MIS Quarterly,* 169–196.

Bose, R. (2009). Advanced analytics: Opportunities and challenges. *Industrial Management & Data Systems, 109*(2), 155–172.

Boyce, C., & Neale, P. (2006). *Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input.* Pathfinder International.

Brinkmann, S. (2014). *Interview encyclopedia of critical psychology.* Springer1008–1010.

Bronzo, M., de Resende, P. T. V., de Oliveira, M. P. V., McCormack, K. P., de Sousa, P. R., & Ferreira, R. L. (2013). Improving performance aligning business analytics with process orientation. *International Journal of Information Management, 33*(2), 300–307. https://doi.org/10.1016/j.ijinfomgt.2012.11.011.

Brown, S. L., & Eisenhardt, K. M. (1995). Product development: Past research, present findings, and future directions. *The Academy of Management Review, 20*(2), 343–378.

Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *The Journal of Economic Perspectives, 14*(4), 23–48.

Cao, G., Duan, Y., & Cadden, T. (2019). The link between information processing capability and competitive advantage mediated through decision-making effectiveness. *International Journal of Information Management, 44*, 121–131. https://doi.org/10.

1016/j.ijinfomgt.2018.10.003.
Cao, G., Duan, Y., & Li, G. (2015). Linking business analytics to decision making effectiveness: A path model analysis. *IEEE Transactions on Engineering Management, 62*(3), 384–395.
Carlsson, S. A. (2003). Knowledge managing and knowledge management systems in inter-organizational networks. *Knowledge and Process Management, 10*(3), 194–206.
Castro, G. M.-d., Delgado-Verde, M., Amores-Salvadó, J., & Navas-López, J. E. (2013). Linking human, technological, and relational assets to technological innovation: Exploring a new approach. *Knowledge Management Research & Practice, 11*(2), 123–132.
Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative research.* London: SagePublications Ltd.
Chatterjee, D., Grewal, R., & Sambamurthy, V. (2002). Shaping up for e-commerce: Institutional enablers of the organizational assimilation of web technologies. *MIS Quarterly,* 65–89.
Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM, 54*(8), 88–98.
Chen, D. Q., Preston, D. S., & Swink, M. (2015). How the use of big data analytics affects value creation in supply chain management. *Journal of Management Information Systems, 32*(4), 4–39.
Chen, H. C., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.
Clark, T. D., Jones, M. C., & Armstrong, C. P. (2007). The dynamic structure of management support systems: Theory development, research focus, and direction. *MIS Quarterly, 31*(3), 579–615.
Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly,* 128–152. https://doi.org/10.2307/2393553.
Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 13*(1), 3–21.
Côrte-Real, N., Oliveira, T., & Ruivo, P. (2017). Assessing business value of Big Data Analytics in European firms. *Journal of Business Research, 70*, 379–390.
Cosic, R., Shanks, G., & Maynard, S. B. (2015). A business analytics capability framework. *Australasian Journal of Information Systems, 19*.
Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory Into Practice, 39*(3), 124–130.
Crouch, M., & McKenzie, H. (2006). The logic of small samples in interview-based qualitative research. *Social Science Information, 45*(4), 483–499.
Danneels, E. (2002). The dynamics of product innovation and firm competences. *Strategic Management Journal, 23*(12), 1095–1121. https://doi.org/10.1002/smj.275.
Davenport, T., & Patil, D. (2012). Data scientist. *Harvard Business Review, 90*, 70–76.
Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning.* Harvard Business Press.
Denzin, N. K., & Lincoln, Y. S. (2005). *Introduction: The discipline and practice of qualitative research the sage handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage Publications Ltd1–32.
Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review, 14*(4), 532–550.
Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. *Academy of Management Journal, 50*(1), 25–32.
Elbashir, M. Z., Collier, P. A., & Davern, M. J. (2008). Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. *International Journal of Accounting Information Systems, 9*(3), 135–153.
Elbashir, M. Z., Collier, P. A., & Sutton, S. G. (2011). The role of organizational absorptive capacity in strategic use of business intelligence to support integrated management control systems. *The Accounting Review, 86*(1), 155–184.
Elbashir, M. Z., Collier, P. A., Sutton, S. G., Davern, M. J., & Leech, S. A. (2013). Enhancing the business value of business intelligence: The role of shared knowledge and assimilation. *Journal of Information Systems, 27*(2), 87–105.
Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research, 69*(2), 897–904.
Fallik, D. (2014). For big data, big questions remain. *Health Affairs, 33*(7), 1111–1114. https://doi.org/10.1377/hlthaff.2014.0522.
Fan, S., Lau, R. Y., & Zhao, J. L. (2015). Demystifying big data analytics for business intelligence through the lens of marketing mix. *Big Data Research, 2*(1), 28–32.
Fink, L., Yogev, N., & Even, A. (2017). Business intelligence and organizational learning: An empirical investigation of value creation processes. *Information & Management, 54*(1), 38–56. https://doi.org/10.1016/j.im.2016.03.009.
Flatten, T. C., Engelen, A., Zahra, S. A., & Brettel, M. (2011). A measure of absorptive capacity: Scale development and validation. *European Management Journal, 29*(2), 98–116.
Forsgren, N., & Sabherwal, R. (2015). *Business intelligence system use as levers of control and organizational capabilities: Effects on internal and competitive benefits.* https://doi.org/10.2139/ssrn.2687710.
Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.
Gao, S., Yeoh, W., Wong, S. F., & Scheepers, R. (2017). A literature analysis of the use of Absorptive Capacity construct in IS research. *International Journal of Information Management, 37*(2), 36–42.
Gartner, I. (2013). *Gartner predicts business intelligence and analytics will remain top focus for CIOs through 2017.* [Press release]. Retrieved fromhttps://www.gartner.com/newsroom/id/2637615.
Glasser, B., & Strauss, A. (1967). *The development of grounded theory.* Chicago, IL: Alden.
Grant, R. M. (1991). The resource-based theory of competitive advantage: Implications for strategy formulation. *California Management Review, 33*(3), 114–135.
Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal, 17*(S2), 109–122.
Hannula, M., & Pirttimaki, V. (2003). Business intelligence empirical study on the top 50 Finnish companies. *Journal of American Academy of Business, 2*(2), 593–599.
Holstein, J. A., & Gubrium, J. F. (1995). *The active interview, Vol. 37.* Sage Publications.
Hou, C.-K. (2012). Examining the effect of user satisfaction on system usage and individual performance with business intelligence systems: An empirical study of Taiwan's electronics industry. *International Journal of Information Management, 32*(6), 560–573.
Ireland, R. D., Hitt, M. A., & Vaidyanath, D. (2002). Alliance management as a source of competitive advantage. *Journal of Management, 28*(3), 413–446.
IşıK, Ö., Jones, M. C., & Sidorova, A. (2013). Business intelligence success: The roles of BI capabilities and decision environments. *Information & Management, 50*(1), 13–23.
Jaklič, J., Grublješič, T., & Popovič, A. (2018). The role of compatibility in predicting business intelligence and analytics use intentions. *International Journal of Information Management, 43*, 305–318. https://doi.org/10.1016/j.ijinfomgt.2018.08.017.
Jansen, J. J., Van Den Bosch, F. A., & Volberda, H. W. (2005). Managing potential and realized absorptive capacity: how do organizational antecedents matter? *Academy of Management Journal, 48*(6), 999–1015.
Kimball, R., & Ross, M. (2011). *The data warehouse toolkit: The complete guide to dimensional modeling.* John Wiley & Sons.
Kiron, D., & Shockley, R. (2011). Creating business value with analytics. *MIT Sloan Management Review, 53*(1), 57.
Kogut, B., & Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science, 3*(3), 383–397.
Kowalczyk, M., & Buxmann, P. (2014). Big Data and information processing in organizational decision processes. *Business & Information Systems Engineering, 6*(5), 267–278.
Kowalczyk, M., Buxmann, P., & Besier, J. (2013). *Investigating business intelligence and analytics from a decision process perspective: A structured literature review: Darmstadt Technical University, Department of Business Administration, Economics and Law.* Institute for Business Studies (BWL).
Lane, P. J., Koka, B. R., & Pathak, S. (2006). The reification of absorptive capacity: A critical review and rejuvenation of the construct. *Academy of management Review, 31*(4), 833–863. https://doi.org/10.5465/amr.2006.22527456.
Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management, 36*(5), 700–710.
Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), 21–32.
Lavie, D., & Rosenkopf, L. (2006). Balancing exploration and exploitation in alliance formation. *Academy of Management Journal, 49*(4), 797–818.
Lee, H., & Choi, B. (2003). Knowledge management enablers, processes, and organizational performance: An integrative view and empirical examination. *Journal of Management Information Systems, 20*(1), 179–228.
Levinthal, D. A., & March, J. G. (1993). The myopia of learning. *Strategic Management Journal, 14*(S2), 95–112.
Lichtenthaler, U. (2009). Absorptive capacity, environmental turbulence, and the complementarity of organizational learning processes. *Academy of Management Journal, 52*(4), 822–846.
Lim, A. H., & Lee, C.-S. (2010). Processing online analytics with classification and association rule mining. *Knowledge-Based Systems, 23*(3), 248–255.
Lin, Y., & Wu, L.-Y. (2014). Exploring the role of dynamic capabilities in firm performance under the resource-based view framework. *Journal of Business Research, 67*(3), 407–413.
Locke, K. (2002). *The grounded theory approach to qualitative research measuring and analyzing behavior in organizations: Advances in measurement and data analysis.* San Francisco, CA, US: Jossey-Bass17–43.
Lonnqvist, A., & Pirttimaki, V. (2006). The measurement of business intelligence. *Information Systems Management, 23*(1), 32–40. https://doi.org/10.1201/1078.10580530/45769.23.1.20061201/91770.4.
Lord, M. D., & Ranft, A. L. (2000). Organizational learning about new international markets: Exploring the internal transfer of local market knowledge. *Journal of International Business Studies,* 573–589.
Luhn, H. P. (1958). A business intelligence system. *IBM Journal of Research and Development, 2*(4), 314–319.
Lund Vinding, A. (2006). Absorptive capacity and innovative performance: A human capital approach. *Economics of Innovation and New Technology, 15*(4–5), 507–517.
Mangematin, V., & Nesta, L. (1999). What kind of knowledge can a firm absorb? *International Journal of Technology Management, 18*(3-4), 149–172.
Mantere, S., & Ketokivi, M. (2013). Reasoning in organization science. *Academy of Management Review, 38*(1), 70–89.
March, J. G. (1978). Bounded rationality, ambiguity, and the engineering of choice. *The Bell Journal of Economics,* 587–608.
Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013). Does sample size matter in qualitative research?: A review of qualitative interviews in IS research. *Journal of Computer Information Systems, 54*(1), 11–22.
Matzler, K., Bailom, F., & Mooradian, T. A. (2007). Intuitive decision making. *MIT Sloan Management Review, 49*(1), 13.
McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: The management revolution. *Harvard Business Review, 90*(10), 60–68.
Melville, N., Kraemer, K., & Gurbaxani, V. (2004). Review: Information technology and organizational performance: An integrative model of IT business value. *MIS Quarterly, 28*(2), 283–322.
Mithas, S., Lee, M. R., Earley, S., Murugesan, S., & Djavanshir, R. (2013). Leveraging big data and Business Analytics [guest editors' introduction]. *IT Professional, 15*(6), 18–20.

Muswazi, M., & Nhamo, E. (2013). Note taking: A lesson for novice qualitative researchers. *Journal of Research & Method in Education, 2*(3), 13–17.
Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change.* Cambridge, MA: Harvard University Press.
Nonaka, I. (2007). The knowledge-creating company. *Harvard Business Review, 85*(7–8), 162.
Pare, G. (2004). Investigating information systems with positivist case research. *The Communications of the Association for Information Systems, 13*(1), 57.
Pavlou, P. A., & El Sawy, O. A. (2010). The "Third hand": IT-Enabled competitive advantage in turbulence through improvisational capabilities. *Information Systems Research, 21*(3), 443–471. https://doi.org/10.1287/isre.1100.0280.
Pawlowski, S. D., & Robey, D. (2004). Bridging user organizations: Knowledge brokering and the work of information technology professionals. *MIS Quarterly,* 645–672.
Popovič, A., Hackney, R., Tassabehji, R., & Castelli, M. (2016). The impact of big data analytics on firms' high value business performance. *Information Systems Frontiers,* 1–14.
Popovič, A., Turk, T., & Jaklič, J. (2010). Conceptual model of business value of business intelligence systems. *Management: Journal of Contemporary Management Issues, 15*(1), 5–30.
Raisch, S., & Birkinshaw, J. (2008). Organizational ambidexterity: Antecedents, outcomes, and moderators. *Journal of Management, 34*(3), 375–409. https://doi.org/10.1177/0149206308316058.
Ramamurthy, K. R., Sen, A., & Sinha, A. P. (2008). An empirical investigation of the key determinants of data warehouse adoption. *Decision Support Systems, 44*(4), 817–841.
Ranjan, J. (2009). Business intelligence: Concepts, components, techniques and benefits. *Journal of Theoretical and Applied Information Technology, 9*(1), 60–70.
Ransbotham, S., Kiron, D., & Prentice, P. K. (2016). Beyond the hype: The hard work behind analytics success. *MIT Sloan Management Review, 57*(3).
Ravichandran, T., & Lertwongsatien, C. (2005). Effect of information systems resources and capabilities on firm performance: A resource-based perspective. *Journal of Management Information Systems, 21*(4), 237–276.
Real, J. C., Roldán, J. L., & Leal, A. (2014). From entrepreneurial orientation and learning orientation to business performance: Analysing the mediating role of organizational learning and the moderating effects of organizational size. *British Journal of Management, 25*(2), 186–208.
Roberts, N., Galluch, P. S., Dinger, M., & Grover, V. (2012). Absorptive capacity and information systems research: Review, synthesis, and directions for future research. *MIS Quarterly, 36*(2), 625–648.
Rosenkopf, L., & Nerkar, A. (2001). Beyond local search: Boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal, 22*(4), 287–306.
Ross, J. W., Beath, C. M., & Goodhue, D. L. (1996). Develop long-term competitiveness through IT assets. *Sloan Management Review, 38*(1), 31.
Ross, J. W., Beath, C. M., & Quaadgras, A. (2013). You may not need big data after all. *Harvard Business Review, 91*(12), 90.
Rothaermel, F. T., & Alexandre, M. T. (2009). Ambidexterity in technology sourcing: The moderating role of absorptive capacity. *Organization Science, 20*(4), 759–780.
Ryu, H.-S., & Lee, J.-N. (2013). *Effect of IT capability on the alignment between business and service innovation strategies. Paper Presented at the PACIS*.
Sammut, G., & Sartawi, M. (2012). Perspective-taking and the attribution of ignorance. *Journal for the Theory of Social Behaviour, 42*(2), 181–200.
Santhanam, R., & Hartono, E. (2003). Issues in linking information technology capability to firm performance. *MIS Quarterly,* 125–153.
Schwandt, T. A., Lincoln, Y. S., & Guba, E. G. (2007). Judging interpretations: But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Evaluation, 2007*(114), 11–25.
Sharma, R., Mithas, S., & Kankanhalli, A. (2014). Transforming decision-making processes: A research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems, 23*(4), 433–441.
Shehzad, R., Khan, M., & Naeem, M. (2013). Integrating knowledge management with business intelligence processes for enhanced organizational learning. *International Journal of Software Engineering and Its Applications, 7*(2), 83–91.
Sheremata, W. A. (2000). Centrifugal and centripetal forces in radical new product development under time pressure. *Academy of Management Review, 25*(2), 389–408.
Shollo, A., & Galliers, R. D. (2016). Towards an understanding of the role of business intelligence systems in organisational knowing. *Information Systems Journal, 26*(4), 339–367. https://doi.org/10.1111/isj.12071.
Shuradze, G., & Wagner, H.-T. (2016). *Towards a conceptualization of data analytics capabilities. Paper Presented at the 2016 49th Hawaii International Conference on System Sciences (HICSS)*.
Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization Science, 2*(1), 125–134.
Simsek, Z. (2009). Organizational ambidexterity: Towards a multilevel understanding. *Journal of Management Studies, 46*(4), 597–624.
Spall, S. (1998). Peer debriefing in qualitative research: Emerging operational models. *Qualitative Inquiry, 4*(2), 280–292.
Staggers, N., & Nelson, R. (2015). *Data, information, knowledge, wisdom Routledge international handbook of advanced quantitative methods in nursing research.* Routledge.
Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Procedures and techniques for developing grounded theory.* Thousand Oaks, CA: Sage.
Teece, D. J. (1998). Capturing value from knowledge assets: The new economy, markets for know-how, and intangible assets. *California Management Review, 40*, 55–79.
Todorova, G., & Durisin, B. (2007). Absorptive capacity: Valuing a reconceptualization. *Academy of Management Review, 32*(3), 774–786.
Trelles, O., Prins, P., Snir, M., & Jansen, R. C. (2011). Big data, but are we ready? *Nature Reviews Genetics, 12*(3) 224-224.
Trieu, V.-H. (2017). Getting value from Business Intelligence systems: A review and research agenda. *Decision Support Systems, 93*, 111–124. https://doi.org/10.1016/j.dss.2016.09.019.
Trkman, P., McCormack, K., De Oliveira, M. P. V., & Ladeira, M. B. (2010). The impact of business analytics on supply chain performance. *Decision Support Systems, 49*(3), 318–327.
Urquhart, C., & Fernandez, W. (2016). *Using grounded Theory method in information systems: The researcher as blank slate and other myths enacting research methods in information systems, Vol. 1*, Springer129–156.
Verona, G., & Ravasi, D. (2003). Unbundling dynamic capabilities: An exploratory study of continuous product innovation. *Industrial and Corporate Change, 12*(3), 577–606.
Vidgen, R., Shaw, S., & Grant, D. B. (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research, 261*(2), 626–639.
Wade, M., & Hulland, J. (2004). Review: The resource-based view and information systems research: Review, extension, and suggestions for future research. *MIS Quarterly, 28*(1), 107–142.
Wang, C. L., & Ahmed, P. K. (2007). Dynamic capabilities: A review and research agenda. *International Journal of Management Reviews, 9*(1), 31–51.
Wang, H.-C. (2014). Distinguishing the adoption of business intelligence systems from their implementation: The role of managers' personality profiles. *Behaviour & Information Technology, 33*(10), 1082–1092.
Watson, H. J. (2009). Tutorial: Business intelligence-Past, present, and future. *Communications of the Association for Information Systems, 25*(1), 39.
Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer, 40*(9), 96–99.
Wieder, B., & Ossimitz, M.-L. (2015). The impact of Business Intelligence on the quality of decision making–a mediation model. *Procedia Computer Science, 64*, 1163–1171.
Wieneke, A., & Lehrer, C. (2016). Generating and exploiting customer insights from social media data. *Electronic Markets,* 1–24.
Wixom, B. H., Watson, H. J., & Werner, T. (2011). Developing an enterprise business intelligence capability: The norfolk southern journey. *MIS Quarterly Executive, 10*(2).
Wixom, B. H., Yen, B., & Relich, M. (2013). Maximizing value from business analytics. *MIS Quarterly Executive, 12*(2).
Yeoh, W., & Koronios, A. (2010). Critical success factors for business intelligence systems. *Journal of Computer Information Systems, 50*(3), 23–32.
Yeoh, W., & Popovič, A. (2016). Extending the understanding of critical success factors for implementing business intelligence systems. *Journal of the Association for Information Science and Technology, 67*(1), 134–147.
Zahra, S. A., & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. *Academy of Management Review, 27*(2), 185–203.

International Journal of Medical Informatics 116 (2018) 1–8

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Interactive data visualization based on conventional statistical findings for antihypertensive prescriptions using National Health Insurance claims data

Inseok Ko[a], Hyejung Chang[b,*]

[a] Division of Biomedical Informatics, Seoul National University College of Medicine, Seoul, South Korea
[b] Department of Management, Kyung Hee University School of Management, Seoul, South Korea

ARTICLE INFO

Keywords:
Prescriptions
National Health Insurance claims database
Hypertension
Interactive visualization
Data mining
Decision making
Software
Drug prescriptions
Comorbidity
Antihypertensive agents

ABSTRACT

Background: Interactive visualization is an important approach to help to understand and to explain large amounts of data, particularly in light of decision support. Although data visualization have been introduced in healthcare and clinical fields, analytics has often been performed by data experts, focused on specific subjects, or insufficient statistical evidence. Therefore, this study suggests the procedures of effective and efficient visualization of big data for general healthcare researchers. Specifically, the procedure includes conventional regression analyses followed by interactive data visualization for prescription patterns of antihypertensive drugs.
Methods: As a large-scale nationally representative prescription data, the Korean National Health Insurance claims data were collected. Conventional descriptive and regression analyses were conducted for therapy decision and prescription patterns using the software R. Then, based on the statistically significant findings, dashboards were developed to visualize interactively the patterns of prescriptions using the software Tableau.
Results: Major characteristics (genders, age groups, healthcare institutions, and comorbidities) explained the differences in therapy and the average number of drugs prescribed as well as differences among most commonly prescribed drug classes. Two interactive dashboards were created for visualizing prescription patterns with incorporation of horizontal bar charts, packed bubble charts, treemaps, filled maps, radar charts, box and whisker plots, and filters.
Conclusion: In the current big data era, interactive data visualization offers substantial opportunities to have comprehensive view, extract insights and evidence from the flood of vast amounts of data. This study's interactive visualizations can provide healthcare professionals insight into prescription patterns and demonstrate the value of creating interactive dashboards to support informed and timely decision-making. Exploring big data using interactive visualization is expected to deliver many future benefits in healthcare fields.

What was known before the study

- Healthcare big data have an important role to provide information and knowledge. Understanding and utilizing the big data has the potential to improve care and lower costs by discovering associations and understanding patterns within the data.
- Although data visualization have been introduced in healthcare and clinical fields, analytics has often been performed by data experts, focused on specific subjects, or insufficient statistical evidence.
- Antihypertensive drug prescription patterns and trends have been changed according to time and analyzed with insufficiently explorative in terms of all integration points in Korea.

What the study has added to the body of knowledge

- This study suggested the procedures of effective and efficient visualization of big data for general healthcare researchers. Healthcare professionals can make interactive visualizations with easy and fast interfaces using Tableau.
- As a two-step approach, the analysis procedure includes conventional regression analyses followed by interactive data visualization for a variety of big data in health care fields.
- This study's interactive visualizations can provide healthcare professionals insight into prescription patterns and demonstrate the value of creating interactive dashboards to support informed and prompt decision-making.
- Interactive visualizations of healthcare big data enable healthcare professionals to reason and make sense of vast and complex claims

* Corresponding author at: Department of Management, Kyung Hee University School of Management, 26, Kyungheedae-ro, Dongdaemun-gu, Seoul, 02447, South Korea.
E-mail address: hjchang@khu.ac.kr (H. Chang).

https://doi.org/10.1016/j.ijmedinf.2018.05.003
Received 26 February 2018; Received in revised form 26 April 2018; Accepted 3 May 2018

data, to reveal and retrieve salient information, and ultimately to build fundamental knowledge.

1. Introduction

The volume of various data being digitally collected and stored, especially healthcare data that are very large and complex, is enormous and expanding rapidly. As a result, data management and analysis skills are also evolving to transform big data into information and knowledge [1]. The National Health Insurance (NHI) database is released as open public information in conformity with the "Government 3.0" initiative in Korea. Healthcare professionals are faced with the challenge of extracting salient information from healthcare big data to construct knowledge infrastructure and make appropriate and timely decisions [2]. To take advantage of healthcare big data, big data analytics are needed to create understanding and actionable conclusions. Understanding and utilizing big data has the potential to improve care and lower costs by discovering associations and understanding patterns within the data.

Although big data analytics has often been performed by data experts such as computer scientists or statisticians, healthcare professionals have also been interested in visualization methods to maximize the impact of the study results [3,4]. Therefore, several visualization methods have been introduced in healthcare and medicine fields. Most commonly, visualizations of text data were presented by extracting the frequencies of keywords and topics in the field [5,6], and a visual analytics system was proposed in medical diagnosis and computationally-enhanced analysis system applied to biomedical domain [7,8]. Visual data mining using interactive and scalable network and analytic techniques have also been used for effective exploration and communication of ideas within various biological and biomedical domains [9,10]. However, these methods are still too professional for general healthcare or clinical researchers to conduct big data analytics, especially interactive visualization approach.

Visualization is an overarching theme across applications of healthcare big data analytics [11]. Especially, interactive visualization is an important approach to helping big data analytics get a comprehensive view of data and discover information. It also enables researchers to gain insights and evidence for improved outcomes and supports better informed decision-making through the analysis of large-scale data [12,13]. Moreover, it can facilitate and expedite knowledge translation and dissemination [2]. Big data visualization approaches are used to display more than one view per representation, with dynamic changes in the numbers of factors and filtering [14].

Data visualization software offers the ability to handle large data sets even in situations with limited human and financial resources [15]. With Tableau software using interactive visualization, users can drag and drop, drill down, and filter data to create visualizations easily and make interactive dashboards quickly, connect to more data such as SQL databases, spreadsheets, cloud apps, and so on. Tableau Desktop can be used for free for academic purposes. Also, interactive dashboards which are easy to publish and share on Tableau Public blogs help users find hidden insights with multiple views.

From health care and clinical perspectives, hypertension shows high prevalence worldwide, and it leads to related chronic diseases with a high risk of mortality. It is important to control blood pressure to prevent hypertension-related diseases, and drug therapy has been found to be cost-effective to control blood pressure [16]. Expenditures for patients with hypertension who are treated with medication and antihypertensive drugs have increased, and antihypertensive drug prescription patterns and trends change over time [17–20].

It is vital to explore the healthcare big data and recent trends in antihypertensive medication use and changes in drug use patterns within data. Based on this exploration, clear interactive visualizations can be made through the analysis of big data, insights can be obtained, and conclusions may be drawn based on these insights. Therefore, interactive visualization is an important approach to help to understand and to explain large-scale nationally representative prescription data, particularly in light of decision support.

For effective and efficient visualization of big data, this study followed a two-step approach. In the first step, conventional regression models are constructed to evaluate significant factors on therapy and prescription patterns. In the second step, based on the statistical findings of the first step, interactive dashboards are suggested for visualizing prescription patterns such as volume and composition of drugs.

2. Methods

2.1. Database and patients

This study analyzed the Korean NHI claims data, the 3% patient sample which is extracted from the national population using a stratified randomized sampling method by Health Insurance Review and Assessment Service (HIRA). The HIRA-NPS (national patient sample) is a relational database of five tables: (1) a table for general information of prescription specification (Table 20) containing demographic information, such as gender, age, and indicators for inpatient and outpatient services; (2) a table for specific information on services provided (Table 30); (3) a table for diagnostic information (Table 40); (4) a table for outpatient prescriptions (Table 53) with each unique key ID; and (5) a table for information on healthcare service providers (Table of providers). All tables are linkable using a key ID [21].

Hypertension patients were those diagnosed I10, I11, I12, I13, and I15 according to the International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10). The final database includes 1,550,273 prescriptions of 204,917 hypertensive outpatients prescribed antihypertensive drugs.

2.2. Antihypertensive drugs and their classification

All antihypertensive drugs defined in this study are based on related research and definitions [19,20,22–24]. To extract antihypertensive drugs, we used the Anatomical Therapeutic Chemical (ATC) classification system of the World Health Organization (WHO) Collaborating Centre for Drug Statistics Methodology and a drug list from the Korea Pharmaceutical Information Service (KPIS) center [25,26]. The ATC codes of antihypertensive drugs used in this study are the following: antihypertensives (C02), diuretics (C03), beta-blocking agents (C07), calcium channel blockers (C08), agents acting on the renin–angiotensin system (C09), terazosin (G04CA03), and combinations (C03EA, C07BB, C07CB, C07FB, C09B, C09D). The KPIS drug list includes the ATC codes and the Korean main ingredient codes of Ministry of Health and Welfare. We extracted 6471 drugs using the ATC codes of antihypertensive drugs, but the main-ingredient codes of 1467 drugs were missing. We reviewed and mapped 150 main-ingredient codes manually using the Korean drug information database websites [26,27]. The final antihypertensive drugs list contains 204 main-ingredient codes of 6332 drugs mapped with HIRA-NPS.

We reclassified the extracted ATC codes as the following six drug types called antihypertensive drug classes: angiotensin-converting enzyme inhibitors (ACEIs), angiotensin receptor blockers (ARBs), beta blockers (BB), calcium channel blockers (CCB), diuretics, and others (alpha blockers, vasodilators, and so on). In addition, therapy is defined as the number of antihypertensive drug classes in prescriptions. Prescribing drugs from one drug class is considered monotherapy, whereas prescribing drugs from two or more drug classes is considered combination therapy. Single pill combinations are also considered combination therapy except for thiazide diuretics/aldosterone antagonists, which were categorized as diuretics in monotherapy.

2.3. Classification of comorbidity

The hypertension-related comorbidities were organized into six groups with the 6th Korean standard Classification of Disease (KCD-6) codes on the basis of the Korean hypertension guideline and ICD-10 coding research [16,28,29]. The following comorbidities are used in this study: (1) angina pectoris (I20.x); (2) cerebrovascular disease (G45.x, H34.1, I60.x-I67.x, I69.x) including stroke (G45.x, H34.1, I60.x, I61.x, I63.x, I64.x); (3) chronic kidney disease (I12.0, I13.1, N03.2-N03.7, N05.2-N05.7, N18.x, N19.x, N25.0, Z49.0-Z49.2, Z94.0, Z99.2); (4) congestive heart failure (I11.0, I13.0, I13.2, I25.5, I42.0, I42.5, I42.8, I42.9, I50.x); (5) diabetes (E10, E11, E12, E13, E14), including diabetes type I (E10) and diabetes type II (E11); and (6) other cardiovascular diseases (acute myocardial infarction (I21.x), post myocardial infarction (I22.x, I25.2), other acute ischemic heart diseases (I24), chronic ischemic heart disease (I25), atrial fibrillation (I48.0), left ventricular hypertrophy (I51.7)), carotid atherosclerosis (I65.2), peripheral vascular disease (I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9), and aortic aneurysm and dissection (I71.x)).

2.4. Statistical analysis and interactive visualization tool

All statistical analyses were performed with R software (version 3.3.2) from the R Foundation for Statistical Computing (Vienna, Austria). Descriptive analysis was carried out to identify demographic characteristics, and regression analyses were applied for variables (therapy, most commonly prescribed drug classes, and number of prescribed drug classes) to evaluate their significance. The significant factors were visualized for intuitive results of statistical analysis, rather than showing big data overall. P values of less than 0.05 were regarded as statistically significant.

Tableau Desktop Professional Edition (version 10.1.4) from the Tableau Software (Seattle, WA, USA) was used to visualize prescription patterns based on the statistical analysis results. Tableau is a commercially available (but Tableau Public is free, and Tableau Desktop is free for students and instructors at accredited academic institution) visualization software. It uses the Visual Query Language (VizQL) to represent large data through interactive visual interfaces. VizQL is a visual query language that translates drag-and-drop actions into data queries. A brief visualization procedure for healthcare data was presented using Tableau as a business intelligence tool in healthcare domain [30]. As described in Ko and Chang [30], we designed horizontal stacked and grouped bar charts with lines, packed bubble charts, treemaps, filled maps, radar charts, box and whisker plots with lines, and filters. All visualizations have responsive tooltips to show in more details instantly and are incorporated into interactive visualization dashboards to represent statistical analysis results and prescription patterns (Fig. 1). All the interactive dashboards were published and shared on the Tableau public sites.

3. Results

3.1. Characteristics of sample

The dataset contained a total of 1,550,273 prescriptions of hypertension outpatients (Table 1). The more common prescriptions were for women, adults aged 30 years and over, patients of small institutions, recipients of NHI, non-metropolitan area residents, and patients without comorbidity, and combination therapy was more common than monotherapy. The total number of prescriptions with comorbidity was 348,285 (22.47%), and diabetes (12.15%) was the most common comorbidity, followed by cerebrovascular disease (2.98%), angina pectoris (2.52%), and so on.

3.2. Analyses of prescription patterns

Combination therapy was more likely than monotherapy to be prescribed for men, adults aged 30–64 years, patients of small healthcare institutions, recipients of medical-aid and non-metropolitan area residents (Table 1). Among comorbidities, congestive heart failure had the highest OR (2.52, $p < 0.001$), but cerebrovascular disease was not significant ($p = 0.099$) and diabetes was less likely than other comorbidities to be treated with combination therapy (OR = 1.02, $p < 0.001$).

Dihydropyridine calcium channel blockers (DHPCCB) were more likely to be prescribed to women, older patients, and patients of small institutions (Table 2). Women, patients of small institutions, and those with diabetes were more likely to be treated with 'ARBs + Diuretics'. Among comorbidities, cerebrovascular disease and diabetes were more likely to be treated with 'ARBs + DHPCCB' and ARBs, respectively.

The number of prescribed drug classes was more likely to increase for men, older patients, patients of small institutions, recipients of medical-aid, and non-metropolitan area residents (Table 3). All comorbidities except cerebrovascular disease were more likely to be treated by a higher number of drugs classes.

3.3. Interactive visualizations

We designed two interactive dashboards based on the statistical analysis results to visualize prescription patterns with Tableau software.

The first dashboard shows a comparison between monotherapy and combination therapy, the distribution of drug classes, generic drugs, prescriptions, major characteristics, and therapy, and filters (Fig. 2). The comparison between monotherapy and combination therapy and distribution of the most commonly prescribed drug classes by gender and therapy is visualized as horizontal stacked bar charts. Major drug classes according to age groups are represented as a horizontal grouped bar chart with lines. Drug classes according to therapy are illustrated with a packed bubble chart, and the total drug classes and generic drugs are presented as treemaps, each with mouse-click filters. The latter visualizations also have a parameter filter of the top number of drugs and check box filters in drop-down lists. The total prescription distributions are expressed in a filled map with a mouse-click filter by location. The total counts and proportions of genders, healthcare institutions, therapies, and comorbidities are presented as treemaps with mouse-click filters. Finally, a packed bubble chart is used to present the distribution of age groups with a mouse-click filter.

The second dashboard shows a comparison of the most commonly prescribed drug classes and the average number of prescribed drug classes by major characteristics, distribution of average number by location, and filters (Fig. 3). The comparison of drug classes is visualized with radar charts with axes of the most commonly prescribed drug classes by gender, age group, healthcare institution, and comorbidity. All legends of radar charts highlight the selected items. For comorbidities, a check box filter for choosing a specific comorbidity from a drop-down list enables a clear drug class distribution comparison between or among the selected comorbidities. Furthermore, a radio button filter is used to change the number of axes for detailed comparison of comorbidities in a radar chart. A comparison of the average number of prescribed drug classes is represented as box and whisker plots with lines by gender, age group, and healthcare institution according to the presence of comorbidity. Finally, the average number distributions with standard deviations are also presented as a filled map.

Exploring the prescribed drugs for specific diseases is represented using interactive visual interfaces. For example, the check box filter of a drop-down list makes a prescribed drug class distribution comparison between 'no comorbidity' and 'congestive heart failure' in a radar chart of the second dashboard (Fig. 4). Based on information about the high use of other classes of drugs for congestive heart failure in comparison to 'no comorbidity', it is possible to explore the other prescribed drug

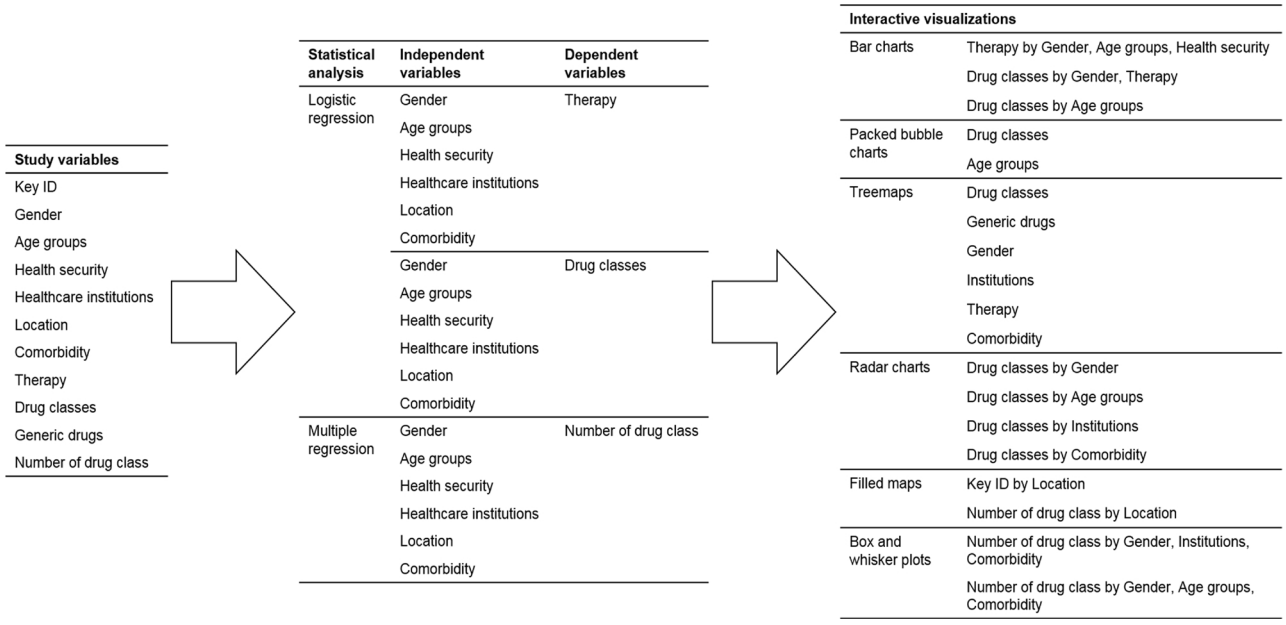


Fig. 1. Variables used in interactive visualizations of dashboards.

Table 1
Results of descriptive analysis and logistic regression analysis on therapy.

Variables	Total (n = 1,550,273)	Therapy[a]	
		OR	95% CI
Gender			
Men	711,687 (45.91%)	–	–
Women	838,586 (54.09%)	0.78	(0.78–0.79)***
Age group			
0-17	678 (0.04%)	–	–
18-29	4,006 (0.26%)	3.95	(3.22–4.84)***
30-64	810,479 (52.28%)	6.40	(5.27–7.77)***
65-	735,110 (47.42%)	6.22	(5.12–7.55)***
Healthcare institutions[b]			
Large	266,285 (17.18%)	–	–
Small	1,283,988 (82.82%)	1.33	(1.31–1.34)***
Health security			
NHI	1,442,574 (93.05%)	–	–
Medical-aid	107,699 (6.95%)	1.13	(1.12–1.15)***
Location			
Metropolitan	711,975 (45.93%)	–	–
Non-metropolitan	838,298 (54.07%)	1.02	(1.02–1.03)***
Comorbidities			
None	1,201,988 (77.53%)	–	–
Angina pectoris	38,995 (2.52%)	1.44	(1.41–1.47)***
Cerebrovascular disease	46,175 (2.98%)	0.98	(0.96–1.00)
Chronic kidney disease	11,707 (0.76%)	1.37	(1.32–1.43)***
Congestive heart failure	29,149 (1.88%)	2.52	(2.45–2.60)***
Diabetes	188,429 (12.15%)	1.02	(1.01–1.04)***
Other cardiovascular diseases	33,830 (2.18%)	1.63	(1.59–1.67)***
Therapy			
Monotherapy	581,634 (37.52%)	–	–
Combination therapy	968,639 (62.48%)	–	–

[a] The reference is 'Monotherapy'.

[b] In healthcare institutions (healthcare service providers), large institutions include tertiary and secondary hospitals and small institutions contain general hospitals and clinics.

*** < 0.001.

classes by using various filters in the first dashboard (Fig. 5). The filtering result is displayed using a mouse-click filter by choosing 'congestive heart failure' in the comorbidity treemap, changing the numbers of the top drugs and comorbidity in the parameter filters, and unchecking 'others' in the drug class or generic drugs of the check box filters in the drop-down lists. The tooltips showing additional information for 'ARBs + DHPCCB + Diuretics' are also displayed in the packed bubble chart when the mouse is hovered over one of the other drug classes.

In the horizontal stacked bar chart, the combination therapy proportion for men is bigger than that for women, except for the 0 to 17 age group. Also, the most commonly prescribed drug classes have the opposite patterns by gender in the horizontal stacked bar chart. Men have higher proportions of ARBs and 'ARBs + DHPCCB', but women have bigger proportions of DHPCCB and 'ARBs + Diuretics'. In the horizontal grouped bar chart, the use of ACEIs is very high in patients aged 17 years and under. The use of DHPCCB increases with age, but that of ARBs decreases from 18 years old. The most commonly prescribed drug class is 'ARBs + DHPCCB', followed by 'ARBs + Diuretics', DHPCCB, ARBs, and so on. In the radar charts, prescriptions show a high proportion of ARBs and other classes in large institutions, but DHPCCB, 'ARBs + DHPCCB', and 'ARBs + Diuretics' are used more often in small institutions. Among comorbidities, other classes of drugs were more frequently prescribed to patients with comorbidities, except cerebrovascular disease and diabetes, than those without comorbidities. As seen in the box and whisker plots, the average numbers of gender and healthcare institutions differ according to presence of comorbidity. A difference between men and women is observed for prescriptions without comorbidity. In the case of comorbidity, there is little difference in gender and healthcare institutions, but the difference between men and women is large, especially, the difference between large institutions and small institutions for men is bigger than that for women in case of no comorbidity. Also, the average drug class number of men increases with age for patients with comorbidity, but for those with no comorbidity, it decreases from 65 years old. On the other hand, the average drug class number for women increases with age regardless of comorbidity. Men without comorbidity have higher average numbers in the 18–29 and 30–64 age groups than those with comorbidity, whereas women have higher average number in the overall subgroups of comorbidity than those of no comorbidity except for the 18–29 age group.

4. Discussions

Exploring healthcare big data using visualization approaches has significant value for healthcare professionals in that it can enable them to find information hidden within data, gain insights and evidence,

Table 2
Results of logistic regression analysis on the most commonly prescribed drug classes.

Variables	Most commonly prescribed drug classes[a]							
	ARBs + DHPCB		ARBs + Diuretis		DHPCCB		ARBs	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Gender								
Men	–	–	–	–	–	–	–	–
Women	0.81	(0.80–0.81)***	1.27	(1.26–1.29)***	1.38	(1.37–1.39)***	1.13	(1.12–1.14)***
Age group								
0-17	–	–	–	–	–	–	–	–
18-29	12.78	(7.32–22.33)***	7.03	(3.72–13.29)***	2.23	(1.53–3.24)***	3.13	(2.46–3.99)***
30-64	21.70	(12.51–37.64)***	15.81	(8.46–29.58)***	4.62	(3.25–6.58)***	3.28	(2.61–4.12)***
65-	13.43	(7.75–23.30)***	11.50	(6.15–21.51)***	5.01	(3.52–7.14)***	1.94	(1.54–2.43)***
Healthcare institutions								
Large	–	–	–	–	–	–	–	–
Small	1.07	(1.06–1.09)***	1.87	(1.84–1.90)***	1.52	(1.49–1.54)***	0.83	(0.82–0.85)***
Health security								
NHI	–	–	–	–	–	–	–	–
Medical-aid	1.03	(1.01–1.05)***	0.94	(0.92–0.96)***	0.83	(0.82–0.85)***	0.88	(0.87–0.90)***
Location								
Metropolitan	–	–	–	–	–	–	–	–
Non-metropolitan	0.93	(0.92–0.94)***	0.99	(0.99–1.00)***	0.95	(0.94–0.96)***	0.92	(0.91–0.93)***
Comorbidities								
None	–	–	–	–	–	–	–	–
Angina pectoris	0.19	(0.18–0.20)***	0.20	(0.19–0.21)***	0.19	(0.18–0.20)***	0.22	(0.21–0.23)***
Cerebrovascular disease	1.49	(1.45–1.53)***	1.15	(1.11–1.19)***	1.39	(1.35-1.43)***	1.42	(1.38–1.46)***
Chronic kidney disease	0.60	(0.57–0.63)***	0.61	(0.57–0.65)***	0.29	(0.27–0.32)***	0.84	(0.80–0.88)***
Congestive heart failure	0.23	(0.22–0.24)***	0.68	(0.66–0.70)***	0.13	(0.12–0.14)***	0.21	(0.20–0.22)***
Diabetes	1.10	(1.09–1.12)***	1.29	(1.27–1.30)***	0.59	(0.58–0.61)***	1.78	(1.76–1.81)***
Other cardiovascular diseases	0.21	(0.20–0.22)***	0.35	(0.34–0.37)***	0.16	(0.15–0.17)***	0.31	(0.30–0.32)***

[a] The reference is 'Other classes' except four most commonly prescribed classes.
*** < 0.001.

Table 3
Results of multiple regression analysis on the number of prescribed drug classes.

Variables	Number of prescribed drug classes	
	Regression coefficient	95% CI
Gender		
Men	0.11	(0.11 to 0.11)***
Women	–	–
Age group		
0-17	−0.61	(−0.66 to −0.55)***
18-29	−0.11	(−0.13 to −0.08)***
30-64	−0.01	(−0.02 to −0.01)***
65-	–	–
Healthcare institutions		
Large	−0.08	(−0.08 to −0.07)***
Small	–	–
Health security		
NHI	−0.06	(−0.07 to −0.06)***
Medical-aid	–	–
Location		
Metropolitan	−0.01	(−0.01 to 0.00)***
Non-metropolitan	–	–
Comorbidities		
Angina pectoris	0.18	(0.18 to 0.19)***
Cerebrovascular disease	−0.01	(−0.02 to 0.00)**
Chronic kidney disease	0.29	(0.28 to 0.31)***
Congestive heart failure	0.43	(0.42 to 0.44)***
Diabetes	0.03	(0.03 to 0.04)***
Other cardiovascular diseases	0.23	(0.22 to 0.24)***
None	–	–

** < 0.01.
*** < 0.001.

improve healthcare outcomes, and make better informed health-related decisions. Clearly, the NHI claims data have a powerful potential to help realize the aforementioned advantages. Interactive visualizations enable healthcare professionals to understand the vast and complex claims data, to retrieve salient information, and ultimately to construct useful knowledge that is essential to make correct decisions [2].

This study was conducted to verify the significant characteristics, to analyze and visualize prescription patterns using NHI claims data, and to propose interactive dashboards to identify prescription patterns from various perspectives using visualization approaches with Tableau software. Tableau software provides a practical solution to common problems regarding how to explore and understand large datasets efficiently. It enables healthcare professionals to customize the analysis and presentation of information according to the user's analysis purpose. The program has various functions to interact with data significantly: to quickly add or exclude variables and to segment, sort, highlight, and perform other actions, such as zooming and filtering. These functions enhances the users' capability to analyze large volumes of quantitative information through the use of interactive visual interfaces. Drag-and-drop functionality also makes it easy to change visualizations. Furthermore, Tableau is able to extract and visualize data in real-time from a large database. However, although Tableau is relatively simple to use, researchers may need considerable time and efforts in learning and mastering the software for more layered, complex visualizations [15].

Previous studies have analyzed prescription patterns with a single aspect and have been insufficiently explorative in terms of all integration points. They usually have explained only the facts that significant factors affect research variables or have displayed static graphs based on analysis results. For these reasons, previous methods have limitations to gain insights and improve the decision-making process through traditional statistical analysis. However, this study was conducted to find significant characteristics and then create and propose visualizations and interactive visualization dashboards that represent prescription patterns with holistic and multiple views on the basis of statistical analysis results.

These interactive visualization dashboards are helpful for providing evidence for research and policy making. For example, there are

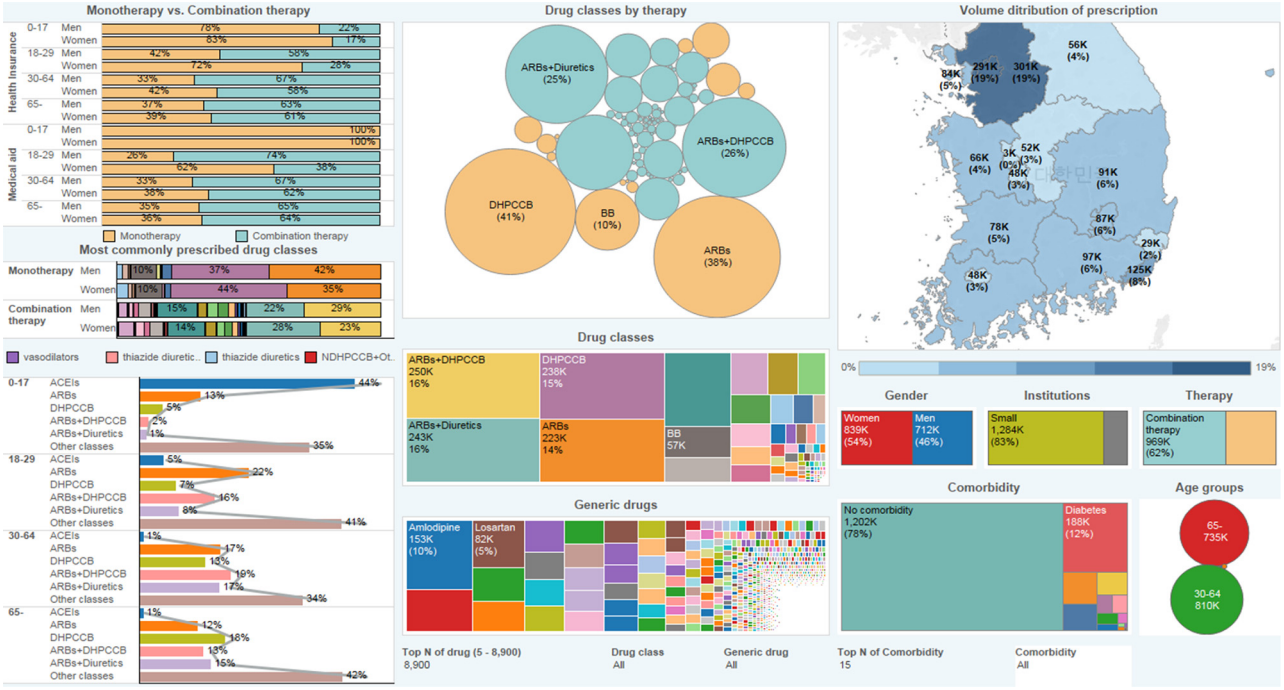


Fig. 2. The first dashboard for presenting distributions of prescriptions and comparing therapy types https://public.tableau.com/profile/sweetino#!/vizhome/Thefirstdashboard/1.

prescription concentrations in the capital region, the highest average number of prescribed drug classes in Gangwon-do, the high average number of drug classes prescribed to men with or without comorbidity in small institutions, and so on. Furthermore, various explorations of prescription patterns can be facilitated by use of the interactive visualization dashboards. For instance, explorations of antihypertensive drug patterns with multiple perspectives, such as specific location, major characteristics, and therapy, as well as the exploration of drugs for specific diseases are suggested by the results. In addition, specific exploration is feasible through customization, such as modification of the number of axes in a radar chart (Fig. 6). Such interactive visualizations can be used for various datasets of surveys, cohorts, and other previous studies that have been performed to evaluate prescription patterns using traditional statistical analysis [17,31,32].

The prescription patterns show differences according to major characteristics. The average differences can be seen among genders, age groups, healthcare institutions, and presence of comorbidities. We should also consider differences regarding the characteristics of the NHI

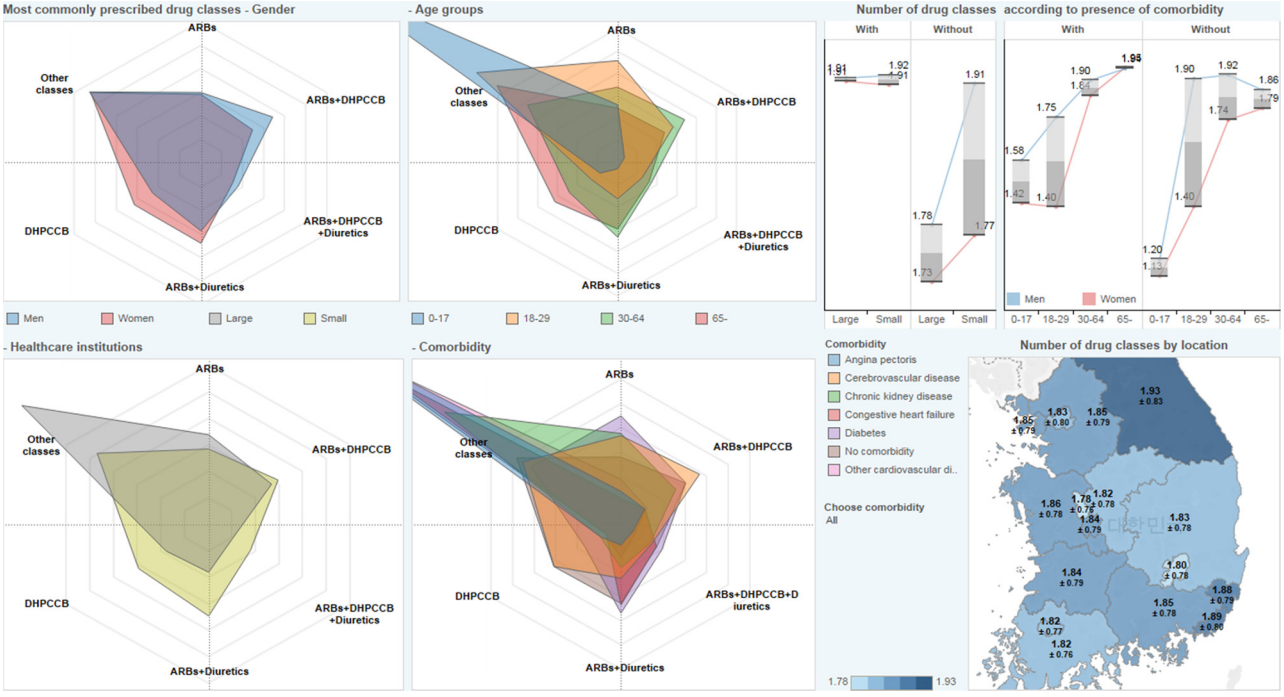


Fig. 3. The second dashboard for the most commonly prescribed drug classes and the number of prescribed drugs https://public.tableau.com/profile/sweetino#!/vizhome/Theseconddashboard/2.

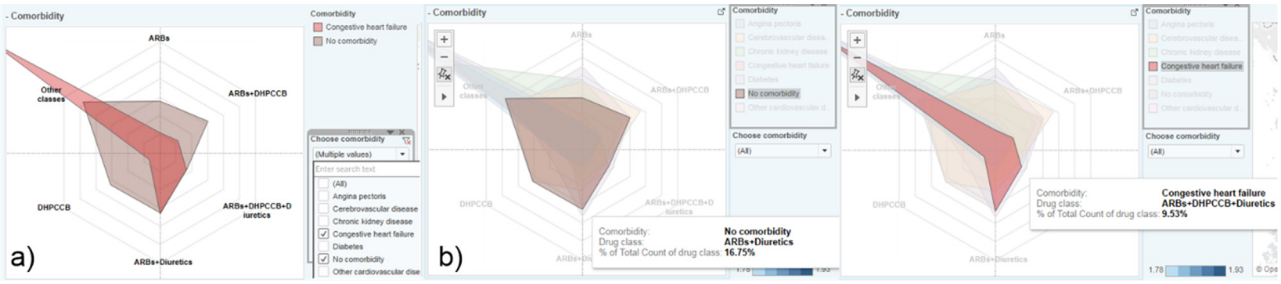


Fig. 4. Filters and tooltips for comorbidity: a) check box filters in drop-down list, b) highlighted items in legends.

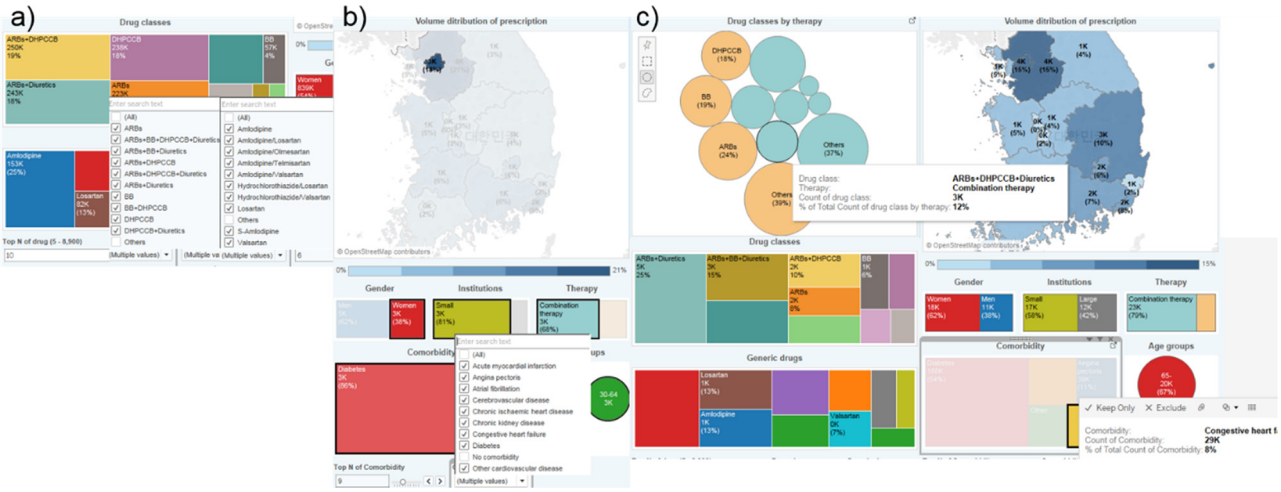


Fig. 5. Filters and tooltips for major characteristics and drugs: a) check box filters in drop-down lists for drugs, b) mouse-click filters for major characteristics and drop-down list for comorbidity, and c) filtered results of congestive heart failure.

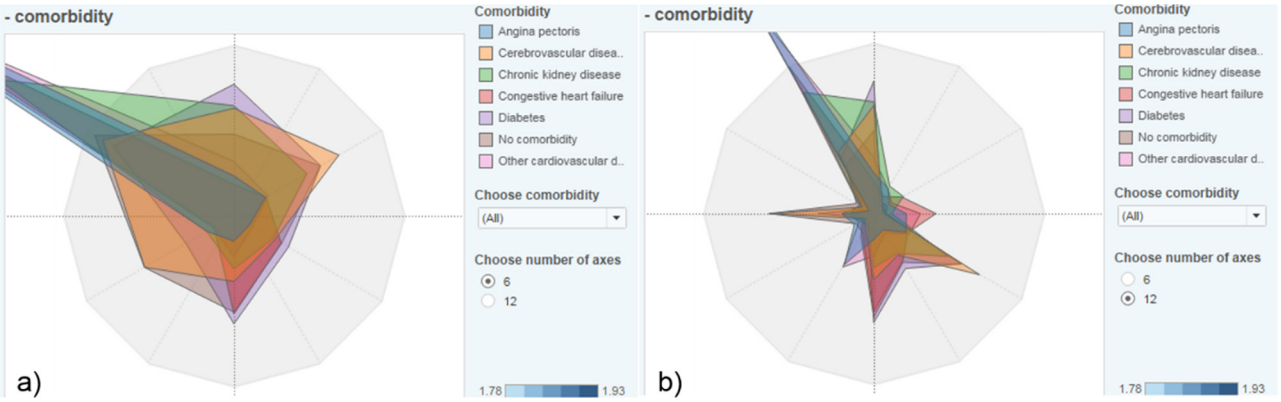


Fig. 6. Modification of axes in radar chart of comorbidity: a) six axes of drug classes, b) twelve axes of drug classes.

claims data. Actually, the proportions are high, with values of 83% and 78% for small institutions and no comorbidity, respectively. The proportion of no comorbidity is 84% for small institutions, and the proportion of comorbidity is 54% for large institutions. All average numbers of age groups of men are higher than those of women; in particular, men have the highest average number for 30–64 years old.

This study focused on the visualization of prescription patterns, but it is also important to provide exact information within healthcare big data. In terms of accuracy and reliability, the results of this study are consistent with those of previous clinical studies. For example, the difference between the prescription patterns of men and women is related to several factors. First, regarding the adherence rate of antihypertensive medication, the adherence rate of male patients is higher than that of female patients attending National Health Service hospitals [33]. Second, low-dose combinations decrease blood pressure significantly in women considering the antihypertensive dose-response relationship [34]. Finally, calcium channel blockers are particularly effective in women, and women more often than men are prescribed diuretics, but men are more frequently treated with ARBs [35,36]. Another example is the high use of ACEIs in children. ACEIs are the agents most commonly prescribed to adolescents with primary hypertension in monotherapy and to children with primary and secondary hypertension [37,38].

There were a few limitations in this study. The present study compared the prescription patterns of monotherapy and combination therapy and identified average numbers and most commonly prescribed drugs, but it could not examine the impact of medication on blood pressure control in hypertension. The Korean NHI claims data are nationally representative, but they do not include clinical information. Another constraint may be inaccuracy of diagnosis information because

insurance claims data is collected for administrative purposes in reimbursement process of healthcare services.

In further study, we can achieve better performance of interactive visualizations using data from EHR and PHR systems. Also, we can evaluate the improvement of blood pressure using advanced visualization techniques if national representative data including blood pressure are available. In addition, we can verify changes in antihypertensive prescription patterns over a period of many years by adopting a time series visualization approach.

5. Conclusions

Exploring big data using visualizations is expected to deliver many future benefits, so it is necessary to use interactive visualization approaches to understand prescription patterns in healthcare big data. This study was carried to verify the significant characteristics of prescription patterns and propose interactive dashboards based on statistical analysis results using large-scale nationally representative data. We found that major characteristics (genders, age groups, healthcare institutions, and comorbidities) explained the differences in therapy, number of prescribed drug classes, and most commonly prescribed drug classes. This study's interactive visualizations can provide healthcare professionals insight into prescription patterns and demonstrate the value of creating interactive dashboards to support informed and prompt decision-making. Such interactive visualizations can be used for various studies conducted using traditional statistical analysis.

Conflict of interest

There is no potential conflict of interests relevant to this manuscript. This manuscript has not been published nor is under simultaneous considerations for publication elsewhere.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A5A2A01014390). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] T.B. Murdoch, A.S. Detsky, The inevitable application of big data to health care, JAMA 309 (2013) 1351–1352, http://dx.doi.org/10.1001/jama.2013.393.
[2] S. Al-Hajj, I. Pike, B. Riecke, B. Fisher, Visual analytics for public health: supporting knowledge construction and decision-making, Syst. Sci. (HICSS), 2013 46th Hawaii Int. Conf., IEEE (2013) 2416–2423, http://dx.doi.org/10.1109/HICSS.2013.599.
[3] R. Osuala, O. Arandjelovic, Ieee, visualization of patient specific disease risk prediction, 2017 Ieee Embs Int. Conf. Biomed. Heal. Informatics (2017) 241–244, http://dx.doi.org/10.1109/BHI.2017.7897250.
[4] J. Li, O. Arandjelovic, Intuitive and interpretable visual communication of a complex statistical model of disease progression and risk, Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS (2017) 4199–4202, http://dx.doi.org/10.1109/EMBC.2017.8037782.
[5] H.A. Park, J.Y. Lee, J. On, J.H. Lee, H. Jung, S.K. Park, 2016 year-in-review of clinical and consumer informatics: analysis and visualization of keywords and topics, Heal. Inf. Res. 23 (2017) 77–86, http://dx.doi.org/10.4258/hir.2017.23.2.77.
[6] A.R. Kim, H.A. Park, T.M. Song, Development and evaluation of an obesity ontology for social big data analysis, Heal. Inf. Res. 23 (2017) 159–168, http://dx.doi.org/10.4258/hir.2017.23.3.159.
[7] B.L.W. Wong, K. Xu, A. Holzinger, Interactive visualization for information analysis in medical diagnosis, Inf. Qual. E-Health, Lect. Notes Comput. Sci. (2011) 109–120, http://dx.doi.org/10.1007/978-3-642-25364-5_11.
[8] C. Turkay, F. Jeanquartier, A. Holzinger, H. Hauser, On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics, Interact. Knowl. Discov. Data Min. Biomed. Inf. (2014) 117–140, http://dx.doi.org/10.1007/978-3-662-43968-5_7.
[9] D. Otasek, C. Pastrello, A. Holzinger, I. Jurisica, Visual data mining: effective exploration of the biological universe, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.) 8401 (2014) 19–33, http://dx.doi.org/10.1007/978-3-662-43968-5_2.
[10] W. Sturm, T. Schreck, A. Holzinger, T. Ullrich, Discovering Medical Knowledge Using Visual Analytics, VCBM, 2015, pp. 71–81, http://dx.doi.org/10.2312/vcbm.20151210.
[11] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Heal. Inf. Sci. Syst. 2 (3) (2014), http://dx.doi.org/10.1186/2047-2501-2-3.
[12] L. Wang, G. Wang, C.A. Alexander, Big data and visualization: methods, challenges and technology progress, Digital Technol. 1 (2015) 33–38, http://dx.doi.org/10.12691/dt-1-1-7.
[13] D.J. Janvrin, R.L. Raschke, W.N. Dilla, Making sense of complex data using interactive data visualization, J. Acc. Educ. 32 (2014) 31–48, http://dx.doi.org/10.1016/j.jaccedu.2014.09.003.
[14] E.Y. Gorodov, V.V. Gubarev, Analytical review of data visualization methods in application to big data, J. Electr. Comput. Eng. 2013 (2013) 22, http://dx.doi.org/10.1155/2013/969458.
[15] S.A. Murphy, Data visualization and rapid analytics: applying tableau desktop to support library decision-making, J. Web Librariansh. 7 (2013) 465–476, http://dx.doi.org/10.1080/19322909.2013.825148.
[16] J. Shin, J.B. Park, K.I. Kim, J.H. Kim, D.H. Yang, W.B. Pyun, Y.G. Kim, G.H. Kim, S.C. Chae, 2013 Korean Society of hypertension guidelines for the management of hypertension: part III-hypertension in special situations, Clin. Hypertens 21 (2015) 3, http://dx.doi.org/10.1186/s40885-014-0014-1.
[17] Q. Gu, V.L. Burt, C.F. Dillon, Trends in antihypertensive medication use and blood pressure control among United States adults with hypertension: the National Health and Nutrition examination survey, 2001–2010, J. Vasc. Surg. 57 (2013) 893, http://dx.doi.org/10.1161/CIRCULATIONAHA.112.096156.
[18] P.H. Liu, J.D. Wang, Antihypertensive medication prescription patterns and time trends for newly-diagnosed uncomplicated hypertension patients in Taiwan, BMC Health Serv. Res. 8 (133) (2008), http://dx.doi.org/10.1186/1472-6963-8-133.
[19] T. Catić, B. Begović, Outpatient antihypertensive drug utilization in canton Sarajevo during five years period (2004–2008) and adherence to treatment guidelines assessment, Bosn. J. Basic. Med. Sci. Basic.nih Med. Znan. Assoc. Basic. Med. Sci. 11 (2011) 97–102, http://dx.doi.org/10.17305/bjbms.2011.2589.
[20] L.Y. Huang, W.Y. Shau, H.C. Chen, S. Su, M.C. Yang, H.L. Yeh, M.S. Lai, Pattern analysis and variations in the utilization of antihypertensive drugs in Taiwan: a six-year study, Eur. Rev. Med. Pharmacol. Sci. 17 (2013) 410–419.
[21] L. Kim, J. Kim, S. Kim, A guide for the utilization of health insurance review and assessment service national patient samples, Epidemiol. Health 36 (2014) e2014008, http://dx.doi.org/10.4178/epih/e2014008.
[22] T. Grimmsmann, W. Himmel, Discrepancies between prescribed and defined daily doses: a matter of patients or drug classes? Eur. J. Clin. Pharmacol. 67 (2011) 847–854, http://dx.doi.org/10.1007/s00228-011-1014-7.
[23] A. Katz, P. Martens, D. Chateau, B. Bogdanovic, I. Koseva, C. McDougall, Understanding the Health System Use of Ambulatory Care Patients, Manitoba Centre for Health Policy, Winnipeg, MB, 2013http://mchp-appserv.cpe.umanitoba.ca/reference/Amb_Care_Deliverable_Web_final.pdf.
[24] M. Chartier, A. Dart, N. Tangri, P. Komenda, R. Walld, B. Bogdanovic, C.A. Burchill, I. Koseva, K.-L. McGowan, L. Rajotte, Care of Manitobans Living With Chronic Kidney Disease, Manitoba Centre for Health Policy, University of Manitoba, 2015, http://mchp-appserv.cpe.umanitoba.ca/reference/ckd_final.pdf.
[25] ATC/DDD Index, WHO, 2016 (Accessed April 19 2018), http://www.whocc.no/atc_ddd_index.
[26] Korea Pharmaceutical Information Service, HIRA, 2016 (Accessed 19 April 2018), http://biz.kpis.or.kr/index.jsp.
[27] BIT Drug Information Database, BIT, 2016 (Accessed 19 April 2018), http://www.druginfo.co.kr.
[28] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L.D. Saunders, C.a Beck, T.E. Feasby, W.a Ghali, Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data, Med. Care (43) (2005) 1130–1139, http://dx.doi.org/10.1097/01.mlr.0000182534.19832.83.
[29] R.A. Kokotailo, M.D. Hill, Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10, Stroke 36 (2005) 1776–1781, http://dx.doi.org/10.1161/01.STR.0000174293.17959.a1.
[30] I. Ko, H. Chang, Interactive visualization of healthcare data using tableau, Heal. Inf. Res. 23 (2017) 349–354, http://dx.doi.org/10.4258/hir.2017.23.4.349.
[31] V. Tandon, S. Sharma, S. Mahajan, A. Mahajan, V. Khajuria, V. Mahajan, C. Prakash, Antihypertensive drug prescription patterns, rationality, and adherence to joint national committee-7 hypertension treatment guidelines among Indian postmenopausal women, J. Midlife Health 5 (2014) 78, http://dx.doi.org/10.4103/0976-7800.133994.
[32] N. Jarari, N. Rao, J.R. Peela, K.A. Ellafi, S. Shakila, A.R. Said, N.K. Nelapalli, Y. Min, K.D. Tun, S.I. Jamallulail, A.K. Rawal, R. Ramanujam, R.N. Yedla, D.K. Kandregula, A. Argi, L.T. Peela, A review on prescribing patterns of antihypertensive drugs, Clin. Hypertens 22 (2015) 7, http://dx.doi.org/10.1186/s40885-016-0042-0.
[33] M.U. Khan, S. Shah, T. Hameed, Barriers to and determinants of medication adherence among hypertensive patients attended National Health Service Hospital, Sunderland, J. Pharm. Bioallied Sci. 6 (2014) 104–108, http://dx.doi.org/10.4103/0975-7406.129175.
[34] M.E. Safar, M.G. Myers, F. Leenen, R. Asmar, Gender influence on the dose-ranging of a low-dose perindopril-indapamide combination in hypertension: effect on systolic and pulse pressure, J. Hypertens 20 (2002) 1653–1661, http://dx.doi.org/10.1097/00004872-200208000-00029.
[35] C. Ljungman, T. Kahan, L. Schiöler, P. Hjerpe, J. Hasselström, B. Wettermark, K.B. Boström, K. Manhem, Gender differences in antihypertensive drug treatment: results from the Swedish primary care cardiovascular database (SPCCD), J. Am. Soc. Hypertens 8 (2014) 882–890, http://dx.doi.org/10.1016/j.jash.2014.08.015.
[36] M.L. Muiesan, M. Salvetti, C.A. Rosei, A. Paini, Gender differences in antihypertensive treatment: myths or legends? High blood press, Cardiovasc. Prev. 23 (2016) 105–113, http://dx.doi.org/10.1007/s40292-016-0148-1.
[37] E.Y. Yoon, L. Cohn, D. Rocchini, G. Kershaw, F. Freed, S. Ascione, Clark, Antihypertensive prescribing patterns for adolescents with primary hypertension, Pediatrics 129 (2012) e1–e8, http://dx.doi.org/10.1542/peds.2011-0877.
[38] W.P. Welch, W. Yang, P. Taylor-Zapata, J.T. Flynn, Antihypertensive drug use by children: are the drugs labeled and indicated? J. Clin. Hypertens 14 (2012) 388–395, http://dx.doi.org/10.1111/j.1751-7176.2012.00656.x.

Visual Informatics 6 (2022) 67–77

Contents lists available at ScienceDirect

Visual Informatics

journal homepage: www.elsevier.com/locate/visinf

A survey of visual analytics techniques for online education

Xiaoyan Kui, Naiming Liu, Qiang Liu, Jingwei Liu, Xiaoqian Zeng, Chao Zhang *

School of Computer Science, Central South University, 932 Lushan South Road, Changsha, 410083, Hunan, China

ARTICLE INFO

Article history:
Received 11 June 2022
Received in revised form 16 July 2022
Accepted 26 July 2022
Available online 1 August 2022

Keywords:
Visual analytics
Online education
Behavior analysis
Content analysis

ABSTRACT

Visual analytics techniques are widely utilized to facilitate the exploration of online educational data. To help researchers better understand the necessity and the efficiency of these techniques in online education, we systematically review related works of the past decade to provide a comprehensive view of the use of visualization in online education problems. We establish a taxonomy based on the analysis goal and classify the existing visual analytics techniques into four categories: learning behavior analysis, learning content analysis, analysis of interactions among students, and prediction and recommendation. The use of visual analytics techniques is summarized in each category to show their benefits in different analysis tasks. At last, we discuss the future research opportunities and challenges in the utilization of visual analytics techniques for online education.

1. Introduction

In recent years, online education has developed rapidly. Students can complete various learning tasks on the learning platforms. A large number of educational data are recorded on these online learning platforms, such as student access learning resource data (Li et al., 2017a), video clickstream data (Shi et al., 2015), forum data (Fu et al., 2016), problem-solving data (Xia et al., 2020c), and course data (Zhao et al., 2018). By analyzing these online education data, instructors and analysts can understand the students' learning process, discover the factors that promote or hinder students' learning, adjust the curriculum structure, and intervene in student behavior to improve the quality of teaching. However, online educational data is huge, complex, heterogeneous and multi-level, which poses a challenge for end users to conduct efficient data analysis.

Visual analytics techniques transform complex data into intuitive representations and provide significant assistance to the data analysis. Based on analysis goals, different visual analytics techniques have been used to facilitate the analysis of online educational data and shown their efficiency in improving learning performance. For instance, Kim et al. (2014) conducted a large-scale analysis of in-video dropout, peaks in viewership and student activity, to explore the rationale behind students' watching video behavior. Schwab et al. (2016) proposed a knowledge concept map to recommend an efficient learning path to students. Several review papers summarize the existing works to give a comprehensive guide to audiences. Qu and Chen (2015) provided a review of progress in MOOC visual analytics in 2015. However, the scope of the review is solely focusing on the MOOC data. Vieira et al. (2018) presented a systematic review of visual learning analytics in educational data, investigating the visualization of educational data and the design of corresponding visual learning analytics tools. However, the distinction of the use of visual analytics tools in different online learning tasks needs to be addressed more clearly. Moreover, emerging analytical tasks in online education are not included in their survey paper. For example, Xia et al. (2020b) used visual analytics techniques to analyze students' actions on online problem-solving platforms.

Taking into consideration various learning materials and recent advances in visual analytics techniques for different tasks, we summarize the state-of-the-art papers and present a comprehensive survey of the visual analytics for online educational data. We propose a novel taxonomy based on different analysis goals. In each category, several representative papers are presented to show visual analytics techniques used on educational data and intuitions brought by visual analytics techniques. Finally, we discuss future research directions and opportunities. This paper can help researchers quickly understand and gain insights into different analysis goals in online education. We hope our survey can promote the development of visual analysis of educational data.

2. Survey landscape

This paper aims to provide an overview of visualization in online education. In order to identify the main uses of visualization in online education, we summarize and categorize the existing visual analytics techniques based on different analysis goals.

* Corresponding author.
E-mail address: chao.zhang@csu.edu.cn (C. Zhang).

https://doi.org/10.1016/j.visinf.2022.07.004

2.1. Paper selection

We carry out the literature search in two manners: search-driven and reference-driven selections. We first use the paper search method to obtain the preliminary paper collection. Considering online education, distance education, problem-solving and other scenarios, we identified the search keywords as online education, MOOC, e-learning, forum, online discussion, problem-solving, engagement, and performance. We used search tools such as IEEE explorer, ACM Digital Library, and Google Scholar to search papers with these keywords combined with "visualization" and "visual analysis" in the past decade (2011 to 2021). In particular, our search covered the high-impact conferences (IEEE VAST, ACM CHI, IEEE EuroVis, and IEEE PacificVis.) and journals (IEEE TVCG, CGF, and CG&A) in the field of visualization. After our first round of search, we recursively searched the papers in the references of the collected papers to make sure that we covered all essential related papers. Next, we browse the abstract and main content of the paper to determine further whether the paper is related to visual analytics techniques for online education data. After collection and refining, we finally selected 60 representative papers for discussion in this survey.

2.2. Taxonomy

After comprehensively analyzing and summarizing the collected papers, we classified visual analytics techniques for online education according to the goal of analysis, including learning behavior analysis, learning content analysis, analysis of interactions among students, and prediction and recommendation. We summarize related works in Table 1.

2.2.1. Learning behavior analysis

There are many teaching modules in online education, such as videos, problems and forums. The multiple choice of learning materials results in significant differences in students' learning behaviors, especially in the access to learning resources. By assessing the all-around performance of students during their studies, visual analytics techniques can help teachers analyze how students access different learning resources, and support teachers in making educational interventions. In addition, for a certain type of learning resources, the diversity of learning behaviors (e.g., pause, play, and backward) make the educational data prohibitive to be analyzed by brute force. Visual analytics techniques could provide intuitive interfaces and visual aids to promote the information extraction from this kind of data.

2.2.2. Learning content analysis

Visual analytics techniques could be used to explore the impact of the content on learning behavior and performance. By exploiting the information conveyed by visual analytics systems, instructors can better evaluate the quality of learning content, promote students' reflection, stimulate students' learning initiative, and further enhance the course content without having expertise in data mining or statistical approaches.

2.2.3. Analysis of interaction among students

Student interaction is an effective way for students to seek help and improve student relationships. Taking into account the interaction data properties (which could be structured or unstructured), visual analytics can be utilized to depict students' interaction, understand the students' characteristics, identify student groups and compare the difference in evolution patterns.

2.2.4. Prediction and recommendation

Educational data recorded by the online learning platform provides sufficient training datasets for prediction. Visual analytics techniques can be used to facilitate the explanation of behavior prediction and explore the correlation between prediction performance and learning behaviors. In addition, it can help teachers review the deficiencies in the learning resources, and help students quickly locate the designated knowledge points. Herein, the objective of visual analytics techniques is to recommend appropriate learning resources and efficient learning paths from various learning materials.

3. Learning behavior analysis

Online learning platforms provide many educational resources (e.g., videos, wikis, questions, and various learning materials) to support knowledge acquisition. One of the main objectives of online education is to improve learners' academic achievements and satisfaction level. To this end, it is crucial to conduct the analysis on the utilization of learning resources. One common way is to use learning behavior data recorded by learning platforms and evaluate the influence of a diverse range of behaviors on the final performance. By exploring the connection between learning behavior data and the achievement, it is beneficial for teachers to make appropriate learning interventions and improve curriculum designs. However, the diversity of the source, the granularity and the multi-modality of data highly increase the complexity of finding out key features relevant to the knowledge acquisition. To tackle this issue, researchers resort to visual analytics techniques to facilitate the analysis, mainly from the following two aspects. At a low level, researchers conduct a detailed analysis on a certain type of learning resource. To explore how learners use the learning material, a deep/comprehensive investigation on the learning behavior has been widely developed. At a high level, researchers focus more on the aggregate/sequence behavior of learners from a summative view, by analyzing time-series data that represents the access of multiple learning materials. For instance, researchers conduct a thorough analysis on the time management of different learning activities to improve the curriculum design.

3.1. Learning behaviors on various types of educational resources

Online learning platforms provide a variety of educational resources for learners. However, the diversity also induces multiple behavior patterns (e.g., access patterns to educational resources) to be discovered. Taking the large size and complex structure of educational data into account, it is prohibitive to analyze behavior patterns by conventional mathematical tools. Instead, visual analytics techniques could explore behavior patterns via event sequence analysis. Through visual analytics methods, teachers can understand the interaction mode between students and various educational resources, flexibly adjust educational resources, and timely intervene in students' learning process. By exploiting the access information to plenty of educational resources, existing works mainly identify learning behaviors from the following two aspects: aggregate analysis and sequence analysis.

As for aggregate analysis, researchers use interactive visual analytics to exhibit the aggregate information of observed events to analyze the activity trend (Dernoncourt et al., 2013) or temporal pattern (Aguilar et al., 2009). Li et al. (2017a) presented an intelligent solution that provides interactive tools to help instructors analyze the learning progress. They mined two learning indicators and visualized the results into the scatter plot. Instructors can understand students' learning progress from multiple scales and discover the students at risk. Calendar charts (Li

Table 1
Categories of visual analytics techniques for online education and representative works in each category.

Goals		Representative Visualization Techniques(Data Types)	Papers
Learning behavior analysis	Learning behaviors on various types of educational resources	Learner trajectory networks chart,tree-like structure bar chart, sunburst chart(user's behavior data)	Li et al. (2017a),Dernoncourt et al. (2013), Aguilar et al. (2009), Li et al. (2015, 2017b), Bueckle and Borner (2017), Chen et al. (2019), Guo et al. (2021), Poon et al. (2017), Ho and Yao (2018), Chen et al. (2018), Mu et al. (2019), Ginda et al. (2019), Coffrin et al. (2014), Davis et al. (2016)
	Learning behaviors on one type of educational resource	Glyph-embedded Sankey diagram,statistical chart,map chart (user's behavior data)	Shi et al. (2015),Xia et al. (2020c),Xia et al. (2020b),Wachtler et al. (2016), Chen et al. (2015), He et al. (2018, 2019b,a), Tsung et al. (2022), Wang et al. (2017), Xia et al. (2020a), Charleer et al. (2013), Li et al. (2021)
Learning content analysis		Keyword cloud,node-link diagram,concept map (original learning source)	Huang et al. (2017)-Wang et al. (2020)
Analysis of interaction among students	Interactions behavior analysis	Parallel coordinate graph,ThreadPulse,calendar graph (user's behavior data)	Fu et al. (2016),Kim et al. (2014),Rei et al. (2017), Wu et al. (2016), Wong (2018), Fu et al. (2018a,b), Wang and Oard (2009), El-Assady et al. (2018)
	Interaction content analysis	ThemeRiver,horizontal stacked bar chart,scatter graph (user's behavior data)	Liu et al. (2018a)-Zheng et al. (2018)
Prediction and recommendation	Prediction of the learning behavior and performance	Discrete graph,Sankey graph,timeline chart (user's behavior data)	Okubo et al. (2015)-Deng et al. (2019)
	Learning resources recommendation	Interactive knowledge concept map,zipper-like learning diagram, tag cloud(original learning source)	Zhao et al. (2018),Schwab et al. (2016),Xia et al. (2019)

et al., 2015) or heat charts (Li et al., 2017b) are used to reveal the temporal patterns of students access information. Bueckle and Borner (2017) presented an interactive heat map analytics dashboard to help instructors explore students' temporal engagement and performance. Recently, more complicated systems with advanced visual analytics techniques are proposed to promote the learning behavior illustration. To help users collect information pieces, Chen et al. (2019) proposed a narrative slideshow system to organize the narrative of information pieces into a data story that helps users explore potential learning patterns. The guided-tour concept is used to guide users to select learning analytic topics, and an interactive drill-down path is presented to guide users to explore specific learning elements or learner groups they are interested in. After exploring the detailed information of different components, users can complete various analysis tasks related to video, forums, assignments, overall action, and learners. The comprehensive system helps users to better understand the whole process of student learning.

As for sequence analysis, since the sequence reflecting students' behavior could be quite long, it is a challenge for teachers to explore the long sequence with classical statistical tools. Current visual analytics works mainly use two methods to explore the behavior sequence pattern: sequential pattern discovery and state transition. Sequential pattern discovery focuses on the sequential relationship between consecutive events (Guo et al., 2021). By analyzing the sequential pattern, teachers can explain the hidden relationship between different learning units and adjust course materials to facilitate students' learning. Frequent pattern mining methods are used to analyze sequential data. Numerous visualization methods are employed to represent the result of sequence pattern mining methods. For example, Poon et al. (2017) employed sequential pattern mining techniques to discover the frequent sequence pattern of students in the log data exported from LMS. As seen in Fig. 1 (1), the results are represented through hierarchical clustering and sunburst visualization, which helps users lower the difficulties in understanding the learning behavior patterns. Ho and Yao (2018) used a tree-like structure bar chart (see Fig. 1 (2)) to show the common navigational patterns mined by integrated sequence analysis methods. Instructors can identify the correlation between navigational patterns and learning outcomes. Chen et al. (2018) introduced a system called Viseq to identify learner groups and understand the reasons behind the learning behavior in MOOC data. They used a novel chord diagram to visualize the learning sequence. Instructors can explore different student groups and discover the correlation between learning sequences and performances. Even though the above-mentioned visual analytics techniques have shown to be efficient in pattern analysis, the identification of abnormal behaviors from various patterns remains to be studied. To fill this gap, Mu et al. (2019) developed a system that

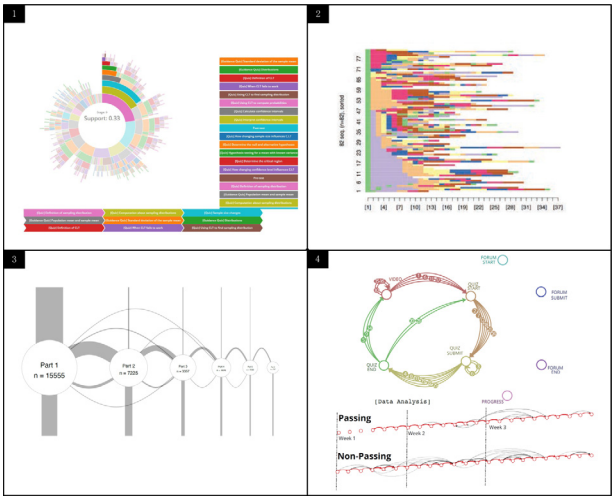


Fig. 1. Visualization of learning behaviors on various types of educational resources. (1) Poon et al. (2017) used a sunburst chart to show the learning pattern. (2) Ho and Yao (2018) used a tree-like structure bar chart to show the common navigational patterns. (3) Coffrin et al. (2014) use a state transition to how students move between different learning states. (4) Davis et al. (2016) use the network diagram and the arc diagram to represent the transfer of students between different types of access events and videos.

helps teachers find students with abnormal behaviors and explore their learning activities in MOOC data. The system utilizes an anomalous detect algorithm to identify the anomalous group and discover their frequent pattern for understanding their behavior. Then they used a set of coordinated statistical views to help instructors analyze the anomalous behavior between and within student groups.

Another manner of conducting sequence analysis is to recognize state transition patterns when students have the access to different learning resources. Ginda et al. (2019) proposed a system to analyze learner engagement, performance, and course trajectories in logged events data. They used a tree chart and a chord chart to analyze the course structure and patterns of learner interactions with course materials, activities, and assessments. Besides, a learner trajectory networks chart is designed to represent the learning path. This work helps teachers optimize the order of curriculum design and formulate effective curriculum intervention strategies. Coffrin et al. (2014) investigated student's learning progress in two MOOCs with different course structures to understand learning behavior patterns in these courses (see Fig. 1 (3)). A state transition diagram represents how students move between different learning modules, which helps instructors understand and compare students' engagement patterns in two courses. Similarly, Davis et al. (2016) proposed an approach to visualize the learning path. As seen in Fig. 1 (4), they used an arc diagram to represent the order in which they watched the video. They also used a network graph to show the state transition pattern of different event types.

3.2. Learning behaviors on one type of educational resource

In addition to a comprehensive exploration of access information to various educational resources, a more detailed behavior analysis on a given type of educational resource is also required. Visual analytics techniques are widely used in the analysis of learning behavior for a given resource. In this part, visual analytics tools used in analyzing video resources and online problem-solving resources are summarized, respectively.

Furthermore, we note that online classrooms cannot achieve the effect of real-time interactive feedback in face-to-face classrooms, visual analytics is exactly based on interactive visual interface for analysis and reasoning, to make up for the lack of interactive gap. So we will briefly introduce some studies on how to perceive students ' participation from facial expressions, emotional characteristics and other aspects to increase teachers ' understanding of students' learning.

3.2.1. Video resources

Video is one of the primary forms of online education. The online platform records the video click stream data. Each data includes the timestamp of students' clicks, the location in the video and the type of click events, such as play or pause. By using visual analytics tools to analyze the video click stream mode, teachers can gain insight into students' behavior of watching video and thus improve the video quality. By exploiting video clickstream data, Wachtler et al. (2016) visualized a violin plot showing the students' status and the delay of response time to the multiple-choice questions, and analyzed students' video learning behavior patterns and identified particular parts in videos. As illustrated in Fig. 2 (1), Shi et al. (2015) developed an analytics system to help instructors understand how students behave on watching videos. They used an event graph to show the distribution of click events and a novel seek graph to show the jump pattern in videos. The correlation analysis of the two views can determine the parts of the video that students pay attention to or ignore, which is helpful for teachers to improve the video content in the future. Chen et al. (2015) studied peak properties in MOOC video clickstream. They proposed an analytics system called PeakVizor to help instructors analyze the peaks in video clickstreams in MOOC (see Fig. 2 (2)). They first applied a peak detection algorithm to detect the peaks in video clickstreams, and then designed a glyph diagram to display valuable statistics of peaks. Finally, they presented a flow view and a correlation view to help instructors analyze the spatio-temporal information of different peaks and the correlation between different learner groups and the peaks. Wachtler et al. (2016) presented a system to monitor the active pattern within videos of a MOOC. They analyzed the behavior of students answering multiple-choice questions and communicating with teachers in the video. A series of statistical charts present reaction delay patterns to multiple-choice questions and video drop patterns.

Another common method to discover learning behaviors is investigating the video resource utilization. He et al. (2018) recently studied the utilization pattern of video resources. They first presented a system called VUSphere (He et al., 2018) to analyze the video utilization pattern and the temporal pattern of watching the video in online distance education. VUSphere provides a spherical layout to depict the distribution of video usage in all courses. They used map charts to display a course's video utilization pattern or an individual student in a learning center. A calendar map is used to reveal the temporal pattern of students learning videos. By designing a video viewing calendar, they presented a system called VUC (He et al., 2019b) to help students check their viewing progress for all videos. He et al. (2019a) also provided a system called LearnerVis to analyze the time management pattern of students learning videos. A scatter chart is implemented to display the overall engagement distribution of all students, and classify different learner groups by comparing a set of box charts across multiple indicators. Moreover, the system provides calendar charts for instructors to explore the overall or detailed time management pattern of the learning video process.

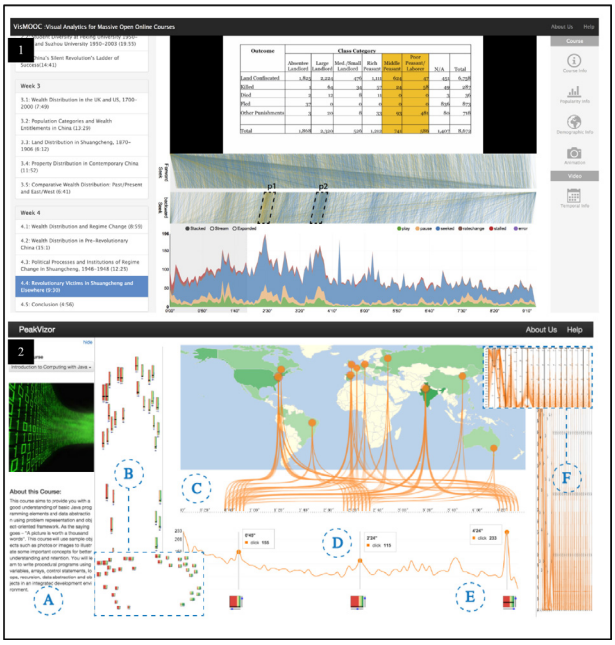


Fig. 2. Two visual analytics systems for clickstream data in video. (1) VisMOOC aims to help instructors understand students' learning behavior by watching the video (Shi et al., 2015). (2) PeakVizor aims to analyze peaks in video clickstreams in MOOCs from different aspects (Chen et al., 2015).

3.2.2. Problem-solving

Online problem-solving behavior refers to students solving problems on the online platform and obtaining the final results. On the one hand, students can consolidate their knowledge by solving problems online; On the other hand, teachers can check students' learning effects through online examinations.

Visual analytics techniques allow researchers to discover students' problem-solving behavior patterns. It can help instructors understand the logical and cognitive changes in the process of students' problem-solving to improve the problem design and provide better guidance to students. Existing works can be divided into two categories based on the data types: multi-step questions problem-solving behaviors pattern and problem submission pattern.

For multi-step questions, researchers analyze the detailed process of students' problem-solving, judging whether students' thinking logic conforms to the design intention of the topic, and obtain insight into problem improvement. It is also conducive to finding out students' difficulties in solving the problems so that teachers can provide timely guidance. For example, Xia et al. (2020b) recently developed two advanced visual analytics systems to analyze the fine-grained problem-solving behavior data that includes the trajectories, the drag/click actions of the mouse, time spent on questions and grades, etc. One is QLens (Xia et al., 2020b), which helps question designers analyze the problem-solving trajectories in multi-step questions. They modeled problem-solving behavior as a hybrid state transition graph. As seen in Fig. 3 (1), a novel glyph-embedded Sankey diagram is employed to visualize the model result. Question designers can get insight into students' problem-solving logic, engagement, and encountered difficulties. Tsung et al. (2022) proposed BlockLens, a novel visual analytics system to assist instructors and platform owners in analyzing students' block-based coding behaviors, mistakes, and problem-solving patterns. BlockLens allows the interactive exploration of student coding behaviors at multiple levels of detail. A student checkpoint warning plot and two distribution plots used to show how students are distributed based on performance metrics. A Sankey-based visualization summarized the paths and snapshot transition information of multiple students.

Researchers can understand students' cognitive ability for problem submission behavior and find students' areas of confusion by discovering students' submission patterns. Wang et al. (2017) adapted the snapshot method to collect data each time students run the code when they solve the programming problem. They used a Sankey chart to identify the state transition pattern in students' problem-solving steps and discover where students are confused. Xia et al. (2020c) developed a system, called SeqDynamics, to analyze students' cognitive and non-cognitive patterns in the process of solving a series of problems of students over time. They visualized the dynamics problem-solving behavior through multiple coordinated views(see Fig. 3 (2)), which helps instructors evaluate students' behavior and provide personalized guidance.

One of the important reasons to conduct analysis on the problem-solving behaviors is to promote students' reflection. Some studies promote student self-reflection and make positive feedback by visualizing the learning information of their peers. For example, there is a bad phenomenon called game the system, which means students directly get the answer rather than think about how to solve the problem. To persuade students not to game the system when solving online programming problems, Xia et al. (2020a) used visualization methods to convey the reasons for not gaming, which is conducive change students' learning behavior and attitude. As seen in Fig. 3 (3), students can compare their learning behaviors or processes with their peers, which promoted students' self-reflection. Charleer et al. (2013) proposed a visualization of badges that provides the detailed learning progress to promote group reflection. As seen in Fig. 3 (4), through the diagram, students can recognize the gap with their peers and motivate themselves to learn hard.

At last, because of the outbreak of Covid-19, online examinations have been developed to facilitate the completion of various examinations. However, due to the lack of face-to-face interaction, there is a big challenge for teachers to detect whether students cheat or not. Even if online exams are accompanied by open cameras, it is still difficult to find out cheating events. To address the challenge, Li et al. (2021) proposed a visual analytics approach to facilitate the proctoring of online exams (see Fig. 3 (5)). They detected the abnormal head and mouse movements and visualized them with other related information from different levels, which supports teachers to identify the risk location quickly, and judge whether there is cheating by viewing the original video.

4. Learning content analysis

Apart from learning behavior analysis, the content analysis is also useful for enhancing the quality of online courses. Taking the large amount of information contained in the online education materials into account, it is necessary for both instructors and learners to understand the degree of knowledge acquisition and assess the quality of learning content. Due to the complicated structure of online educational data, it is time-consuming to check the learning content by brute force. Visualization analytics techniques have shown their efficiency in knowledge exploration and concept extraction, and is beneficial to facilitating the content analysis and enhancing the curriculum design.

Video is the main resource in online education that enables professional instructors to share their knowledge at a large scale. Existing efforts help instructors analyze the video content mainly from three aspects: text, action, and voice. Huang et al. (2017)

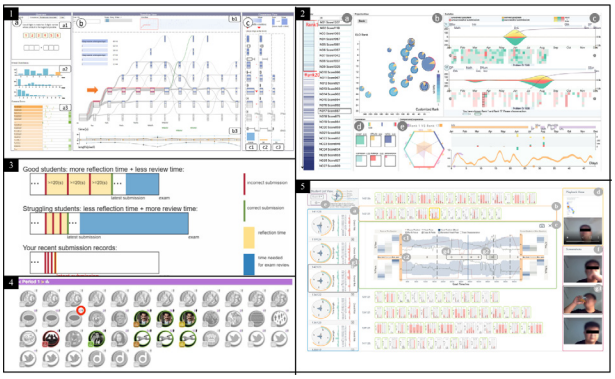


Fig. 3. Five examples for problem-solving. (1) QLens, a visual analytics system that analyzes students' multi-step problem-solving behavior in terms of their problem-solving logic, engagement, and difficulties encountered (Xia et al., 2020b). (2) SeqDynamics, a visual analytics system, evaluates the online problem-solving dynamics from cognitive and non-cognitive perspectives (Xia et al., 2020c). (3) A visual approach that helps students compare different learning patterns of students with different performances (Xia et al., 2020a). (4) A visualization of badges that shows students their learning progress (Charleer et al., 2013). (5) A visual analytics system that monitors students' engagement when they attend an online exam (Li et al., 2021).

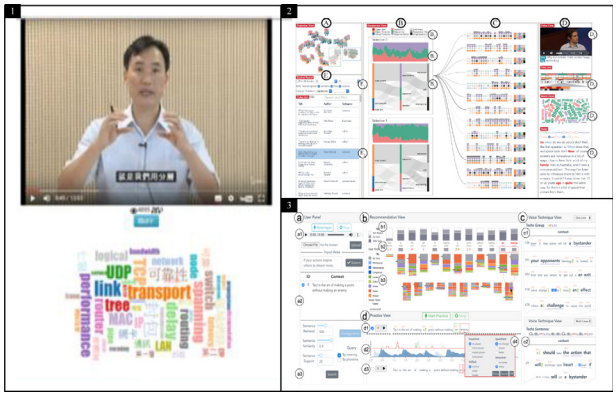


Fig. 4. Three visual analytics systems that analyze the content, posture, and voice, respectively. (1) VideoMark, a visual analysis approach that enables learners to review concepts in MOOC videos and quickly locate the position of concepts in videos (Huang et al., 2017). (2) A visual analytics system that helps users explore the speaker's content and behavior in TED Presentations (Wu and Qu, 2018). (3) VoiceCoach, a visual analysis approach that helps users train their voice modulation skills (Wang et al., 2020).

presented an approach that allows learners to quickly view learning concepts in videos. As seen in Fig. 4 (1), a keyword cloud is employed to visualize the content of the video. They divided the video into several segments based on the students' interaction in the video. When users select one keyword from the keyword cloud, the video will jump to the corresponding video clip. Liu et al. (2018b) presented a system called ConceptScape, which uses a concept map to help users locate parts of the video linked to specific concepts. A node-link diagram is used to construct the concept map. Each concept has a time anchor that links to a specific time point of the video. Thus, users can quickly explore the content of the video and locate the time point of the video of one concept. In addition to the text information in the video, some works analyze the posture and voice of the speaker, which help instructors rethink the content and further improve their teaching level. Wu and Qu (2018) presented a system for users to explore the presentation techniques used in TED Talks videos (see Fig. 4 (2)). They analyzed multimodal content in video collections, including frame images, text, and metadata. After resorting to computer vision and natural language process methods to capture presentation techniques, they used a projection view, Sankey charts, and matrix diagrams to reveal and compare the pattern of different groups using these presentation techniques. Wang et al. (2020) presented a system to analyze the voice modulation skills of speakers in TED Talk videos, which helps speakers effectively train their voice modulation skills (see Fig. 4 (3)). The system provides a panel view for users to submit a query sentence via audio or text input. Then the system will recommend a good voice modulation example in the recommendation view. Finally, a line chart provides real-time and quantitative feedback, which helps users practice their voice modulation techniques.

5. Analysis of interactions among students

Interactions among students refer to students' collaborative discussion, which helps students acquire knowledge and enhance friendship with classmates. A deep analysis of interactions, especially with visual analytics techniques, can lead to a better understanding of online education assessment. Learners' interaction frequently varies with time, and also highly depends on the personality of participants. Assisted by visual analysis, the representation of the interaction data could be more intuitive and easy to be interpreted, which allows instructors to identify the general characteristics of interactions and the rationale behind it. There are mainly three types of interaction in online education: forum, team conversation, and time-anchored comments in the video. First, in the forum, different students can discuss the same topic by posting and replying. Forum data is network structure data, including the sender, receiver and content of the post. Second, team conversation refers to students discussing behavior in a chat room. Team conversation data is a kind of time series data, including time, student ID and conversation content. Third, time-anchored comments in the video refer to the comment with a timestamp when students watch the video. Video comment data belongs to Spatio-temporal data, and each data records the location and time in the video. We review the use of visual analytics techniques to explore interactions from the following two aspects: interaction behavior analysis and interaction content analysis.

5.1. Interaction behavior analysis

Analyzing students' interaction behaviors helps teachers find out different groups of students and students who actively or passively participate in the discussion. The current visual analytics works of online interaction focus on three kinds of interactive data: forum data, conversation data, and video comments. These interactive data are divided into two categories: structured interaction data and unstructured interaction data. Forum data is structured interaction data, a network structure containing the sender and replicator of the post. Unstructured data include conversation data, time-anchored comment data in video, and discussion data. For example, there is only the speaker information but no listener information in the conversation data.

For forum data, due to a large number of students and complex network structure, it is difficult for teachers to perceive the interaction between students. Some works use the node-link diagram to reveal the relationship among students (Rei et al., 2017). Wu et al. (2016) proposed a visual analytics system to help instructors understand the interaction of different student groups in the MOOC forum. They used the parallel coordinate graph to filter different groups of students, and the node-link graph to show the interactions and topics of the selected group. Wong (2018) developed a system called Messagelens that supports the

multifaceted analysis of forum discussion, including students' general discussion topics, the student's attitude, and the communication networks. This system also supports analysis at different levels, including global, groups and individual. A network view is used to understand the interaction among students. Users can uncover the sub-group related to the selected topic, and analyze students' attitude attribution in the sub-network. Some works focus on analyzing the user groups on the forum. Fu et al. (2018a) presented a visual analytics system called VisForum, which helps instructors explore, compare, and track student groups in a forum. They transformed the forum post data into reply sequences and proposed a group detecting sort algorithm to cluster posts of the same groups. A novel group glyph design is provided to present group details information, including group size and relative frequencies of group occurrence and user activeness. The system supports quickly identifying multiple student groups and comparing the difference in evolution patterns among these groups. iForum (Fu et al., 2016) is a visual analytics system that helps instructors discover and understand the dynamic patterns of MOOC forums. They used a set of novel visualizations to present information about posts, users, and threads in MOOC forums. The system can help users comprehensively understand the temporal pattern of student interaction.

For conversation data, there are two challenges in analyzing conversation data. One is how to disentangle interleaving conversational threads in discussions. The other is how to judge the beginning and the end of a conversation. Fu et al. (2018b) proposed a system called T-cal to analyze team communication and collaboration on the modern team messaging platforms. They used a conversation disentangling method (Wang and Oard, 2009) to split interleaving conversations into threads. Then they used a novel ThreadPulse design to show the overall temporal trends of a conversational thread and the detailed information of the message. A graph combining the calendar graph and packing graph is used for users to explore the relationships among team members, messages, and conversational threads from different levels. El-Assady et al. (2018) presented a system to reveal the untangled reply chains in massive online conversations and verbatim text transcripts. A thread diagram is designed to represent and compare the reply chains generated by different models. The system also supports analyzing the decision spaces to understand the inner-work of models, including all considered candidate relations.

5.2. Interaction content analysis

Interaction content refers to the text information generated during the discussion of students. By analyzing the interaction content, instructors can recognize topics attracting more attention and confusing parts of the course. Researchers analyze the content of interaction mainly from emotions and topics. A common analysis process uses the text mining method to obtain topics and emotions, and visualize the results to enhance teachers' understanding of students' interactive content. Liu et al. (2018a) presented a node-link graph to show the social network and encoded the color of nodes to discover the emotional distribution of students on the forum. Sung et al. (2016) provided a system to analyze the emotion and topic evolution of time-anchored comments in online educational videos. As seen in Fig. 5 (1), a river diagram is employed to reveal the evolution of the topic. They presented a river diagram and a novel timeline view to represent the evolution of topics of users' time-anchored comments. Wong et al. (2016) and Atapattu et al. (2016) used the LDA topic mining method to obtain the discussion topics of the forum. Wong et al. (2016) projected topics into a scatter graph according to their similarities. The size of the point represents the heat of the topic.

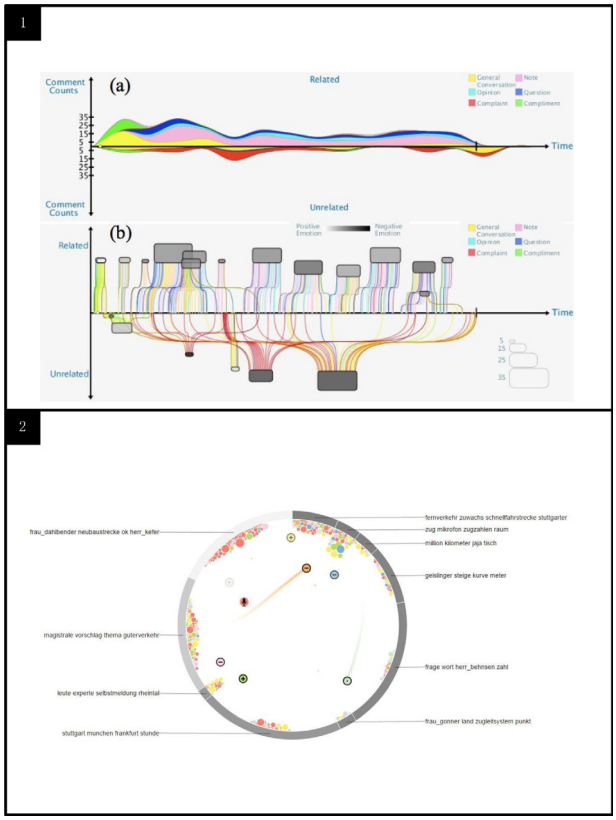


Fig. 5. Visualization examples for interaction content analysis. (1) A ThemeRiver and a timeline view to show learners' time-anchored comments on a philosophy course (Sung et al., 2016). (2) A topic-space view that shows the movement of the speaker's topics in a conversation (El-Assady et al., 2016).

When the user selects a point, a bar chart will update to show all the words under the selected topic. El-Assady et al. (2016) proposed a visual analytics approach for political science scholars to explore the dynamics of a conversation over time. As seen in Fig. 5 (2), they presented a novel radial plot to represent the detailed information of all topics of a conversation.

In addition, the information in forums or blogs could be overloaded so that desired messages are hard to be extracted. To tackle this issue, Hoque and Carenini (2014) presented a visual text analytic system to help users quickly explore the blog data. Several horizontal stacked bar charts are stacked vertically to show the summary of the whole conversation. Each stacked bar represents a comment and encodes comment length, position in the thread, and depth of the comment within the thread and the text analysis results. Moreover, Zheng et al. (2018) presented a system to analyze the students' knowledge elaboration in groups' online discussions. Knowledge elaboration refers to how learners organize, restructure, interconnect, and integrate knowledge, thereby promoting knowledge acquisition and retention. They used a key-term-based automated analysis approach to measure the level of knowledge elaboration in terms of three indicators: coverage, activation, and equitability. A radar graph is used to overview all indicators of knowledge elaboration. They used the node-link graph, heat graph, bar chart, and scatter plot to represent the group discussion process's coverage, activation, equitability, and time series. The system helps teachers monitor students' online discussions and provides more targeted feedback.

6. Prediction and recommendation

The large-scale and multi-dimensional data recorded by the online learning platform provides sufficient training datasets for the prediction of students' learning behavior and performance. After discovering the overall behavior patterns, it is possible to predict students ' learning behavior and learning performance. Moreover, necessary actions can be adopted to improve teaching quality and appropriate learning resources can be recommended for target learners to enhance their knowledge acquisition.

6.1. Prediction of the learning behavior and performance

Due to the high complexity of online educational data, there are two main challenges for the behavior prediction. First, the current prediction model is challenging to achieve very high accuracy. Thus, how to validate the result of prediction is a problem. Second, it is difficult for teachers to determine the factors that affect students' performance. In order to verify the accuracy of prediction results and understand the reasons behind students ' behavior and performance, we can better explain the results of behavior prediction with visual analytics techniques, and then explore the relationship between prediction performance and behavior.

Some works use visual analytics techniques to interpret results of behavior prediction. Okubo et al. (2015) presented a system to explore the learning process of students and prediction results. They used a discrete graph to visualize the four types of learning logs stored in the learning management system, including attendance, time spent for browsing slides, submission of a report, and quiz score. In the graph, nodes are arranged into a matrix. Each row represents a state, and each column represents a course. The line between nodes represents the transition of the active state. The last state represents the predicted result. Later, they presented a system (Okubo et al., 2017) that uses a Sankey graph as the state transition graph to visualize the learning progress together with predictions of the final grades of students and following learning activities. This system enables teachers to understand the features of students' learning activities and identify the students at risk. Chen et al. (2016) proposed a system called DropoutSeer that allows instructors better understand the reasons for the dropout behavior. They used MOOC log data to predict dropout behavior and grouped learners based on the dropout prediction results. Then they used a timeline chart and a flow chart to show the clickstream behavior, performance over time, and learners' posts on each group's forum (see Fig. 6 (1)). The system helps instructors infer the reason for dropout and identify crucial features that significantly affect the prediction results. In order to discover the crucial factors of dropout, Schaffer et al. (2016) used a bar chart to show the correlation between dropout and features, which helps instructors discover which behavior features are highly associated with dropout.

Another important research interest is to explain the correlation between predicted performance and behavior. He et al. (2019c) presented a system called LearnerExp, for visualizing the temporal patterns of learning activities. They used a convolutional neural network model to predict the grade. Instructors can comparatively explore the temporal activity patterns of learner groups with the result of the different predictions of learning performance, and explain the reason behind the prediction result. Deng et al. (2019) proposed the visual analytics tool Performancevis to study the correlation between assignments and exams, and using machine learning techniques to predict student performance. Performancevis helps teachers identify inconsistencies between expected difficulty and students' perceived difficulty and adjust lesson plan design accordingly. Performancevis can also help administrators identify students at risk of failing a course early and measure whether specially designed courses are meeting their goals.

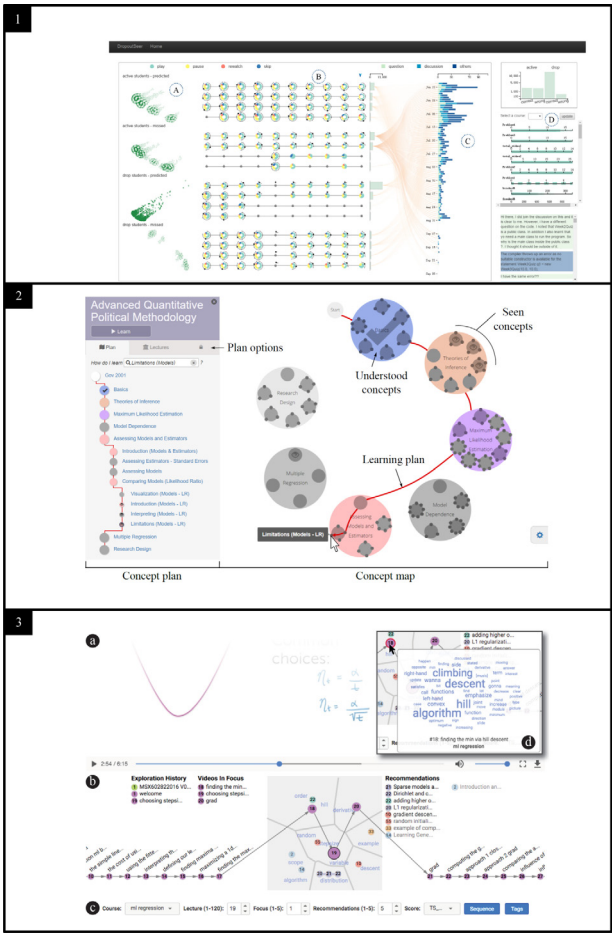


Fig. 6. Visualization examples for prediction and recommendation. (1) DropoutSeer, a visual analytics system to explain the dropout reason behind the learning behavior (Chen et al., 2016). (2) booc.io, a visual analytics system that uses a hierarchical concept map to represent learning materials and supports the exploration of non-linear learning paths (Schwab et al., 2016). (3) MOOCex, a visual recommendation system for users to explore the recommendations of lecture videos (Zhao et al., 2018).

6.2. Learning resources recommendation

Online learning platform provides large-scale learning resources, such as video courses and online programming problems. Students need to choose their own learning resources from the vast amount of learning resources, and flexibly arrange learning routes according to their actual situation through the syllabus provided by the platform. But as a novice, it is difficult to choose the most efficient learning path. In order to recommend more reasonable learning resources and plan better learning paths, visual analytics is utilized based on certain association rules.

Some efforts provide an interactive knowledge concept map for students to customize their learning path. Schwab et al. (2016) presented booc.io, which used a knowledge concept map to present the linear and nonlinear navigation of educational concepts and materials (see Fig. 6 (2)). They used a packaging diagram to represent the hierarchical structures of materials. When the user selects a material in the map, the concept map will display a learning path to guide the user from the beginning of the course to the selected materials. However, the learning resources contained in the concept map are limited due to interactivity deficiency. To this end, some works support the interactive recommendation of learning resources. Zhao et al. (2018) developed a recommendation system called MOOCex. MOOCex recommends

videos according to the content and sequential inter-topic relationships. As seen in Fig. 6 (3), they used the node-link graph to present the recommended videos. Each node represents a recommending video. The link between nodes indicates that it conforms to the order of the syllabus. When the user hovers a video, a tooltip with a tag cloud shows the content of the video. Users can interactively explore the MOOCex to understand the content of different recommendation videos, which helps them decide to watch which video next. In addition, the online question pool provides a large number of questions, which makes it difficult for students to know the hidden knowledge points behind each problem and find the appropriate problem. To solve this problem, Deng et al. (2019) presented a visual analytics system called PeerLens, which helps students peer-inspired plan their learning path in online question pools. The system can recommend a learning path based on the history of peers. They used radar charts to show the typical peer learners' learning attributes and allow students to customize their learning scenarios. Then students can explore details of recommended learning paths with different difficulties in this scenario through a novel zipper-like learning diagram and decide which questions to do next.

7. Research opportunities

In recent years, visual analytics research for online education data has received extensive attention and obtained many results, but many works are still worthy of further investigation. In this section, we discuss some promising research opportunities.

First, due to the diversity of learning resources, the online educational data is large, high-dimensional (with many event types), and heterogeneous. Even though the visualization for unimodal data is prevalent, the multimodal property of educational data poses a challenge in visual representations. For example, to analyze and extract key features of online learning resources that might cause the sudden change in learner's clickstream data, the visual representation of video, audio and text data could provide insights to reveal the inherent connections between these features and clickstream data. However, the multimodal property increases the difficulty to provide sufficient visual aids such that instructors can easily infer the main features to cause the sudden change. Regarding the design of visual analytics systems with multimodal data, the tradeoff between the degree of abstraction and the degree of detail, the scalability, the usability, and the interpretability needs to be precisely considered. A framework for analyzing data with complex structures requires further investigation such that it could fit the need of end users.

Second, there is still a gap between identifying learning behaviors and making teaching decisions. Although many visual analytics systems can discover the learning pattern or predict the learning performance, it may still be difficult for instructors to decide how to improve the teaching or whether an intervention is effective. For instance, Aslan et al. (2019) presented a real-time student engagement visual analytics to help teachers understand student engagement in online classrooms. Through the proposed system, teachers can find students' abnormal participation behaviors in time and realize that some teaching decisions should be taken. However, due to the lack of students' prior information, how to make appropriate teaching decisions to improve student engagement requires a deeper investigation. How to provide efficient visual aids to teachers's decision-making process is worthy of further discussion.

Third, most visual analytics research for interaction analyzes the forum data's student relationship, topic, and emotion. However, few works focus on the visualization of group discussions. Group discussion refers to the process in which students communicate and discuss problems in groups under the guidance of teachers. Teachers need to track the discussion content, role allocation, and interaction of the discussion group. They also need to guide students in discussing and evaluating students' performance. Some studies use the visual analytics method to analyze meeting discussions, which brings inspiration for group discussions. Shi et al. (2018) proposed a visual analytics system called MeetingVis that uses a storyline diagram to present the meeting topic to help users recall the content of the meeting and improve the efficiency of discussion. However, in teaching practice, many discussion groups are often carried out simultaneously. Therefore, teachers need to switch between groups to quickly understand the process of discussion. More importantly, they need real-time review, diagnosis, and comparative analysis of multiple groups of discussions. However, in the real-time discussion environment, teachers can only participate in different discussion groups linearly, making it difficult for teachers to switch quickly and supervise at the same time. Therefore, how to help teachers track and compare the topic evolution, role allocation, and interaction of multiple discussion groups in real-time is a meaningful research direction.

Finally, it is a significant analysis task to discover the students' problem-solving logic or thinking logic. Recently HKUST VisLab E-Learning Group (Anon, 0000) presented a survey for online learning of math and computational thinking, which introduced a method to analyze students' cognition using fine-grained behavior data of the problem-solving process. However, how to promote students' self-reflection or improve their metacognitive ability with visual aids is still an open problem. There is an opportunity for researchers to use visualization methods to enhance students' self-regulation ability based on their problem-solving performance.

8. Conclusion

In this paper, we presented a survey of visual analytics techniques for online educational data. Our motivation for this work is to provide an overview of existing visual analytics techniques that have been applied in online education data, which can help researchers quickly understand the related works of visual analytics of online educational data and gain insights for future research. We build a novel taxonomy for existing visual analytics works of online education data. Then we present details of using visual analytics tools in each category and introduce a set of examples. At last, this paper discusses several challenges for online educational data analysis and proposes future research directions. We hope this survey can spur more interdisciplinary work at the intersection of online education and visualization, and promote the development of visual analytics techniques on online educational data.

Ethical Approval

This study does not contain any studies with human or animal subjects performed by any of the authors. All data used in the study are taken from public databases that were published in the past.

CRediT authorship contribution statement

Xiaoyan Kui: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Visualization, Writing – review & editing. **Naiming Liu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Validation, Writing – original draft. **Qiang Liu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Validation, Writing – review & editing. **Jingwei Liu:**

Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Validation, Writing – review & editing. **Xiaoqian Zeng:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Validation, Writing – review & editing. **Chao Zhang:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Supervision, Visualization, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work described in this paper was supported by the grant from the National Natural Science Foundation of China (No. 62177047); The first batch of new liberal arts research and reform practice projects of the Ministry of Education, China (Teaching Hall Letter [2021] No. 31); research Project of Teaching Reform in Colleges and Universities in Hunan Province, China (Xiangjiaotong [2020] No. 232); Hunan graduate education teaching reform research project, China (2020JGZD010); Central South University Graduate Education Teaching Reform Research Project, China (2020JGA007).

References

Aguilar, D.A., Theron, R., Garcia-Penalvo, F.J., 2009. Semantic spiral timelines used as support for e-learning. J. Univ. Comput. Sci. 15 (7), 1526–1545.

Anon, 0000. E-LEARNING. http://vis.cse.ust.hk/groups/e-learning/.

Aslan, S., Alyuz, N., Tanriover, C., Mete, S.E., Okur, E., D'Mello, S.K., Arslan Esme, A., 2019. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12.

Atapattu, T., Falkner, K., Tarmazdi, H., 2016. Topic-Wise Classification of MOOC Discussions: A Visual Analytics Approach. International Educational Data Mining Society.

Bueckle, M.G., Borner, K., 2017. Empowering instructors in learning management systems: Interactive heat map analytics dashboard. Retrieved Nov. 2017 Mar;2.

Charleer, S., Klerkx, J., Santos, J.L., Duval, E., 2013. Improving Awareness and Reflection Through Collaborative, Interactive Visualizations of Badges. ARTEL EC-TEL, Sep;1103:69-81.

Chen, Q., Chen, Y., Liu, D., Shi, C., Wu, Y., Qu, H., 2015. Peakvizor: Visual analytics of peaks in video clickstreams from massive open online courses. IEEE Trans. Vis. Comput. Graphics 22 (10), 2315–2330.

Chen, Y., Chen, Q., Zhao, M., Boyer, S., Veeramachaneni, K., Qu, H., 2016. DropoutSeer: Visualizing learning patterns in massive open online courses for dropout reasoning and prediction. In: 2016 IEEE Conference on Visual Analytics Science and Technology. VAST, IEEE, pp. 111–120.

Chen, Q., Li, Z., Pong, T.C., Qu, H., 2019. Designing narrative slideshows for learning analytics. In: 2019 IEEE Pacific Visualization Symposium. PacificVis, IEEE, pp. 237–246.

Chen, Q., Yue, X., Plantaz, X., Chen, Y., Shi, C., Pong, T.C., Qu, H., 2018. Viseq: Visual analytics of learning sequence in massive open online courses. IEEE Trans. Vis. Comput. Graphics 26 (3), 1622–1636.

Coffrin, C., Corrin, L., Barba, P.de., Kennedy, G., 2014. Visualizing patterns of student engagement and performance in MOOCs. In: Proceedings of the Fourth International Conference on Learning Analytics and Knowledge. pp. 83–92.

Davis, D., Chen, G., Hauff, C., Houben, G.J., 2016. Gauging MOOC Learners' Adherence to the Designed Learning Path. International Educational Data Mining Society.

Deng, H., Wang, X., Guo, Z., Decker, A., Duan, X., Wang, C., Ambrose, G.A., Abbott, K., 2019. Performancevis: Visual analytics of student performance data from an introductory chemistry course. Vis. Inf. 3 (4), 166–176.

Dernoncourt, F., Taylor, C., O'Reilly, U.M., Veeramachaneni, K., Wu, S., Do, C., Halawa, S., 2013. MoocViz: A large scale, open access, collaborative, data analytics platform for MOOCs. In: NIPS Workshop on Data-Driven Education, Lake Tahoe, Nevada, USA. Dec 5.

El-Assady, M., Gold, V., Acevedo, C., Collins, C., Keim, D., 2016. ConToVi: Multi-party conversation exploration using topic-space views. In: Computer Graphics Forum. Vol. 35. (3), pp. 431–440.

El-Assady, M., Sevastjanova, R., Keim, D., Collins, C., 2018. ThreadReconstructor: Modeling reply-chains to untangle conversational text through visual analytics. In: Computer Graphics Forum. Vol. 37. (3), pp. 351–365.

Fu, S., Wang, Y., Yang, Y., Bi, Q., Guo, F., Qu, H., 2018a. Visforum: A visual analysis system for exploring user groups in online forums. ACM Trans. Interact. Intell. Syst. (TiiS) 8 (1), 1–21.

Fu, S., Zhao, J., Cheng, H.F., Zhu, H., Marlow, J., 2018b. T-cal: Understanding team conversational data with calendar-based visualization. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–13.

Fu, S., Zhao, J., Cui, W., Qu, H., 2016. Visual analysis of MOOC forums with iforum. IEEE Trans. Vis. Comput. Graphics 23 (1), 201–210.

Ginda, M., Richey, M.C., Cousino, M., Borner, K., 2019. Visualizing learner engagement, performance, and trajectories to evaluate and optimize online course design. PLoS One 14 (5), e0215964.

Guo, Y., Guo, S., Jin, Z., Kaul, S., Gotz, D., Cao, N., 2021. A survey on visual analysis of event sequence data. IEEE Trans. Vis. Comput. Graphics Jul 27.

He, H., Dong, B., Zheng, Q., Di, D., Lin, Y., 2019a. Visual analysis of the time management of learning multiple courses in online learning environment. In: 2019 IEEE Visualization Conference. VIS, IEEE, pp. 56–60.

He, H., Dong, B., Zheng, Q., Li, G., 2019b. VUC: Visualizing daily video utilization to promote student engagement in online distance education. In: Proceedings of the ACM Conference on Global Computing Education. pp. 99–105.

He, H., Zheng, O., Dong, B., 2018. Vusphere: Visual analysis of video utilization in online distance education. In: 2018 IEEE Conference on Visual Analytics Science and Technology. VAST, IEEE, pp. 25–35.

He, H., Zheng, Q., Dong, B., 2019c. LearnerExp: exploring and explaining the time management of online learning activity. In: The World Wide Web Conference. pp. 3521–3525.

Ho, J.C., Yao, M.Z., 2018. Sequence analysis in distributed interactive learning environments: Visualization and clustering of exploratory behavior. J. Educ. Online (2), n2.

Hoque, E., Carenini, G., 2014. Convis: A visual text analytic system for exploring blog conversations. In: Computer Graphics Forum. Vol. 33. (3), pp. 221–230.

Huang, N.F., Hsu, H.H., Chen, S.C., Lee, C.A., Huang, Y.W., Ou, P.W., Tzeng, J.W., 2017. VideoMark: A video-based learning analytic technique for MOOCs. In: 2017 IEEE 2nd International Conference on Big Data Analysis. ICBDA, IEEE, pp. 753–757.

Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K.Z., Miller, R.C., 2014. Understanding in-video dropouts and interaction peaks in online lecture videos. In: Proceedings of the First ACM Conference on Learning Scale Conference.

Li, X., Men, C., Zhang, F., Du, Z., 2017. A smart visual analysis solution for MOOC data. 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, DASC/PiCom/DataCom/CyberSciTech, IEEE, pp. 101–106.

Li, H., Tsuchiya, T., Suehiro, D., Taniguchi, Y., Shimada, A., Suzuki, Y., Ohashi, H., Ogata, H., 2017. Behavioral Analysis and Visualization on Learning Logs from the CALL Course. The 31st Annual Conference of the Japanese Society for Artificial Intelligence.

Li, H., Xu, M., Wang, Y., Wei, H., Qu, H., 2021. A visual analytics approach to facilitate the proctoring of online exams. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–17.

Li, X., Zhang, X., Liu, X., 2015. A visual analytics approach for e-learning education. In: 2015 9th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. IEEE, pp. 34–40.

Liu, Z., Kang, L., Rüdian, S., Su, Z., Liu, S., Sun, J., 2018a. Dynamics of emotions and network structures in a course forum: An empirical investigation in the last four weeks before the exam. In: 2018 International Joint Conference on Information, Media and Engineering. ICIME, IEEE, pp. 177–182.

Liu, C., Kim, J., Wang, H.C., 2018b. ConceptScape: Collaborative concept mapping for video learning. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–12.

Mu, X., Xu, K., Chen, Q., Du, F., Wang, Y., Qu, H., 2019. MOOCad: Visual analysis of anomalous learning activities in massive open online courses. In: EuroVis. Short Papers, pp. 91–95.

Okubo, F., Shimada, A., Taniguchi, Y., 2017. A Visualization System for Predicting Learning Activities Using State Transition Graphs. International Association for Development of the Information Society.

Okubo, F., Shimada, A., Yin, C., Ogata, H., 2015. Visualization and prediction of learning activities by using discrete graphs. In: Proceedings of the 23rd International Conference on Computers in Education. ICCE, Hangzhou, China, pp. 739–744, November 30-December 4, 2015.

Poon, L.K., Kong, S.C., Yau, T.S., Wong, M., Ling, M.H., 2017. Learning analytics for monitoring students participation online: Visualizing navigational patterns on learning management system. In: International Conference on Blended Learning. Springer, Cham, pp. 166–176.

Qu, H., Chen, Q., 2015. Visual analytics for MOOC data. IEEE Comput. Graphics Appl. 35 (6), 69–75.

Rei, A., Figueira, A., Oliveira, L., 2017. A system for visualization and analysis of online pedagogical interactions. In: Proceedings of the 2017 International Conference on E-Education, E-Business and E-Technology. pp. 42–46.
Schaffer, J., Huynh, B., O'Donovan, J., Höllerer, T., Xia, Y., Lin, S., 2016. An analysis of student behavior in two massive open online courses. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ASONAM, IEEE, pp. 380–385.
Schwab, M., Strobelt, H., Tompkin, J., Fredericks, C., Huff, C., Higgins, D., Strezhnev, A., Komisarchik, M., King, G., Pfister, H., 2016. Booc. io: An education system with hierarchical concept maps and dynamic non-linear learning plans. IEEE Trans. Vis. Comput. Graphics 23 (1), 571–580.
Shi, Y., Bryan, C., Bhamidipati, S., Zhao, Y., Zhang, Y., Ma, K.L., 2018. Meetingvis: Visual narratives to assist in recalling meeting context and content. IEEE Trans. Vis. Comput. Graphics 24 (6), 1918–1929.
Shi, C., Fu, S., Chen, Q., Qu, H., 2015. VisMOOC: Visualizing video clickstream data from massive open online courses. In: 2015 IEEE Pacific Visualization Symposium. PacificVis, IEEE, pp. 159–166.
Sung, C.Y., Huang, X.Y., Shen, Y., Cherng, F.Y., Lin, W.C., Wang, H.C., 2016. ToPIN: A visual analysis tool for time-anchored comments in online educational videos. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. pp. 2185–2191.
Tsung, S., Wei, H., Li, H., Wang, Y., Xia, M., Qu, H., 2022. BlockLens: Visual analytics of student coding behaviors in block-based programming environments. In: Proceedings of the Ninth ACM Conference on Learning and Scale. pp. 299–303.
Vieira, C., Parsons, P., Byrd, V., 2018. Visual learning analytics of educational data: A systematic literature review and research agenda. Comput. Educ. 122, 119–135.
Wachtler, J., Khalil, M., Taraghi, B., Ebner, M., 2016. On using learning analytics to track the activity of interactive MOOC videos. In: SE VBL LAK. pp. 8–17.
Wang, L., Oard, D.W., 2009. Context-based message expansion for disentanglement of interleaved text conversations. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 200–208.
Wang, Y., White, W.M., Andersen, E., 2017. Pathviewer: Visualizing pathways through student data. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 960–964.
Wang, X., Zeng, H., Wang, Y., Wu, A., Sun, Z., Ma, X., Qu, H., 2020. Voicecoach: Interactive evidence-based training for voice modulation skills in public speaking. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–12.
Wong, J.S., 2018. MessageLens: a visual analytics system to support multifaceted exploration of MOOC forum discussions. Vis. Inf. 2 (1), 37–49.
Wong, G.K., Li, S.Y., Wong, E.W., 2016. Analyzing academic discussion forum data with topic detection and data visualization. In: 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering. TALE, IEEE, pp. 109–115.
Wu, A., Qu, H., 2018. Multimodal analysis of video collections: Visual exploration of presentation techniques in ted talks. IEEE Trans. Vis. Comput. Graphics 26 (7), 2429–2442.
Wu, T., Yao, Y., Duan, Y., Fan, X., Qu, H., 2016. NetworkSeer: Visual analysis for social network in MOOCs. In: 2016 IEEE Pacific Visualization Symposium. PacificVis, IEEE, pp. 194–198.
Xia, M., Asano, Y., Williams, J.J., Qu, H., Ma, X., 2020a. Using information visualization to promote students' reflection on gaming the system in online learning. In: Proceedings of the Seventh ACM Conference on Learning Scale. pp. 37–49.
Xia, M., Sun, M., Wei, H., Chen, Q., Wang, Y., Shi, L., Qu, H., Ma, X., 2019. Peerlens: Peer-inspired interactive learning path planning in online question pool. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12.
Xia, M., Velumani, R.P., Wang, Y., Qu, H., Ma, X., 2020b. Qlens: Visual analytics of multi-step problem-solving behaviors for improving question design. IEEE Trans. Vis. Comput. Graphics 27 (2), 870–880.
Xia, M., Xu, M., Lin, C.E., Cheng, T.Y., Qu, H., Ma, X., 2020c. SeqDynamics: Visual analytics for evaluating online problem-solving dynamics. In: Computer Graphics Forum. Vol. 39. (3), pp. 511–522.
Zhao, J., Bhatt, C., Cooper, M., Shamma, D.A., 2018. Flexible learning with semantic visual exploration and sequence-based recommendation of MOOC videos. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–13.
Zheng, Y., Xu, C., Li, Y., Su, Y., 2018. Measuring and visualizing group knowledge elaboration in online collaborative discussions. J. Educ. Technol. Soc. 21 (1), 91–103.

Xiaoyan Kui is a professor in the School of Computer Science and Engineering at Central South University, China. She received the B.Sc., M.Sc., and Ph.D. degrees, all in Computer Science, from Central South University, China, in 2003, 2008, and 2012, respectively. Her research interest includes information visualization, data visualization, visual analytics, mobile computing and vehicular network. She has published more than 40 technique papers in international journals and conferences.

Naiming Liu received the B.E. degree in China University of Petroleum (Beijing) in 2016. He is working toward the graduate degree at the Central South University, Hunan, China. His current research interests include data visualization and human computer interaction.

Qiang Liu received the B.E. degree in Computer Science and Technology from Harbin Engineering University,Harbin , China, in 2019. He is working toward the graduate degree at the Central South University, Hunan, China. His current research interests include data visualization and data mining.

Jingwei Liu received the B.E. degree in software engineering from Xiangtan University,Hunan, China, in 2017. He is working toward the graduate degree at the Central South University, Hunan, China. His current research interests include data visualization and human computer interaction.

Xiaoqian Zeng received the B.E. degree in IoT engineering from Hunan University of Technology, Hunan, China, in 2020. She is studying for the graduate degree at the Central South University, Hunan, China. Her current research direction is data visualization and visual analysis.

Chao Zhang received the B.E. degree in Huazhong University of Science and Technology, Wuhan, China in 2012, and M.Sc., and Ph.D. degree from University of Paris-Saclay in 2014 and 2017 respectively. Before joining Central South University as an assistant professor, he was an postdoc research fellow in CentraleSupelec, Princeton University, respectively. He has published more than 20 papers in international journals and conferences. His current research interests include resource allocation, semantic communication, information visualization and visual analytics.

Visual Informatics 6 (2022) 1–10

Contents lists available at ScienceDirect

Visual Informatics

journal homepage: www.elsevier.com/locate/visinf

MDIVis: Visual analytics of multiple destination images on tourism user generated content

Changlin Li [a], Mengqi Cao [a], Xiaolin Wen [a], Haotian Zhu [b], Shangsong Liu [a], Xinyi Zhang [a], Min Zhu [a,*]

[a] College of Computer Science, Sichuan University, Chengdu 610065, China
[b] Department of Computing, Xi'an Jiaotong-Liverpool University, Suzhou 215000, China

ARTICLE INFO

Article history:
Received 17 January 2022
Received in revised form 21 May 2022
Accepted 16 June 2022
Available online 30 June 2022

Keywords:
Tourism destination images
Visual analysis
Sentiment visualization
User-generated content

ABSTRACT

Abundant tourism user-generated content (UGC) contains a wealth of cognitive and emotional information, providing valuable data for building destination images that depict tourists' experiences and appraisal of the destinations during the tours. In particular, multiple destination images can assist tourism managers in exploring the commonalities and differences to investigate the elements of interest of tourists and improve the competitiveness of the destinations. However, existing methods usually focus on the image of a single destination, and they are not adequate to analyze and visualize UGC to extract valuable information and knowledge. Therefore, we discuss requirements with tourism experts and present MDIVis, a multi-level interactive visual analytics system that allows analysts to comprehend and analyze the cognitive themes and emotional experiences of multiple destination images for comparison. Specifically, we design a novel sentiment matrix view to summarize multiple destination images and improve two classic views to analyze the time-series pattern and compare the detailed information of images. Finally, we demonstrate the utility of MDIVis through three case studies with domain experts on real-world data, and the usability and effectiveness are confirmed through expert interviews.

1. Introduction

According to the cognition-emotion model defined by Baloglu and McCleary (1999), the tourism destination image is separated into cognitive themes and emotional experiences, with the cognitive themes referring to tourists' knowledge and the emotional experiences exposing the emotional tendency (Huang et al., 2021). In addition, tourism user-generated content(UGC) consists mainly of online travel notes and comments on tourism platforms. With the richer and more convenient Internet applications, online travel platforms have become prevalent in people's daily lives, and they have become an important means for tourists to make travel plans and share their experiences. These platforms contain a wealth of user-generated content reflecting visitors' practical sentiments. Similarly, the tourism UGC may play a significant role in portraying the destination image, providing valuable opportunities for tourism research. By constructing and examining multiple destination images with tourists' perceptions, tourism managers can explore the commonalities and differences of multiple destination images to investigate the elements of interest to tourists and improve the competitiveness of the destinations (Agus et al., 2020).

The existing research on destination images based on tourism UGC mainly uses text mining combined with text description for single destination image construction (Sheng et al., 2020; Garay, 2019). Visual analysis of tourism destination images currently has not drawn adequate research attention, with visualization mainly used to present data processing results. Although a few studies use visual analysis for single destination image exploration, studies that focus on destination images comparison are still lacking (Li et al., 2016). Therefore, tourism managers lack specific techniques and tools to process and visualize essential data to extract valuable information and knowledge (Barroso et al., 2020), and the discovery of commonalities and differences between multiple destination images is considered a challenging task in this context.

In this paper, we introduce MDIVis, an interactive visual analytics system for tourism UGC, to assist analysts in exploring and analyzing multiple destination images. From the travel notes and comments, MDIVis extracts the cognitive entities and emotional descriptions, and the cognitive entities that make up the

* Corresponding author.
E-mail addresses: lichanglin@stu.scu.edu.cn (C. Li), zhumin@scu.edu.cn (M. Zhu).

https://doi.org/10.1016/j.visinf.2022.06.001

cognitive image are divided into five categories to facilitate the exploration and comparison of destination images from multiple perspectives. Meanwhile, MDIVis allows users to compare at both overview and detail levels. At the overview level, we design and employ the Sentiment Matrix View to help analysts intuitively compare the characteristics of multiple destination images. Analysts can investigate and compare the temporal evolution and detailed information of multiple destination images at the detail level. To evaluate the usability and effectiveness of MDIVis, we performed case studies with real-world data, followed by interviews with domain experts. The results show that the system can identify the commonalities and differences between multiple destination images, reveal the temporal pattern of images, and explain the elements of tourists' concerns that change the images. The main contributions of our work include:

- A set of novel visualization designs are proposed to support the interactive comparison of destination images.
- A system based on linkage views is developed. The system provides users the exploration of tourism destination images at two levels of overview and detail.
- Three case studies and an expert interview based on real-world data that demonstrate the usability and effectiveness of MDIVis.

2. Related works

Previous studies related to our work can be divided into two parts: techniques for tourism destination image analysis and visualization of user-generated content.

2.1. Techniques for tourism destination image analysis

Destination image refers to a person's overall beliefs, ideas, and impressions of the destination (Woosnam et al., 2020). Researchers primarily use three methods to identify a destination image: questionnaire surveys, UGC mining, and picture recognition methods. Jeng et al. (2019) and Han et al. (2021) conducted destination image questionnaire surveys to assist marketers in studying visitor behavior patterns and developing publicity strategies. However, the destination image research was limited by investigation time and questionnaire design. Tourism UGC data provides a reliable way for destination image research by the characteristics of easy access, popularity, authenticity, and direct participation of tourists. Qi and Chen (2019) classified the collected tourism comment texts and built destination images using analysis software. They analyzed and summarized tourists' attention to various aspects, including destination leisure, culture, etc. Since UGC is usually presented in text, some content analysis methods (LDA, etc.) are often used to transform UGC into a structured topic model, and emotional experience contained in the text can also be extracted easily (Wang et al., 2020; Gkritzali et al., 2018). Furthermore, in recent years, some studies have investigated and broadened the use of artificial intelligence technology in tourism destination research, focusing on the construction of destination images with tourism picture content mining (Zhang et al., 2019; Xiao et al., 2020). Sheng et al. (2020) combined images and text descriptions to construct a tourist destination image. They found that the destination image revealed by text descriptions is clearer than images.

The cognition-emotion paradigm is frequently used in destination image analysis and research, and it has a favorable impact on tourist satisfaction and loyalty (Chiu et al., 2016). At the same time, not all cognitive factors influence tourists' desire to return, and categorizing cognitive entities can make destination image analysis easier (Triantafillidou et al., 2019; Leković et al., 2020). Through quantitative content analysis, Garay (2019) investigated the distribution of cognitive themes and emotional experiences of the destination image. Huang et al. (2021) established a research framework from a cognitive perspective, investigated how cognitive-behavioral characteristics and emotional experiences under various cognitive themes serve destination image and proposed novel suggestions for improving tourist experiences.

The current research focuses on constructing destination images by combining word frequency statistics with text descriptions, but the results are too complex for tourism analysts to recognize and interpret. Furthermore, most studies focus on constructing a specific destination image without considering image contrast and time-series impact.

2.2. Visualization of user-generated content

User-generated content refers to blogs, comments, notes, and other forms that include user experiences, sentiments, and opinions. In this section, we present the current state of visual analytics studies of UGC on social media and travel websites.

Social media is a growing source of user-generated content. Similar to our work to explore the destination images, many researchers focus on the abstraction and construction of hot topics using visual analytics. Knittel et al. (2021) used a clustering method to update the visualization of the topic. They integrated familiar and highly relevant visual metaphors to summarize methods for visualizing details about a specific topic of interest. Troudi et al. (2019) employ visual analytics to undertake multidimensional research of hot events, collecting data from numerous social media sources to identify events that have occurred. Kucher et al. (2020) built a text visualization analysis tool to explore and analyze sentiments and positions in social media UGC. In addition, mining temporal features has been a research focus in UGC visual analysis in recent years. TagNet was created by Chen (2018) for tag-based sentiment analysis. It combines a traditional tag cloud with an upgraded node-link graph to represent the temporal evolution of emotions through simple and intuitive visual expressions. Furthermore, the targeted view design aids users in comprehending the potential information of UGC. Hu et al. (2016) created a visualization approach for unstructured social media text that incorporates word cloud and tree cloud principles, which display keywords in social media and keep the sentence structure of the texts, allowing readers to grasp significant concepts and perspectives rapidly.

A few visual analytics of tourism UGC studies utilize the exploratory power of visual analytics (Kim et al., 2017; Zhang and Koshijima, 2019; Yuan et al., 2016; Xu et al., 2015). Francalanci and Hussain (2015) combined with k-shell analysis theory to propose a novel visual peripheral layer graphical representation to help travel experts explore and analyze the most competitive locations or events in social networks. Li et al. (2016) used visual analytics to investigate the social network relationships and uncover the tendency of hot tourism areas. Cao et al. (2020) suggested a multi-attribute dual-relationship technique to investigate the relationship between knowledge and pictures but did not consider the difference of multiple destination images.

Based on the above work, we focus on the cognitive and affective elements of destination images and design a visual analysis framework to support interactive exploration and comparative analysis of multiple destination images.

3. Scenario and task analysis

To better identify the scenario for the need to investigate the commonalities and differences of multiple destination images, we

invite four domain experts (E1–E4) to gather the requirements and find the design candidate. E1 is a strategy analyst working in the tourism management department. E2 and E3 are product managers of the tourism industry. E4 is a professor at the school of tourism.

In this section, we first discuss and organize analysis tasks with experts, then introduce the research scenario data and formulate the requirements accordingly based on the tasks.

3.1. Task abstraction

Different tourist attractions have different attraction elements, and people perceive them differently. Our overall analysis goal is to explore and compare multiple destination images, which will assist tourism managers in discovering the most interesting destinations for tourists and the competitive elements of each destination. Furthermore, an exploratory analysis approach from overview to detail is well accepted. The characteristics of each destination image in terms of overview need to be intuitively discovered to narrow down the set of candidates of interest. At the same time, the user needs to explore the details of the destination image in terms of its temporal characteristics and the factors that are of widespread interest from multiple perspectives. After a roundtable with tourism experts, we formed the following specific tasks:

T1 Summarize multiple destination images. To investigate the multiple destination images comprised in the dataset, experts need to summarize and examine the general situation of these destination images. Three main areas are included as follows:

- **T1.1 Summarize the overall images of the destination set.** The analysts first need to generalize the overall images of the destination set and discover the distribution of cognitive and affective images.
- **T1.2 Summarize the cognitive themes of the destination set.** Cognitive themes are a form of expression of people's perception of destinations. The analysts need to perceive the richness of cognitive themes of destinations and further analyze the information about each cognitive theme category.
- **T1.3 Summarize the emotional experiences of the destination set.** Emotional experiences point to the emotional tendency of tourists toward destinations, which usually presents a positive or negative state. For example, experts want to know which destination image of landscapes performs more positively than others.

T2 Explore the time-series evolution of destination images. Over the years, multiple destinations have changed in popularity in a competitive environment with each other. Analysts need to analyze and compare the temporal patterns of different destinations with the temporal characteristics of travel UGC data.

T3 Analyze and compare detailed information of individuals. After the overview analysis, the experts can find some destinations of interest as a subset to be dissected to develop a comprehensive understanding of the relevant destination images and compare them. The following requirements are considered:

- **T3.1 Compare the cognitive themes of individuals.** Each destination image contains its exclusive cognitive themes, and comparing cognitive themes is looking for differences in the cognitive entities that make up the cognitive themes. For example, experts want to know the main elements of the attractiveness of some destinations and different values for tourists.
- **T3.2 Compare the emotional experiences of individuals.** Similar to T1.3, detailed comparisons for sentiment analysis are needed to be combined with cognitive themes, and the analysts need to compare specific sentiment descriptions of cognitive entities.
- **T3.3 Compare the perceived/projected image of a single destination.** A single destination usually includes two images, a perceived image based on visitor feedback and a projected image constructed by the official portrayal. The analysts need to compare these two images to understand their differences, which helps propagandists optimize their propaganda strategies.
- **T3.4 Exhibit raw user-generated contents.** In the raw user-generated content, travelers share their perceptions of destination images in detail. Therefore, analysts need to incorporate the complete descriptions of the UGC during the analysis to better comprehend the destination images.

3.2. Data description

Travel UGC is the main source of data for building destination images. Various travel community platforms have emerged in daily life, providing people with rich channels to exchange travel experiences and generate different UGC data forms. One type of UGC data is travel notes information, in which travelers publish travel notes which are long-form content after visiting a city or province, such as www.mafengwo.cn, and www.youxiake.com, etc. The other type is the comments information, which is the short-form content tourists post after visiting a specific site attraction, such as www.ctrip.com, www.tripadvisor.cn, and www.qyer.com. After comparing various travel websites, this work selects travel notes from 'mafengwo' (long text) and comments from 'tripadvisor' (short text) as the primary research data. The travel notes information contains records and experiences of people and objects experienced by tourists during the tour. Each paragraph of travel notes has different description objects, which is a comprehensive embodiment of the image of a destination. The comments contain more explicit time information and describe the visitor's feelings after visiting a destination.

First, we use a network crawler to resolve the UGC data from 2014 to 2020, collecting 560,000 tourism comments and approximately 1.57 million comments. Second, we refer to the type of destination in the 'mafengwo' website and classify destinations into five categories museums, religious sites, city parks, ancient towns, ecological sites, and others, so that users can make an initial selection of destinations based on their preferences, while the commonalities and differences in the images of similar destinations better indicate tourist concerns and potential competitiveness. After that, we remove the deactivation words from the UGC and use the TextRank algorithm to extract nouns and adjectives as cognitive and emotional elements of the destination image. Furthermore, we associate cognitive entities with emotions with textual contexts to help users understand what visitors are really thinking. In inspiration by Beerli and Martin (2004), we divide the cognitive themes into five categories: foods, scenes, landscapes, facilities, and atmospheres. Finally, we calculate the emotional score of key phrases using SNOWNLP, with negative to positive degrees mapped from -1 to 1. This value will be used to visualize visitors' real emotional tendencies towards destinations.

3.3. Design requirements

To address the above analysis tasks, we combine the data characteristics of UGC to formulate the following design requirements.

Fig. 1. MDIVis allows users to investigate and compare multiple destination images on UGC from multiple aspects. Setting panel (A) helps users to roughly filter destinations of interest, the Sentiment Matrix View (B) presents an overview of images of multiple candidate destinations, the Timing View (C) describes changes in the ranking of several destinations over the years, the Keywords Radar View (D) provides detailed information to explore and compare the destination images, and the Auxiliary View (E, F) offer extra information for users.

DR1. Provides an overview of multiple destination images. The system should support effective identification and interpretation of the cognitive themes and emotional experiences of the destination images. The visualization of a destination image encodes its emotional tendency and the richness of the cognitive entities (**T1**).

DR2. Visualize the temporal changes of destination images. The changes of tourist destination images per year should be visualized. The views should provide an overview and comparison of how the destination images have changed over the years (**T2**).

DR3. Compare the detailed information of destination images. In order to discover and comprehend the common and different elements between destination images, the views should show the detailed features of the image from different aspects and support the independent selection of destinations for comparative exploration (**T3**).

4. MDIVis system design

4.1. System pipeline

We design MDIVis, an interactive visual analytics system to explore multiple tourism destination images based on UGC, integrating the requirements and data characteristics. The pipeline of the system is shown in Fig. 2. After acquiring UGC and processing the dataset, we save the formed structured data in the database.

We design and implement a series of visual linkage views, which are combined with rich interactive means to assist users in analyzing and exploring destination images from various perspectives.

The view components and interactions of MDIVis follow an analysis model from overview to detail. The overview level views serve as the entry point for analysis, with the Sentiment Matrix View providing the user with an overview of multiple destination images in the dataset, and the Cognitive View shows comprehensive statistical information on cognitive entities (**T1**). The detailed level views are designed to provide more specific image information to help users compare destinations. Specifically, the Timing View presents the destinations' ranking in recent years (**T2**), the Keywords Radar View illustrates the differences in the images of the destination under each category (**T3**), and the UGC view shows the original UGC that supports the destination image.

4.2. Sentiment Matrix View

With multiple destination images in a dataset, users first need a quick overview of these destination images (DR1). We summarize the destination image into two parts: cognition and emotion. To better comprehend and compare them, we provide classification and overall visual design.

As shown in Fig. 1(B), multiple destination images are displayed in a matrix. Each row represents a destination image, including image units of five cognitive types (food, attraction, scenery, service, and atmosphere) and the overall image unit. Analysts can compare the distribution of the same destination image across different cognitive categories horizontally and multiple destination images within a single cognitive type or the overall situation vertically.

An image unit identifies the emotional tendency of visitors to a destination in a specific cognitive category. As shown in Fig. 3(a), each image unit is encoded with four rectangles, two large rectangles $LArea_1$ and $RArea_1$ and two small rectangles $LArea_2$ and $RArea_2$ embedded inside. Color depths used by $LArea_1$ and $RArea_1$ encode positive and negative emotional degrees, respectively. The area size by $LArea_2$ and $RArea_2$ map the number of positive and negative cognitive entities.

An overall image unit (Fig. 3(b)) provides general information of the image, including the integrated distribution of cognitions and emotions of the destination. In travel notes or comments, not all cognitive entities are associated with emotional expression. We mine the distribution of entities with emotional descriptions and non-emotional entities in the text to help users identify the credibility of destination images with the overall image unit. The pie charts at the top of the unit show the distribution of cognitive entities with emotional descriptions after classification, while the horizontal stack chart shows the overall distribution. The length of the line segment with two points encodes the difference between negative emotion value and positive emotion value. The distance *negNum* between the starting point of the line segment and the left border of the rectangle encodes negative emotion value, and the corresponding distance *posNum* encodes positive emotion value.

4.3. Timing View

Each tourist destination is visited by numerous tourists every year, and its popularity varies annually. In this view, we follow

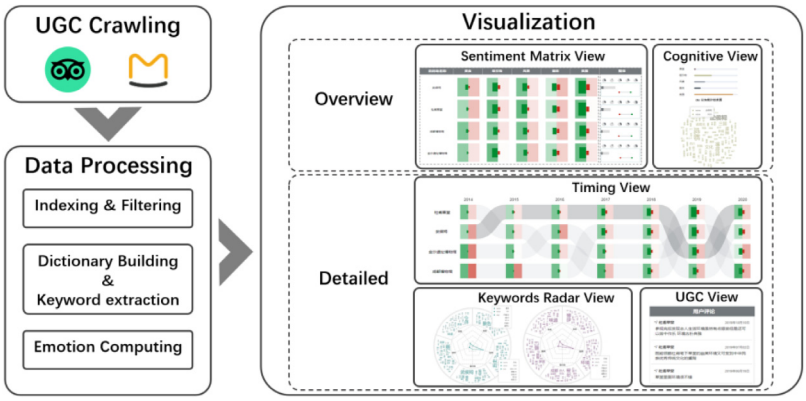


Fig. 2. The pipeline of MDIVis, contains UGC crawling, data processing, and visualization.

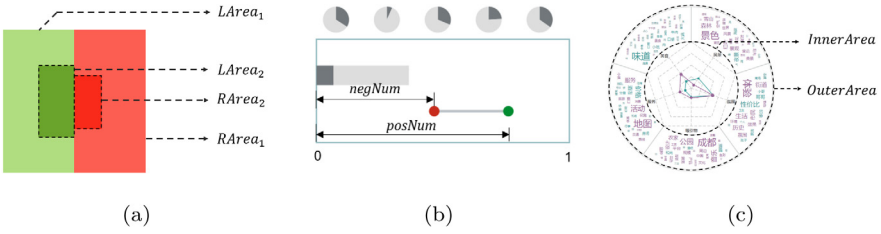


Fig. 3. Design of multiple views. (a) A cognitive image unit consisting of $LArea_1$, $LArea_2$, $RArea_1$ and $RArea_2$. (b) An overall image unit with pie charts, bar charts, and a points-and-line graph. (c) A Keywords Radar View contains *InnerArea* and *OuterArea*.

the sentiment matrix units to represent the destination's image each year. In addition, in conjunction with the work of Zhang et al. (2020) on visual ranking channels, we use visual channel chains showing the ranking of multiple destination images over the years (DR2).

As shown in Fig. 1(C), the time-series design adds a temporal dimension to the dimensional image units of the sentiment matrix, with the destination name text arranged vertically on the left and the temporal information juxtaposed horizontally above. The central part comprises image units arranged by a matrix-like layout, in which represents the overall image in a specific year. For a destination, the image units per year are connected by a chain of channels on a light background, and each chain reflecting the change in the ranking of images over the years. When the experts hover over a destination or image unit, the associated image chain will be highlighted, to help experts to focus on the selected destination and mine its temporal evolution pattern. For example, in Fig. 1(C), the images of the destination Temple of Marquis Wu are highlighted.

4.4. Keywords Radar View

The Keywords Radar View has been designed to explore and compare the detailed characteristics of multiple destination images. As shown in Fig. 1(D), we add the visual mapping of the points' area on the axes to the classical radar map to present the sentiment image of the cognitive dimensions. We combine it with the annular word cloud to provide a detailed comparison of the images (DR3). The view uses a radial layout and contains *InnerArea* and *OuterArea* parts, as shown in Fig. 3(c). In the *InnerArea* part, a modified radar map is used to encode the image information of the destination in multiple cognitive dimensions to facilitate visual comparison. The corresponding cognitive attribute values are encoded by the distance between the intersection on the axis and the axis center in the radial axes. The area of points encodes the emotional attributes of the destination in that dimension. In the *OuterArea* part, the word clouds represent the cognitive entities of the destination image, and different colors are used to distinguish the destinations. This view supports the comparative analysis between different dimensions (different sectoral word clouds) and enables the content comparison of a single dimension (the same sectoral word cloud).

4.5. Interactions

In this section, we describe the interactions between the visual components involved in MDIVis. Interactions are designed to assist users in exploring and comparing multiple destination pictures and completing relevant analysis tasks.

Before conducting a formal analysis, users need to roughly filter destinations by category in the Setting panel (Fig. 1(E)), or manually search for destinations to add to the candidate list to be analyzed. While the user generates or changes the set of candidate destinations, the Cognitive View presents comprehensive cognitive information about them, and the Sentiment Matrix View provides overall emotional images of the multiple destinations in each cognitive category. Then, users can interactively sort the candidate destinations in the matrix to facilitate the selection of specific destinations of interest, and switch to the detailed level views for detailed comparison and exploration by clicking on the tabs in the setting panel. The Timing View shows how multiple destination images have changed in ranking over time, allowing users to hover over an image to notice which destination it belongs to and explore the context of that destination with the highlighted bar. Moreover, users can investigate more information by hovering over the visible elements with the mouse to expand the bubble tooltips in the Keywords Radar View. In addition, when the user explores a specific destination image by interacting with the view component, the view will automatically update the UGC information that supports the destination image.

5. Evaluation

Destination image is one of the popular subjects in tourism, and tourism managers are concerned about the differences in

multiple destination images. To evaluate the effectiveness and availability of MDIVis, we conduct three case studies and an expert interview in real-world data.

5.1. Case studies

In the following, we present three case studies and highlight insights gathered from real-world data, with an example from Chengdu, Sichuan Province, China. The three case studies jointly cover all tasks described in Section 3.1.

5.1.1. Summarize and compare multi-destination images

Summarizing tourists' emotional tendencies and the richness of cognitive entities towards each destination in Chengdu is an important prerequisite for understanding the commonalities and differences in destination images. In this case, we describe how our system helps experts understand multiple destination images from various perspectives.

First, we select the destinations in the Museum category in the setting panel to get an overview of the destination. Fig. 1(E) shows the distribution of cognitive themes for these destinations, with atmospheres being the most popular, followed by attractors, and foods being the least. Then, we analyze the cognitive word cloud by selecting the atmosphere cognitive histogram. In the bottom, some terms, such as history, characteristics, and children, have been discussed extensively, which means that tourists prefer to bring their children to broaden horizons and experience the sense of history and culture.

The destination images in the museum category are shown in the Sentiment Matrix View (Fig. 1(B)). Then we can find that light gray takes up more area than dark gray in overall image units generally, indicating that the cognitive entities of visitors' rich experience are unrelated to emotion. In addition, the emotional line segments in the bottom all deviate to the right, indicating that tourists have good experiences of the destinations (**T1.1**). The emotional matrix depicts the classified emotional situation of the various destinations, and the most popular destinations are Temple of Marquis Wu, Du Fu Thatched Cottage, Chengdu Museum and Jinsha Site Museum, according to the overall ranking. For each cognitive category, the green embedded rectangle on the left of the image units occupies more area than the red rectangle on the right, indicating that tourists develop more positive impressions of the destinations, and the overall destination images are pleasant (**T1.3**). In addition, it is noticeably that almost every destination has the highest frequency of cognitive themes in the category of atmosphere, followed by attraction, and services least, which is consistent with the distribution of cognitive classification word cloud (**T1.2**). In general, all destinations present positive images, i.e., more positive feelings among visitors. Also, the museum destinations focus on enhancing attractions and landscapes to create a historical atmosphere but lack reputations of service (**T1**).

5.1.2. Explore the temporal characteristics and details

Exploring the development of destination images over the years, combined with detailed information on destination images, can help experts identify specific differences in destination images and further provide a basis for improving the competitiveness of destinations. In this case study, we describe how MDIVis helps experts explore image differences and potential competitiveness of destinations.

In the preceding analysis, the experts are interested in the destinations of the Temple of Marquis Wu, Du Fu Thatched Cottage, Chengdu Museum, and Jinsha Site Museum, which occur with high frequencies of appearance, and we investigate the temporal patterns of these destination images. Fig. 1(C) illustrates that the image units for all destinations show a similar trend from 2014 to 2019. The areas of inner rectangles grow increasing over the chain, which means that the impression of tourists grows richer over year. It is worth noting that the areas of the inner rectangles decline significantly from 2019 to 2020. Based on actual events, it may be due to the social impact of COVID-19 in early 2020, which reduced the number of tourist trips and weakened their perception of the destinations. It is also worth mentioning that, while the ranking of Temple of Marquis Wu fluctuates in the overall ranking, the general trend is improving. The ranking rose from the second in 2014 to the first in 2015 and remained in 2018 before dropping in 2019, but it climbed back to the top two in 2020. Next, we examine the temporal evolution of sentiment images of the atmosphere category and sort them by positive images, as shown in Fig. 4a. Similar to the overall situation, in the image units of each year, the green rectangular area of the embedded left side is larger than the right red rectangle, indicating that the destinations had left more positive impressions for the tourist. When we hover over the name text of the Du Fu Thatched Cottage, the relevant image units are connected by a chain. Since 2014, its ranking has increased year by year. It rose to the first place in 2017 and remained its place until 2020, indicating that the destination positively impacts tourists.

We then sort them by negative images, and the result is shown in Fig. 4b. The negative ranking of Du Fu Thatched Cottage varies greatly, increasing year after year and decreasing year after year. It returned to fourth place in 2019 and maintained a certain level. It has been discovered that while tourists' positive impressions of Du Fu Thatched Cottage are increasing year by year, there are also more negative impressions. However, the increase in negative impact has gradually decreased since 2017. It can also be seen that the color depth of the units' right rectangles declines to about 0 from 2019 to 2020, indicating that tourists' negative perception of this destination is decreasing. As a result of the above observation and analysis, due to environmental changes and competition among various tourist destinations, the images of multiple destinations have undergone big or small changes over the years (**T2**).

We continue to compare and analyze the images of the Temple of Marquis Wu and Du Fu Thatched Cottage in the Keywords Radar View (**T3**). In the positive perception of tourists, as shown in Fig. 5(a, b), both destinations have the same cognitive entities such as *history*, *culture*, and some synonyms, which indicates that tourists prefer to feel the influence of the Chinese excellent traditional culture (**T3.1**). In negative perception, as shown in Fig. 5(c, d), the two destinations have similar entities per category, including *scenery*, *taste*, *ticket*, *price*, *cost performance*, indicating that the destinations leave some similar negative impression on tourists (**T3.2**). To support this insight, we browsed through the relevant original user-generated content, and many tourists said they would go there frequently if the tickets were not very expensive (**T3.4**). The above examples show that museum destinations perform well in terms of historical and cultural heritage. However, the service elements, mainly the entrance fee, may become the key competitive element for the destination in the subsequent development.

5.1.3. Compare the perceived/projected images

Destination image can be divided into the perceived image and projected image, where perceived image refers to tourists' overall impression and feeling of the destinations, and the projected image is defined as the ideal image assigned to the destination by tourism management. Tourists' impressions of destinations in UGC provide valuable data to construct perceived images. In this case, we take Chengdu city as an example with publicity data to explore the divergence between perceived and projected images.

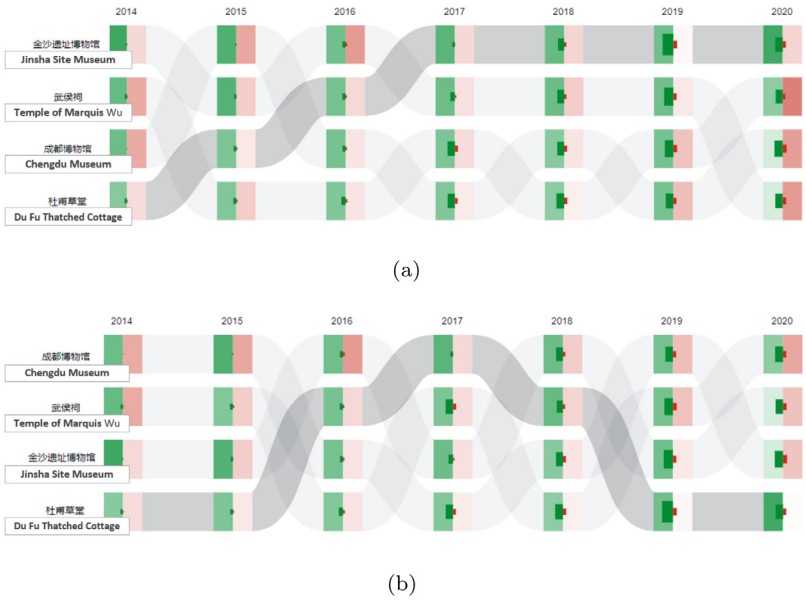


Fig. 4. Timing view. The images are sorted by their positive degree (a) or negative degree (b).

Fig. 5. Detailed comparative analysis with the Keywords Radar View. The destination image of Temple of Marquis Wu is divided into two parts: a positive image (a) and a negative image (c). And the destination image of Du Fu Thatched Cottage is divided into two parts: a positive image (b) and a negative image (d).

From the perspective of tourists (Fig. 6(a)), it is observed that the terms mentioned more often by tourists for the landscape category are *scenery* and *landscape*. The most frequently mentioned entity of food category is *flavor*, and the words *network* and *service* are often mentioned in the service category. In the cognitive type of atmosphere, travelers often mention *life* and *environment*, and *Chengdu* and *culture* are frequently used in the attraction category. Fig. 6(b) shows the details of the projected image. There are more high-frequency entities, including *map*, *event*, *transportation* for services, and *park*, *music* for attractions,

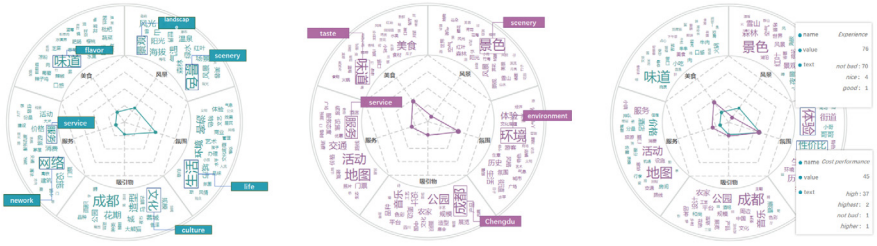


Fig. 6. Detailed comparative analysis of perception and projection. The perceived image (a) and the projected image (b) are displayed under the juxtaposition layout, respectively. And the perceived image and the projected image are displayed together under a overlay layout (c).

Table 1
Questionnaire of expert evaluation.

Usability	Q1	Is it easy (difficult) to choose some destinations for comparison analysis?
	Q2	With the sentiment matrix view, is it easy (difficult) to compare the images of the destination subset globally?
	Q3	Is it easy (difficult) to understand the temporal evolution of images between destinations?
	Q4	Is it easy (difficult) to select a subset of destinations of interest and compare them in detail?
Effectiveness	Q5	Is it easy (difficult) to analyze the difference between the perceived image of visitors and the official projected image of Chengdu?
	Q6	Overall, is it easy (or difficult) for you to use MDIVis to compare multi-destination images?
	Q7	Is it easy (difficult) to learn and use MDIVis?
	Q8	Is it easy (difficult) to understand the visual designs in MDIVis?

etc. It indicates that the projected image is more prosperous and dedicated to promoting destination diversity, which still needs to be experienced and felt by tourists in-depth.

To further compare the perceived image of tourists with the official projected image, the juxtaposition layout is replaced with an overlay layout. As shown in Fig. 6(c), the green is used to indicate tourist perceptions and purple with official projections. The internal radar diagram compares the information of the perceived image and projected image in five dimensions. It is observed that tourists mention more about the scenery category. The officials make a lot of publicity in the food, service, and attraction types, and they are more consistent in the atmosphere category with tourists. We further focus on the detailed descriptions of each dimension. The tourists feel more strongly about *taste*. In contrast, there are more cognitive entities in the projective image, such as *scenery*, *park*, *music*, etc. Then, we focus on the detailed comparison of images in the atmosphere category (Fig. 6(b,c)) and find that tourists are more concerned about *cost-effectiveness*, which is mainly described as *high* and *highest* while the official focus more on *experience* (**T3.3**). This case illustrates differences between the perceived image and the projected image, and it may be a challenge or an opportunity to shape the image according to the interests of tourists.

5.2. Expert evaluation

The above case studies have validated the utility of the MDIVis proposed in this paper. We develop the following expert evaluation to demonstrate the system's effectiveness and usability.

There are 10 exports invited to participate in the expert evaluation phase, including 2 tourism managers, 4 visual analysis researchers, and 4 researchers with tourism research backgrounds. We designed a questionnaire (Table 1), where Q1–Q4 correspond to general requirements to verify the usability of MDIVis, and Q5–Q8 involve the overall evaluation of MDIVis in the comparative analysis of destination images to evaluate the effectiveness of MDIVis. Second, we briefly introduce the background of our work and the user interface of MDIVis, followed by an explanation of MDIVis's function via an operation example. Finally, participants were encouraged to explore the MDIVis freely and respond to relevant evaluation questions.

Fig. 7 shows the results of an expert evaluation of problems Q1–Q8, demonstrating the usability and effectiveness of MDIVis. In terms of destination analysis subset selection, all participants believe that MDIVis makes it very simple to determine a subset of destinations via destination search or type selection for comparative analysis (Q1). In the overview analysis (Q2), participants by the emotional matrix view carry on the preliminary analysis to the selected destination collection, believe the design of image unit can depict the cognitive and emotional information related to intuition, and easily reflect the differences in the set of destinations in different dimensions. Furthermore, they can perform the comparative analysis of the destination collection with the type of interaction. After the overview analysis, participants interactively select destinations of interest and further analyze temporal and detailed features. In the aspect of time sequence comparison (Q3), participants prefer the Timing View to discover the trend of image's annual evolution over the year intuitively. For example, participants chose the overall ranking method, and they found that the image of Du Fu Thatched Cottage fluctuated wildly, which is difficult to obtain directly through questionnaires or raw UGC data in previous studies. In terms of detailed comparison (Q4), nine of the 10 participants thought they could complete a comparative analysis of detailed information of destination images. They agreed that it is beneficial for a complete contrast to multiple destination images that the Keywords Radar View presented classified detailed information. When comparing the visitor's perceived image with the official projected image, participants agreed that this feature was very effective for a comprehensive understanding of the Chengdu destination image (Q5). For example, participants mentioned that they could identify apparent differences between the images only by the image overview view. Participants agreed that MDIVis is effective for comparative destination images based on the above exploration and analysis experience (Q6). According to Q7 and Q8, it is easy for experts to use MDIVis and understand its visual design. Users who have never used the visual analysis system can also efficiently conduct a comparative analysis of the destination images. Moreover, most experts agree that the Sentiment Matrix View presents an overview of destination images in a table-like

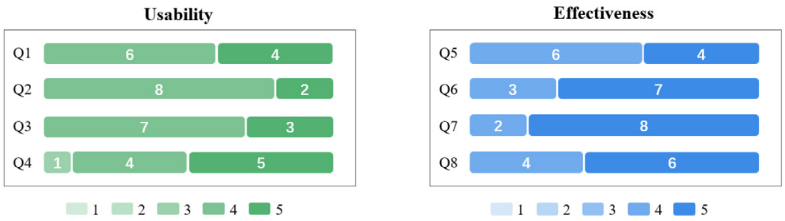


Fig. 7. Expert evaluation result.

format, reducing learning costs, and visualizes the data in a clear form, while the custom graphic design of the matrix cells offers obvious advantages in performing comparisons of multiple destination images. Experts also think that the Timing View and the Keywords Radar View support a detailed comparison of selected destination images and can effectively address users' analytical needs. However, some experts suggested that destination image indicators such as transportation, accommodation, and surrounding facilities, can be added in future work to make the analysis results more comprehensive and specific.

6. Discussion

Although the usability and effectiveness of MDIVis have been confirmed in our evaluation, there are some limitations that may serve as meaningful references for future studies.

Scalability. In the above case studies, we demonstrated the effectiveness of MDIVis for task exploration, but its scalability might be improved. The Sentiment Matrix View presents destination images in the form of a matrix, and the users get limited information about the destination at each observation. If there are massive destination images in the dataset, the overall distribution information is challenging to capture. In this work, experts tend to investigate the outstanding destinations under each category after classification, and the matrix layout is a suitable way to meet that need.

Generalization. In the current research work, MDIVis has been applied only to explore multiple destination images, where some of the visual analysis methods and views can be referenced to other domains. The Keywords Radar View and the Sentiment Matrix View are not limited to comparing destination images. They are also suitable for demand for fine-grained classification and comparison involving keywords, emotions, and time series in other fields related to text visualization, such as education. For example, we can analyze student comments on online courses to find differences by adapting the proposed methods in such a context.

7. Conclusion

We propose MDIVis for tourism destination images analysis to help users explore and understand destination image features from UGC and discover competitive elements of destinations in comparative analysis. Specifically, we firstly combine literature review and expert interviews to extract the system requirements and analysis tasks. Then, we design and implement a novel sentiment matrix view and improve two classic views, which assist users in comparing the destination images from various perspectives at the overview and detail levels. Finally, we use UGC in the actual environment for case analysis and expert evaluation to verify the usability and effectiveness of MDIVis.

In the future, we plan to add data forms to capture more information about the travel experience. The tourism destination image contains a variety of contents. This paper only analyzes the travel notes and comment information from the tourism platform. The other forms are not considered, such as transportation, accommodation, social environment, and other aspects of destination images. Also, geographical features are equally meaningful for destination image analysis, and some top-rated tourist attractions are likely to popularize the surrounding tourist places.

CRediT authorship contribution statement

Changlin Li: Writing – original draft, Methodology, Visualization, Writing – reviewing. **Mengqi Cao:** Data curation, Investigation. **Xiaolin Wen:** Investigation, Resources, Software. **Haotian Zhu:** Writing – review & editing, Formal analysis. **Shangsong Liu:** Writing – review & editing, Formal analysis. **Xinyi Zhang:** Writing – review & editing, Formal analysis. **Min Zhu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Chengdu Science and Technology Bureau, China (Grant No. 2019-YF05-02121-SN).

Ethical approval

This study does not contain any studies with human or animal subjects performed by any of the authors. All data used in the study are taken from public databases that were published in the past.

References

Agus, S., Gamal, S., et al., 2020. The impact of Islamic destination attributes on Saudi Arabians' decision to visit Jakarta: tourism destination image as a mediating variable. Int. J. Religious Tour. Pilgr. 8 (3).

Baloglu, S., McCleary, K.W., 1999. A model of destination image formation. Ann. Tourism Res. 26 (4), 868–897.

Barroso, C.M., Santos, C.Q., Espindola, L.S., Silveira, M.S., 2020. Reflections on data visualization design by professionals in the tourism field. In: International Conference on Human-Computer Interaction. Springer, pp. 538–554.

Beerli, A., Martin, J.D., 2004. Factors influencing destination image. Ann. Tour. Res. 31 (3), 657–681.

Cao, M.-q., Liang, J., Li, M.-z., Zhou, Z.-h., Zhu, M., 2020. Tdivis: visual analysis of tourism destination images. Front. Inf. Technol. Electron. Eng. 21 (4), 536–557.

Chen, Y., 2018. Tagnet: toward tag-based sentiment analysis of large social media data. In: 2018 IEEE Pacific Visualization Symposium (PacificVis). IEEE pp. 190–194.

Chiu, W., Zeng, S., Cheng, P.S.-T., 2016. The influence of destination image and tourist satisfaction on tourist loyalty: a case study of Chinese tourists in Korea. Int. J. Culture, Tour. Hospital. Res..

Francalanci, C., Hussain, A., 2015. A visual analysis of social influencers and influence in the tourism domain. In: Information and Communication Technologies in Tourism 2015. Springer, pp. 19–32.

Garay, L., 2019. Visitspain. Breaking down affective and cognitive attributes in the social media construction of the tourist destination image. Tour. Manag. Perspect. 32, 100560.
Gkritzali, A., Gritzalis, D., Stavrou, V., 2018. Is Xenios Zeus still alive? Destination image of Athens in the years of recession. J. Travel Res. 57 (4), 540–554.
Han, P., Zhang, M., Yao, M., Hui, Z., Chen, M., 2021. An empirical study on the tourism image of nanjing from the perspective of international students. In: International Conference on Artificial Intelligence and Security. Springer pp. 476–485.
Hu, M., Wongsuphasawat, K., Stasko, J., 2016. Visualizing social media content with sententree. IEEE Trans. Vis. Comput. Graphics 23 (1), 621–630.
Huang, W., Zhu, S., Yao, X., 2021. Destination image recognition and emotion analysis: evidence from user-generated content of online travel communities. Comput. J. 64 (3), 296–304.
Jeng, C.-R., Snyder, A.T., Chen, C.-F., 2019. Importance–performance analysis as a strategic tool for tourism marketers: The case of Taiwan's destination image. Tour. Hospital. Res. 19 (1), 112–125.
Kim, K., joung Park, O., Yun, S., Yun, H., 2017. What makes tourists feel negatively about tourism destinations? application of hybrid text mining methodology to smart destination management. Technol. Forecast. Soc. Change 123, 362–369.
Knittel, J., Koch, S., Tang, T., Chen, W., Wu, Y., Liu, S., Ertl, T., 2021. Real-time visual analysis of high-volume social media posts. IEEE Trans. Vis. Comput. Graphics.
Kucher, K., Martins, R.M., Paradis, C., Kerren, A., 2020. Stancevis prime: visual analysis of sentiment and stance in social media texts. J. Vis. 23 (6), 1015–1034.
Leković, K., Tomić, S., Marić, D., Ćurčić, N.V., 2020. Cognitive component of the image of a rural tourism destination as a sustainable development potential. Sustainability 12 (22), 9413.
Li, Q., Wu, Y., Wang, S., Lin, M., Feng, X., Wang, H., 2016. Vistravel: visualizing tourism network opinion from the user generated content. J. Vis. 19 (3), 489–502.
Qi, S., Chen, N., 2019. Understanding Macao's destination image through user-generated content. J. China Tour. Res. 15 (4), 503–519.
Sheng, F., Zhang, Y., Shi, C., Qiu, M., Yao, S., 2020. Xi'an tourism destination image analysis via deep learning. J. Ambient Intell. Humaniz. Comput. 1–10.
Triantafillidou, A., Yannas, P., Lappas, G., 2019. The impact of the destination image of Greece on tourists' behavioral intentions. In: Economic and Financial Challenges for Eastern Europe. Springer, pp. 345–359.
Troudi, A., Jamoussi, S., Zayani, C.A., Amous, I., 2019. Multidimensional analysis of hot events from social media sources. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, pp. 2112–2119.
Wang, J., Li, Y., Wu, B., Wang, Y., 2020. Tourism destination image based on tourism user generated content on internet. Tour. Rev..
Woosnam, K.M., Stylidis, D., Ivkov, M., 2020. Explaining conative destination image through cognitive and affective destination image and emotional solidarity with residents. J. Sustain. Tour. 28 (6), 917–935.
Xiao, X., Fang, C., Lin, H., 2020. Characterizing tourism destination image using photos' visual content. ISPRS Int. J. Geo-Inf. 9 (12), 730.
Xu, H., Yuan, H., Ma, B., Qian, Y., 2015. Where to go and what to play: Towards summarizing popular information from massive tourism blogs. J. Inf. Sci. 41 (6), 830–854.
Yuan, H., Xu, H., Qian, Y., Li, Y., 2016. Make your travel smarter: Summarizing urban tourism information from massive blog data. Int. J. Inf. Manage. 36 (6), 1306–1319.
Zhang, K., Chen, Y., Li, C., 2019. Discovering the tourists' behaviors and perceptions in a tourism destination by analyzing photos' visual content with a computer deep learning model: The case of Beijing. Tour. Manag. 75, 595–608.
Zhang, K., Koshijima, I., 2019. Trend analysis of online travel review text mining over time. J. Model. Manag..
Zhang, H., Xu, H., Xiao, Y., Guo, X., Ma, J., 2020. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, pp. 12797–12804.

Telematics and Informatics Reports 11 (2023) 100070

Contents lists available at ScienceDirect

Telematics and Informatics Reports

journal homepage: www.elsevier.com/locate/teler

Navigating the acceptance of implementing business intelligence in organizations: A system dynamics approach

Mehrdad Maghsoudi [a,*], Navid Nezafati [b]

[a] Department of Industrial and Information Management, Faculty of Management and Accounting, Shahid Beheshti University, Tehran, Iran
[b] BPP University, Business School, London, England

ARTICLE INFO

Keywords:
Business intelligence
System dynamics
Implementation of business intelligence
Information technology organizations
Vansim

ABSTRACT

The rise of information technology has transformed the business landscape, with organizations increasingly relying on information systems to collect and store vast amounts of data. To stay competitive, businesses must harness this data to make informed decisions that optimize their actions in response to the market. Business intelligence (BI) is an approach that enables organizations to leverage data-driven insights for better decision-making, but implementing BI comes with its own set of challenges. Accordingly, understanding the key factors that contribute to successful implementation is crucial.

This study examines the factors affecting the implementation of BI projects by analyzing the interactions between these factors using system dynamics modeling. The research draws on interviews with five BI experts and a review of the background literature to identify effective implementation strategies. Specifically, the study compares traditional and self-service implementation approaches and simulates their respective impacts on organizational acceptance of BI. The results show that the two approaches were equally effective in generating organizational acceptance until the twenty-fifth month of implementation, after which the self-service strategy generated significantly higher levels of acceptance than the traditional strategy. In fact, after 60 months, the self-service approach was associated with a 30% increase in organizational acceptance over the traditional approach. The paper also provides recommendations for increasing the acceptance of BI in both implementation strategies. Overall, this study underscores the importance of identifying and addressing key factors that impact BI implementation success, offering practical guidance to organizations seeking to leverage the power of BI in today's competitive business environment.

Introduction

The speed of data generation and accumulation has increased with the increasing use of information technology solutions by organizations, which are now using digital tools to store and analyze vast amounts of data in real-time [28,37,63]. Companies and organizations can gain a comparative advantage and overtake their competitors through the analysis and continuous and effective use of data and information [12,30,68].

Business intelligence is one of the popular and welcomed solutions for organizations to use data analysis for decision-making and data-oriented business [67,74]. For example, a retail company can use business intelligence to monitor customer behavior and preferences and adjust its marketing strategy accordingly [48]. Likewise, a manufacturing company can use BI to optimize its supply chain, reduce production costs, and improve product quality [27]. This solution offers managers and experts of organizations the possibility of making smart and updating analyzes and decisions. A business intelligence system is usually defined as a set of technological solutions [12] that facilitate organizations to collect, integrate, and analyze large data stores in order to understand their opportunities, strengths, and weaknesses [23].

According to researchers [1], business intelligence systems are quite close to the concept of decision support systems, these systems expand the categories of users and support a wide range of decisions. Business intelligence systems are designed to reduce uncertainty in the decision-making process and support decision-makers efficiently and effectively [51].

The size of the business intelligence market in 2020 was valued at $23.1 billion and is expected to reach $33.3 billion by 2025 and experience a compound annual growth rate of 6.6% [70]. However, despite growing investments and market expansion, evidence shows that many organizations cannot take advantage of the benefits of implemented business intelligence systems [5]. More than 70% of business intelligence projects cannot bring the expected returns [1] or result in little or no benefits for organizations [72]. To find the best

* Corresponding author.
E-mail address: M_Maghsoudi@sbu.ac.ir (M. Maghsoudi).

https://doi.org/10.1016/j.teler.2023.100070
Received 29 October 2022; Received in revised form 20 May 2023; Accepted 7 June 2023

way to use the value of business intelligence systems and succeed in their implementation [69].

Individual and organizational acceptance is one of the main challenges in the successful implementation of business intelligence systems in organizations, which is very complex due to their human nature and requires careful monitoring and control [65]. The acceptance of users is crucial for the successful utilization of business intelligence (BI) systems over an extended period. When users embrace BI and use it consistently, it becomes compatible with other organizational processes. Besides, BI can facilitate organizational change that leads to improvements in coordination and control processes. User acceptance is fundamental to implementing information system projects as a whole [14,24]. Specifically, user acceptance is critical when it comes to BI systems [15,21,32,55,61,71]. Many researchers have talked about the importance of organizational acceptance in the successful implementation of the business intelligence system in the organization. For example, [6] have used the Technology Acceptance Model (TAM) method to accept business intelligence, [22] have used the exploratory approach to conceptualize the acceptance of business intelligence, [39] has used motivation theory to analyze the two modes of routine use and innovative use. have benefited [21] is also a mixed study with the aim of investigating the factors that affect the acceptance behavior in the field of business intelligence. However, it is worth noting that, despite the extensive research on organizational acceptance of business intelligence systems, so far no article has used system dynamics to analyze this phenomenon. Therefore, this article aims to fill this research gap by using a dynamic system method to analyze the organization's acceptance of BI systems.

System dynamics is a powerful approach that can help researchers gain a deeper understanding of the complex interrelationships between various factors influencing organizational acceptance. System dynamics is an approach that enables researchers to model the complex interactions between different factors [38] influencing organizational acceptance. By developing simulation models that capture the dynamic behavior of the system over time, researchers can identify the root causes of problems and test different policies and strategies to improve user acceptance. By employing a system dynamics approach, researchers can develop simulation models that capture the dynamic behavior of the system over time, which can provide valuable insights into the long-term consequences of different policies and strategies. Therefore, there is a need for more research that utilizes system dynamics to investigate organizational adoption of business intelligence systems [52]. After stating the generalities of the research and the statistics of the factors affecting the successful implementation of business intelligence projects, this research has drawn the cause-effect relationships of these factors based on the opinions of experts and simulated the dynamic behavior of organizational acceptance in two traditional and self-service strategies and based on The results of that analysis provide.

This article is divided into seven sections. In Section 2, we provide an overview of existing literature on business intelligence and system dynamics, including strategies for implementing business intelligence systems, and review associated works. Section 3 discusses the research methodology used in this study. In Section 4, we present our findings on the identification of factors that affect the implementation of business intelligence systems. We also showcase the results of our dynamic system modeling and simulation of different scenarios. Section 5 is dedicated to the Conclusion and Outlook of the Research. Section 6 is dedicated to discussing the limitations of our study. While we have made every effort to conduct a comprehensive study, it is possible that some factors were not accounted for due to these limitations. Finally, in Section 7, we outline future research directions that could build upon our work.

Research literature review

Business intelligence

The term business intelligence was first proposed in 1989 by the Gartner Group. They introduced business intelligence as a set of concepts and methods to develop business decisions through reality-based systems [57]. Business intelligence includes all the processes of collecting, storing, accessing, analyzing, and extracting quality information or knowledge in different business fields [50]. A business intelligence system enables employees and organizations to better understand their business or market and make timely strategic decisions [47]. Business intelligence is a concept that has evolved over time. In the beginning, business intelligence was mainly focused on data analysis, but today it also includes organizational processes and strategies because business intelligence affects not only technology but also the organization that applies business intelligence [54]. Business intelligence has also been developed and used at different levels of the organization, from the strategic level to the operational level, in order to make more decisions based on data [2].

The importance of business intelligence lies in its ability to provide organizations with valuable insights that can enhance their competitiveness and profitability [47]. BI can help organizations improve their operational efficiency, reduce costs, increase revenue, and identify new business opportunities [46]. By having access to timely and accurate data, organizations can make better decisions, mitigate risks, and respond to changing market conditions more effectively. In today's fast-paced and data-driven business environment, companies that fail to leverage the power of business intelligence risk falling behind their competitors and missing out on growth opportunities [10].

System dynamics

System dynamics appeared in the late 1950s as a result of focusing on the behavior of complex systems in a specific period [29]. The main features of system dynamics simulation are feedback loops, state-flow functions, and time delays which are used to model the nonlinearity of systems' behavior [60].

In system dynamics, when a set of variables are connected to each other in a connected path, they form a feedback loop, which includes positive feedback loops and negative feedback loops. Positive feedback loops are circles in which if a factor is changed in one direction, the circle reinforces the changes in that direction. Negative feedback loops are circles that, if a factor is changed in one direction, the circle opposes

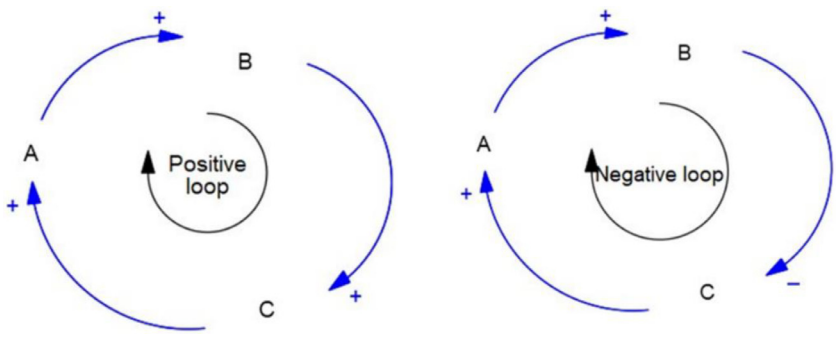


Fig. 1. Positive feedback loop (left) and negative feedback loop (right) [9].

changes in the factor in that direction [17]. Fig. 1 shows positive and negative feedback loops.

In system dynamics, dynamic variables move through currents in the system and accumulate in the state variable. A state variable calculates and reports the quantitative state of dynamic variables at any given time [25]. Feedback loops and delays in system dynamics can simulate a dynamic system, that is, if a change occurs at any point in the system, it will lead to a chain reaction throughout the system [53]. System dynamics allows for an iterative process of model-building, testing, and refinement to gain insights into the behavior of the system over time [7]. The approach is widely used in various fields such as engineering, economics, and management to understand and optimize the performance of complex systems.

Compared to other methods such as Delphi, system dynamics has several advantages. Firstly, it allows for the explicit modeling of feedback loops, which are often present in complex systems but difficult to capture using traditional modeling approaches. Secondly, system dynamics incorporates the concept of time delays, which can have significant impacts on the behavior of the system over time. Thirdly, the iterative nature of system dynamics modeling allows for continuous improvement and refinement of the model based on new data or insights. Finally, system dynamics provides a visual representation of the system, which can facilitate communication and understanding among stakeholders with varying levels of expertise.

The general process of system dynamics involves the following steps [41,43,56]:

1 Identifying the problem: The first step in system dynamics is to identify the problem or issue that needs to be analyzed. This could be something like a declining sales trend in a business or an increase in unemployment in a community.
2 Building a conceptual model: Once the problem has been identified, a conceptual model is built to represent the system being analyzed. This model includes all the relevant variables, relationships between variables, and feedback loops.
3 Quantifying the model: The next step is to quantify the conceptual model by assigning numerical values to the variables and relationships in the model. This allows for mathematical analysis of the system and prediction of how it will behave over time.
4 Simulation and testing: After the model has been quantified, it is simulated using computer software to test its behavior under different conditions. This helps to identify the causes of the problem and potential solutions.
5 Policy design and implementation: Based on the results of the simulation, policy recommendations are developed to address the problem. These policies are implemented, and their impact is monitored over time to assess their effectiveness.

Types of strategies in implementing business intelligence

In general, there are two methods and strategies for deploying business intelligence in organizations. Implementation of business intelligence in the traditional way and self-service business intelligence [2,34,35], each of these two methods has its own characteristics:

Traditional BI implementation

Traditional BI implementations follow a centralized approach, where IT departments take complete control over the BI system development process [26]. This implementation method involves various stages such as data warehousing, ETL (Extract, Transform, Load), data modeling, report generation, analysis, and distribution. These stages require specialized skills and expertise, which are generally available in the IT department [3]. The following are some of the critical elements of traditional BI implementation:

· **Data Model Design**: Traditional BI implementation relies heavily on the IT team to develop the data model. This is done by mapping the business requirements to the data warehouse schema. The IT team uses tools like ER diagrams and dimensional modeling techniques to create the data model(David [13]).
· **Data Extraction, Transformation, and Loading**: Data extraction involves collecting data from various sources and transforming them into a standard format. The transformed data is then loaded into the data warehouse. This process requires significant effort and expertise [40].
· **Report Generation**: Once the data is ready, reports are generated using specialized tools like Crystal Reports, Business Objects, or Cognos [75]. The IT team creates custom reports based on the business requirements, which can be published and distributed to end-users [33].
· **Analysis and Distribution**: Traditional BI implementation involves the creation of static reports that are distributed to end-users for analysis. End-users have limited options to manipulate the data, and any changes require IT involvement [20].

Table 1
Differences between traditional and self-service BI.

Task	Traditional BI Implementation	Self-Service BI Implementation
Data Model	Developed by IT team	Developed by End-users
ETL	Done by IT team	Done by End-users
Report Gen.	Done by IT team	Done by End-users
Analysis	Limited options for end-users	Flexible options for end-users
Control	Centralized	Decentralized
Expertise	Requires specialized skills	No specialized skills required

Self-service BI implementation

Self-Service BI implementation follows a decentralized approach, where end-users have more control over the BI system development process. Self-service BI provides end-users with more flexibility and agility in accessing and analyzing data [36,49]. The following are some of the critical elements of self-service BI implementation:

· **Data Model Design**: Self-service BI implementation allows end-users to create their own data models using simple drag-and-drop interfaces. End-users can create data models based on their business requirements without any IT involvement [49].
· **Data Extraction, Transformation, and Loading**: Self-service BI implementation allows end-users to collect data from various sources and transform them into a standard format. This process does not require any specialized skills or knowledge [4].
· **Report Generation**: Self-service BI implementation provides end-users with tools like Tableau, QlikView, or PowerBI to create custom reports. These tools have user-friendly interfaces that allow end-users to create reports based on their specific needs [16].
· **Analysis and Distribution**: Self-service BI implementation enables end-users to analyze and manipulate data on their own. They can create custom dashboards, drill-downs, and filters to explore data in real time. End-users can share their findings with others using interactive reports that can be accessed on any device [59].

Table 1 summarizes the differences between traditional and self-service BI implementation methods:

Associated works

Nalchiger et al., in a research they conducted in 2014, presented an approach based on systems dynamics modeling to support decisions and choose the best alternative actions for the organization through the outputs of the business intelligence system. These researchers modeled the results of the business intelligence system with the decision-making process in the organization and actually presented a combination of business intelligence and a decision support system [45].

Ain et al. also investigated the reason why organizations do not achieve the benefits of implementing a business intelligence system and

Table 2
Summary of related works.

Year	Authors	Subject	Key Findings
2014	Nalchiger et al.	Systems Dynamics Modeling and Business Intelligence for Decision Support	Systems dynamics modeling can support decision-making in organizations through business intelligence system outputs. Combination of BI and decision support system presented.
2019	Ain et al.	Factors Affecting Adoption, Use and Success of Business Intelligence Systems	Effective factors related to the adoption, use and success of business intelligence identified through a systematic review of past research.
2020	Kumar & Krishnamoorthy	Adoption of Business Analysis Systems in India	Data quality and human resource competencies are main challenges; technology assets and competitive pressure are driving factors for organizations to adopt business analytics systems.
2020	Müller et al.	Success Factors of Implementing Business Intelligence in Medium and Large Organizations	Top management support, information technology infrastructure and system quality are of highest importance for success of business intelligence systems in medium and large organizations in food industry based on data analysis using structural equation modeling.
2021	Mehri	Critical Success Factors of Business Intelligence Projects in Public Sector	Information systems and data quality are most important factors among fourteen critical success factors of business intelligence projects in public sector categorized into organization, process and technology using hierarchical analysis.
2022	Fu et al.	Important Factors Considered by Companies to Introduce a Business Intelligence System	Company information is most important factor when introducing a business intelligence system in companies followed by system performance integrity, closeness to company's strategy, license cost and technology maturity after analyzing vital factors using fuzzy hierarchical analysis and VIKOR techniques.

using a systematic review of past research, provided comprehensive knowledge about what has been stated in the field of acceptance, use, and success of a business intelligence system. They say These researchers reviewed 11 related articles and identified the effective factors related to the adoption, use, and success of business intelligence [1].

Kumar and Krishnamoorthy have made the adoption of business analysis systems the subject of their research and have studied the technological capabilities and adoption of business analysis systems in India. These two researchers obtained their data through semi-structured interviews and came to the conclusion that data quality and human resource competencies are the main challenges to the adoption of business analytics systems in India, and technology assets and competitive pressure are the main driving factors for organizations to pay attention to business analytics systems. are business analysis [31].

In another study, researchers have investigated the success factors of implementing business intelligence systems in medium and large organizations. In this research, the researchers went to the food industry and conducted their studies based on the data of 69 companies active in this industry. These researchers used structural equation modeling for data analysis based on which top management support, information technology infrastructure, and system quality are of the highest importance for the success of business intelligence systems in this industry [44].

Mehri has also evaluated the critical success factors of business intelligence projects in the public sector using hierarchical analysis. In this article, he identified fourteen main factors and after dividing them into three categories of organization, process, and technology, he prioritized the factors using the opinions of nine experts in this field. According to the findings of this project management researcher, information systems and data quality are the most important factors among the fourteen critical success factors [42].

In their research, Fu et al. investigated the important factors considered by companies to introduce a business intelligence system. This study has collected and analyzed the critical factors considered by companies when introducing a business intelligence system. By studying the research literature, these researchers have calculated all the vital factors before, during, and after the introduction of the business intelligence system and by using the two techniques of fuzzy hierarchical analysis and VIKOR, four factors of system performance integrity, closeness to the company's strategy, license cost, and technology maturity. Company information has been selected as the most important factor when introducing a business intelligence system in companies [18]. Overall, these studies reveal that the successful adoption and implementation of BI systems depend on various factors such as organizational support, technological capabilities, data quality, and strategic alignment. The findings of these studies can guide organizations in making informed decisions about implementing BI systems to enhance decision-making processes and improve overall business performance. Table 2 provides a summary of related works:

Methodology

In terms of both purpose and methodology, this study is pragmatic and falls under the category of survey research. Fig. 2 illustrates the approach taken in conducting the study. The simulation was conducted at a granular level on the surface of the moon with a time horizon of 5 years. This timeframe was determined based on expert opinion, as various factors within organizational environments often impact one another with some lag time, resulting in delayed effects. System dynamics modeling allows for incorporating such delays into the analysis [73].

In this study, we undertake a comprehensive process to identify the factors that influence the implementation of business intelligence systems. This involves an extensive review of the subject literature and previous studies to compile a list of potential factors. We present the extracted factors to a panel of five experts who provide feedback and confirmation on their relevance. To ensure the validity of the identified factors, the selection criteria require approval from a minimum of three out of the five experts present.

Following the determination of effective factors, close collaboration between the research team and experts leads to the establishment of relationships between these factors, as well as the initial values associated with the exogenous factors for each of the implementation strategies. We use this information to create a system dynamics model, which we evaluate and simulate in both traditional and self-service strategies. In the final stage of the research, we compare and analyze the simulation results from both strategies.

Choosing a panel of five experts to provide feedback and confirmation on the relevance of extracted factors is necessary for this study for several reasons. Firstly, it ensures that multiple perspectives are considered when selecting the factors that influence the implementation of business intelligence systems. With more people involved in the selection process, there is a greater chance of identifying factors that may have been overlooked by one person. Secondly, requiring approval from a minimum of three out of the five experts present ensures the validity of identified factors. This criterion provides a higher level of confidence in the selected factors, as they have been evaluated and approved by a majority of experts. Lastly, close collaboration between the research team and the chosen experts enables the establishment of relationships between the identified factors and the initial values associated with exogenous factors for each implementation strategy. This collaboration facilitates a better understanding and interpretation of the data and results. In turn, this improves the accuracy of the system dynamics model created, which is essential for evaluating and simulating both traditional and self-service strategies.

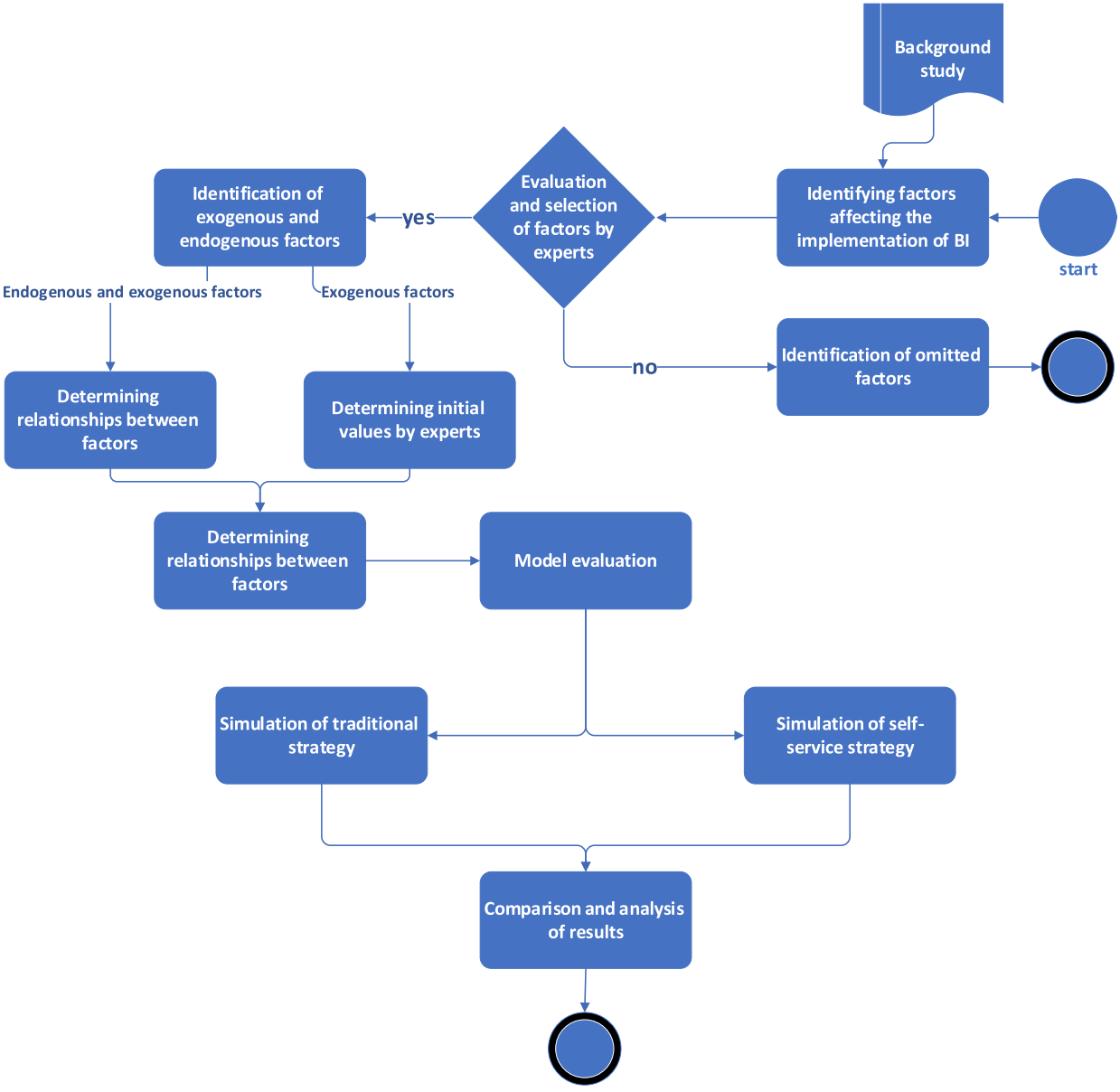


Fig. 2. Research model.

Table 3
Profile of experts.

gender	position	education	work experience	Organization size
Female	Head of BI Department	Masters	16 years	More than 2000
Man	Data engineering team manager	P.H.D	6 years	More than 400
Man	BI project manager	Masters	5 years	More than 500
Man	technical manager	Masters	18 years	More than 200
Female	BI specialist	PhD candidate	5 years	Different

It is worth noting that additional information about the participating experts can be found in Table 3, providing further insight into their contributions towards the identification of the factors affecting the implementation of business intelligence systems.

Result

Identifying factors affecting the implementation of BI

Many researchers [1,14,19] have tried to identify factors affecting the implementation of business intelligence systems. The set of indicators available in the literature is described in Table 4:

Evaluation and selection of factors by experts

Since the high number of factors affecting the implementation of business intelligence, which leads to the high complexity of the dynamic model, the list of factors in the form of a questionnaire was presented to five research experts to select the main factors affecting the implementation of the business intelligence system, and it was established that if at least three The effect of the factor has been confirmed and that factor

Table 4
List of factors affecting the implementation of business intelligence.

Factors	References	Factors	References	Factors	References	Factors	References
Project Management Skills	1, 2, 3, 4	Domain knowledge of BI team	2, 3	Project leadership power	4	Service quality	1, 2, 4
Net Benefits	1, 2, 3, 4	Ability to manage change	1, 2, 3, 4	External environment	4	Ease of use of the system	1
User participation	1, 2, 3, 4	Communication strength of the BI team	2,3	Technical readiness of BI	4	User satisfaction	2, 3, 4
Timely response	1	System integrity	1, 2, 3, 4	Third-Party Interactions	3,4	Intention To Use	1
Data quality	1, 4	Hope for performance	3, 4	Developer Skills	4	Technology Experience	1
Ease of use	1, 2, 4	Organizational Learning	2, 4	Development Approach	2, 4	Attitude Towards Change	1
Training users	2, 4	BI system maturity	2, 3, 4	Organizational Structure	2,4	Trust	2, 4
Top management commitment	1, 4	Data maturity of the organization	3,4	Organizational Competence	2,4	User Expectations	2, 4
IT Infrastructure	1, 2, 4	Manager's social influence	1,2	Organizational Size	4	Subjective Norms	2, 3, 4
System quality	1, 2	Organization acceptance	2,3,4	Organizational Culture	1,2,4	Teamwork & composition	2, 4
Access to organization data	1	Definition of clear vision	2,3	Competency Development	4	Management support	1, 2, 3, 4
Ability to integrate with other systems	1	Fear of losing organizational status	2, 3, 4	Human Resources	1	Knowledge and technical capabilities of the BI team	2, 4

1: [19], **2**: [1], **3**: [14], **4**: [65]

Table 5
Factors selected by experts.

Definition of clear vision	IT Infrastructure
Top management commitment	Ability to integrate with other systems
Ease of use of the system	Hope for performance
User participation	Organizational Learning
Training users	BI system maturity
Data quality	Data maturity of the organization
Access to organization data	Project management ability
System quality	Ability to manage change
Timely response	Communication strength of the BI team
System integrity	Organization acceptance
Management support	Domain knowledge of BI team
Knowledge and technical capabilities of the BI team	Fear of losing organizational status
	Manager's Social influence
	Organization size

should be presented in the model. The result of the selection of experts is described in Table 5:

Identification of exogenous and endogenous factors

In the field of economics and social sciences, it is essential to understand the dynamics of a system and how different variables interact with each other [64]. In this context, three categories of variables are often used to analyze and model complex systems: endogenous, exogenous, and omitted variables [58,62].

- Endogenous variables are those that are within the system being analyzed and are influenced by other variables within that same system. These variables are typically the focus of the analysis, as they represent the dependent or outcome variable(s) of interest. These variables are important because they help us understand how changes in one part of the system can affect other parts of the system [43].
- Exogenous variables, on the other hand, are external to the system being analyzed and are not influenced by other variables within that same system. Instead, these variables are typically considered to be independent or causal factors that affect the behavior of the endogenous variables. These variables are important because they help us understand the broader context in which the system operates and how it might be affected by external forces [66].
- Finally, there are omitted variables, which are simply those that are not included in the analysis of the system. These variables may be relevant to the behavior of the system, but for various reasons (such as lack of data or the complexity of the system) they are not considered in the analysis [62].

Table 6
boundary of the model.

Endogenous variable	Exogenous variable	Omitted variables
User participation	Knowledge and technical capabilities of the BI team	Organization size
Organization acceptance	Domain knowledge of BI team	Organizational Learning
Data quality	Communication strength of the BI team	Hope for performance
Ease of use of the system	Fear of losing organizational status	Ability to integrate with other systems
Access to organization data		BI system maturity
IT Infrastructure		Ability to manage change
System quality		Project management ability
Training users		Data maturity of the organization
Timely response		Manager's Social influence
System integrity		
Management support		
Top management commitment		
Definition of clear vision		

Complex dynamic systems can be effectively analyzed and predicted by understanding the role of these different types of variables [11]. By carefully considering the relationships between endogenous and exogenous variables, and striving to identify any potentially relevant omitted variables, analysts can gain a deeper understanding of how the system works and make more accurate predictions about its future behavior [43].

Table 6 shows the boundary diagram of the research. Endogenous factors are factors whose behavior changes in the model according to other variables. Exogenous factors are factors that are not affected by other variables and their behavior is not affected by the model. The reason for leaving out some factors is to prevent the enlargement of the model and based on the opinions of experts.

Causal loop diagram

System dynamics models have a crucial feature in feedback loops where positive feedback loops are referred to as reinforcing, symbolized by + or R, and negative feedback loops are called balancing, and represented by - or B. This is because positive loops amplify changes while negative loops self-correct.

Fig. 3 's cause-effect diagram displays the connections between the factors listed in Table 6 and their respective positive or negative impacts on each other.

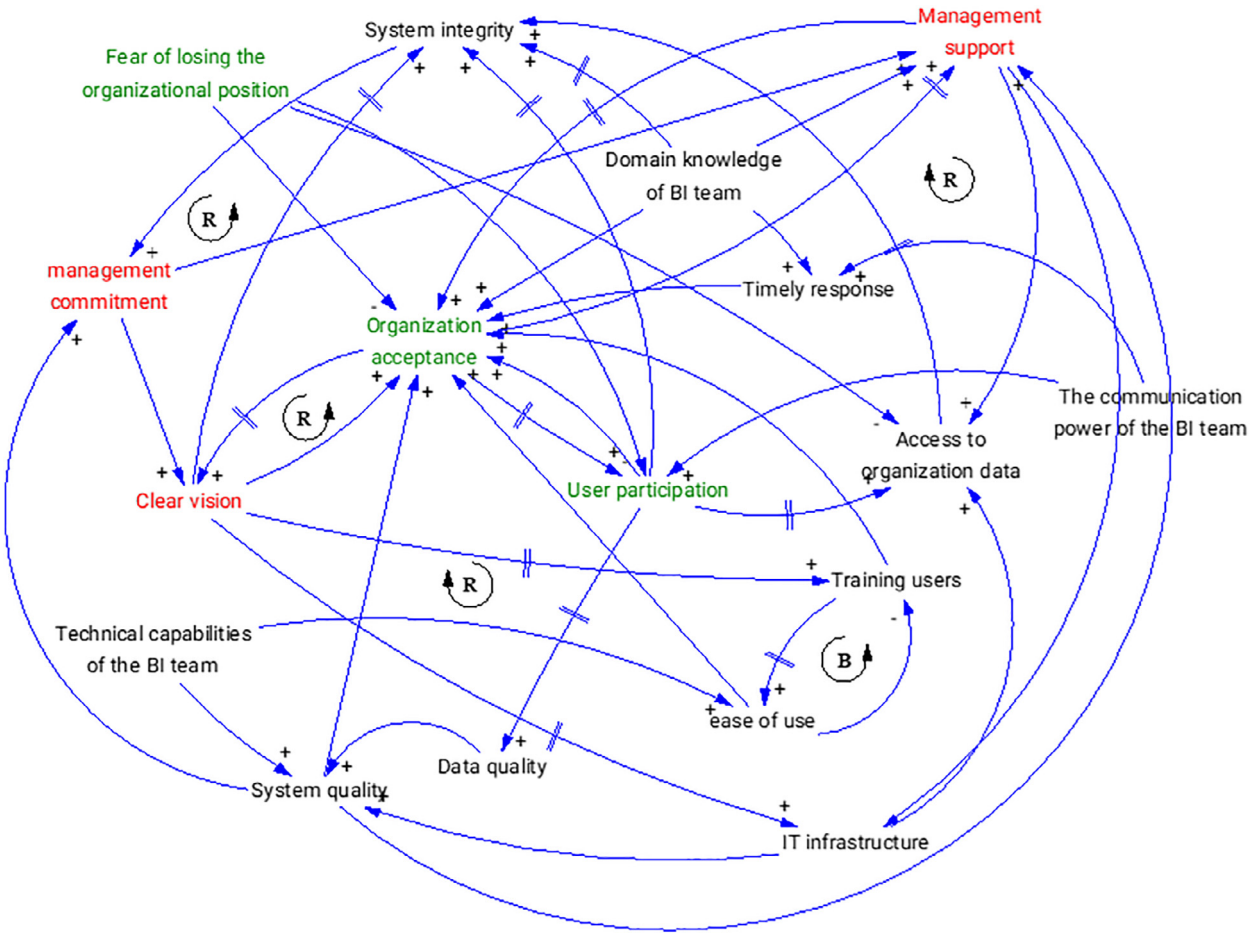


Fig. 3. Cause and effect diagram.

Stock-flow diagram

The flow diagram of the model highly relies on stock and flow variables.

Fig. 4 illustrates the prepared stock-flow diagram for the problem, based on the determined cause-and-effect diagram and five expert opinions. Stocks are depicted as rectangles while streams take on the form of arrows entering or leaving the stocks. In system dynamics modeling, stock and flow concepts are fundamental, with all system equations being written based on these concepts. Stocks denote the state of the system at a specific time and possess properties of aggregation [8]. Their values can increase or decrease through input and output currents, with the stock value obtained from the difference between the two. The remaining variables are covariates that impact not only each other (blue arrows) but also flows and stock.

Formulas were assigned to each variable using the problem flowchart. At this point, expert opinions served as the determining factor. In the initial stage, experts' views were consulted to determine the state variables' values and their rates, based on current Iranian organization conditions. Closed interviews were conducted to gather these opinions, and Table 7 displays the average results of these sessions.

Validation of the model

The model's validation is a prerequisite for utilizing and analyzing the results, as failure to verify its accuracy renders it unusable. In this research, validation is determined through a sentiment analysis test involving variable modifications. Specifically, Fig. 5 depicts changes in the "Fear of Losing Organizational Position" variable, revealing that organizational acceptance decreases significantly when fear is quadrupled (blue line). This natural behavior in the model serves as evidence of its validity.

Table 7
Default conditions considered for exogenous variables based on experts' opinion.

Variable name	The initial numerical value is considered according to the opinion of experts	
	Self-service	Traditional
Technical knowledge of the BI team	0.8	1
Domain knowledge of the BI team	0.9	0.6
Communication strength of the BI team	0.8	0.4
Fear of losing position	0.3	0.8

Results of model execution

Once the model is validated through sensitivity testing, it becomes reliable for analysis following implementation. The system's dynamics are taken into account during the model's implementation, with a focus on the positive and negative interactions between all factors involved. Fig. 6 depicts the results of implementing the model in both traditional and self-service strategies. The model operates over a five-year time horizon and at the monthly level of granularity.

As it is clear from Fig. 6, the behavior of the organizational acceptance variable in traditional and self-service strategies is very close to each other and the direction of the numbers of this variable in each of the months is shown in Table 8.

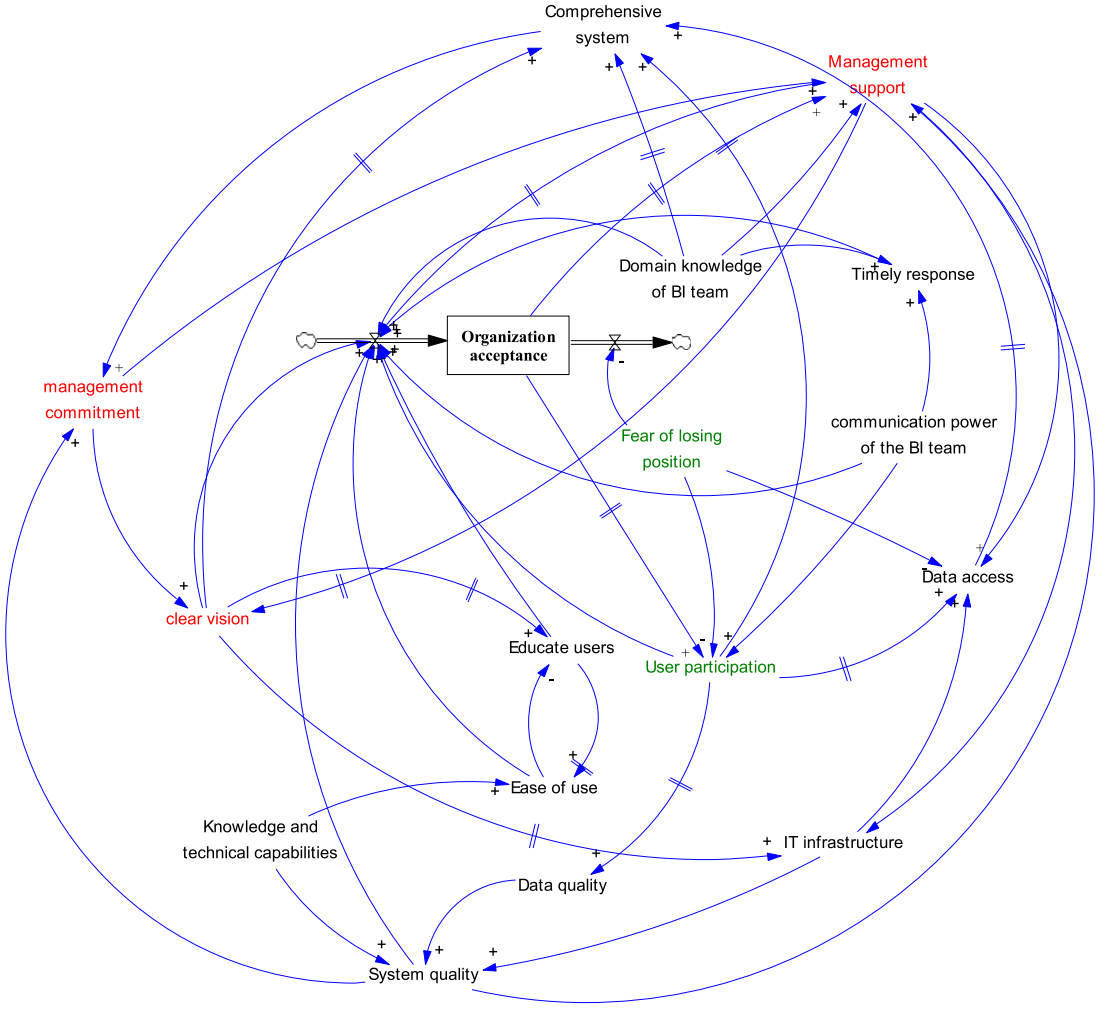


Fig. 4. Stock-flow diagram.

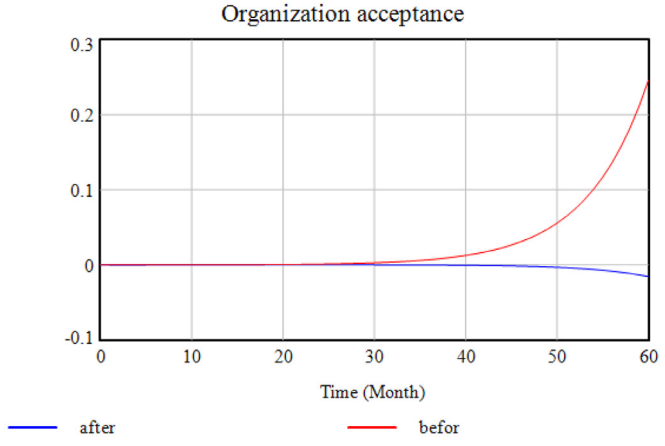


Fig. 5. Diagram of sensitivity analysis of organizational acceptance behavior with changes in the variable of fear of losing organizational position.

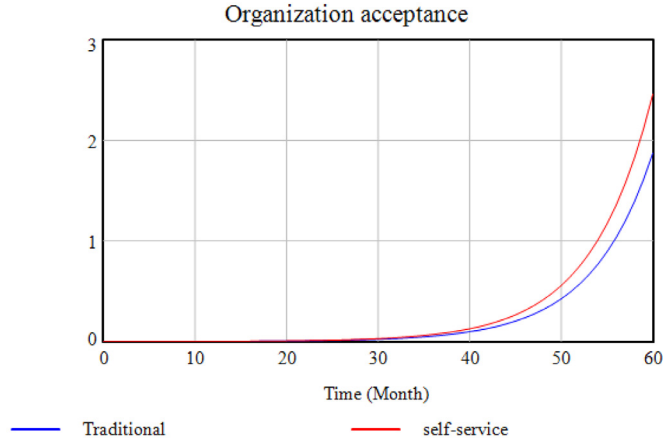


Fig. 6. The variable behavior of organizational acceptance in traditional strategies (red) and self-service (blue).

As it is clear from Fig. 5 and its values in Table 8., the amount of organizational acceptance variable was the same until the twenty-fifth month from the start of project implementation in both traditional and self-service strategies, but after that, the distance between these two strategies increased significantly. He found that the self-service strategy in the 60th month attracted 30% more organizational acceptance than the traditional strategy and is actually more successful.

Check the scenarios

Based on the initial results of the simulation of the two strategies, the self-service strategy has been more successful in attracting organizational acceptance and has overtaken the traditional strategy. However, many IT organizations use this strategy to implement and develop

Table 8
The amount of changes in organizational acceptance variable in traditional and self-service strategies.

Month	Self-service	Traditional	Month	Self-service	Traditional	Month	Self-service	Traditional
1	0.0000	0.0000	**21**	0.0060	0.0056	**41**	0.1460	0.1117
2	0.0001	0.0001	**22**	0.0071	0.0066	**42**	0.1694	0.1296
3	0.0001	0.0001	**23**	0.0083	0.0076	**43**	0.1966	0.1503
4	0.0002	0.0002	**24**	0.0097	0.0089	**44**	0.2282	0.1744
5	0.0003	0.0002	**25**	0.0113	0.0103	**45**	0.2648	0.2023
6	0.0004	0.0003	**26**	0.0132	0.0120	**46**	0.3073	0.2346
7	0.0005	0.0004	**27**	0.0153	0.0140	**47**	0.3566	0.2722
8	0.0007	0.0005	**28**	0.0179	0.0162	**48**	0.4138	0.3157
9	0.0009	0.0007	**29**	0.0208	0.0188	**49**	0.4802	0.3663
10	0.0011	0.0009	**30**	0.0242	0.0218	**50**	0.5572	0.4249
11	0.0013	0.0010	**31**	0.0281	0.0253	**51**	0.6465	0.4929
12	0.0016	0.0013	**32**	0.0327	0.0294	**52**	0.7501	0.5718
13	0.0019	0.0015	**33**	0.0380	0.0341	**53**	0.8704	0.6633
14	0.0022	0.0018	**34**	0.0442	0.0395	**54**	1.0099	0.7695
15	0.0027	0.0022	**35**	0.0513	0.0459	**55**	1.1717	0.8927
16	0.0032	0.0025	**36**	0.0596	0.0532	**56**	1.3595	1.0356
17	0.0037	0.0030	**37**	0.0692	0.0617	**57**	1.5773	1.2014
18	0.0044	0.0035	**38**	0.0804	0.0716	**58**	1.8301	1.3938
19	0.0052	0.0041	**39**	0.0933	0.0830	**59**	2.1233	1.6170
20	0.0060	0.0048	**40**	0.1083	0.0963	**60**	2.4635	1.8759

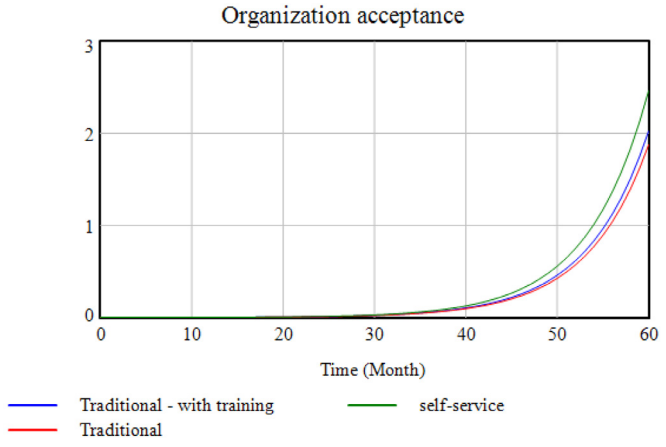


Fig. 7. Changing behavior of organizational acceptance with increasing domain knowledge.

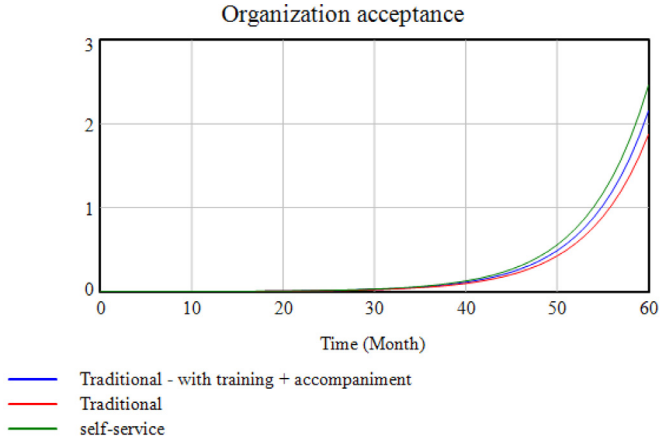


Fig. 8. Organizational acceptance variable behavior with increasing domain knowledge along with companionship.

their business intelligence system. For this reason, in this section, based on the opinions of experts, several scenarios for changes in the initial conditions of the model are discussed:

First scenario: acquisition of domain knowledge

According to the opinion of experts, one of the main shortcomings of the traditional method is the lack of sufficient Domain knowledge in the technical unit (business intelligence) regarding the activity of the unit requesting the dashboard and business intelligence system. To solve this shortcoming, experts suggest the solution of participating in training courses to acquire Domain knowledge of the system applicant. According to experts, this method increases people's knowledge by 20%. The results of this scenario are shown in Fig. 7.

Based on the results obtained from the implementation of the model in this scenario (blue line), the improvement of this situation compared to before after 60 months of project implementation is 11%, which makes the distance between traditional implementation and autonomous implementation reduced to 20%. And the traditional method will be more successful to gain organizational acceptance.

The second scenario: acquiring domain knowledge along with increasing companionship

Another scenario that was suggested by the experts is to increase cooperation with the requesting team (or the team for which the system is developed). One of the serious challenges in the implementation of the business intelligence system is the failure of experts to accompany it due to the fear of losing their position and organizational position. In this case, experts resist using the system by citing reasons such as business intelligence being useless, technical problems of having the system, inability to use the system, etc. Experts have recommended the solution of holding briefing meetings with the manager of the relevant team or the direct technical manager and have stated that this method will improve resistance by 25%. If these conditions are established together with the past scenario and the acquisition of domain knowledge of the team, the results of Fig. 8 will be created.

Based on the results of this scenario (blue line), the amount of gaining organizational acceptance in this scenario has improved by 19% compared to the initial state, and in this case, the difference between the traditional strategy and the self-service strategy in gaining organizational acceptance is 11%. has reduced.

Third scenario: acquiring domain knowledge along with acquiring soft skills

Another challenge mentioned by the experts, who have emphasized it, is the lack of communication skills in the dashboards development team, so these people have a major problem in fully delivering the system to the requester and also responding to their problems, which leads to challenges in the project. becomes Experts have stated the

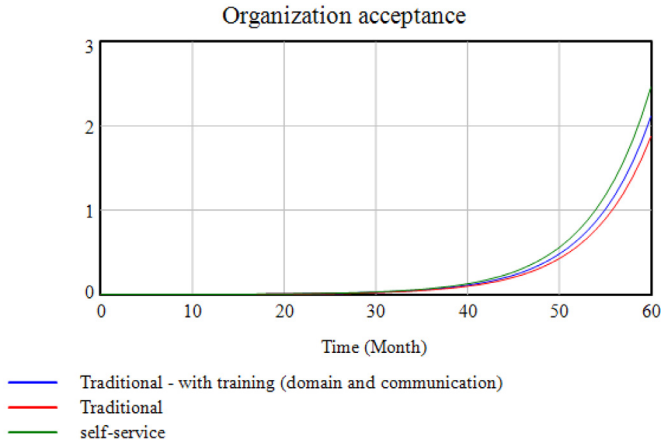


Fig. 9. Changing behavior of organizational acceptance with increasing domain knowledge and increasing communication skills.

reason for this issue is the basic reliance of the business intelligence system development team on technical issues and lack of attention to soft skills. The solution provided by the experts to reduce this problem is holding soft skills and negotiation art courses for the development team. Experts have acknowledged that the effect of these courses (if held as a workshop and problem-oriented) improves conditions by 50%. In this scenario, the changes in the model will be as shown in Fig. 9.

In this scenario (blue line), organizational acceptance has experienced a 17% growth compared to the initial state of the traditional strategy, and this scenario reduces its distance from the self-service strategy to 13%.

Discuss results

The results of the model execution show that implementing the self-service strategy leads to higher levels of organizational acceptance compared to the traditional approach. The model operates over a five-year period at a monthly level of granularity, and it is evident from Fig. 6 and Table 8 that the organizational acceptance variable in both strategies was similar until the twenty-fifth month, after which the distance between the two increased significantly. Specifically, the self-service strategy attracted 30% more organizational acceptance than the traditional method and proved more successful overall.

However, to check the validity of the initial results, several scenarios were discussed based on expert opinions. The first scenario suggested acquiring domain knowledge, and the results showed an improvement of 11% in gaining organizational acceptance compared to the initial state, reducing the difference between the traditional and self-service strategies to 20%. The second scenario involved increasing cooperation with the requesting team, which improved resistance by 25%, and combining this with the acquisition of domain knowledge resulted in a further increase in gaining organizational acceptance by 19% compared to the initial state. The third scenario recommended acquiring soft skills through workshops and problem-oriented training, resulting in a 17% growth in organizational acceptance compared to the initial state of the traditional strategy, which reduced its difference from the self-service strategy to 13%.

Overall, these scenarios suggest that while the self-service strategy initially appeared more successful, there are ways to improve the traditional approach and make it more competitive. By addressing the lack of domain knowledge, increasing cooperation with the requesting team, and developing soft skills in the development team, organizations can increase their chances of successful implementation and gain organizational acceptance for their business intelligence systems.

Conclusion and outlook of research

The purpose of this research was to investigate the impact of business intelligence system implementation strategies in IT organizations on organizational acceptance. The study identified effective factors in the implementation of business intelligence systems through expert consultation and created a dynamic model to simulate the self-service and traditional approaches. The results showed that the self-service strategy led to higher levels of organizational acceptance compared to the traditional approach, with a difference of 30%.

To validate the initial results, three scenarios were proposed by experts: acquiring domain knowledge, increasing cooperation with the requesting team, and developing soft skills in the development team. These scenarios showed that the traditional approach can become more competitive and gain more organizational acceptance by addressing the lack of domain knowledge, improving cooperation with the requesting team, and developing soft skills in the development team. Each scenario resulted in an increase in organizational acceptance by 17–25% compared to the initial state of the traditional strategy.

This research emphasizes the importance of considering different implementation strategies for business intelligence systems and highlights the potential benefits of utilizing a dynamic model to simulate and improve these strategies. By implementing the recommendations from this study, organizations can increase their chances of successful implementation and gain more support for their business intelligence systems. The system dynamics tool provides researchers with the possibility of modeling and simulating complex peripheral phenomena, leading to more logical and accurate decisions based on the results. Thus, this study creates a framework for future research in the area of business intelligence system implementation strategies.

Limitations of the study

There are a few limitations to this study that should be acknowledged. First, the research is based on interviews with only five BI experts. While their opinions and insights were valuable, they may not represent the full range of perspectives on BI implementation strategies. Future research could expand the sample size to include a broader range of stakeholders, including IT managers, business analysts, and end-users.

Second, the simulation model used in this study is based on certain assumptions about the behavior of users and the organization over time. The accuracy of the model depends on the validity of these assumptions, which may not always hold in real-world settings. Future research could use alternative modeling approaches or validate the assumptions through field studies or experiments.

Third, this study focuses exclusively on organizational acceptance as an outcome variable. Other important outcomes of BI implementation, such as user satisfaction, system performance, and business impact, were not considered. Future research could examine these additional outcomes and explore the relationships between them and organizational acceptance.

Future research directions

This study provides a foundation for future research in the area of BI implementation strategies. One potential direction for future research is to investigate the role of data quality and data governance in BI adoption. Data quality and governance are critical factors that can directly affect the usefulness and reliability of BI systems. Understanding how these factors interact with implementation strategies and organizational acceptance can provide valuable insights for improving BI adoption.

Another direction for future research is to explore the impact of organizational culture on BI implementation success. Organizational culture can shape attitudes and behaviors toward BI adoption, and understanding its role can help organizations develop more effective implementation strategies. Future research could examine the relationship between

organizational culture, implementation strategies, and organizational acceptance.

Finally, future research could further explore the potential benefits of using system dynamics modeling in the context of BI implementation. System dynamics can enable researchers to capture the complexity of organizational processes and interactions and simulate different scenarios to test and refine implementation strategies. Further research could investigate the effectiveness of system dynamics modeling in other contexts and explore ways to enhance its accuracy and validity.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.teler.2023.100070.

References

[1] N. Ain, G. Vaia, W.H. DeLone, M. Waheed, Two decades of research on business intelligence system adoption, utilization and success–a systematic literature review, Decis. Support Syst. 125 (2019) 113113.

[2] P. Alpar, M. Schulz, Self-service business intelligence, Bus. Inf. Syst. Eng. 58 (2) (2016) 151–155.

[3] A.L. Antunes, E. Cardoso, J. Barateiro, Incorporation of ontologies in data warehouse/business intelligence systems-a systematic literature review, Int. J. Inf. Manag. Data Insights 2 (2) (2022) 100131.

[4] H. Ashok, S. Ayyasamy, A. Ashok, V. Arunachalam, E-business analytics through ETL and self-service business intelligence tool, 2020S International Conference on Inventive Research in Computing Applications (ICIRCA), 2020.

[5] A. Audzeyeva, R. Hudson, How to get the most from a business intelligence application during the post implementation phase? Deep structure transformation at a UK retail bank, Eur. J. Inf. Syst. 25 (1) (2016) 29–46.

[6] M.P. Bach, A. Čeljo, J. Zoroja, Technology acceptance model for business intelligence systems: preliminary research, Procedia Comput. Sci. 100 (2016) 995–1001.

[7] E. Ballard, A. Farrell, M. Long, Community-Based system dynamics for mobilizing communities to advance school health, J. Sch. Health 90 (12) (2020) 964–975.

[8] M. Bastan, M. Zarei, R. Tavakkoli-Moghaddam, A new technology acceptance model: a mixed-method of grounded theory and system dynamics, Kybernetik 51 (1) (2022) 1–30.

[9] A. Blouchoutzi, G. Tsaples, D. Manou, J. Papathanasiou, Investigating public-private cooperation in migrant labor market integration: a system dynamics study to explore the challenge for Greece, Economies 11 (2) (2023) 38 https://www.mdpi.com/2227-7099/11/2/38.

[10] C. Brewis, S. Dibb, M. Meadows, Leveraging Big Data for Strategic Marketing: a dynamic capabilities model for incumbent firms, Technol. Forecast. Soc. Change 190 (2023) 122402.

[11] Y. Cao, Z.J. Zhou, C.H. Hu, S.W. Tang, J. Wang, A new approximate belief rule base expert system for complex system modelling, Decis. Support Syst 150 (2021) 113558.

[12] N. Caseiro, A. Coelho, Business intelligence and competitiveness: the mediating role of entrepreneurial orientation, Compet. Rev. 28 (2) (2018) 213–226.

[13] M. David Stone, N. David Woodcock, Interactive, direct and digital marketing: a future that depends on better use of business intelligence, J. Res. Interact. Market. 8 (1) (2014) 4–17.

[14] N.A. El-Adaileh, S. Foster, Successful business intelligence implementation: a systematic literature review, J. Work-Appl. Manag. (2019).

[15] L.Y. Fang, N.F.M. Azmi, Y. Yahya, H. Sarkan, N.N.A. Sjarif, S. Chuprat, Mobile business intelligence acceptance model for organisational decision making, Bull. Electric. Eng. Inf. 7 (4) (2018) 650–656.

[16] J. Fjermestad, S. Kudyba, K. Lawrence, in: Business Intelligence and Analytics Case Studies, Taylor & Francis, 2018, pp. 77–78. Vol. 28.

[17] D.N. Ford, A system dynamics glossary, Syst. Dyn. Rev. 35 (4) (2019) 369–379.

[18] H.-P. Fu, T.-H. Chang, Y.-H. Teng, C.-H. Liu, H.-C. Chuang, Critical factors considered by companies to introduce business intelligence systems, Axioms 11 (7) (2022) 338.

[19] R. Gaardboe, T.S. Jonasen, Business intelligence success factors: a literature review, J. Inf. Technol. Manag. 29 (1) (2018) 1–15.

[20] S. Goundar, K. Okafor, A. Cagica, P. Chand, S. Singh, Using business intelligence in organizations, in: Enterprise Systems and Technological Convergence: Research and Practice, 2021, p. 99.

[21] T. Grublješič, P.S. Coelho, J. Jaklič, The shift to socio-organizational drivers of business intelligence and analytics acceptance, J. Organ. End User Comput. 31 (2) (2019) 37–64.

[22] T. Grublješič, J. Jaklič, Business intelligence acceptance: the prominence of organizational factors, Inf. Syst. Manag. 32 (4) (2015) 299–315.

[23] R. Harrison, A. Parker, G. Brosas, R. Chiong, X. Tian, The role of technology in the management and exploitation of internal business intelligence, J. Syst. Inf. Technol. (2015).

[24] L. Harst, H. Lantzsch, M. Scheibe, Theories predicting end-user acceptance of telemedicine use: systematic review, J. Med. Internet Res. 21 (5) (2019) e13117, doi:10.2196/13117.

[25] J. Hou, L. Wang, Z. Lin, M. Gao, Research on operational effectiveness evaluation of network information system based on system dynamics, 2021 International Conference on Electronics, Circuits and Information Engineering (ECIE), 2021.

[26] Z. Hussain, A. Jabbar, K. Kong, Power, Dominance and Control: Implementing a New Business Intelligence System, Digital Transformation and Society, 2023.

[27] T. Jafari, A. Zarei, A. Azar, A. Moghaddam, The impact of business intelligence on supply chain performance with emphasis on integration and agility–a mixed research approach, Int. J. Prod. Perform. Manag. (2021).

[28] M. Jalilvand Khosravi, M. Maghsoudi, S Salavatian, Identifying and clustering users of VOD platforms using SNA technique: a case study of cinemamarket, New Market. Res. J. 11 (4) (2022) 20–21.

[29] Z. Jamshidi, S.M. Sajadi, K. Talebi, S.H. Hosseini, Applying system dynamics approach to modelling growth engines in the international entrepreneurship era, in: Empirical International Entrepreneurship, Springer, 2021, pp. 491–513.

[30] M.A. Kermani, M. Maghsoudi, M.S. Hamedani, A. Bozorgipour, Analyzing the interorganizational collaborations in crisis management in coping with COVID-19 using social network analysis: case of Iran, J. Emerg. Manag. 20 (3) (2022) 249–266.

[31] A. Kumar, B. Krishnamoorthy, Business analytics adoption in firms: a qualitative study elaborating TOE framework in India, Int. J. Glob. Bus. Compet. 15 (2) (2020) 80–93.

[32] C. Lennerholt, Facilitating the Implementation and Use of Self Service Business Intelligence, University of Skövde, 2022].

[33] C. Lennerholt, J.v. Laere, E. Söderström, Success factors for managing the SSBI challenges of the AQUIRE framework, J. Decis. Syst. (2022) 1–22, doi:10.1080/12460125.2022.2057006.

[34] C. Lennerholt, J. van Laere, Data access and data quality challenges of self-service business intelligence, 27th European Conference on Information Systems (ECIS), 2019.

[35] C. Lennerholt, J. van Laere, E. Söderström, Implementation challenges of self service business intelligence: a literature review, 51st Hawaii International Conference on System Sciences, 2018.

[36] C. Lennerholt, J. Van Laere, E. Söderström, User-related challenges of self-service business intelligence, Inf. Syst. Manag. 38 (4) (2021) 309–323, doi:10.1080/10580530.2020.1814458.

[37] P.M. Leonardi, COVID-19 and the new technologies of organizing: digital exhaust, digital footprints, and artificial intelligence in the wake of remote work, J. Manag. Stud. 58 (1) (2021) 249.

[38] Q. Li, L. Zhang, L. Zhang, S. Jha, Exploring multi-level motivations towards green design practices: a system dynamics approach, Sustain. Cities Soc. 64 (2021) 102490.

[39] X. Li, J.P.-A. Hsieh, A. Rai, Motivational differences across post-acceptance information system usage behaviors: an investigation in the business intelligence systems context, Inf. Syst. Res. 24 (3) (2013) 659–682.

[40] M.R. Llave, Data lakes in business intelligence: reporting from the trenches, Procedia Comput. Sci. 138 (2018) 516–524.

[41] E. Malbon, J. Parkhurst, System dynamics modelling and the use of evidence to inform policymaking, Policy Stud. (2022) 1–19.

[42] M.I. Merhi, Evaluating the critical success factors of data intelligence implementation in the public sector using analytical hierarchy process, Technol. Forecast. Soc. Change 173 (2021) 121180.

[43] A. Mousavi, M. Mohammadzadeh, H. Zare, Developing a system dynamic model for product life cycle management of generic pharmaceutical products: its relation with open innovation, J. Open Innov. 8 (1) (2022) 14 https://www.mdpi.com/2199-8531/8/1/14.

[44] J. Müller, G. Schuh, D. Meichsner, G. Gudergan, Success factors for implementing Business Analytics in small and medium enterprises in the food industry, 2020 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), 2020.

[45] S. Nalchigar, E. Yu, S. Easterbrook, Towards actionable business intelligence: can system dynamics help? IFIP Working Conference on The Practice of Enterprise Modeling, 2014.

[46] N. Nithya, R. Kiruthika, Impact of Business Intelligence Adoption on performance of banks: a conceptual framework, J. Ambient Intell. Humaniz Comput. 12 (2021) 3139–3150.

[47] Y. Niu, L. Ying, J. Yang, M. Bao, C. Sivaparthipan, Organizational business intelligence and decision making using big data analytics, Inf. Process. Manag. 58 (6) (2021) 102725.

[48] C.I. Papanagnou, O. Matthews-Amune, Coping with demand volatility in retail pharmacies with the aid of big data exploration, Comput. Oper. Res. 98 (2018) 343–354.

[49] J. Passlick, L. Grützner, M. Schulz, M.H. Breitner, Self-service business intelligence and analytics application scenarios: a taxonomy for differentiation, Inf. Syst. e-Bus. Manag. (2023), doi:10.1007/s10257-022-00574-3.

[50] M.D. Peters, B. Wieder, S.G. Sutton, J. Wakefield, Business intelligence systems use in performance measurement capabilities: implications for enhanced competitive advantage, Int. J. 21 (2016) 1–17.

[51] G. Phillips-Wren, M. Daly, F. Burstein, Reconciling business intelligence, analytics and decision support systems: more data, deeper insight, Decis. Support Syst. 146 (2021) 113560.
[52] I. Pluchinotta, A. Pagano, T. Vilcan, S. Ahilan, L. Kapetas, S. Maskrey, V. Krivtsov, C. Thorne, E O'Donnell, A participatory system dynamics model to investigate sustainable urban water management in Ebbsfleet Garden City, Sustain. Cities Soc. 67 (2021) 102709.
[53] M.J. Radzicki, System dynamics and its contribution to economics and economic modeling, in: System Dynamics: Theory and Applications, 2020, pp. 401–415.
[54] J. Ranjan, C. Foropon, Big data analytics in building the competitive intelligence of organizations, Int. J. Inf. Manage. 56 (2021) 102231.
[55] A.Z. Ravasan, S.R. Savoji, An investigation of BI implementation critical success factors in Iranian context, Int. J. Bus. Intell. Res. 5 (3) (2014) 41–57, doi:10.4018/ijbir.2014070104.
[56] F. Ricciardi, P. De Bernardi, V. Cantino, System dynamics modeling as a circular process: the smart commons approach to impact management, Technol. Forecast. Soc. Change 151 (2020) 119799.
[57] K. Saeed, A. Sidorova, A. Vasanthan, The bundling of business intelligence and analytics, Int. J. Comput., Inf., Syst. Sci., Eng. (2022) 1–12.
[58] G.B. Sajons, Estimating the causal effect of measured endogenous variables: a tutorial on experimentally randomized instrumental variables, Leadersh Q 31 (5) (2020) 101348.
[59] D. Schuff, K. Corral, R.D. St. Louis, G Schymik, Enabling self-service BI: a methodology and a case study for a model management warehouse, Inf. Syst. Front. 20 (2018) 275–288.
[60] S. Shafiee, S. Jahanyan, A.R. Ghatari, A. Hasanzadeh, Developing sustainable tourism destinations through smart technologies: a system dynamics approach, J. Simul. (2022) 1–22.
[61] F. Sönmez, Technology acceptance of business intelligence and customer relationship management systems within institutions operating in capital markets, Int. J. Acad. Res. Bus. Soc. Sci. 8 (2) (2018) 400–422.
[62] V. Srikrishnan, D.C. Lafferty, T.E. Wong, J.R. Lamontagne, J.D. Quinn, S. Sharma, N.J. Molla, J.D. Herman, R.L. Sriver, J.F. Morris, Uncertainty analysis in multi-sector systems: considerations for risk analysis, projection, and planning for complex systems, Earth's Fut. 10 (8) (2022) e2021EF002644.
[63] C.A. Tavera Romero, J.H. Ortiz, O.I. Khalaf, A. Ríos Prado, Business intelligence: business evolution after industry 4.0, Sustainability 13 (18) (2021) 10026.
[64] J.R. Turner, R.M. Baker, Complexity theory: an overview with potential applications for the social sciences, Systems 7 (1) (2019) 4.
[65] N. Ul-Ain, G. Vaia, W. DeLone, Business intelligence system adoption, utilization and success-A systematic literature review, in: Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.
[66] P.D. Vecchio, G. Secundo, Y. Maruccia, G. Passiante, A system dynamic approach for the smart mobility of people: implications in the age of big data, Technol. Forecast. Soc. Change 149 (2019) 119771, doi:10.1016/j.techfore.2019.119771.
[67] N. Vella, Business Intelligence and Data-Driven Decision-Making: A Management Accounting Perspective, University of Malta, 2021.
[68] M.A. Villanthenkodath, M.A. Ansari, P. Kumar, Y.N. Raju, Effect of information and communication technology on the environmental sustainability: an empirical assessment for South Africa, Telemat. Inform. 7 (2022) 100013, doi:10.1016/j.teler.2022.100013.
[69] L.L. Visinescu, M.C. Jones, A. Sidorova, Improving decision quality: the role of business intelligence, Int. J. Comput., Inf., Syst. Sci., Eng. 57 (1) (2017) 58–66.
[70] M. Wee, Business Intelligence & Analytics Adoption in Australian SMEs: Identified Processes, Decision-Making, and Leadership Skills, Swinburne University of Technology, 2021].
[71] P. Weichbroth, J. Kowal, M. Kalinowski, Toward a unified model of mobile Business Intelligence (m-BI) acceptance and use, in: Proceedings of the 55th Hawaii International Conference on System Sciences, 2022.
[72] W. Yeoh, A. Popovič, Extending the understanding of critical success factors for implementing business intelligence systems, J. Assoc. Inf. Sci. Technol. 67 (1) (2016) 134–147.
[73] F. Zare, S. Elsawah, A. Bagheri, E. Nabavi, A.J. Jakeman, Improved integrated water resource modelling by combining DPSIR and system dynamics conceptual modelling techniques, J. Environ. Manage. 246 (2019) 27–41.
[74] M. Zhang, M. Cheng, Big data, social media, and intelligent communication, Telemat. Inform. 8 (2022) 100026, doi:10.1016/j.teler.2022.100026.
[75] X. Zhang, H. Ding, Y. Wang, Study on business intelligence products supporting knowledge management, in: Frontiers in Enterprise Integration, CRC Press, 2020, pp. 213–222.

International Journal of Information Management Data Insights 2 (2022) 100129

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

Context-aware OLAP for textual data warehouses

Santanu Roy [a,*], Agostino Cortesi [b], Soumya Sen [c]

[a] Future Institute of Engineering and Management, Kolkata, India
[b] DAIS, Ca' Foscari University, Venice, Italy
[c] University of Calcutta, Kolkata, India

ARTICLE INFO

Keywords:
OLAP
Text data warehouse
Information System
Concept hierarchy
Word embedding
Agglomerative hierarchical clustering

ABSTRACT

Decision Support Systems (DSS) that leverage business intelligence are based on numerical data and On-line Analytical Processing (OLAP) is often used to implement it. However, business decisions are increasingly dependent on textual data as well. Existing research work on textual data warehouses has the limitation of capturing contextual relationships when comparing only strongly related documents. This paper proposes an Information System (IS) based context-aware model that uses word embedding in conjunction with agglomerative hierarchical clustering algorithms to dynamically categorize documents in order to form the concept hierarchy. The results of the experimental evaluation provide evidence of the effectiveness of integrating textual data into a data warehouse and improving decision making through various OLAP operations.

1. Introducton

With the incessant growth of textual information in a variety of business systems, it has become more desirable for organizations to analyze both structured data records and unstructured text data simultaneously. In recent times in order to automate the process of organizational data analysis for extractions of business intelligence, the enterprises apply Information System (IS) based work systems (Struijk, Ou, Davison, & Angelopoulos, 2022). IS is a system in which human participants and/or machines perform work (processes and activities) using information, technology, and other resources to produce informational products and/or services for internal or external customers. Performing On-line Analytical Processing (OLAP) operations on data warehouses have been the most widely used technique by the organizations to implement IS enabled decision support systems.

While OLAP tools have been proven very useful for handling structured data, they face challenges in handling text data. Usually, data warehousing technologies and OLAP tools are unable to analyze textual data. Moreover, as OLAP queries of a decision-maker are generally related to a context, contextual information must be taken into account during the exploitation of data warehouses. OLAP systems allow navigation through multiple dimensions from one view to another which can be effectively used to analyze big data. In order to deal with textual data information retrieval (IR) techniques are generally used to evaluate the relevance of data to a query composed of simple keywords expressing needed information. Most often this relevance is based on the terms' frequency in the document. But in a text-OLAP system, the interest is in navigational analysis which may be based on operations corresponding to the analysis of text context at different levels in a data warehousing model.

1.1. The research question

Traditionally data analysis focuses on business data managed by a decision support system with data being mostly stored in data warehouses or structured files. In the era of digitalization and the prolific rise of big data, business analytics must evolve constantly. The volume of unstructured data is more rapidly growing in comparison to the growth of structured data. According to Gartner's magic quadrant of 2019, unstructured data is growing by 30% to 60% year over year. According to the figures from the ITC research firm, the volume of unstructured data is set to grow from 33 zettabytes in 2018 to 175 zettabytes, or 175 billion terabytes by 2025. In many complex fields, such as academics, research communities, company Human Resource activities, medical diagnosis, social media feedbacks, online customer feedback and customer support, decision-makers require helpful indicators and tools to make analyze text data and make business decision. Over the years, data warehouses and OLAP tools have emerged as most useful Information Systems of managing his huge volume of data assist users in the process of business decision making. Data warehouses can be implemented using several data models. Multidimensional database, represented by Multidimensional Data Model (MDM) is often a part of a data warehouse. This model is defined using set of dimensions and facts. The indicators to assess the facts are known as measures. Dimensions are the perspectives or entities based on which an organization wants to perform analytical

* Corresponding author.
E-mail addresses: santanuroy84@gmail.com (S. Roy), cortesi@unive.it (A. Cortesi), iamsoumyasen@gmail.com (S. Sen).

https://doi.org/10.1016/j.jjimei.2022.100129
Received 21 September 2021; Received in revised form 1 October 2022; Accepted 4 October 2022

processing. Each dimension may be associated with a hierarchy known as concept hierarchy. For the navigation and the visualization OLAP uses operations such as roll-up, drill-down, slice, and dice (Sen, Roy, Sarkar, Chaki, & Debnath, 2014).

The traditional OLAP tools are effective when data are numerical but they are not suitable for unstructured data such as text. Because of the fast growth of textual data there is a need for new approaches that take into account the textual content of data in OLAP analysis and it is called text-OLAP. However, this involves not only dealing with the heterogeneity of representations and granularities but also dealing with large volume of data. Large volume of text documents are generated everyday in every organization. Consequently, documents should be integrated into the decision support system. The perfect process of integrating unstructured data on the context of data warehouses is to manage, query and visualize information in a way that is as effective and meaningful with structured data.

In order to capture the notion of text-OLAP it is important to propose OLAP operations to process and analyze textual data and summarize them into an OLAP cube (Cuzzocrea, 2020) for fast and effective decision making. The long-established OLAP operations can't be applied in their conventional form due to the complex nature of unstructured text data (Zhang, Wang, & Feng, 2018).To use the OLAP with textual data, text mining provides the necessary techniques for textual aggregation. Research study reveals there have been few attempts to perform OLAP operations on text data but even the current state-of-the art algorithms on text-OLAP are unable to extract the semantic information from the texts with immaculate precision and accuracy. Embedding the semantics and context in textual data in OLAP analysis (Oukid, Benblidia, Asfari, Bentayeb, & Boussaid, 2015) and aggregating them to enhance the decision-making is a challenge in business intelligence systems. Therefore, it is imperative to modify the traditional data warehousing models and introduce new aggregation techniques (Sen et al., 2014) appropriate for text-OLAP. Most of the existing works employ information retrieval (IR) techniques (Kosmopoulos, Androutsopoulos, & Paliouras, 2015; Lin, Ding, Han, Zhu, & Zhao, 2008; Oukid et al., 2015) to evaluate the Semantic Textual Similarity (STS) between a set of text documents and an OLAP aggregation query containing simple keywords to express the desired information. Often this context analysis is based upon Term Frequency and Inverse Document Frequency (TF-IDF) or Bag-of-Word (BOW) methods (Chakrabarty, Roy, & Roy, 2018; Kim & Gil, 2019; Oukid et al., 2015; Ravat, Teste, Tournier, & Zurfluh, 2008). However these techniques are inadequate to capture the similar contexts across different levels of a dimension table. Thus, the results generated from IR systems suffer from the limitation of extracting contextual information in the development of decision support system (Sarkar & Shankar, 2021) from text data warehouse. Moreover, for a dimension having concept hierarchy (Sen et al., 2014), these feature based Vector Space Models (VSM) are often not suitable to extract the hierarchical relationships among documents due to their frequent near-orthogonality and inability to capture the semantic similarity as a metric of distance between different words having similar meanings or contexts.

This study identifies the possible opportunities for business analysis on text data warehouses by embedding context into the model and subsequently performing OLAP operations. These textual data can attribute to the different decision making process for any organization. In this study the authors propose IS based context-aware work system model that integrate word embedding with agglomerative hierarchical clustering algorithm to perform OLAP operations on textual data warehouses to generate IT enabled corporate reports that may aid in fast and effective business decision making.

1.2. Summary of the proposed methodology

The proposed model presents a novel methodology for the creation of a textual data warehouse with textual dimensions organized by contexts (set of topics) named as contextual dimensions and its implementation in a real OLAP system. This study uses star schema (Sen et al., 2014) to build the conceptual textual data warehousing model.

The proposed methodology processes the text documents and constructs a data cube around a central theme of analysis called fact table F defined by several dimensions Dim_R where $R \in [1, *]$. A set measure(s) of a fact table F is denoted by M, stores values to be aggregated. A Fact F with its dimensions Dim_R and set of measure(s) M, form a star schema model which is formalized as: $\$F$; Dim_1, Dim_2,..., Dim_* ; M_1, M_2,..., M_*).

After arranging the documents according to the star schema, proposed methodology combines word embeddings (De Miranda, Pasti, & de Castro, 2019; Ángel González, Hurtado, & Pla, 2020; Maas et al., 2011; Mikolov, Chen, Corrado, & Dean, 2013a; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013b) (De Miranda et al., 2019) in conjunction with agglomerative hierarchical clustering method to group similar documents by extracting their contextual similarities. At first, the text documents are represented by their word embedding based centroid word vectors and then the hierarchical agglomerative clustering algorithm is applied to word centroid vectors to cluster the documents according to a concept hierarchy existing among the contextual dimensions. This overall approach is dynamic as it is not necessary to declare the number of clusters in the design hierarchy at the beginning of algorithm execution. The dynamically constructed hierarchy of concepts will comprise a sequence of hierarchical mappings ranging from a set of low-level concepts to a broader higher-level concept. The contextual dimension will allow the decision makers to analyze and query on the set of documents after selecting a context that has been automatically extracted during the formation of the concept hierarchy.

1.3. Validation of the proposed methodology

In order to validate our proposed model experimental studies have been carried out on huge sets of publicly available resumes (bio-data) collected from different job portals that can facilitate search using skill set, the domain of specialization, location of a person, and experience as contextual dimensions in the multi-dimensional text data warehousing model. The idea of working on resume dataset has been adopted from the research work carried out in the paper by Oukid et al. (2015). However, the collection of the resumes and the corresponding customization of the resumes into suitable format has been done by the authors of this paper. Therefore the authors work with the resume data set preprocessed and prepared by themselves. As an example, in the resume dataset, the dimension Topic (Dim_T) contains a concept hierarchy on the skill-set specialization domain. An example of concept hierarchy is illustrated in Fig. 1.

1.4. Novelty and the findings of the proposed methodology

The novelty and the findings of the proposed methodology may be listed as follows:

- The proposed methodology uses word embedding algorithm to represent each document by its centroid word vectors. Experiments have been carried out to show the effectiveness of this approach in extracting contextual similarity between documents having very few terms in common. In the result analysis section, it has been shown the proposed word embedding based approach is superior in extracting contextual similarity in comparison to the state-of-the-art VSM models using TF-IDF approaches. The enhanced performance by capturing contextual similarity of the proposed method is measured by the cosine based similarity measure.

- The agglomerative hierarchical clustering algorithm categorizes the text documents (resumes) according to a concept hierarchy. The proposed method shows highly improved performance in dynamically forming the concept hierarchy based on contextual dimensions in comparison to the state-of-the-art methods. The novelty of the proposed

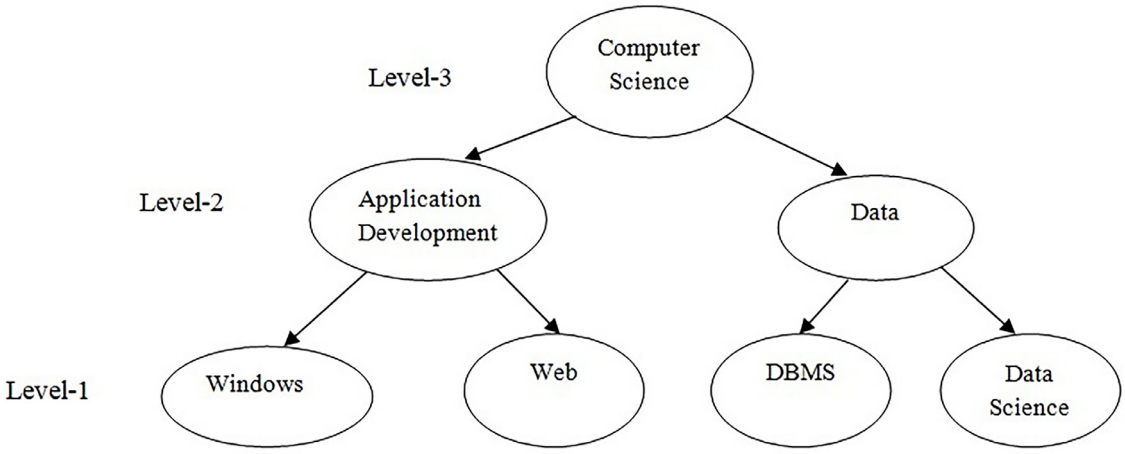


Fig. 1. Example of concept hierarchy of the dimension-Topic (Dim_T).

method of the formation of the concept hierarchy is understood by the following two facts: a) Number of clusters (concepts) is not needed to be declared before the execution of the algorithm. According to the proposed methodology, different linkage criteria of the agglomerative algorithm are applied and accordingly Silhouette Score(s) is measured by varying the number of clusters. The number of clusters and the corresponding linkage criterion producing the highest Silhouette score are chosen dynamically by the proposed algorithm. b) During the formation of the concept hierarchy new concepts are automatically added to the concept hierarchy based on their contextual similarity. This is a major improvement over the state-of-the-art algorithms which try to capture contextual similarity based on Vector Space Model (VSM). Existing works require the leaf concepts in the concept hierarchy to be supplied (as input) statically as a collection of a few terms or words belonging to the different concepts (topics). Thus the existing approaches are heavily limited in dynamically adding new concepts represented by new words or terms during the processing of the documents. c) After the formation of the concept hierarchy a case study (related to the H.R. Manager's activity in correspondence to a job advertisement and subsequent job applications by candidates) is carried out to perform OLAP aggregation operations on the set of documents. The proposed model in this research work can extract useful information to support business intelligence. d) The state-of-the-art algorithms represent the text documents by TF-IDF based vectors. With the increase in the number of documents, these sparse TF-IDF based vectors can be very high dimensional. In contrast, the proposed model represents each document by less dimensional dense centroid word vectors. Therefore, processing the less dimensional word vectors for contextual similarity computation with a posed OLAP query is much faster than the processing time required to scan through the high dimensional TF-IDF vectors. Hence, during experimental evaluation proposed model shows considerable improvement in speeding up the execution time of OLAP operations.

This paper can be seen as a further contribution in the recent discussions of the IJIM Data Insights on issues related to the application of data analytics techniques for decision support systems. In particular, it is worth mentioning that our contribution on analytical processing of textual data can be combined in an effective way with the recent advancements on Natural Language Processing (NLP) and Big Data Analytics (Atkinson & Escudero, 2022; Georgiadou, Angelopoulos, & Drake, 2020), as well as on security management of textual contents (Fujii, Sakaji, Masuyama, & Sasaki, 2022; Wadud et al., 2022), business intelligence (Unhelkar et al., 2022) and complex decision-making problems (Razavisousan & Joshi, 2022).

This paper aims to provide a model to perform OLAP operations in textual data warehouses. The proposed model can capture the contextual similarity between the documents and thus categorizes the documents according to a dynamically formed concept hierarchy existing among contextual dimensions.

1.5. Organisation of the paper

The rest of the paper is organized as follows: In Section 2, we discuss the related works in the domain of text-OLAP. The limitations of the existing works is discussed in Section 3. Section 4 presents the foundation concepts associated with the proposed methodology. The proposed methodology is described in Section 5. Materials and methods needed for the experimental evaluation are discussed in Section 6. Experimental results and the performance analysis of the proposed methodology is discussed in Section 7. Discussion on the effectiveness work is presented in Section 8. Section 9 concludes.

2. Related work

In this digital era, data warehouses are extensively used in the industry for organizing and analyzing large amounts of data. A survey work presented in Bouakkaz, Ouinten, Loudcher, & Strekalova (2017) broadly classifies the text-OLAP and aggregation techniques into two major categories, approaches based on the data structure such as the proprieties of the data cube, and approaches that are not based on the data structure. Approaches that are not based on the data structure are further classified into four subcategories, approaches based on linguistic knowledge, approaches based on external knowledge, approaches based on graphs and approaches based on statistical information. Details concerning these approaches are developed next.

2.1. Approaches based on data structure and data models

The X-OLAP: (XML-OLAP) proposed by Park, Han, & Song (2005) is based on the text mining approach. XML-OLAP is based on the text mining technique that aggregates the text content of XML documents. This approach to analyzing XML documents stored in a data warehouse is represented by a multidimensional model.

The DocCube: DocCube was introduced by Mothe, Chrisment, Dousset, & Alaux (2003). It treats several facts of a document as dimensions. These dimension tables are similar to the standard of OLAP systems. Nevertheless, the major characteristic of DocCube lies like the content of a fact table that contains links.

Topic Cube: Zhang, Zhai, & Han (2009) proposed an approach called Topic Cube, the main idea of a topic cube is to use the hierarchical topic tree as the hierarchy for the text dimension. This structure allows a user to drill-down and roll-up along this tree and discovers the content of the text documents.

Text Cube: In order to introduce the semantic aspect in the textual aggregation Lin et al. (2008) proposed an approach for data cube called Text Cube. The main idea is to give the user the possibility to make semantic navigation in the data dimension. To achieve that, two OLAP operations such as the pull-up and push-down.

The R-Cube: Perez, Aramburu, Berlanga, & Pedersen (2007) focus on the task of integrating structured and textual data in the same data warehouse. The authors proposed an architecture for a decision support system called contextualized warehouse that allows a user to obtain knowledge from heterogeneous data and documents by analyzing data under different contexts.

The Cube Index: Azabou, Khrouf, Feki, Soulé-Dupuy, & Vallès (2015) proposed a model called Cube Index based on a hierarchical description of each document. This hierarchy specifies relationships between words with respect to one document. It is used for the analysis of words in various levels of abstraction in a document. It supports TF-IDF (Term Frequency-Inverse Document Frequency) to facilitate information retrieval techniques.

2.2. Approaches based on content

The approaches that, describe document warehousing through the most representative keywords without using the structure of data or the proprieties of cube, found in the literature can be classified into four categories. The first one is based on linguistic knowledge, the second one is based on the use of external knowledge, the third one is based on graphs, and the last uses statistical methods.

2.2.1. Approaches based on linguistic knowledge

The approaches based on linguistic knowledge consider a corpus as a set of the vocabulary mentioned in the documents but the results are sometimes ambiguous. To overcome this obstacle, techniques based on lexical knowledge and syntactic knowledge previews have been introduced. Kohomban & Lee (2007) described a classification of textual documents based on scientific lexical variables of discourse. Among these lexical variables, they chose nouns because they are more likely to emphasize scientific concepts, rather than adverbs, verbs, or adjectives.

2.2.2. Approaches based on external knowledge

The approaches based on the use of external knowledge select certain keywords that represent a domain. These approaches often use models of knowledge such as ontology. Ravat, Song, Teste, & Trojahn (2020) proposed an aggregation function that takes as input a set of keywords extracted from documents of a corpus and outputs another set of aggregated keywords. They assumed that both the ontology and the corpus of documents belong to the same domain. Oukid et al. (2015) proposed an aggregation operator Orank (OLAP rank) that aggregates a set of documents by ranking them in a descending order using a vector space representation. The same concept propagation technique has been used in the research work discussed in Chakrabarty et al. (2018). The work (Chakrabarty et al., 2018) uses a context-aware Fuzzy Classification based technique to capture the semantic ontology from the text documents and classifies them to aggregate in terms of their relevant concepts.

2.2.3. Approaches based on graph

The approaches based on graphs use keywords to construct graphs where each node represents a keyword obtained after preprocessing and candidate selection. An edge represents the strength or relatedness (or semantic relatedness) between two keywords. After the graph representation step, different types of keyword-ranking approaches have been tried. The first proposed is an approach called TextRank (Mihalcea & Tarau, 2004) where the edges represent the co-occurrence relations between the keywords. Two successive research works by Bouakkaz, Loudcher, & Ouinten (2016) focus on textual aggregation techniques. In their earlier work Bouakkaz et al. (2016) proposed a method that performs aggregation of keywords of documents based on the construction of a graph using the affinities between keywords. Term Frequency (TF) based keyword extraction technique has been used in this work. The following work (Bouakkaz et al., 2017) tries to capture semantic aggregation of the keywords by applying k-means algorithm using Google Similarity Distance Measure.

2.2.4. Approaches based on statistical methods

The approaches based on statistical methods use the occurrence frequencies of terms and the correlation between terms. Landauer, Foltz, & Laham (1998) proposed a method called the Latent Semantic Analysis (LSA) in which the corpus is represented by a matrix where the rows represent the documents and the columns represent the keywords. Ravat et al. (2008) proposed a second aggregation function called TOP-Keywords to aggregate keywords. They computed the frequencies of terms using the TF-IDF function, and then they selected the first k most frequent terms.

The papers discussed in this Section offer quite a large choice of methodologies applicable to a variety of datasets to perform OLAP operations on text data. Most of the techniques conglomerate text mining approaches with OLAP aggregation operations.

3. Limits of the state-of-the-art methodologies

It is identified from the literature survey of Section 2 that the existing research works on text-OLAP suffer from the following limitations. 1 summarizes a few of the works which deal with context-aware textual data warehouses.

1. Works described in Azabou et al. (2015), Bouakkaz et al. (2016), Chakrabarty et al. (2018), Manuel Pérez-Martínez, Berlanga-Llavori, Aramburu-Cabo, & Pedersen (2008), Oukid et al. (2015), Ravat et al. (2008) try to focus on capturing contextual information during text-OLAP analysis. However, in all of these schemes, the documents are represented using either of the models between the BOW model, TF calculation, or using TF-IDF feature vectors. These techniques are often not suitable to grasp the semantic relationships between contexts during the comparison of related parts of different documents. In BOW representations each word of the vocabulary is represented as a 'one-hot' vector with as many components (features) as the size of the vocabulary, and only one non-zero component (corresponding to the particular word). Thus the resulting vector is a high dimensional sparse vector (mostly zero components). Standard feature selection algorithms can be used to reduce the dimension. However, if the concept hierarchy is formed in bio-medical text document datasets for OLAP analysis then the number of concepts (class) may extend up to the order of a few thousands (Kosmopoulos et al., 2015). In these kinds of scenarios even with the least number of features per class (after application of the feature selection algorithm), the total number of features representing each document may contain a significantly huge number of features in the Vector Space Model (VSM). OLAP query processing in these high dimensional feature vectors can be very slow.
2. Regarding text-OLAP, very little number of works have addressed the concept hierarchy (Sen et al., 2014) existing in a certain contextual domain. Studies suggested in Chakrabarty et al. (2018) and Oukid et al. (2015) highlight the contextual dimensions having concept hierarchy. Both of the approaches use the relevance/concept propagation technique to calculate contextual term weights of the documents across the different levels of the concept hierarchy. However, both of these assume a static structure of the concept hierarchy with a few arbitrary terms related to a concept. This method is highly inefficient as any concept (other than the statically mentioned concepts at the beginning) discovered with the increasing size of the dataset does not get categorized into a proper domain of specialization topic (class). Study

Table 1
Comparison of the works on context-aware text-OLAP.

Works	Data Format	Approach used	Formation of Concept Hierarchy	Nature of Concept hierarchy
(Ravat et al., 2008)	XML	TF-IDF	No	NA
(Azabou et al., 2015)	Text	TF-IDF	No	NA
(Lin et al., 2008)	Text	Cube	No	NA
(Bouakkaz et al., 2016)	Text	Graph/TF-IDF	No	NA
(Bouakkaz et al., 2017)	Text	Graph/TF-IDF	No	NA
(Oukid et al., 2015)	Text	TF-IDF	Yes	Static
(Chakrabarty et al., 2018)	Text	TF-IDF	Yes	Static
Proposed Work	Text	Word Vector	Yes	Dynamic

suggested in Bouakkaz et al. (2017) tries to aggregate the keywords by using Google Similarity Distance Measure. However, this study also suffers from the problem of static declaration of the number of clusters as it uses k-Means algorithm to find the similarity between keywords.

4. Preliminaries and theoretical foundations

Our proposal (discussed in Section 5) is based on a list of theoretical results already introduced in the literature. In this section we list them systemically, giving credit to who introduced them and we indicate how they constitute a determining element in our solution.

An overview of the word embedding technique and associated Word2Vec algorithm is presented in Section 4.1. The use of word embedding based centroid vector has also been highlighted in Section 4.1. The proposed methodology uses hierarchical agglomerative clustering algorithm applied over the centroid vectors to categorize the documents according to the concept hierarchy. An extensive discussion is made in Section 4.2 on agglomerative hierarchical clustering algorithm with different linkage criteria. The utility of dendrograms in the proposed methodology is also explained.

4.1. Word embedding

In recent few years, word embeddings have generated a lot of interest in the text analysis (Ángel González et al., 2020) research domain ever since two very simple log-linear models (Mikolov et al., 2013a; Mikolov et al., 2013b) were proposed that outperformed all previous of NLP models. Word2Vec has become the most reliable technique to be used as the basis of all NLP models. Of course, there have been proposals on improvement using Deep Learning Recursive Neural Networks (RNN) based on Long-Short Term Memory (LSTM) (Alcamo, Cuzzocrea, Bosco, Pilato, & Schicchi, 2020) nodes and also the very recent BERT algorithm (Devlin, Chang, Lee, & Toutanova, 2019), but in the last five years word embedding has proven to be a strong baseline. Both the Deep Neural-Net LSTM model and word embedding based models are scalable to very large corpus sizes and produce accurate results. However, the word embedding model is very simple in architecture. Word embedding based algorithms also have the advantage of drastically reduced time complexity. Therefore recent works (Krishna & Sharada, 2019; Periñán-Pascual, 2021) on capturing semantic context in the text are still employing the word embedding technique as one of the state-of-the-arts in the text mining task.

We briefly explain the working principle of the Skip-gram model. Let there be a corpus, a sequence of words w_1, w_2,..., w_T. The window is defined by parameter c, where c words at the right and left of the target are taken. For Skip-gram, each context is predicted independently given the target. The objective function to be maximized is defined as :

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j}\mid w_t) \tag{1}$$

Skip-gram models the probability of word w_{t+j} being observed within word w_t's context window as a probability. The probability $p(w_{t+j}\mid w_t)$ is defined as a softmax, where $u_w$is a target embedding vector for w and v_w is a context embedding vector. The u_w embeddings are the ones that are kept, v_w is a side product. The following definition is used for Skip-gram:

$$p(W_c\mid W_t)=\frac{\exp v_{wc}^{T}u_{wt}}{\sum_{w=1}^{W}\exp v_w^{T}u_{wt}} \tag{2}$$

Finally, using these equations word embedding vectors of the text documents are generated.

The similarity between any two embedding vectors represented as $\vec{w}_i$ and $\vec{w}_j$ respectively, is measured by the Cosine Similarity distance (Oukid et al., 2015) value and is calculated as:

$$CSM(\vec{w}_i,\vec{w}_j)=\frac{\vec{w}_i.\vec{w}_j}{||\vec{w}_i||.||\vec{w}_j||} \tag{3}$$

4.1.1. Document centroids

Having computed the dense vectors of all the vocabulary words, the simplest method to obtain a dense vector (of the same dimensionality) for a text document $d=\langle w_1, w_2, ..., w_n\rangle$ of n consecutive word occurrences is to simply compute the centroid $\vec{d}$ of the dense vectors $\vec{w}_i$ of the word occurrences:

$$\vec{\mathbf{d}}=\frac{1}{n}\sum_{i=1}^{n}\vec{w}_i=\frac{\sum_{j=1}^{|v|}\vec{w}_j.TF(w_j,d)}{TF(w_j,d)} \tag{4}$$

However, as suggested by Kosmopoulos et al. (2015), we hereby compute the document centroid vectors using Eq. (5) by taking its IDF scores of tokens/words into consideration. As shown in Kosmopoulos et al. (2015), this modification results in improved document categorization performance.

$$\vec{\mathbf{d}}=\frac{\sum_{j=1}^{|v|}\vec{w}_j.TF(w_j,d).IDF(w_j)}{\sum_{j=1}^{|v|}TF(w_j,d).IDF(w_j)} \tag{5}$$

Here $|v|$ is the vocabulary size, w_j is the j-th vocabulary word, $\vec{w}_j$ represents its embedding, TF(w_j,d) is the term frequency of w_j in d and IDF(w_j) represents the inverse document frequency of w_j.

4.2. Agglomerative hierarchical clustering algorithm

The agglomerative clustering technique performs a hierarchical clustering using a bottom up approach to form the concept hierarchy existing in the contextual dimension. The distance between two clusters has been computed based on the length of the straight line drawn from one cluster to another. We have represented the documents using word vectors, therefore, as discussed in Mikolov et al. (2013a, 2013b) if they are mapped into the Euclidean Space, it may be observed that the similar pair of words tend to exhibit similar displacement vectors. Such that the straight line distance between 'DBMS' and 'MS-Access' will be equal to the straight line distance between 'Data Science' and 'Python'. Keeping this property of the word vectors in mind, in the proposed methodology we use the Euclidean Distance as the distance metric to compute the distance between two clusters. After selecting a distance metric, it

Table 2
Parameters of the Lance-Williams update formula for different agglomeration methods with the definition of dissimilarity measure.

Method	α_i	α_j	β	γ	Dissimilarity Measure
Single Linkage	0.5	0.5	0	0.5	d_{ij}
Complete Linkage	0.5	0.5	0	0.5	d_{ij}
Average Linkage	$\frac{N_i}{N_i+N_j}$	$\frac{N_j}{N_i+N_j}$	0	0	d_{ij}
Ward	$\frac{N_i+N_m}{N_i+N_j+N_m}$	$\frac{N_i+N_j}{N_i+N_j+N_m}$	$\frac{-N_m}{N_i+N_j+N_m}$	0	d_{ij}^2

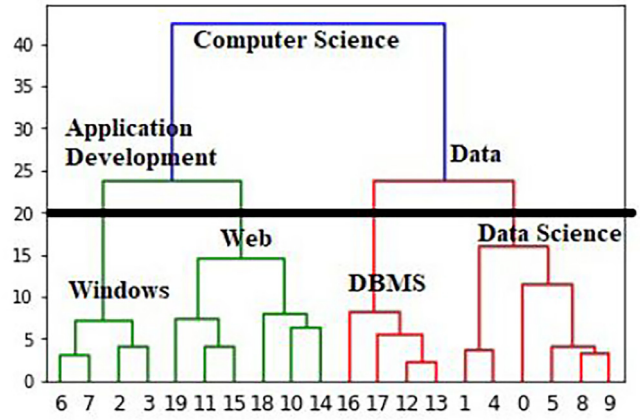


Fig. 2. Dendrogram using average linkage agglomerative clustering algorithm.

is necessary to determine from where the distance is computed (linkage criterion). For example, it can be computed between the two most similar parts of a cluster (single-linkage), the two least similar bits of a cluster (complete-linkage), the center of the clusters (mean or average-linkage), or some other criterion. Many linkage criteria have been developed. All hierarchical methods described in this work can be easily implemented through the widely used Lance-Williams dissimilarity update formula (Theodoridis & Koutroumbas, 2009). Lance-Williams updated formula allows us to calculate this distance straightforwardly, according to the following equation:

$$d_{km} = \alpha_i.d_{im} + \alpha_j.d_{jm} + \beta.d_{ij} + \gamma.|d_{im} - d_{jm}| \quad (6)$$

The coefficients α_i, α_j, β and γ define the agglomerative criterion. The two different types of dissimilarity measures are also described in Table 2. The dissimilarity measure can either be the Euclidean distance or its squared value. N_i, N_j and N_m are the number of documents in clusters i, j and m respectively.

4.2.1. Use of dendrogram in the formation of concept hierarchy

During the formation of the concept hierarchy the dendrogram (Ángel González et al., 2020) structure has been referred to get a visual representation for working out the number of concepts (clusters) of the formed hierarchy. It is created as an output of hierarchical clustering algorithm and displayed graphically as a tree diagram. The use of a dendrogram is to determine the number of clusters that best fits the data in terms of compactness and closeness. The different parts of a dendrogram are demonstrated in Fig. 2. This dendrogram is achieved by experimenting with 20 random documents (CVs) selected from our dataset. The horizontal axis indicates the number of documents and the vertical axis corresponds to the dissimilarity measure between the documents.

In the proposed methodology we have used the Silhouette Coefficient(s) (Shahapure & Nicholas, 2020) to more accurately select the optimal number that best fits the documents. The cut-off method has also been used in dendrograms to visually represent the concepts in the hierarchy. The Silhouette Coefficient (s) is defined for each sample and is composed of two scores: a- mean distance between a sample and all other points in the same class and b- mean distance between a sample and all other points in the next nearest cluster.

The Silhouette Coefficient s for a single sample is then given as:

$$s = \frac{b-a}{max(a,b)} \quad (7)$$

The best value is 1 and the worst value is -1. For agglomerative hierarchical clustering, the Silhouette Coefficient can be computed for several cuts (k=2,3,..,N-1). The user selects the k with the maximum Silhouette Coefficient value.

5. Proposed methodology

In this Section, we introduce and discuss a novel context-aware model that uses word embedding in conjunction with agglomerative hierarchical clustering algorithms to dynamically categorize documents in order to form the concept hierarchy.

The proposed methodology can be broadly divided into 4 steps: (i) The raw text files are first represented as document centroid vectors (see Section 5.1), (ii) The agglomerative hierarchical clustering algorithm is applied for the formation of the concept hierarchy (see Section 5.2). Two novel algorithms to perform the first two tasks are proposed in this paper. (iii) Clusters are labelled according to the relevant concepts (see Section 5.3). Finally, (iv) OLAP aggregation operations are performed according to the business requirement (see Section 7.4). A schematic diagram representation of the methodology is provided in Fig. 3.

5.1. Computation of document centroid vectors

The initial preprocessing of raw text involves the removal of stop words present in the documents to prepare a corpus of meaningful terms. Upon the large corpus of documents (resumes), Skip-gram algorithm is executed to form the word embedding vectors. Subsequently document centroids are computed using Eq. (5) presented in Section 4.1.1. As a result the text documents are represented as a N-dimensional [N=50,100,200] dense word vectors. The steps for the formation of document centroid vectors from the raw text documents are formalized in the form of a novel algorithm and presented as Algorithm 1.

5.2. Categorization of documents using hierarchical agglomerative clustering

After representing the text documents as N-dimensional document centroid vectors (output of Algorithm 1), we apply the agglomerative hierarchical clustering algorithm to the centroid vectors to categorize them into a set of clusters. Each cluster represents a concept of a dimension having concept hierarchy. We have used the state-of-the-art standard agglomerative hierarchical clustering algorithm (Theodoridis & Koutroumbas, 2009) keeping the linkage criterion generic. Lance-Williams dissimilarity update formula (Eq. (6)) has been used as the generic dissimilarity measure. However, the final choice of the linkage criterion and selection of the number of clusters has been decided based on the Silhouette Coefficient (s) (Eq. (7)) score being obtained during the experiment. The linkage method producing the highest Silhouette score value is considered to be the suitable agglomeration linkage criterion for a particular dataset. The methodology of the formation of the concept hierarchy is presented as Algorithm 2.

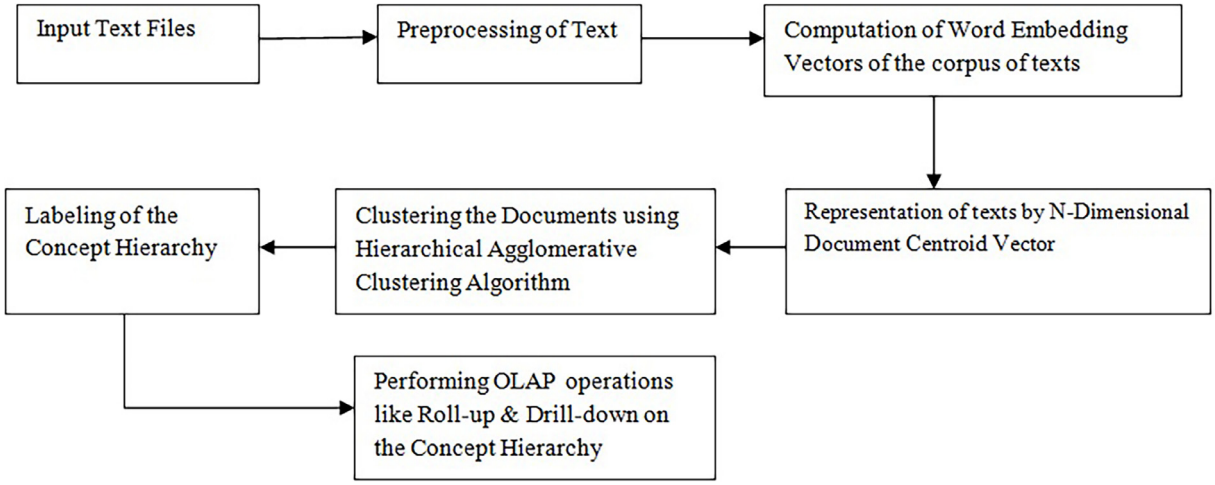


Fig. 3. Schematic diagram of the proposed methodology.

Algorithm 1 Construction of the word embedding based document centroid vectors.

Input: $\{d_i : d_i \in D\}$ - Documents (in raw text format) in the corpus D,
N - Number of documents,
c - Window size of the Word2Vec model,
D' - Dimensionality of the embedding
$|v|$ - Vocabulary size
Result: $\{d_i : d_i \in D\}$ - Centroid vectors of each document
Preprocessing: Each document d_i is preprocessed by performing stop word removal and tokenization. Each preprocessed document d_i is represented as a sequence of meaningful words. $d_i = \langle w_1, w_2, ..., w_j \rangle$
Repeat $\forall d_i : d_i \in D$
Repeat $\forall w_j : w_j \in d_i$
- Compute Term frequency $TF(w_j, d_i)$ of the j^{th}vocabulary word w_j in document d_i by counting the number of occurrences of w_j in d_i.
- Compute the Inverse Document Frequency $IDF(w_j)$ of the j^{th} vocabulary word in document d_i as: $log\frac{N}{N(w_j)}$, $N(w_j)$ is the number of documents containing the word w_j.
- Compute the word embedding vector w_j of the word w_j belonging to d_i by executing the Skip-gram algorithm with window size c.
- Compute the document centroid vector d_i of document d_i by combining the results of Term Frequency, Inverse document Frequency and word embedding vector of each word w_j belonging to document d_i by using Eq. (5) as:

$$\boldsymbol{d_i} = \frac{\sum_{j=1}^{|v|} \vec{w}_j . TF(w_j, d_i) . IDF(w_j)}{\sum_{j=1}^{|v|} TF(w_j, d_i) . IDF(w_j)}$$

End{Repeat}
End{Repeat}
Return$\{\vec{d}_i : d_i \in D\}$

5.3. Labeling of the concept hierarchy

After applying the agglomerative hierarchical clustering algorithm the documents are clustered in a hierarchy of concepts. Labeling is the task of selecting descriptive and human-readable labels/names for the clusters that summarize the concept or topic of the clusters. Labels distinguish the clusters from each other. This is the only step that requires human intervention.

In the proposed methodology we pick the twenty top most frequently occurring terms belonging to a particular cluster concept. After that, we consulted domain experts and referred to query logs on related topics from the FAQs in different job portals. Thereafter we were suggested with one or more concept names as labels for that particular cluster but finally, we chose the one which has the highest average cosine similarity value with the twenty top most frequently occurring terms. The predecessor concepts in the hierarchy are marked by the clades of the produced dendrogram. Combining the descendant clusters, the connecting clades are also labeled with appropriate concept names. Finally, the root of the hierarchy is labeled with the most generalized concept that covers all the descendant concepts.

Algorithm 2 Dynamic formation of concept hierarchy.

Input: $\{\vec{d}_i : d_i \in D\}$ - Document centroid vectors (Output of Algorithm-1),
N - Number of documents,
Linkage criterion (selected based on maximum Silhouette score s),
$d_{(c_i,c_j)}$ - Distance metric (using Eq. (6)) between any two clusters c_i and c_j, $\forall c_i, c_j \in C$
α_i, α_j, β and γ - Values of coefficients defining linkage criterion
Result: Assignment of documents in the concept hierarchy
Repeat $i = 1$ to N
- $c_i = \{\vec{d}_i\}$
End{Repeat}
$C = \{c_1, c_2, ..., c_N\}$
Repeat $C.size > 1$
- $\{c_{min1}, c_{min2}\} =$ minimum $d_{(c_i,c_j)}$, $\forall c_i, c_j \in C$
- Remove c_{min1}, c_{min2} from C
- Add $\{c_{min1}, c_{min2}\}$ to C
End{Repeat}

6. Materials and methods for experimental evaluation

We accomplished the experiments to illustrate the effectiveness of the proposed methodology by using the following components.

6.1. Benchmark datasets creation

The selection of the dataset is crucial to explain the functionalities of the proposed method. Here resume dataset is chosen for the following reasons: (i) Resumes are generally structured in nature and hence different contextual dimensions can be well defined. In a text-OLAP environment users or decision makers usually pose queries based on the notion of context. For example the candidature of a candidate can be shortlisted by referring to several contextual dimensions together, like- expertise in skill-set,years of experience,location of work,qualification,etc. Thus contextual information during the exploitation of data warehouses

Table 3
Dataset description.

Dataset	Characteristics	No. of Samples	No. of Features	Value range	Missing values	Nature	No. of classes
Dataset-I	Multivariate	850	200	0.0 - 1.00	No	Dense	5
Dataset-II	Multivariate	80	50	0.0 - 1.00	No	Dense	4
Dataset-III	Multivariate	80	200	0.0 - 1.00	No	Sparse	4

must be taken into account. With the structured format of the resume dataset, it is easier to segregate the different partitions of a resume based on contextual factors. The contextual factors can be represented as contextual dimensions. For example, a resume can be easily formatted and segmented into a collection of contextual dimensions like, skill-set, location, etc. (ii) Some of these contextual dimensions such as skill-set and location maintain concept hierarchy, therefore the OLAP operations such as roll-up, drill-down can be performed. OLAP analysis on the contextual concept hierarchy allows a decision-maker to navigate through the OLAP cube and to observe the data along several analysis axes organized in different hierarchical levels. For instance, the decision-maker can observe the competencies in 'Computer Science' for the year 2022 in India and then, by a drill down operation he observes those for the city Kolkata, etc. Further drill down can also be done on skill-set by observing 'Computer Science' in a more detailed view of specializations in 'DBMS', 'Web Technology', 'Machine Learning' etc.

As a workbench, we have considered the resume dataset (Chakrabarty et al., 2018; Oukid et al., 2015) for experimental purposes. The dataset was obtained from different resume portals from where 850 resumes of candidates belonging to various job specializations were collected from different publicly accessible job portals available online at websites like: www.freeresumesites.com, www.resumeworld.com, www.eresumex.com, www.freshersworld.com, www.linkedin.com, etc. Thereafter, the raw text documents have been preprocessed to prepare the datasets with the desired format.

Experiments have been carried out on three benchmark datasets, referred as: Dataset-I,[1] Dataset-II[2] and Dataset-III.[3] Table 3 presents a description of the dataset.

- Preprocessing:

After the collection of resumes, we manually extracted and formatted the documents according to the dimensions in our proposed star schema for text-OLAP analysis. In the proposed OLAP model, we have experimented with three dimensions: Skill-set specialization Topic (Dim_T), Location of candidate (Dim_L), and Experience (Dim_E). After extraction of the relevant sections from the resumes, data cleaning techniques such as Text Tokenization, and Stop-words removal are performed to create the final text corpus. Job profile resumes are notably different text documents with a higher density of specialized terminology. Hence, we haven't performed stemming techniques on our dataset. For example, terms like 'Data-Mining', 'Deep-Learning', 'Encapsulation' remain unaltered in our methodology, However, as a result, terms like 'Program' and 'Programming' turned out to be very similar in the Word2Vec Model. Similarly, the text is also not converted in lower case as certain terms like 'R', 'RDBMS', 'MVC', 'SVM' are always written in upper case.

6.2. Software and libraries

To validate our model, we have developed our prototype application written in Python-3.6 with Application Software Spyder (64-bit) and IDE Anaconda-3. For text preprocessing, developing the word-embedding and centroids of documents by importing the Word2Vec model and executing agglomerative clustering algorithm, the following libraries have been used- nltk, numpy, scikit-learn, scipy, tensorflow, gensim etc. We test skip-gram neural architecture by varying the embedding sizes. To find the best parameters configuration, we run a grid search using this setting: embedding size [50,100,200] and finally settle with 200, topic and similarity thresholds respectively in [0,0.5] and [0.5,1] with a step of 0.01. Window size is taken as 5. The centroid of the documents is stored in a numpy array of shape: Number of documents × Embedding size. Later we store the values (centroid vectors) of the numpy array in a.csv file. The hierarchical agglomerative clustering is executed on the.csv file.

7. Experimental results

In this section some experimental results are discussed.

7.1. Performance of document centroid vector in capturing semantic text similarity

In this section, the effectiveness of the document centroid vector in capturing semantic text similarity is discussed. The Word2Vec Skip-gram algorithm is applied over the large corpus of text documents to form the word vectors. Each text document is then represented by its document centroid vector. Word2Vec allows learning complex semantic relationships using simple vectorial operators (Mikolov et al., 2013a). For example, it may be written as: $vec(DBMS) - vec(Access) + vec(Python) = vec(DataScience)$.

In Fig. 4, we can see the Word2Vec t-SNE (Van der Maaten & Hinton, 2008) visualization of our implementation, using the resume dataset and a window size of 2. The t-SNE algorithm reduces the dimensionality of the vectors and plots the high dimensional word vectors into 2-dimensions. For a better visibility, we have partially shown the right hand side figure with the zoomed portion of the diagram having a vocabulary size $|v|$ of 700 words. On the right hand side figure, it can be seen that words with semantic similarity are in close proximity with each other. The diagram has been grouped into three clusters: marked as Black, Red, and Green. The words enclosed in the black coloured group are: Association, Data Science, R, Python, Classification and Data Warehouse. They signify that they belong to the concept-Data Science and Machine Learning. The blue marked cluster has the highest number of contextually related words, like: C, C++, VB, VC++, PHP, HTML, ASP.NET, MVC, Java, Hibernate, Struts, API, JSON, Tomcat etc. Based on the context, it can be said that these highlight the specialization of Application development Programming. The third cluster marked in red contains words related to job specialization in Database handling with keywords like: DBMS, RDBMS, SQL, PL/SQL, Access, Stored Procedure and Oracle etc. This observation proves how Word2Vec can group the semantically similar words into close distanced vector space vicinity.

7.2. Evaluation of proposed centroid based method over TF-IDF models in capturing semantic similarity

In order to illustrate the effectiveness of the method let's consider two small pieces of paragraphs extracted from two separate resumes with specialization in Application development. **Text-1**: *"Experience in*

[1] https://drive.google.com/file/d/1t96dYfcMTnAkx7c_4iMLlHlJ8baHng60/view?usp=sharing.
[2] https://drive.google.com/file/d/1lRmuGiY0OpI6yoYQ09X60sZMLgROTuG9/view?usp=sharing.
[3] https://drive.google.com/file/d/1R_Uaf2cFmpWsgYSkyfQ8yv7Ovne38576/view?usp=sharing.

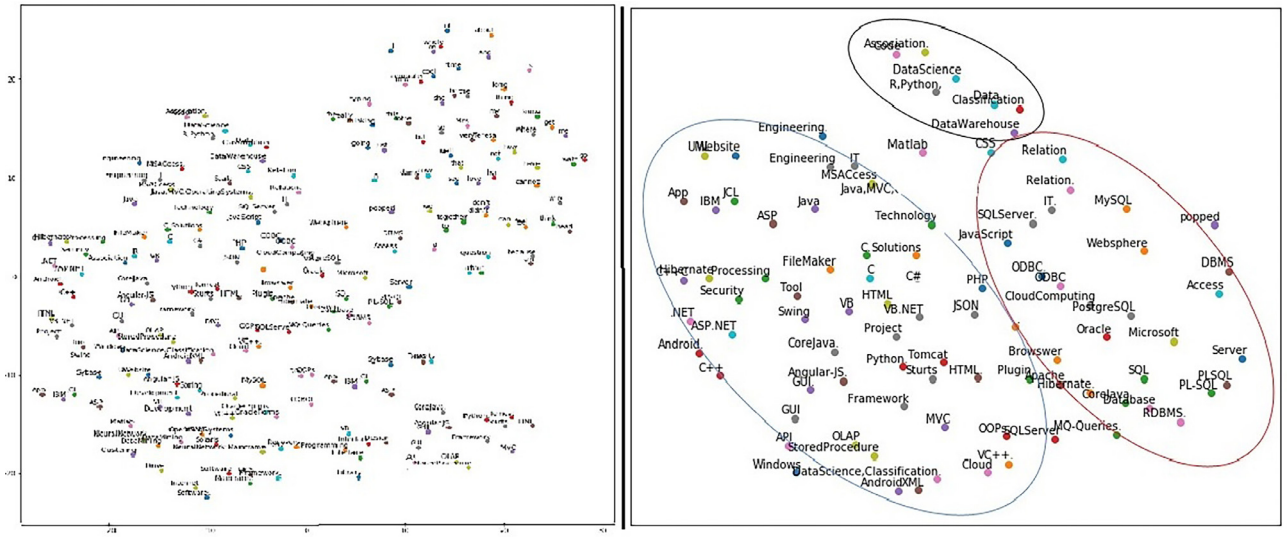


Fig. 4. Left: Word relationships using dimensionality reduction by t-SNE algorithm Right:Zooming in of the rectangle in the left figure.

Table 4
Semantic similarity between word vectors and document centroid.

Experience	ASP.NET	C#	coding	Centroid embedding
Work (0.872)	.NET (0.913)	ASP.NET (0.986)	code (0.985)	Working
Working (0.868)	C# (0.901)	.NET (0.978)	Programming (0.864)	C#
years (0.712)	VB.NET (0.834)	VB.NET(0.902)	programming(0.845)	.NET
Professional(0.644)	framework(0.77)	programming(0.837)	language(0.733)	programming
Industry (0.605)	MVC (0.654)	application (0.745)	skill (0.662)	-

Table 5
Cosine similarity between the word vectors of Tex-1 and Text-2.

Word Vector of Text-1	Word Vector of Text-2	Cosine Similarity between Text-1 and Text-2
Experience	Working	0.868
ASP.NET	Java	0.563
C#	Hibernate	0.474
coding	framework	0.558
Average Similarity		0.616

ASP.NET in C# coding" and **Text-2:** *"Working in Java Hibernate framework"*. After cleaning (excluding stop words) the modified texts look like: **Text-1:** *"Experience ASP.NET C# coding"* and **Text-2:** *"Working Java Hibernate framework"*.

Since Text-1 and Text-2 have no words in common, thus the document similarity between these is zero according to BOW or TF-IDF models. Experimental results depicted in Table 4 prove the superiority of the proposed centroid based method. The table header corresponds to the words of Text-1 and the last column contains the most similar word to the centroid embedding computed using Eq. (5).

In Table 5, we show the pairwise cosine similarity between Text-1 and Text-2 calculated using the Skip-gram word vector model. We find the average pairwise word similarity between Text-1 and Text-2 is 0.616 which is a major improvement over the TF-IDF model where the similarity value turned out to be zero between the two pieces of texts.

7.3. Evaluation of the agglomerative hierarchical clustering algorithm

In this section, we present the experimental result of deciding the number of clusters in the hierarchy by applying different linkage criteria, such as Single, Complete, Average, and Ward. The total number of clusters (k) in the hierarchy is chosen based on the maximum Silhouette score (s).

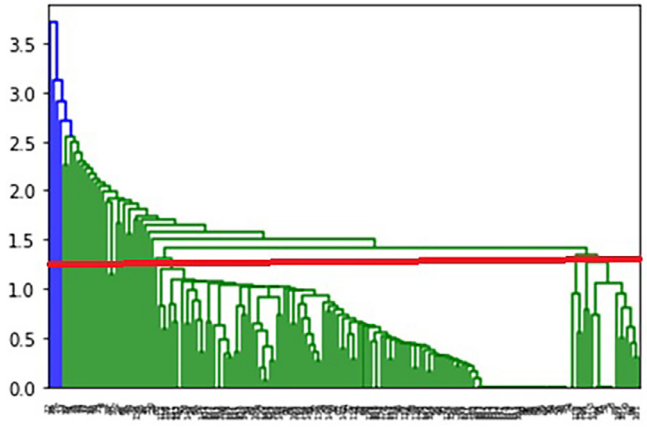


Fig. 5. Dendrogram for average linkage criterion agglomerative clustering algorithm applied on 375 documents.

When the corpus is made up of 375 text documents (resumes), it can be seen in Table 6 that the maximum Silhouette score ($s = 0.419$) is achieved with the Average Linkage method for the number of clusters (k)=4. Table 6 also confirms that with 850 documents maximum Silhouette score ($s = 0.467$) is obtained by using Ward linkage criterion for the number of clusters=5. Hence the number of clusters (k) is chosen as 5 for 850 documents.

The result of Table 6 can be visualized in Fig. 5 which represents the dendrogram produced by applying the agglomerative algorithm on 375 text documents. If the cut-off line is set at value 1.25 (marked in red) then we get 2 distinct sets of separated clusters. The large numbered left hand side clusters are similar to each other and represent the

Table 6
Linkage criteria wise Silhouette scores with varying number of documents and clusters.

Number of Documents	No. of Cluster	Silhouette Score (s)			
		Single Linkage	Complete Linkage	Average Linkage	Ward Linkage
375	2	0.283	0.346	0.290	0.349
	3	0.324	0.357	0.325	0.378
	4	0.362	0.403	0.419	0.405
850	3	0.392	0.364	0.391	0.367
	4	0.380	0.407	0.380	0.424
	5	0.407	0.466	0.406	0.467

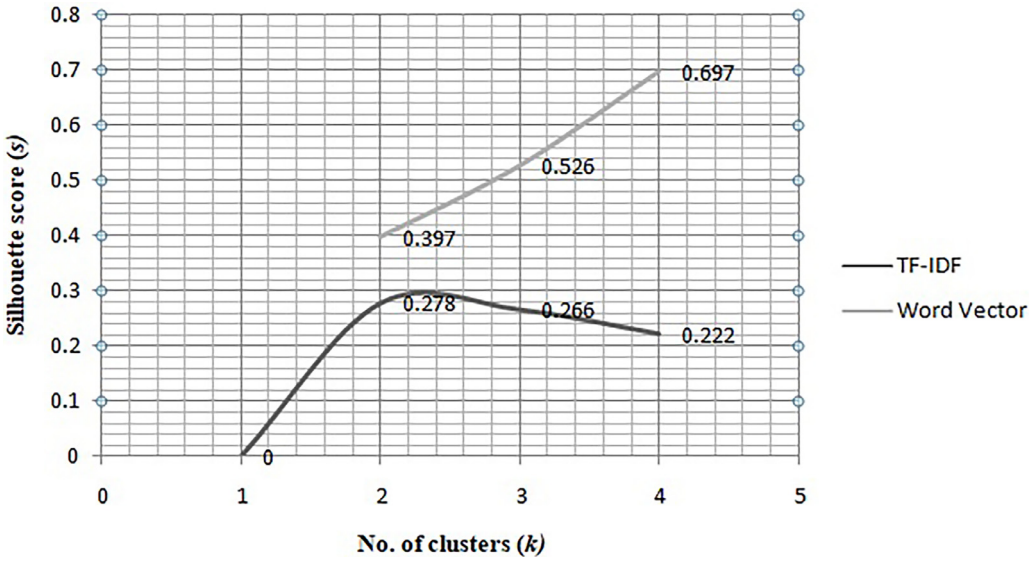


Fig. 6. Comparison of Silhouette scores (s) between proposed model and TF-IDF vector models.

concept 'Application Development'. On the other hand, the right hand side clusters represent the concept 'Data'. The formed concept hierarchy has already been shown in Fig. 1. If we drill-down even further in the concept hierarchy then it can be seen that 'Level-1' of Fig. 1 consists of 4 concepts. Therefore, the documents should be ideally categorized in 4 clusters. However, with bare eyes it is quite difficult to set a cut-off value in the dendrogram of Fig. 5 to select 4 clusters. In such scenarios, the determination of the optimal number of clusters is made by referring to the Silhouette score (s).

7.3.1. Performance comparison of centroid vectors over the TF-IDF models in the formation of clusters

A comparative study (in terms of Silhouette scores (s) between the centroid word vectors over the TF-IDF based models is depicted in Fig. 6. The agglomerative hierarchical clustering algorithm is executed on both Dataset-II and Dataset-III containing word vector and TF-IDF vector respectively. Fig. 6 confirms for all three values of the number of clusters (2,3 and 4), the Silhouette scores achieved from the centroid word vector are better than the ones achieved from the TF-IDF vector.

7.4. Case study on OLAP operations

The final concept hierarchy formed using Dataset-I (875 documents), is represented in Fig. 7. With the inclusion of new documents, the cluster (concept) 'Data Science' of Fig. 1 (375 documents) has been further divided into two separate clusters 'Machine Learning' and 'Data Warehouse'. Therefore the total number of clusters/concepts is 5 (Number of leaf nodes) according to Fig. 7. In this section we focus on a case study where an H.R. Manager wants to study the resumes that are submitted in response to a recruitment advertisement. Traditional methods of resume screening by the H.R. managers are not a standardized process. Earlier it has been done manually. Now-a-days, they have a computerised tool where they put the requirement by the organization by specifying the parameters as per the requirement of the different departments of the organization. The computerized tools are developed as per the requirement of the organization and parameters specific. These tools work by just matching the terms of the given parameters. Therefore if the terms are different but belong to the same category or domain it fails to identify the required skill level. These tools lack the intelligence as no context awareness are incorporated.

In the proposed model, since resumes sharing similar context (domain) are hierarchically clustered, thus searching the documents from the highest level of generalized requirement to the most specific requirement is much more efficient. Various OLAP operations (like-roll-up, drill-down) are performed which will extract knowledge and aid the manager in making decisions.

In order to include the other two dimensions along with the dimension Topic (D_T), each document is tagged with the working Location of the candidate (Dim_L) and job Experience (Dim_E). After clustering the documents according to the concept, we have represented each document according to its dimension values to carry out further OLAP operations. For example, a random sample under the concept 'Web Application Development' (member of cluster 'Web') may be expressed as $\langle id-678, \}Kolkata', 8\rangle$. This means a resume of a candidate with document id-678 has been categorized with specialization in web application development having 8 years of industry experience and is currently located in Kolkata.

If the H.R. manager of a company wants to analyze the resumes for candidates working in Kolkata then he would like to perform the slice operation on the dimension Location (Dim_L) using the selection criteria Location= 'Kolkata'. On the other hand, a dice operation by using se-

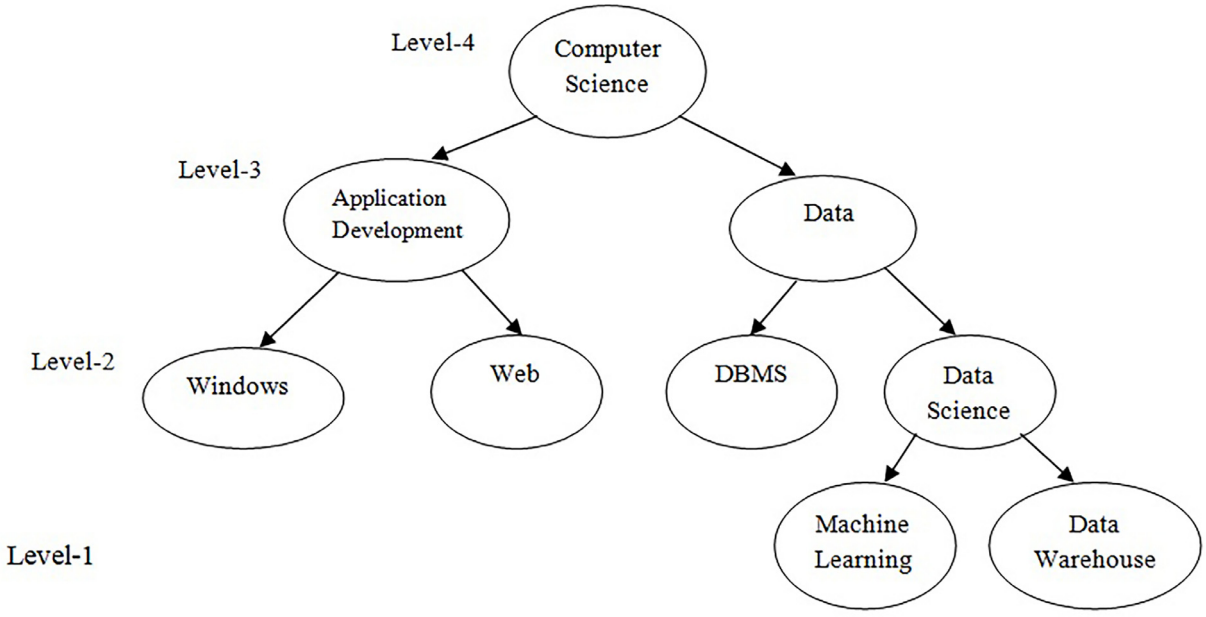


Fig. 7. Concept hierarchy formed with 850 Documents.

Table 7
Drill-down operation on the concept hierarchy on Dimension- Topic (D_T).

Level-4	No. of Files	Level-3	No. of Files	Level-2	No. of Files	Level-1	No. of Files
Computer Science	850	Application Development	583	Web	462	-	-
				Windows	121	-	-
		Data	267	DBMS	94	-	-
				Data Science	173	Machine Learning	132
						Data Warehouse	41

Table 8
Dice operation on the resume dataset.

Selection Criteria	Count (No.of Resumes)	(%) of total number of documents
Topic='Data Science' & Experience='2 years'	16	9.25
Topic='Data Science' & Experience='5 years'	67	38.72
Topic='Web Application' & Experience='2 years'	248	42.54
Topic='Web Application' & Experience='5 years'	184	31.56

lection conditions on the dimensions Topic and Experience is shown in Table 8.

Results investigated during the case study prove the importance of performing OLAP operations on organizational text data repositories for business oriented improvised decision making purpose.

7.5. Comparison of OLAP query execution time between proposed model and TF-IDF models

At level-2 (with the maximum number of clusters), we have performed the various OLAP operations such as: roll-up, drill-down, slice, and dice on documents represented using both 200-dimensional centroid word vector and also as TF-IDF vector format. Fig. 8 shows the query execution time on the centroid word vector is much faster than the higher dimensional TF-IDF word vectors.

The existing methods which address text-OLAP, use the TF-IDF vectors to represent the text documents in the vector space model. With a large dataset having a huge vocabulary set, these TF-IDF vectors can be very sparse and very high dimensional. However, the proposed model uses much lesser dimensional dense word vectors for document representation. Processing the less dimensional word vectors for contextual similarity computation with a posed OLAP query is much faster than the

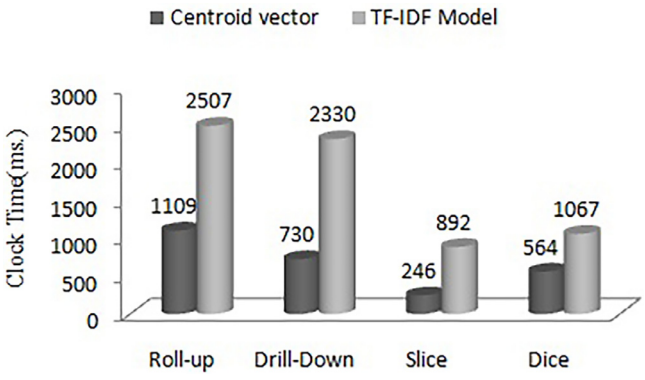


Fig. 8. Comparison of OLAP operation time in wall clock (milliseconds) between proposed centroid vector based model and TF-IDF based models.

processing time required to scan through the high dimensional TF-IDF vectors. Hence, during experimental evaluation proposed model shows considerable improvement in speeding up the execution time of OLAP operations. Results in Fig. 8 confirm that the proposed centroid vector

model significantly reduces the OLAP query processing time compared to the TF-IDF based models for all four OLAP operations. Roll-up operation is almost 2.3 times faster in the proposed model compared to the existing models. Dice operation is performed even faster with a speed-up of almost 3.2 times. The improvements in the query execution time of the proposed model are visible for the slice(3.6 times faster) and dice (1.9 times faster) operations.

7.6. Computational complexity

The proposed model first converts the corpus of text documents into word embedding based N-dimensional documents centroid vectors. After that, the hierarchical agglomerative clustering algorithm is applied to the centroid vectors to form the concept hierarchy. Once the concept hierarchy is performed, documents are aggregated and retrieved according to different dimensions using OLAP queries on the concepts (clusters). The computational complexity of the Word2Vec Skip-gram algorithm is: $Q = O(C \times (D + D \times log_2(V)))$ (Mikolov et al., 2013a), C is the window or context size, D is the dimensionality of the embeddings and V is the vocabulary size. In addition, the computational complexity of the hierarchical clustering algorithm is $R = O(N^2)$, where N is the number of documents. Therefore, in total the computational complexity in forming the concept hierarchy is:

$O(Q + R) = O(C \times (D + D \times log_2(V))) + N^2)$.

This analysis shows the computational cost of the proposed method is an expensive one. However, in this paper we have prioritized extracting semantic information from the text documents to form the concepts (clusters) with better accuracy. Since this entire analysis is performed in off-line mode, thus for the sake of contextual accuracy this compromise can be made at the cost of heavy computational complexity. Accurate knowledge is critical in making managerial decisions and empowering organizations with business intelligence. Having said that, once the documents are clustered according to the concept hierarchy then the concept wise real time (on-line mode) document retrieval time is improved. Then the searching time complexity will be $O(log_2 k)$, where k is the number of clusters.

8. Discussion

The synthesis and analysis of the resume datasets considered for the simulation of our proposed model underline the importance of maintaining textual data warehouses. Findings from the analysis of the result implicate the benefits of leveraging managerial decision making by performing OLAP operations on the textual data warehouses. Our experimental model serves as a decision support making tool for the HR managers to analyze and manage the huge volume of the contextually variant sets of resumes that are received in response to a job offer. By the implementation of the proposed model and making a case study on the resume datasets, the following analytics driven benefits were established:

8.1. Key findings

In particular, the results of our experiments are contrasting with that of existing TF-IDF based models dealing with textual data warehouses. During the empirical evaluation, the findings of this research work turn out to be promising in performing OLAP aggregation operations on textual data warehouses. The findings more specifically reveal the following facts: i) Text documents represented as dense word embedding based centroid vectors perform significantly better than the sparse TF-IDF vectors in capturing the contextual similarities between documents ii) With fewer number of documents TF-IDF models are better than word vectors but as the number of documents increases the performance of the word centroid vectors drastically improves in capturing contextual similarities iii) Agglomerative hierarchical clustering algorithm forms the concept hierarchy dynamically. Based on the Silhouette Coefficient value the linkage criterion of the clustering algorithm is chosen. Experiments with varying Silhouette Coefficient values ensure the better formation of the clusters in terms intra cluster similarity and inter cluster separability. iii) Automatic extraction of the underlying contexts in the concept hierarchy allows decision makers to query upon the set of documents without knowing the contexts in advance. iv)Execution time of text OLAP aggregation operations on the concept hierarchy using centroid vectors is considerably faster than that of TF-IDF based word vectors.

8.2. Implications for theory

With the emergence of digitalization, the generation of unstructured data is more than ever evident. In the business sector, virtually all organizations are operating with the huge voluminous text data generated every day. The rapid advancement of data storage hardware and the growth of the generation of digital unstructured data on the web has solved the problem of availability of data for every organization. However, it has become harder than ever for organizations to keep up with this data. In the competitive market scenario, it has become a mandate for every organization to efficiently manage the data and get access to the right data at the right time for effective business decision making purposes. Management of organizations is finding it very hard to have a grasp on the hundreds of emails, reports, scholarly articles, medical diagnosis reports, customer feedback, product documentation, forum discussion, transaction management report, customer profiling, resume submission for job recruitment, and other plethora of applications. Related study reveals that organizations have already opted for comprehensive IT enabled data retrieval tool sets for quick access to the data needed for managerial analysis leading to business intelligence. However, these IT enabled tools often suffer from the accuracy of the retrieved data. With the emergence of AI and NLP and high volume of distributed computing, organizations are increasingly looking for alternative Information System enabled work systems. Literature survey (Carvalho, 2000) on Information System reveals there are four categories of Information Systems, such as: IS1, IS2, IS3 and, IS4, all with different purpose of handling information to communicate with different stakeholders of an organization. According to Alter (2008) the four types of objects 'all deal with information; they all are somewhat related to organizations or the work carried out in organizations; and they all are related to information technology, either because they can benefit from its use or because they are made with computers or computer-based devices.'

In this paper, the authors have proposed a model which is a solution for managing a huge repository of textual documents and goes beyond the retrieval of documents by mere keyword-matching with the user query. The proposed model can dive into the contextual semantics of the documents and performs hierarchical clustering of text documents according to their contextual similarity. This work is intended to be one of the small first steps in the direction of making an Information System which will ensure effective reporting of data between the management and operation modules of an organization. According to the definition of Information Systems, this proposed framework can be conceptualized as an IS2 work system. The sharing of information from operational subsystems allows the management of an organization to analyze and discover hierarchical relationships existing among different actors. In this paper, we presented a framework for integrating text into multi dimensional data model capable of OLAP text analysis. The proposed conceptual star schema is surrounded by contextual dimensions. In the proposed measure each document is represented by word embedding based centroid vectors of weighted concepts. Next, the agglomerative hierarchical clustering algorithm categorizes the documents into a concept hierarchy based on their contextual similarities. Documents arranged in a concept hierarchy allow users to query the corpus of documents from different levels of granularity. The H.R. manager of a company is highly benefited from this system in terms of discovering new business facts through fast OLAP querying on hierarchically organized resumes based on the sim-

ilarity of the domain of expertise of different candidates. The existing works on text-OLAP represent data using high dimensional sparse TF-IDF vectors. Processing of OLAP queries on these high dimensional vectors results in more execution time than the processing of the same set of OLAP queries on much smaller dimensional word vectors. The dense nature of the word vectors also increases the accuracy of the hierarchical clustering. Increase in the accuracy of clusters means more number of relevant documents are categorized in a group representing a particular context. The notion of extracting hierarchical contextual similarity between documents through OLAP queries is highly beneficial for the H.R. managers of a company. As already shown, in the proposed model, two separate resumes with specialization in 'ASP.NET C#' and 'Java Hibernate framework' are considered to be under the same genre or context 'Object Oriented Programming' despite sharing very few or no keywords between them. Therefore, clustering accuracy increases. Since OLAP queries are executed on these clusters, thus accuracy of clustering in turn results in the enhancement of accuracy during the retrieval of relevant documents in correspondence to an OLAP query.

To summarize the impact of this research work, the proposed Information System based context-aware work system helps in taking organizational managerial decisions through fast and accurate OLAP query processing.

8.3. Implications for practice

The implications for practice have been broadly classified into three sections, such as A) Descriptive analytics, B) Discovery analytics and C) Predictive analytics.

A) Descriptive analytics

Descriptive analytics deals with reporting and visualizations. In the different documents based on the underlying contextual information, concept hierarchy may be inferred. For instance in the given example Fig. 1 provides a descriptive visualization of the notion of concept hierarchy with the associated theory on the contextual dimensions that are extracted from the resumes based on the skill-set or domain of specialization of the candidates. The experimentally formed context-aware concept hierarchy depicted in Fig. 7 confirms our claim.

B) Discovery analytics

Discovery analytics can forecast early signals through text summarization OLAP operations. Results investigated from the case study depicted in Table 8, epitomize the importance of early knowledge discovery for developing a decision support system. The case study contributes to the knowledge of knowing the candidature of different candidates with respect to their area of specialization and job experience. This may serve as an aid to figure out the demand and supply of skilled human resource according to the latest industry needs.

In the case of our dataset the following observations can be witnessed from the analysis carried upon Table 8. It can be seen that there is a huge difference in the percentage of the total number of resumes between the candidates with 2 years of experience (9.25%) and candidates with 5 years of experience (38.72%) for the job specialization in Data Science. In contrast, this difference is not that drastic for the job profiles in 'Web Application Development' for the 2 years (42.54%) and 5 experience (31.56%) holders respectively. The inference that can be taken out from this data is a) opportunity to work on the projects in Data science domain for the entry level candidates is still relatively less. However, candidates are enhancing their skills in Data Science Projects when they are deployed into the domain after gaining some years of experience and b) there is a uniform distribution of entry-level and moderately experienced candidates in the domain of web application development.

The proposed model claims to fasten the OLAP query processing time by reducing the number of features of the text documents. Experimental results show drastic improvements in speeding up aggregation operations like- roll-up, drill-down, slice, and dice. The supporting results are given in Tables 7, 8 and Fig. 8.

C) Predictive analytics:

The proposed model uses word embedding based technique combined with agglomerative hierarchical clustering algorithm to predict the class of the documents during the formation of the concept hierarchy. Application of word embedding technique on the resume dataset ensures the contextually similar text documents are categorized with higher accuracy in terms of cosine similarity. Results reported in Tables 4 and 5 confirm the superiority of the proposed method over TF-IDF based models. The context-aware ensemble clustering method is dynamic as it does not require the number of clusters to be defined at the beginning. This methodology further ensures the better formation of clusters in terms of intra-cluster cohesion and inter-cluster separation. The Silhouette scores (s) shown in Fig. 6 confirm our claim.

In general, this research work serves as a potential tool for the development of Information Retrieval (IR) systems at a time when several existing organizations and start-ups are investing in developing text data analytics models with the motive of deriving actionable insights from the large volume of text data. The practice of text summarization technique and OLAP operations provide an excellent facility to the management of the organizations in keeping-up and processing high volume of text data which are hard to manage manually. The simulated model on the resume dataset enables the managers to drastically reduce the longer recruitment processing time by automating the resume short-listing time as per the requirement of skill-set specialization, job location, and preferred experience.

8.4. Limitations

There remains room for improvement in the proposed model. This model demands the documents be formalized in a certain format so that inherent contextual dimensions can be extracted by the algorithm. The experiments performed in this paper can be further performed on a bigger sized data set. Alternate methodologies can be further investigated which may result in lesser computational complexity. Query optimization strategies have not been applied in this work, and hence remain as a scope for further work.

8.5. Future work

Further work can be carried out to compare the performance of the proposed model with deep learning based algorithms using RNN or BERT. This algorithm can be also explored in other domains like scientific journal, bio-medical engineering etc. Satisfactory performance of the proposed algorithm with a higher sized dataset and higher numbers of features is another challenge. Another research challenge is the formation of the lattice of cuboids over textual data warehouse to enable a holistic decision support system.

9. Conclusion

This article proposed a novel methodology to construct a textual data warehouse and perform OLAP operations on the textual data after categorizing the text documents in a concept hierarchy by capturing the contextual similarity between the text documents. The proposed model outperforms existing models in capturing contextual similarity between texts. The proposed model also proves its supremacy with respect to the ability to add new clusters during the formation of the concept hierarchy. Despite being computationally expensive during the initial word vector formation time, the proposed model is able to retrieve documents in logarithmic time once the concept hierarchy is formed. Experimental results show the efficiency of the proposed model in faster processing of OLAP queries in comparison to the existing techniques. In the end, the findings from a case study render the necessity of OLAP tools over textual documents of an organization to explore and analyze the business facts.

Acknowledgement

Work partially supported by iNEST (Interconnected NordEst Innovation Ecosystem), funded by PNRR (Mission 4.2, Investment 1.5), NextGeneration EU (Project ID: ECS 00000043).

References

Alcamo, T., Cuzzocrea, A., Bosco, G. L., Pilato, G., & Schicchi, D. (2020). Analysis and comparison of deep learning networks for supporting sentiment mining in text corpora. In *Proceedings of the 22nd international conference on information integration and web-based applications & services*. In *iiWAS '20* (pp. 91–96). New York, NY, USA: Association for Computing Machinery. 10.1145/3428757.3429144.

Alter, S. (2008). Defining information systems as work systems: Implications for the is field. *European Journal of Information Systems, 17*(5), 448–469.

Ángel González, J., Hurtado, L.-F., & Pla, F. (2020). Transformer based contextualization of pre-trained word embeddings for irony detection in twitter. *Information Processing & Management, 57*(4), 102262. 10.1016/j.ipm.2020.102262.

Atkinson, J., & Escudero, A. (2022). Evolutionary natural-language coreference resolution for sentiment analysis. *International Journal of Information Management Data Insights, 2*(2), 100115. 10.1016/j.jjimei.2022.100115.

Azabou, M., Khrouf, K., Feki, J., Soulé-Dupuy, C., & Vallès, N. (2015). Diamond multidimensional model and aggregation operators for document olap. In *2015 ieee 9th international conference on research challenges in information science (rcis)* (pp. 363–373). IEEE.

Bouakkaz, M., Loudcher, S., & Ouinten, Y. (2016). Olap textual aggregation approach using the google similarity distance. *International Journal of Business Intelligence and Data Mining, 11*(1), 31–48.

Bouakkaz, M., Ouinten, Y., Loudcher, S., & Strekalova, Y. (2017). Textual aggregation approaches in olap context: A survey. *International Journal of Information Management, 37*(6), 684–692.

Carvalho, J. A. (2000). Information system? which one do you mean? In *Information system concepts: An integrated discipline emerging* (pp. 259–277). Springer.

Chakrabarty, A., Roy, S., & Roy, S. (2018). A context-aware fuzzy classification technique for olap text analysis. In *Recent findings in intelligent computing techniques* (pp. 73–85). Springer.

Cuzzocrea, A. (2020). Sppolap: Computing privacy-preserving olap data cubes effectively and efficiently algorithms, complexity analysis and experimental evaluation. *Procedia Computer Science, 176*, 3831–3842.

De Miranda, G. R., Pasti, R., & de Castro, L. N. (2019). Detecting topics in documents by clustering word vectors. In *International symposium on distributed computing and artificial intelligence* (pp. 235–243). Springer.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. 10.18653/v1/N19-1423.

Fujii, M., Sakaji, H., Masuyama, S., & Sasaki, H. (2022). Extraction and classification of risk-related sentences from securities reports. *International Journal of Information Management Data Insights, 2*(2), 100096. 10.1016/j.jjimei.2022.100096.

Georgiadou, E., Angelopoulos, S., & Drake, H. (2020). Big data analytics and international negotiations: Sentiment analysis of brexit negotiating outcomes. *International Journal of Information Management, 51*, 102048. 10.1016/j.ijinfomgt.2019.102048.

Kim, S.-W., & Gil, J.-M. (2019). Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences, 9*(1), 1–21.

Kohomban, U., & Lee, W. (2007). *Optimizing classifier performance in word sense disambiguation by redefining sense classes* (pp. 1635–1640).

Kosmopoulos, A., Androutsopoulos, I., & Paliouras, G. (2015). Biomedical semantic indexing using dense word vectors in bioasq. *J BioMed Semant Suppl BioMedl Inf Retr, 3410*, 959136040–1510456246.

Krishna, P. P., & Sharada, A. (2019). Word embeddings-skip gram model. In *International conference on intelligent computing and communication technologies* (pp. 133–139). Springer.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes, 25*(2-3), 259–284.

Lin, C. X., Ding, B., Han, J., Zhu, F., & Zhao, B. (2008). Text cube: Computing ir measures for multidimensional text database analysis. In *2008 eighth ieee international conference on data mining* (pp. 905–910). IEEE.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142–150).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research, 9*(11).

Manuel Pérez-Martínez, J., Berlanga-Llavori, R., Aramburu-Cabo, M. J., & Pedersen, T. B. (2008). Contextualizing data warehouses with documents. *Decis. Support Syst., 45*(1), 77–94. 10.1016/j.dss.2006.12.005.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404–411). Barcelona, Spain: Association for Computational Linguistics. https://aclanthology.org/W04-3252.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mothe, J., Chrisment, C., Dousset, B., & Alaux, J. (2003). Doccube: Multi-dimensional visualisation and exploration of large document sets. *Journal of the American Society for Information Science and Technology, JASIST, Special, 54*, 650659.

Oukid, L., Benblidia, N., Asfari, O., Bentayeb, F., & Boussaid, O. (2015). Contextualized Text OLAP Based on Information Retrieval. *International Journal of Data Warehousing and Mining (JDWM), 11*(2), 1–21. 10.4018/ijdwm.2015040101.

Park, B.-K., Han, H., & Song, I.-Y. (2005). Xml-olap: A multidimensional analysis framework for xml warehouses. *Dawak*.

Perez, J., Aramburu, M., Berlanga, R., & Pedersen, T. (2007). R-cubes: Olap cubes contextualized with documents. In *Proceedings of the 2007 ieee 23rd international conference on data engineering*. IEEE Press. Null; Conference date: 15-04-2007 Through 20-04-2007.

Periñán-Pascual, C. (2021). Measuring associational thinking through word embeddings. *Artificial Intelligence Review*, 1–38.

Ravat, F., Song, J., Teste, O., & Trojahn, C. (2020). Efficient querying of multidimensional rdf data with aggregates: Comparing nosql, rdf and relational data stores. *International Journal of Information Management, 54*, 102089.

Ravat, F., Teste, O., Tournier, R., & Zurfluh, G. (2008). Top_keyword: An aggregation function for textual document olap. In *International conference on data warehousing and knowledge discovery* (pp. 55–64). Springer.

Razavisousan, R., & Joshi, K. P. (2022). Building textual fuzzy interpretive structural modeling to analyze factors of student mobility based on user generated content. *International Journal of Information Management Data Insights, 2*(2), 100093. 10.1016/j.jjimei.2022.100093.

Sarkar, B. D., & Shankar, R. (2021). Understanding the barriers of port logistics for effective operation in the industry 4.0 era: Data-driven decision making. *International Journal of Information Management Data Insights, 1*(2), 100031. 10.1016/j.jjimei.2021.100031.

Sen, S., Roy, S., Sarkar, A., Chaki, N., & Debnath, N. C. (2014). Dynamic discovery of query path on the lattice of cuboids using hierarchical data granularity and storage hierarchy. *Journal of Computational Science, 5*(4), 675–683. 10.1016/j.jocs.2014.02.006.

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. In *2020 ieee 7th international conference on data science and advanced analytics (dsaa)* (pp. 747–748). IEEE.

Struijk, M., Ou, C., Davison, R., & Angelopoulos, S. (2022). Putting the is back into is research. *Information Systems Journal, 32*(3), 1–4. http://dro.dur.ac.uk/33883/.

Theodoridis, S., & Koutroumbas, K. (2009). clustering algorithms ii: Hierarchical algorithms. *Pattern Recognition (Fourth Edition): Academic Press*.

Unhelkar, B., Joshi, S., Sharma, M., Prakash, S., Mani, A. K., & Prasad, M. (2022). Enhancing supply chain performance using rfid technology and decision support systems in the industry 4.0–a systematic literature review. *International Journal of Information Management Data Insights, 2*(2), 100084. 10.1016/j.jjimei.2022.100084.

Wadud, M. A. H., Kabir, M. M, Mridha, M. F., Ali, M. A., Hamid, M. A., & Monowar, M. M. (2022). How can we manage offensive text in social media-a text classification approach using LSTM-BOOST. *International Journal of Information Management Data Insights, 2*(2), 100095. 10.1016/j.jjimei.2022.100095.

Zhang, D., Zhai, C., & Han, J. (2009). Topic cube: Topic modeling for olap on multidimensional text databases. In *Proceedings of the 2009 siam international conference on data mining* (pp. 1124–1135). SIAM.

Zhang, Z., Wang, H., & Feng, X. (2018). Olap on multidimensional text databases: Topic network cube and its applications. *Filomat, 32*(5), 1973–1982.

Information Processing and Management 59 (2022) 103086

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Design and Implementation of an Efficient Electronic Bank Management Information System Based Data Warehouse and Data Mining Processing

Jia Luo [a], Junping Xu [b,*], Obaid Aldosari [c], Sara A Althubiti [d], Wejdan Deebani [e]

[a] Hunan Key Laboratory of Macroeconomic Big Data Mining and its Application, School of Business, Hunan Normal University, Changsha, Hunan 410081, P.R. China
[b] The People's Bank of China Changsha Central Sub-branch, Changsha, Hunan 410005, P.R. China
[c] College of Engineering at Wadi Addawaser, Prince Sattam Bin Abdulaziz University, Wadi Al-Dawaser, 11991, Saudi Arabia
[d] Department of Computer Science, College of Computer and Information Sciences, Majmaah University. Al-Majmaah, 11952, Saudi Arabia
[e] King Abdulaziz University, College of Science & Arts, Department of Mathematics, Rabigh, Saudi Arabia

ARTICLE INFO

Keywords:
Data mining
Management information system
Index view
Marketing
Data warehouse

ABSTRACT

The quantity of electronic bank data grows exponentially with development of Information Technology (IT). The size of these data is impossible for traditional database and human analyst to come up with interesting information that will help in process of decision making. Management Information System (MIS) based Data warehouse (DW) and Data Mining (DM) techniques support the development of IT and process of management decision-making. But the traditional DW size make the query complex, which may cause unacceptable delay in decision support queries. Thus, in this paper an Efficient Electronic Bank MIS based DW and Mining Processing (EEBMIS-DWMP) was developed with cluster and non-cluster indexed view to provide decision-makers with both best response time and precise information. Also, analysis of the multilayer perception neural network, naïve Bayes, random forest, logistic regression, support vector machine and C5.0 on a real-world data of bank was done to improve effectiveness for campaign by analyzing the most useful features that influence campaign success. Results offer how the proposed EEBMIS-DWMP developed bank organizations by comparing performance of system with and without index view in terms of balance accuracy, accuracy, precision, recall, mean absolute error, root mean square error, F measure and running time. Conclusions from results offers that EEBMIS-DWMP can construct a database for each customer, a storage system that integrates data from a variety of sources into a single unified framework, decrease errors and time required to prepare financial reports, quickly access for information, analysis of data in multivariate, accurate prediction of competent, profitability segmentation.

1. Introduction

Increased market competitiveness causes a permanent decrease in the ability to react efficiently and quickly to new market trends. Companies are being overburdened with complex data, and those that can convert it into valuable information will have a competitive

* Corresponding author at: The People's Bank of China Changsha Central Sub-branch, Changsha, Hunan 410005, P.R. China.
E-mail address: xjunping@126.com (J. Xu).

https://doi.org/10.1016/j.ipm.2022.103086
Received 28 April 2022; Received in revised form 30 August 2022; Accepted 1 September 2022
Available online 20 September 2022

advantage (Bos et al., 2013, Ebadi et al., 2017, Jafari et al., 2021). It is common knowledge that business information is not used on a daily basis at the strategic level of decision-making, but rather cumulative and derivative data from a given time period. Because the problems that need to be handled in strategic decision making are usually nonstructural, huge volumes of data from the past must be included in the process of decision-making to ensure that the decision-making quality is met (Jaleel and Abbas, 2020, Malekshah et al., 2021, Zhao et al., 2021). As a result, the DW and DM concepts are mandated as a sound foundation for corporate decision-making. Furthermore, unstructured problems frequently accompany strategic corporate decision-making, which is why data warehouses have become a foundation for the creation of instruments for business decision-making, such as decision support systems.

As a current technical concept, DW is responsible for incorporating linked data from critical company operations in a format that is suitable for various analyses implementation. In the commercial world, data warehouses are not a new notion. The continually changing application of DW, on the other hand, is what is constantly being innovated. These apps enable managers to make better strategic decisions and gain a competitive edge through strategic planning (Adnan and Abbas, 2020, Arora and Gupta, 2017). The technology used to execute the DW will aid making decision, analysis design, and would be more cost-effective in the future. As a result, company managers should consider the cost-benefit analysis of deploying the DW in the long run. Bank DW technology is widely employed in the financial business, and the banking sector has seen significant benefits from it, prompting many banks to install or develop bank DW systems (Warners and Randriatoamanana, 2016). Online transaction processing (OLTP) systems handle an organization's operational data needs, which are critical to the day-to-day operation of its business. They are, nevertheless, unsuitable for running a business issues or decision-support queries that managers are frequently faced with. These problems necessitate data analytics like drilldown, slicing/dicing, and aggregation which are best handled by online analytical processing (OLAP) systems (Warners and Randriatoamanana, 2016, Malekshah et al., 2022, Chu et al., 2022).

Data may be stored and managed in a multidimensional fashion, DW aids OLAP applications. Transfer, Extract, and Load (ETL) tools that utilized to extract and load data from variety of OLTP data sources (containing business issues, SQL Server, DB2, flat files, and Oracle) into an OLAP warehouse. Financial management information systems, in general, are automation solutions that help governments plan, implement, and monitor their budgets. Via centralized web-based information and communication technology (ICT) solutions, modern FMIS platforms assist governments in complying with financial legislation and reporting requirements, as well as supporting dispersed budget operations (Sadoghi et al., 2016). Financial Management Information Systems (FMIS) solutions also make it easier for citizens to access public financial (PF) data, which improves engagement, government accountability, and budget transparency. The key FMIS functions and their interrelationships are depicted in Fig. 1 (Dener and Min, 2013, Malekshah et al., 2022, Seliya et al., 2021).

The following platforms are referred to as Integrated FMIS (IFMIS) for FMIS and other Public Financial Management (PFM) information systems (for instance, debt management, payroll, e-Procurement) are connected with a central DW to report and record all daily financial transactions, providing reliable consolidated outcomes for web publishing, budget analysis, performance monitoring, and decision support. FMIS solutions are uncommon in reality, and the phrase should never be used as a substitute for fundamental

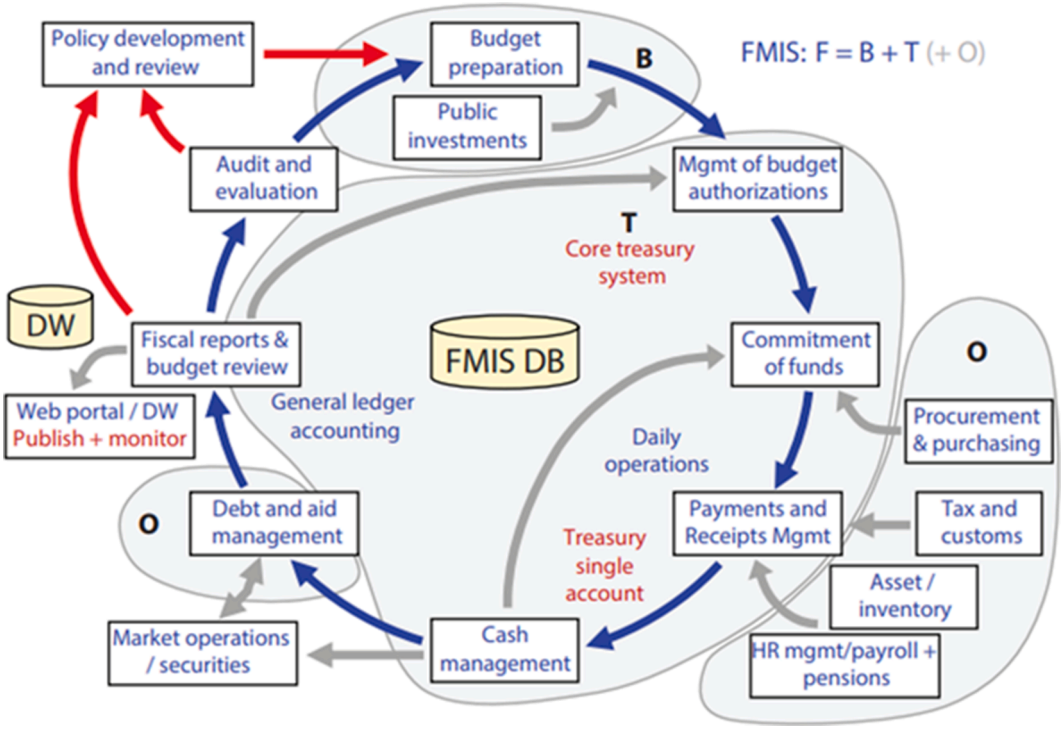


Fig. 1. Central FMIS interfaces and functions with other system of PFM.

FMIS capability to avoid creating false expectations. Operational systems PFM for online transaction processing (OLTP) are combined with capabilities of powerful DW for multidimensional online analytical processing (OLAP) in next generation IFMIS solutions to aid in decision support, performance monitoring, planning, and effective forecasting. Innovative IFMIS solutions also facilitate the dissemination of open budget data and enable more detailed analysis by giving dynamic query possibilities to a wide users' number, either internal (public organizations) and external (businesses, nongovernmental organizations, and citizens). While posting open budget data, take into account the need to safeguard the personal confidentiality or classified information. Other components of the government's financial operations (like, contingent liabilities, additional budgetary funds, fixed assets, quasi-fiscal activities, and tax expenditures) are included in public finance (PF) information (Dener and Min, 2013). Only data regarding the budget data disclosure will be collected in this study. In this paper, we designed an EEBMIS-DWM with cluster and non-cluster index to improve information accuracy and reduce query response time. The following is the rest of the paper: The second section contains related studies, the third section contains methods and materials, the fourth section has findings and discussion, and the fifth section contains conclusions and future works (Fig. 2).

2. Review of Data Warehouse and Data Mining

It can't be overstated the DW importance because of its provided benefits. The level of management decisions would not depend on erroneous and restricted data, it will also assist in avoiding various issues of businesses. As a result, DW becomes a must for each and every enterprise. Around 200 percent more items are predicted to access the internet and share data by 2020. The importance of devices and associated data in DW cannot be overstated.

The more linked devices you have, the more useful and powerful DW becomes. Many companies (Shahid et al., 2016) project that by 2016, the global room will have 6.4 billion linked peers, an increase of 30 percent from 2015. By 2020, Cisco and other research organizations (Shahid et al., 2016) estimate that between 20 and 50 billion of devices will be linked as shown in Fig. 3 (Shahid et al., 2016). The flip side of the coin is that costs will rise as well. If we speak about hardware expenditure, consumer applications will reach $546 billion by the end of 2016; so that, the use of connected things in the workplace will reach $868 billion by the end of 2016 as shown in Fig. 4 (Shahid et al., 2016).

When it comes to the value of DW, it is said that only a few application areas have the occurrence and integration of data across the enterprise, as well as the ability to make quick decisions based on current and historical (previous) data, and provide precise information for the systems that are identified. Fig. 5 depicts the real-world cycle of DW applications in several fields and how they are interconnected based on user preferences. The banking industry is regarded as one of the most information-intensive industries in the world of business. With the growth of information technology, the function of business intelligence (BI) in the banking process is becoming increasingly important (Nithya and Kiruthika, 2021). The requirement for financial intelligence has grown tremendously as company speed has increased and competitiveness has grown.

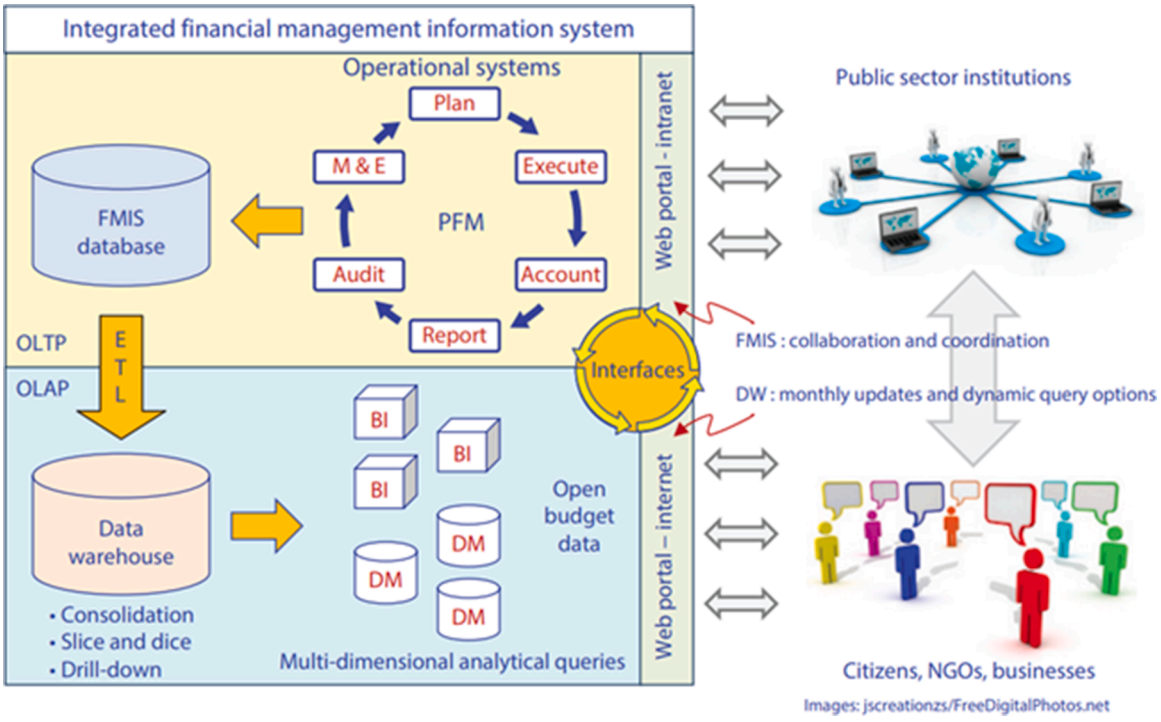


Fig. 2. IFMS with Data warehouse.

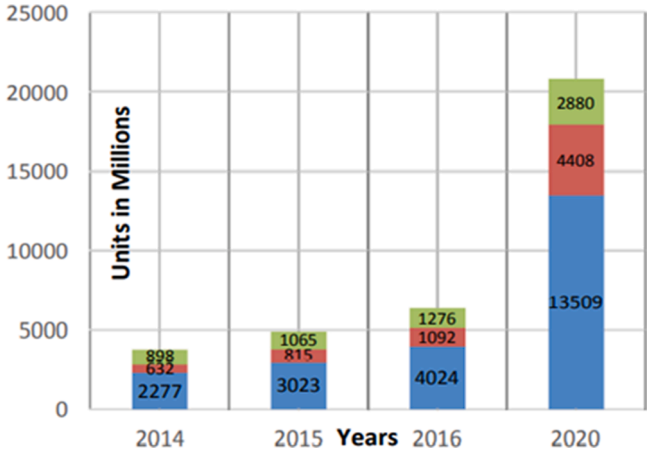


Fig. 3. Number of units.

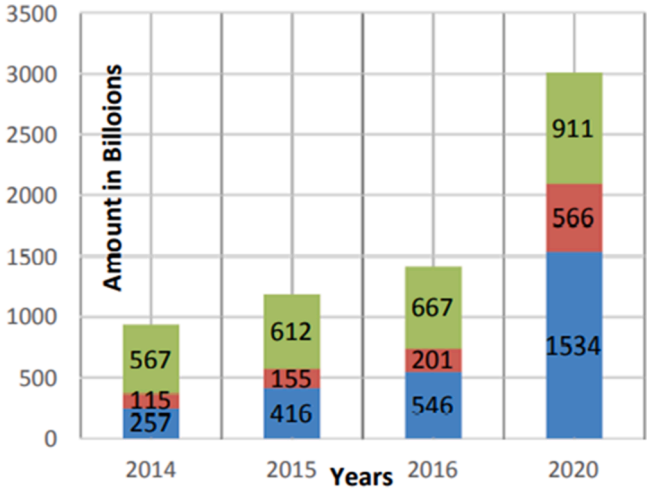


Fig. 4. Cost in Billion.

The ability to collect, handle, and analyze a huge quantity of data on bank all transactions, product, customers, services, operations, partners, and suppliers is known as bank intelligence. As the amount of data grows, it becomes more difficult to gather, manage, and translate it into meaningful information, which is where DW comes in. Many DW varieties are created to aid the banking business. With the growth of technology, particularly in the IT industry, new ways of conducting business, particularly in terms of financial systems, have opened up. In finance, both the government and the private sector play an equal role. Income tax, insurance firms, post offices, Banks and all other tax agencies are examples of financial systems. The use of DW in the financial business provides a number of advantages. It can, for example, maintain opening an account and transaction transparency. Similarly, the government has the authority to make judgments in the event of a financial crisis. These systems are clever enough to detect defaulters and take appropriate action. Although data warehouse was maintaining in this situation, a quick process of decision-making is possible. These data warehouses in financial applications may be used to analyze and forecast various elements of company, including performance of bond and stock analysis (Shahid et al., 2016, Bhedi et al., 2014). Their study question is around the variables that the banking sector should examine prior to and throughout the use of DW technology. Their findings indicated the Taiwanese banks number that have used this technology, as well as these banks architectures have implemented (Chau et al., 2003).

The agency of (Internal Revenue Service) is responsible for revenue collecting and enforcing tax laws in the United States. They built CRIS in the data warehouse since there was no other method to retrieve entities with certain attributes and run analysis on them. The implementation of DW consist of following domains: tax payments details entity, people's income sources entity, connected to taxpayer transactions entity, tax returns entity, and Business entity (Shahid et al., 2016). We begin our inquiry by looking at review papers on the use of data mining technologies to financial market difficulties, which is a topic that does not receive a lot of attention. The majority of them were about the stock market, but some of them also talked about other financial topics including credit rating, financial crisis analysis, interest rate forecasts, and exchange rate projections. The majority of the literature published surveys were focused on the use of bionic computing, text mining, support vector machines, artificial neural networks, ensemble learning, and

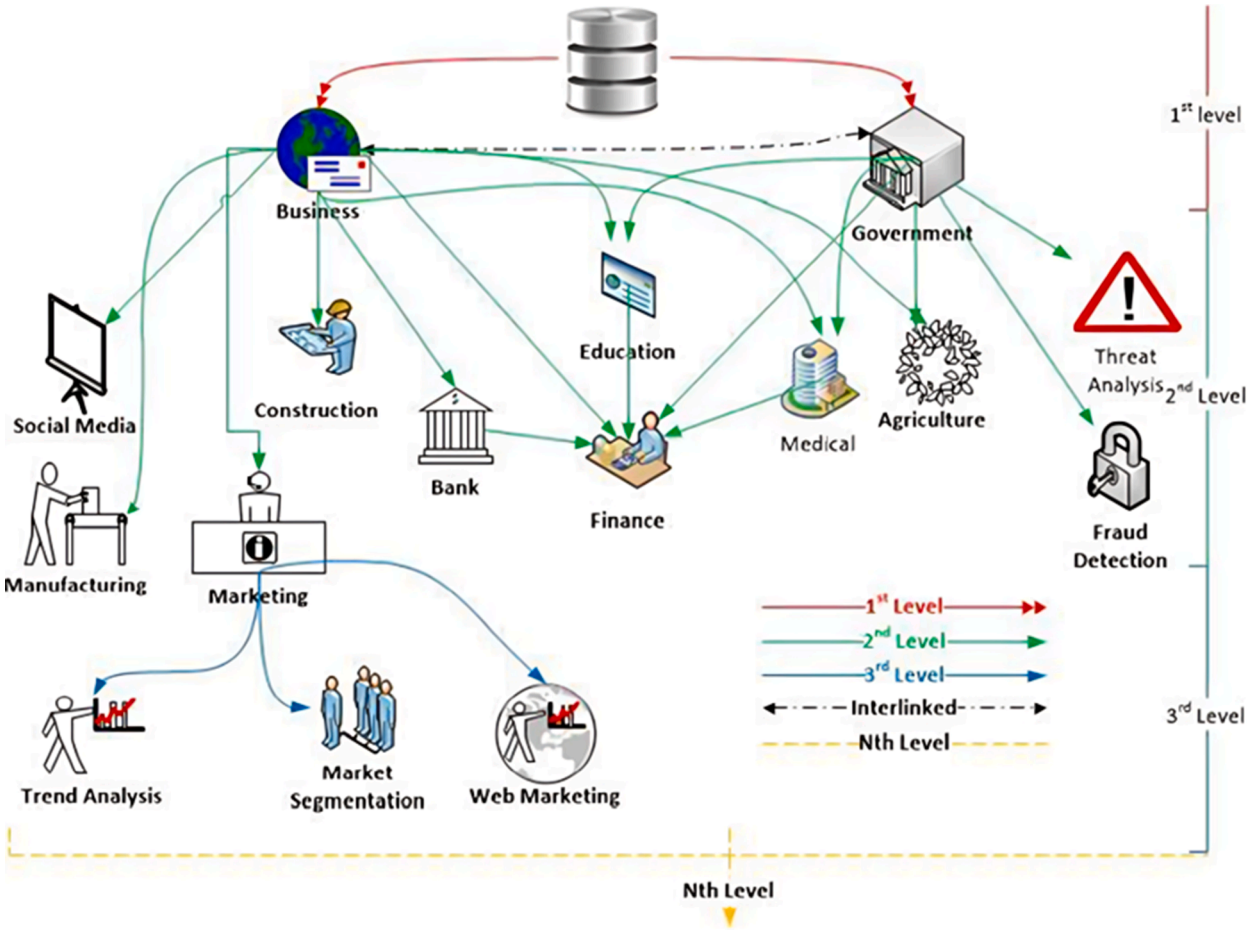


Fig. 5. Real life Application of DW.

decision trees in the financial market.

Some studies compared the most important forecasting techniques. Valavanis and Atsalakis from the four features of performance measures, benchmarks, predictor variables, and modeling techniques. Atsalakis and Valavanis (2009) analyzed about 100 scholarly articles on data-mining strategies to handle stock market forecasting challenges. They described the transfer functions types, membership functions types, training methods, and architecture of network to summarize fuzzy and neural networks. The use of stock market forecasting's artificial neural networks (ANN) was researched by Soni (2011). This paper presents stock market basic concepts and ANN brief history before enumerating several key investigations that employ ANN approaches to tackle stock market forecasting difficulties. Market's text mining forecasting was examined by Nassirtoussi et al. (2014), who classified and summarized the papers based on modeling technology type, preprocessing techniques, market category, and the text nature. Al Nasseri et al. (2015) looked through the literature for studies that linked changes in stock returns. They showed how these works had limits and recommended some modeling strategies for further investigation. Cavalcante et al. (2016) detailed numerous algorithms relating to different financial market concerns in addition to forecasting algorithms.

To address these issues, they suggested feature selection, grouping, segmentation, and outlier identification, as well as an architecture for the real market autonomous transactions. This paper summarizes the most relevant research in the duration from 2009 to 2015, which focuses on the use of various computational intelligence and data mining approaches in a range of financial applications, such as market forecasting, and financial data preprocessing and clustering. Other technologies include financial text mining, trends, and other technologies. Xing et al. (2017) used natural language to choose the most significant financial forecasting publications. They explained the techniques utilized in implementation and modeling details, as well as the sorts of financial language used as predicted input and associated processing methods. Quimbaya and Bustos.

The goal of Bustos and Pomares-Quimbaya (2020) was to update and synthesize the stock market forecasting approaches, including comparison, characterisation, and categorization. The stock market trends analysis forecasting from 2014 to 2018 is the topic of this review. It also examines survey and other assessments of new research published in a similar database and period of time. Despite the fact that these evaluations connect with our study period and themes, their reach is limited since they prefer to focus on a data mining methods single family; This examination is wider.

Major contributions of this work are:

- DW provides a versatile solution to the user, who can explore database effectively.
- User doesn't need to know about relational model or sophisticated query languages.
- This data analysis method allows OLTP systems to be optimized for data analysis.

This review also discusses and presents data mining algorithms, as well as the implementation of significant data mining algorithms, and financial market. This review, in comparison to these remarks, gives an up-to-date debate on the subject. It is assumed that these surveys will fast become outdated due to the enormous new works number produced each year (Liu et al., 2021).

3. Methods

The methodology for the proposed efficient framework starts with loading bank data that taken from Elsalamony (2014), the next step is data has been preprocessed and analyzed; In the proposed EEBMIS-DWM, the critical step of data preparation helps to improve the quality of data and bank data valuable insights facilitates extraction. It is an organizing and cleaning process of raw data to make it suitable for training and creating models of DM. After that Extract Transformation Load (ETL) has been applied; extract is a process in DW responsible for pulling bank data out of the bank source system, transforms means that raw data convert into an understandable and readable format, loading means placing bank data into DW.

It is worth noting that the construction of the DW must be completed before the operation of ETL process; the building of DW with

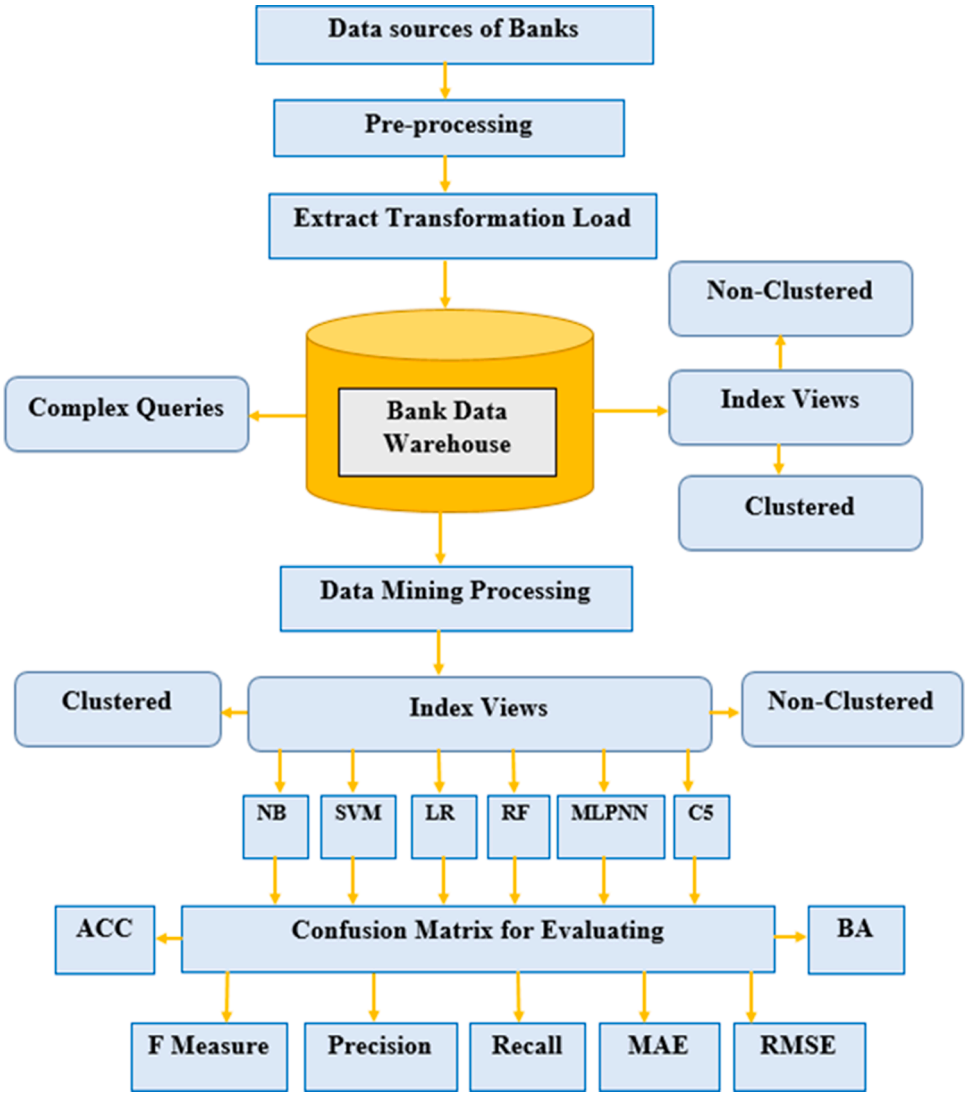


Fig. 6. Proposed EEBMIS-DWMP.

index view shown in section of results and discussion of this paper. Then after the building of DW has been completed; we added cluster and non-cluster index view for the DW to decrease the time response for the quires of decision makers, also the index view has been positive effect on the DM models. Then six DM algorithms has been applied on the DW of bank MIS to test its performance. Finally, the EEBMIS-DW has been evaluated by using important measures (accuracy, mean absolute error, root mean square error, precision, recall, F measure, and balance accuracy) from confusion matrix. Also, the time has been testing. The framework of EEBMIS-DWM shown in Fig. 6.

3.1. Data Collection

This paper used data of marketing of bank direct from the California University at Irvine that found in Elsalamony (2014) to evaluate the performances of NB, SVM, RF, ID3, and C4.5 models. According to a Portuguese bank's direct marketing initiatives, this information is relevant. In the marketing initiatives, phone calls were used. In many cases, it was necessary to make multiple contacts with the same client before learning if the product was subscribed or not. There are (45211) samples in the data of bank direct marketing, and only (17) attributes have missing values. Data set characteristics composed of two kinds: numeral and nominal attributes, as offer in Table 1. Three types of features are shown in this table; Numerical, which are in type of range for all of them like (Balance, Age, Day, Pdays, campaign, Duration, and Previous), Categorical are in type of set as the features (Marital, Job, Education, Month, Contact, Pout-come), and in their classes, binary categories include all of the traits that are either "yes" or "no."; for example, the features (Housing, Default, Output, Loan).

3.2. Pre-Processing and Extract Transformation Load

The data selection process identifies the data types in the data source, as well as the appropriate method for data extraction from a CSV data source. All the data that selected will be placed in a server's staging area table, where it will be processed and loaded into the DW tables. The schema of DW should be install and be ready for loading the staging area processed data, which is presented by a table that holds all of the data that selected with its types and serves as a link between the data source and the storage area of DW (Hamoud et al., 2020).

The essential data integration process is extraction, transformation, and loading (ETL), it is frequently linked to data warehousing. ETL tools extract data from a source, transform it depending on business needs, and then load it into a destinations data structure. Three database functionalities are merged into a single tool that automate the process of transferring data from one database to another. The following are the database functions: The process of reading data from a defined original database and obtaining a desired subset of it is known as extraction. Transform — the transferring process of extracted/acquired data from one type to other one so that it can entered to the database. This transformation done by utilizing lookup tables or rules, or by integrating it with other information. The writing data process in the target database is known as loading. When a question is submitted, data is taken from source data sources on demand and transformed to give a query response, which is an alternate way to information integration (Gour et al., 2010).

3.3. Clustered and Non-Clustered Index View

Since the rows of data themselves can only be sorted in one order, multiple clustered indexes on a single database are impossible (MJ Al Taleb et al., 2020). Usually, it is determined by the primary key (PK). Arrange the records and physically place them on the correct order disk. A table with a clustered index is referred to as a clustered table. If a table lacks a clustered index, the rows of data are stored as a heap, which is an unsorted structure. A clustered index instructs the database to keep values that are near to each other on the disk. This has the advantage of allowing for quick scanning and retrieval of data that fall inside a certain range of index values cluster. Retrieving data is faster with clustered indexes than with non-clustered indexes. Extra space is not required to store the logic structure. Because the table rows are kept on disc in similar precise order of clustered index, clustered indexes shouldn't need to store the actual row reference. The rows storage order in the table is determined by a clustered index, that would not require extra disk

Table 1
Performance of DM without building Data Warehouse.

DM	ACC%	BA%	Precision%	Recall%	F-Score%	MAE%	RMSE%	Time (s)
LPNN	61.21	59.4	63.11	65.22	30.11	42	47	20
SVM	70.15	68.2	72.41	74.55	35.52	39	44	25
RF	72.75	69.11	74.15	77.33	37.42	32	37	22
LR	70.2	68.81	71.18	73.12	33.13	36	40	24
NB	72	70.01	74.22	76.19	37.12	34	38	20
C5	74.2	72.51	77.12	78.02	38.66	30	32	19

space. (MJ Al Taleb et al., 2020) cites the following: The non-clustered index is as follows: non-clustered indexes can be as many as you like in a database since they don't affect the sequence in which entries are saved on disc as cluster indexes do.

Typically, this is done across any key. The values of non-clustered index key are stored in the non-clustered index, and that each key value entry contains a reference to the data row containing the key value. Do not disrupt the natural order. Use references to files of physical data to make a proper logical order for data rows. A row locator is a pointer from an indexing row in a nonclustered index to a data row. The data pages are kept in a heap or a clustered table determines the row structure locator. A row locator is a row pointer in a heap. The index key clustered is the row location for a clustered table. Extra space can be used for logical structure storing. The table's non-clustered index is kept separate from the table.

3.4. Building of Data Warehouse with Cluster and Non-Cluster Index

Initially, a view is generated to fulfill the user's needs for a critical query. The index of unique clustered established on the view when it is created. At first index that produced on the view has to be a unique clustered index, and this index of unique clustered should be built before any further indexes on the view may be formed. After a view has a unique clustered index, additional non-clustered indexes can be added. Views with non-clustered indexes can improve query speed and provide the query optimizer more alternatives to pick from during the process of compilation. On the fact table, a single identification column named column fact key is formed and made PK. Multicolumn PK is an alternate to a single column PK.

A mixture of dimensions key columns is used to produce the PK. Usually, only a few of the dimension key columns are affected. Just those whose combination allows us to identify each fact row individually. Because a multicolumn PK might not have been unique, an identification single column PK is preferable than a multi-column PK. Even though those columns combination is unique right now, that doesn't mean the main key won't be copied in the future. Since it is an identification column, the single-column technique ensures that the unique PK, which is the primary rationale for constructing fact keys. The second purpose for generating a fact key is to increase insert efficiency. When using a multicolumn PK, SQL server must first choose where to put that row depending somewhat on clustered column. Depending on the clustered column values, the row doesn't really automatically move to the table end; instead, it goes to the center. The SQL server spend a significant amount of time determining where each entry should be inserted. The fact key's data type is a BIGINT. The query performance is the most important aspect in DW, followed by the load performance. When it comes to balancing loading and query speed, single column PK is superior, while multicolumn PK is slower in loading and quicker in querying than single column PK. The fact key column clustered index, followed by non-clustered indexes on every one of the dimensional key columns, is a typical approach of index a fact database. Because a non-clustered index in SQL Server utilizes the key of clustered index as the row location, this would enable every one of the non-clustered indexes to be efficiently utilized while also guaranteeing that they are thin. Because they are based on every one of the dimension keys, they are quite effective. As the initial index key, each of them is either a second index key or the third one. Since they're the index key's first column, they will be accessed by queries that utilize that dimension key. If the changed row requires more capacity, repeated updates might degrade performance, necessitating maintenance to build indexed views again (MJ Al Taleb et al., 2020).

3.5. Data Mining Algorithms

Knowledge Discovery (KD) or SDM discovery in databases is defined as the process of locating and extracting new and useful information from massive data sets (Fig. 7). With the use of a variety of methodologies, it is possible to unearth previously unknown qualities, features, or connections within a set of data. Currently, this regulation is being used in a wide range of corporate applications. Following are the steps of the general knowledge discovery process that were defined in 1996 by U. Fayyad and G. Shapiro: understanding field of application, data preprocessing, selecting, integration, transformation, cleaning, reduction, choosing algorithms of

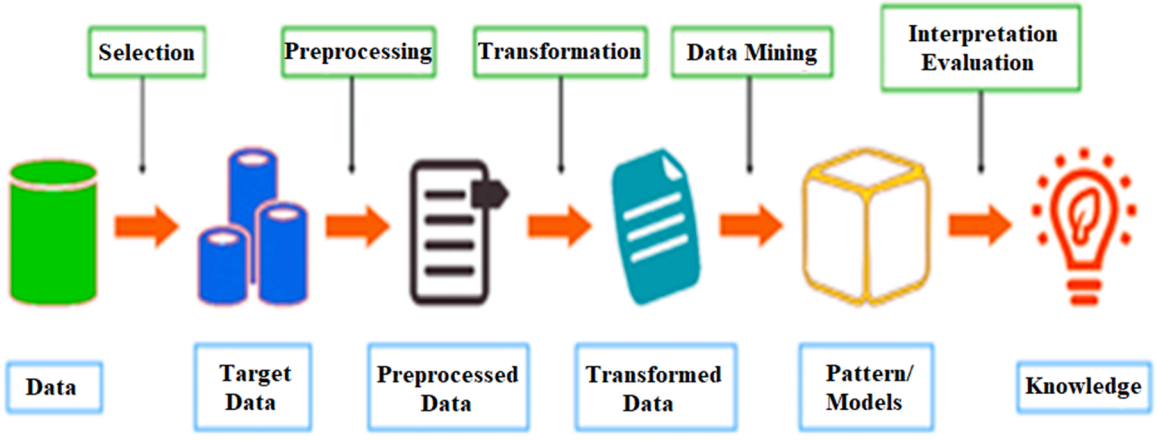


Fig. 7. Data Mining steps.

mining data, description and interpretation of the findings using the discovered knowledge (Nikam, 2015).

3.5.1. NAÏVE BAYES

An efficient and successful algorithm for classification is NB. It is a one particle type of a network for Bayesian. Unrestricted Bayesian networks appear to have a logical topology and parameter set. On a large number of benchmark data sets, Friedman (1997) discovered that NB easily beats an unconstrained Bayesian network classifier. A sample's likelihood of falling into a certain category can be predicted with the use of Bayesian classifiers. Large databases benefit from this method's high classification accuracy and speed, as well as its ease of training using simple models. To estimate the parameters, only a little amount of training data is required. (Variances and means of the variables) handle discrete and real data, necessary for classification. Streaming data is no problem for it. Some of the alleged shortcomings of Bayesian analysis, on the other hand, aren't actually problems in practice. Any ambiguity in a prior choice is often not problematic because the many convenient priors usually do not disagree substantially in the regions of interest. Analysis of bayesian isn't confined to what's generally thought of as statistical data; it can be applied to any model space (Campagner and Ciucci, 2018, Pandya and Pandya, 2015, Xu et al., 2020).

3.5.2. Logistic Regression

The LR algorithm is well suited to dealing with a wide variety of data; it is provided with a suitable number of samples that are evenly distributed. If you want to test hypotheses regarding the link between one or more categories of continuous predictor attributes and an outcome variable, this is a good choice. Furthermore, LR employs maximum probability estimation rather than least squares estimation unlike classic multiple regression. The likelihood that the sample came from a population with the predicted parameters is calculated using the predicted parameter starting values. The calculated parameters' values are modified iteratively until the highest probability value is found. As a result, methods based on maximization of probability seek to identify parameter values that make data seen "most likely". (Zanaty, 2012, Montebruno et al., 2020).

3.5.3. Random Forest

Decision trees are the ancestors of RF, which is a classifier derived from them. This method, as its name implies, produces a forest from a collection of trees. When it comes to classification and regression difficulties, the random forest approach is an excellent choice. A decision tree is a sort of model that can be used for regression as well as classification. A decision tree usually starts with a single node and then divides into different branches to represent different outcomes. RF generates a large number of decision trees from randomly selected data samples, receives predictions from each tree, and votes on the best option. We get a straight line that bisects the log(x) function if we try to design a basic linear model to forecast y using x. However, if we use a random forest, the log(x) curve is much better approximated, and the resulting function looks much more like the real thing (Chen et al., 2022).

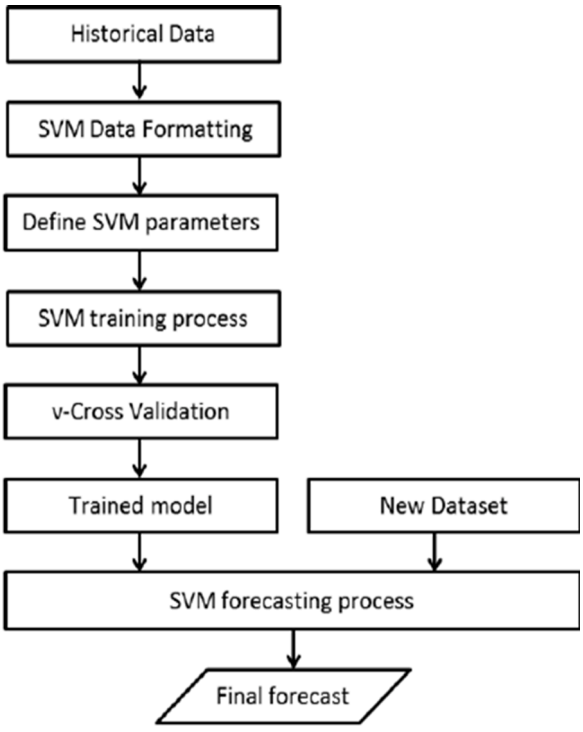


Fig. 8. SVM operation.

3.5.4. Support Vector Machine

As SDM algorithm for problems like classification and regression, SVM was created by Boser, Guyon, and Vapnic. SVMs can handle high-dimensional data with ease and are adaptable when modeling a variety of data sources. Both the training and testing phases make up the SVM modeling framework (Fig. 8). Depend on training data the goal of the SVM is to create a model that can forecast the target values of test data given just the test data features (Choi and Kim, 2013).

3.5.5. Decision Tree (C5)

C5.0, a freshly designed modeling technique that is an upgraded variant of C4.5 and ID3, is the most well-known typical in decision trees. C5.0 is a commercial program developed by Rule Quest Research Ltd Pty to analyse large data sets and is included in the SPSS Clementine data mining package. The C5.0 tree employs standard splitting methods, including entropy-based information gain. The model divides the sample according to the property that delivers the most information gain. Each subsample specified by the initial split is then divided a second time, generally basing on a different feature, and the procedure is repeated until the sub-samples could no longer be split. Lastly, the low-level split is re-examined, and those that do not significantly contribute to the model's value are trimmed or deleted. In the occurrence of issues like as incomplete data and a high number of input fields, the C5.0 model is extremely resilient. Estimation normally does not need extensive training. Furthermore, because the rules produced by the model have a really simple and direct interpretation, C5.0 models are easier to comprehend than other model types (Campagner and Ciucci, 2018, Pandya and Pandya, 2015).

3.5.6. Multilayer Perception Neural Network

The artificial neural network architecture that frequently used is the multilayered perceptual neural network (MLPNN) with back propagation. The MLPNN is a well-known function approximation for classification and prediction applications. In 1943, Neuro-physiologist Warren McCulloch and mathematician Walter Pitts published a paper on how neurons could operate, which launched this discipline. They discovered an electrical circuit-based model for a basic neural network. This paradigm was given the moniker "threshold logic" by the researchers. The model opened the way for the division of neural network research into two distinct methodologies.

The first focused on the brain's biological processes, while the second focused on neural network applications in artificial intelligence. In 1982, the most fascinating in the field was revived. John Hopfield pioneered the use of bidirectional wires to build more usable devices. Three distinct research groups, one of which contained David Rumelhart, suggested the same concepts in 1986 using several layered neural networks, that is known nowadays as back propagation networks since they propagate pattern recognition mistakes across the network. Multiple layers of back propagation networks, whereas hybrid networks have just two. In Craven and

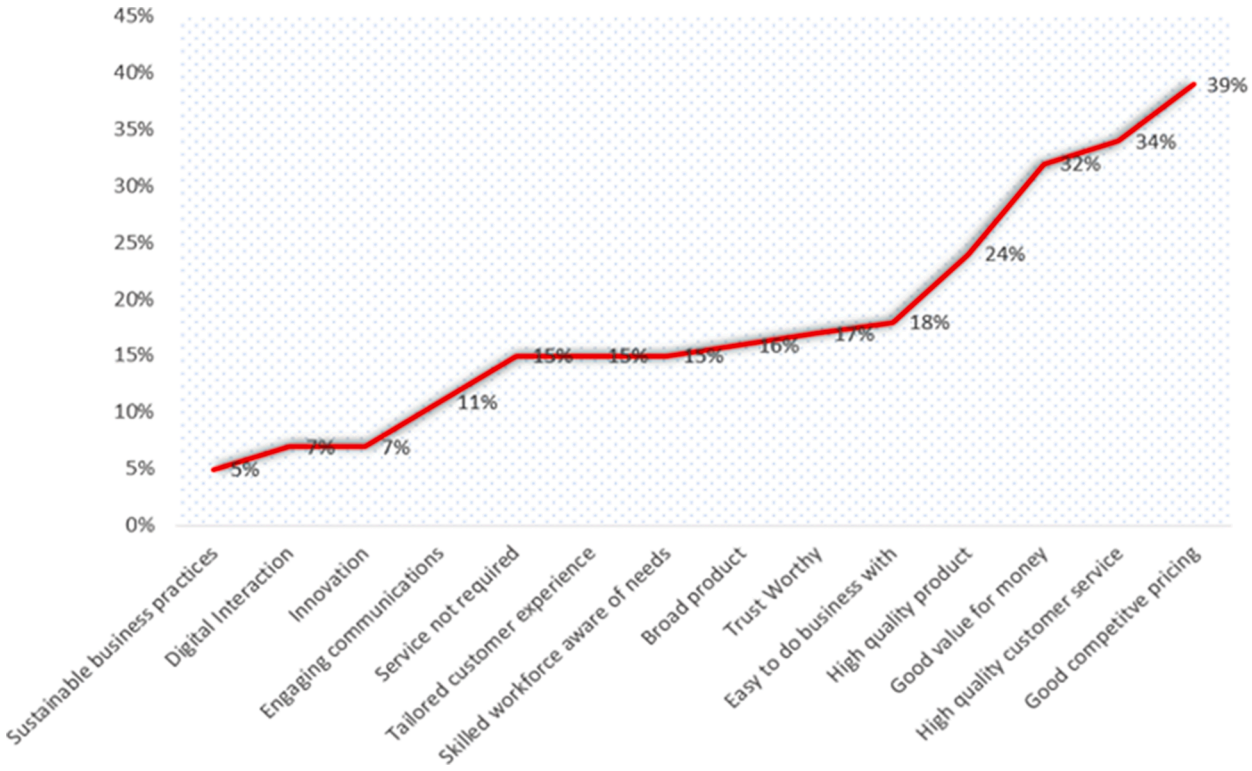


Fig. 9. Reasons.

Sahvlik, neural networks are used to mine data (1997) (Zanaty, 2012).

4. Results and Discussion

The details of the results offer in this section and we divide it into 3 subsections; first subsection for the Bank MIS in IT departments, second subsection for confusion matrix and performance measures, third subsection for the effect of index views for the proposed system models.

4.1. Bank MIS IN IT Departments

The most important aspect of every business is the customer. Organizations must not only attract clients, but also work hard to retain them. This implies that they must improve their offers and services. Customers of stores and banks, according to research in (Pilarczyk, 2016, Mohammadi et al., 2010), had the highest level of disloyalty. The top three prominent criteria that persuade clients to move to another bank are competitive pricing, high-quality customer service, and exceptional value for money (Fig. 9). It's no accident that banking has been one of the most widely diffused sectors in recent years as a result of digital technology. More than 80% of customers who transferred to some other supplier due to bad service claimed they could be kept if their problem had been fixed on the first interaction with the bank. There are several factors that influence the choice to switch or quit conducting business, as indicated in Fig. 6. The installation of MIS in IT departments is one of the keys for improving service and being able to deliver a better offer. It aids in the enhancement of visibility, control, and management. Increases service quality and operator efficiency by combining service monitoring and infrastructure technologies. Managing the experience of the customer (both back-end and front-end) and providing early notifications when issues arise enhances customer retention and service quality. Performance management and consolidated events and also improve efficiency.

Other advantages recognized by IBM after analyzing successful implementation include reduced inventory, supply chain improvement (people and equipment are always on time), standardization and unification of processes, lower expenses and maintenance costs, better equipment utilization, increased ROA factor (return on assets), increased OEE factor (efficiency of operational equipment), and increased resource and service productivity. Banks are forced to adopt IT because they have no other option. Technology must be dependable and nimble in order to win and retain customers. With MIS in place in IT departments, this would be possible to obtain. The World Bank Group can help with the construction or modernization of treasury systems, as well as the implementation of integrated Public Financial Management (PFM) frameworks, because it has cutting-edge expertise and extensive worldwide experience. Since 1984, the World Bank has supported 152 projects in 84 countries (122 completed, +25 ongoing, and +5 in the pipeline), as illustrated in Fig. 10 (Dener and Min, 2013).

4.2. Confusion Matrix and Performance Measures

Six statistical indicators are used to assess the performance of each model; Accuracy (ACC), Balance Accuracy (BA), precision, recall, F-Score, and time of running. In Kohavi and Provost, 1998, these metrics are referred to as a confusion matrix, which contains information on the classifications made by a classification framework and those projected. It is using true negative (TN), true positive (TP), false negative (FN), and false positive (FP). Correct categorization % is the difference between forecasted and actual values for a set of variables. The number of correctly predicted instances that indicate a true positive (TP); When a classifier's positive prediction and a target attribute's positive forecasting occur simultaneously, this occurs. A True Negative (TN) is a prediction that an instance is untrue, (i.e.) a negative prediction is made when the classifier and the target attribute disagree. It is the number of false positives (FP)

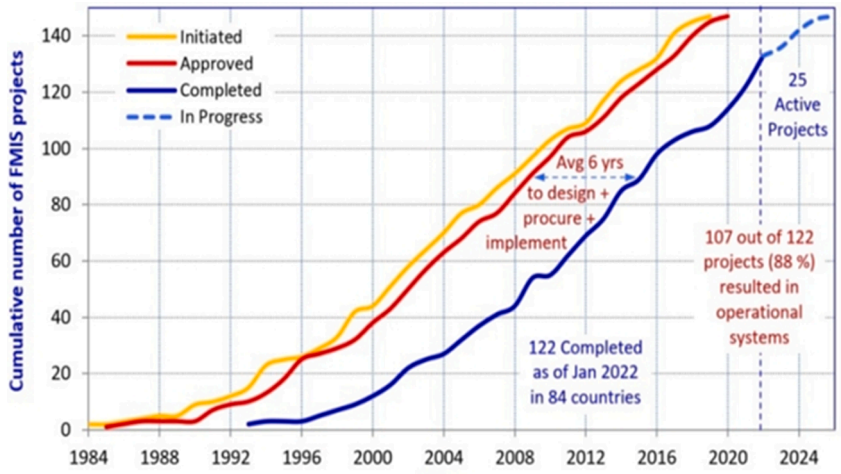


Fig. 10. FMIS Projects funded by WBG.

that an instance is false. True Negative (FN) is how many times an instance has been wrongly predicted to be untrue (Wang et al., 2022, Mehregan and Asadollahi, 2022). Figs 11, 12 and Table 1 offer the confusion matrix for the two classes and its parameters of SDM algorithms respectively without building DW while Figs. 13, 14 and Table 2 are with building DW. ACC is defined as the number of correctly classified cases divided by the total number of cases N, which is the sum of TP and TN, which is calculated using eq. 1. BA is computed using eq. 2. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE): All the absolute value discrepancies between actual and anticipated examples are used to calculate MAE, a measurement. Eq. 7 can be used to calculate the MAE. RMSE computes the average value of all squared distinction between the true and forecasted situations and then progress to calculate the root of square of the finding. Eq. 8 explains how to calculate it. Where b_i: is the true cases $\widehat{b_i}$ the forecasted states and n: are the number of states. Sensitivity (TPR), True Positive Rate or the Recall: The recall of the forecasting algorithm is the number of accurately forecasted negative states in the bank data from all states. For calculating recall, equation5 was used. And the equation of Specificity, or TNR shown in eq. 3. There is a direct link between the bank's direct marketing data set node and an EXCEL file that holds the source data. The data set was looked at in terms of ordinal data types.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

$$BA = \frac{TPR + TNR}{2} \tag{2}$$

$$TNR = \frac{TN}{TN + FP} \tag{3}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\mathrm{F-Measure} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{6}$$

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |b_i| - |\widehat{b_i}| \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (b_i - \widehat{b_i})^2} \tag{8}$$

4.3. Effect Of Cluster and Non-Cluster Index Views

Because the fact table includes only 123502 records and the test results are in milliseconds, the stress on the system is raised by utilizing the SQLQueryStress tool to acquire results in minutes or seconds. For illnesses, an indexed view was constructed to increase the performance of the primary queries that decision maker was frequently asking, and SQLQueryStress was used to assess its performance. This indexed view displays the number of consumers for each transaction in different areas and periods as shown in Figs. 15, 16 and Table 3. Each index (clustered or non-clustered) takes up around 4.718 MB of storage, with a row count of 42211 records

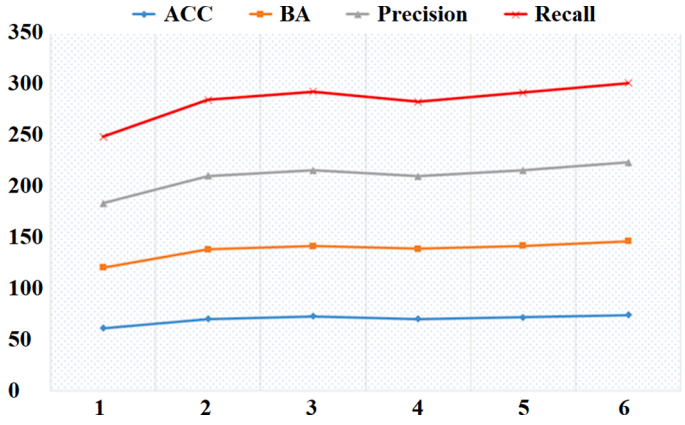


Fig. 11. Performance of DM without building Data Warehouse (Set 1).

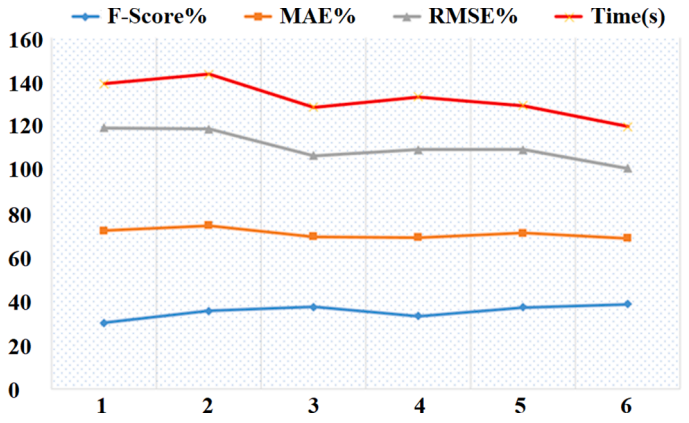


Fig. 12. Performance of DM without building Data Warehouse (Set 2).

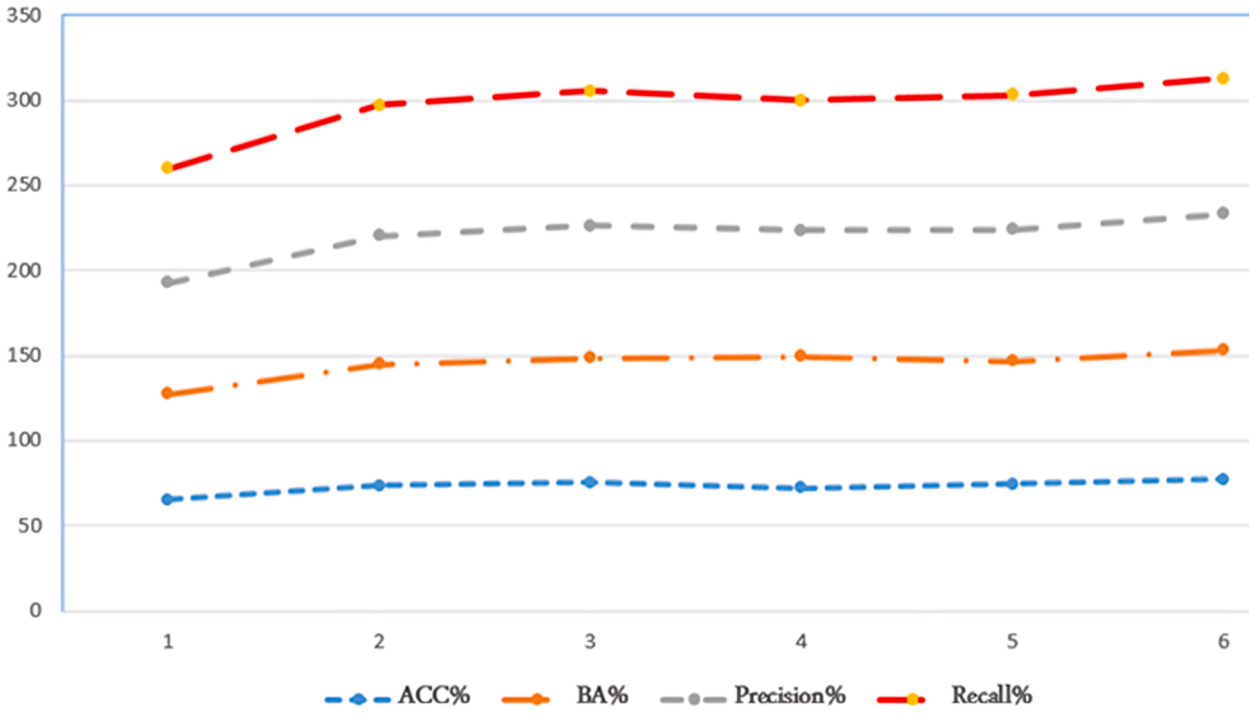


Fig. 13. Performance of DM with building Data Warehouse (Set 1).

connected to finance transactions, and the size grows as the number of data grows. Table 4 compares the speed of the ordinary and indexed views, using the standard query and row number of 42211, 400 threads, and 100 iterations. Speed increased after applying a clustered index, and performance improved even more after adding a non-clustered index, although the size of the indexed view grew with each addition. Although date is significant and utilized in most queries, it is not chosen as a non-clustered index in the indexing view of illness since entries are added in the fact table on a regular basis (see Table 5). If the date is set as a non-clustered index, performance may increase somewhat or decline slightly. Note that various factors can impact the elapsed time, including server load, input/output load, and network bit rates between server and client.

5. Conclusions

The stock market trends analysis forecasting from 2014 to 2018 is the topic of this review. It also examines survey and other assessments of new research published in a similar database and period of time. Despite the fact that these evaluations connect with our study period and themes, their reach is limited since they prefer to focus on a data mining methods single family. Our examination is wider. Our review also discusses and presents data mining algorithms, as well as the implementation of significant data mining

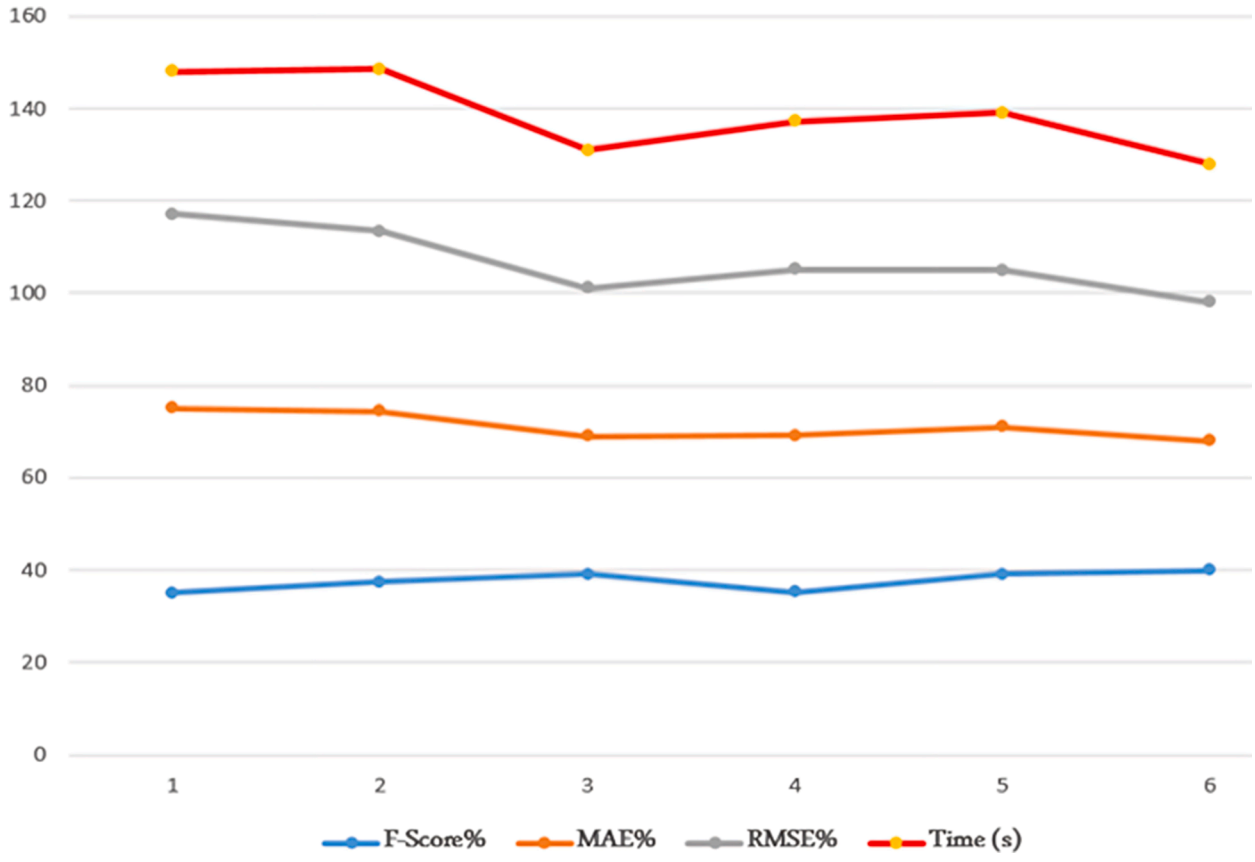


Fig. 14. Performance of DM with building Data Warehouse (Set 2).

Table 2
Performance of DM with building Data Warehouse.

DM	ACC%	BA%	Precision%	Recall%	F-Score%	MAE%	RMSE%	Time (s)
MLPNN	64.92	62.44	65.19	67.11	35.12	40	42	31
SVM	73.44	71.44	75.24	77.17	37.47	37	39	35
RF	75.45	73.15	77.45	79.3	39.05	30	32	30
LR	72.22	77.22	74.18	76.31	35.15	34	36	32
NB	74.55	72.19	77.22	79.28	39.02	32	34	34
C5	77.12	75.91	79.92	80.07	40.02	28	30	30

algorithms, and financial market. This review, in comparison to these remarks, gives an up-to-date debate on the subject.

This paper depicts the processes that lead to the formation of a DM and DW solution. Based on the example, we can infer that DW provides a versatile solution to the user, who can explore the database more effectively using tools like Excel with user-defined queries than with any other OLTP tool. The user doesn't need to know about the relational model or sophisticated query languages, which is a big advantage of this database knowledge and information retrieval solution. The Data warehouse method of data analysis is becoming increasingly popular since it allows OLTP systems to be optimized for their intended function and data analysis to be transferred to OLAP systems. In conclusion, this discourse hopes to have substantially explained the importance and role of the proposed EEBMIS-DWMP in bank and finance organizational development.

Indexed views can help increase query speed and solve the decision support workloads problem by minimizing the work amount that SQL server has to do to retrieve the relevant data. On the bank direct marketing data set, this research evaluated and compared the six different DM classification performance methods' models MPLNN, NB, LR, RF, SVM, and C5.0 to classify for bank deposit subscription. The training and test portions of this data set are 70 percent and 30 percent, respectively. The efficiency of models has been demonstrated by experimental data and offers that C5 is better than TAN, SVM, RF, LR, and MLPNN.

This data analysis method allows OLTP systems to be optimized for data analysis, thus, for the future research, it should be tested in practical environment.

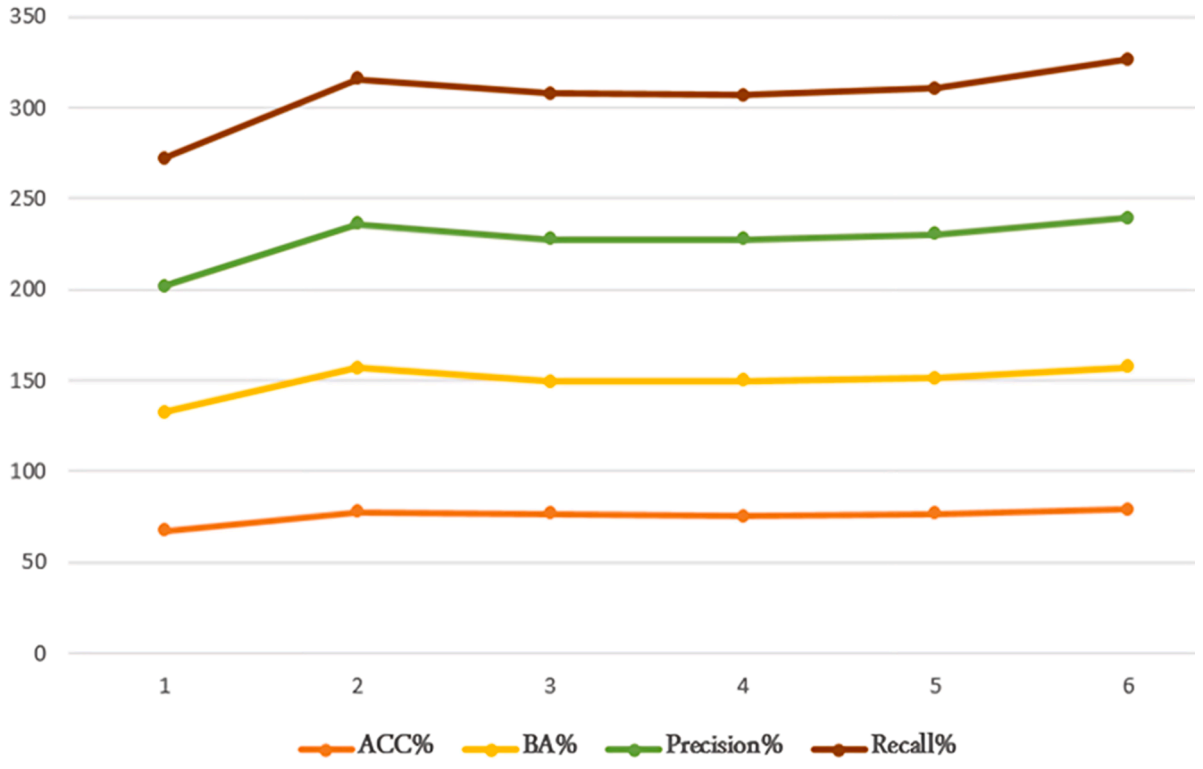


Fig. 15. Performance of DM with building data warehouse and Index views (Set 1).

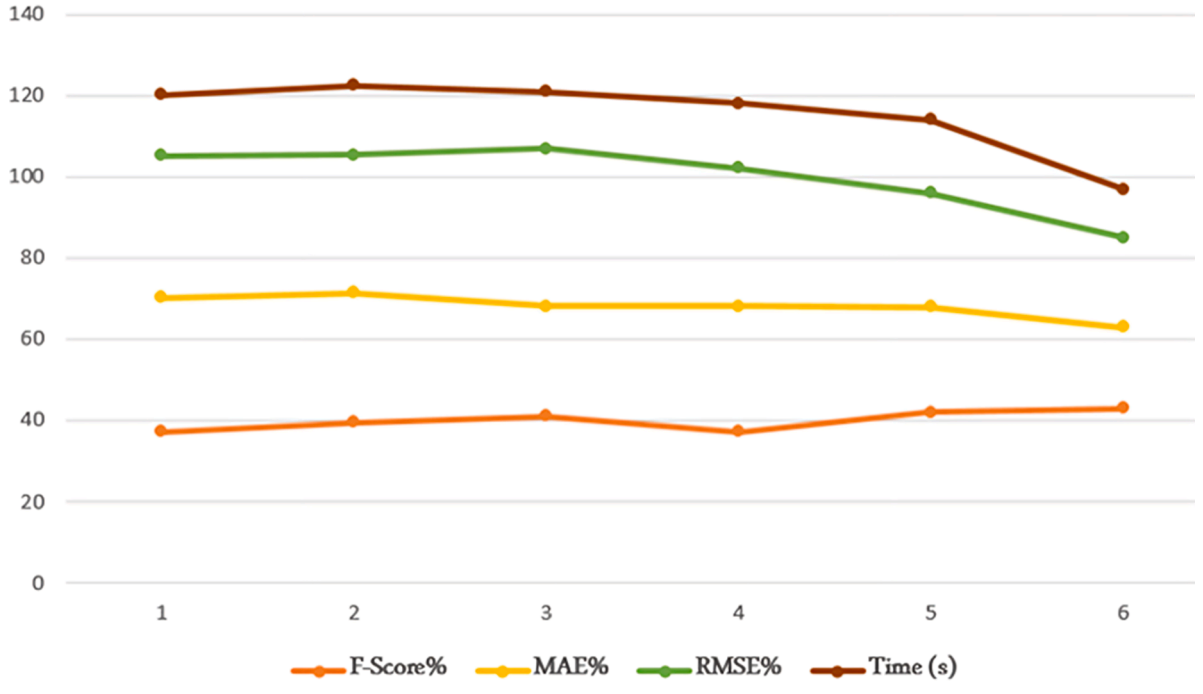


Fig. 16. Performance of DM with building data warehouse and Index views (Set 2).

Table 3
Performance of DM with building data warehouse and Index views.

DM	ACC%	BA%	Precision%	Recall%	F-Score%	MAE%	RMSE%	Time (s)
MLPNN	67.34	65.22	69.19	70.31	37.22	33	35	15
SVM	77.54	79.14	79.21	80.11	39.47	32	34	17
RF	76.45	73.15	78.11	80.23	41.05	27	39	14
LR	75.32	74.32	77.82	79.31	37.15	31	34	16
NB	76.45	74.99	78.92	80.28	42.02	26	28	18
C5	79.22	77.91	82.22	87.07	43.01	20	22	12

Table 4
Compares between index and regular view.

Function	Regular	Clustered Index	Clustered index with Non-clustered Index
Average Client Seconds /Iteration	4	3	2
Average Logical Reads/Iteration	1398	341	341
Average CPU Seconds/Iteration	0.0761	0.0301	0.0102
Average Actual Seconds/Iteration	4	3	1
Current space	Default	1.921 MB	2.921 MB
Elapsed Time	00:05:75.2719	00:03:77.3678	00:02:01.1411

Table 5
Comparing between Traditional system and the proposed EEBMIS-DWMP.

Traditional System	Proposed System (EEBMIS-DWMP)
Relational DB Technique	DW and DM Techniques
No Complex Analysis	Complex Analysis
No Dependency Analysis	Dependency Analysis
Data in detailed	Data in aggregated
No security for data	Security for data
Used for analysis and processing	Used for data analysis
No Sophisticated analysis	Sophisticated analysis
No Detection of faulty data	Detection of faulty data
No User defined queries	User defined queries
High Avg. query time	Little Avg. query time

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was supported by National Social Science Foundation of China [Project in 2022：A Research on Synergetic Regulation and Supervision of "Gray Rhino" financial risks]. Also, this publication (for the 3rd author) was supported by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia.

References

Adnan, R., & Abbas, T. M. (2020). Materialized Views Quantum Optimized Picking for Independent Data Marts Quality. *Iraqi Journal of Information and Communications Technology, 3*(1), 26–39.

Al Nasseri, A, Tucker, A, & de Cesare, S (2015). Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Syst Appl, 42*, 9192–9210.

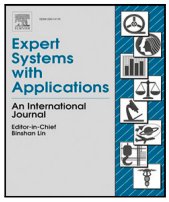
Arora, R. K., & Gupta, M. K. (2017). e-Governance using data warehousing and data mining. *International Journal of Computer Applications, 169*(8), 975–8887.
Atsalakis, GS, & Valavanis, KP (2009). Surveying stock market forecasting techniques – part II: Soft computing methods. *Expert Syst Appl, 36*, 5932–5941.
Bhedi, V. R., Deshpande, S. P., & Ujwal, A. L. (2014). Data Warehouse Architecture for Financial Institutes to Become Robust Integrated Core Financial System using BUID. *International Journal of Advanced Research in Computer and Communication Engineering, 3*(3), 2278-102.
Bos, J. W., Kolari, J. W., & van Lamoen, R. C. (2013). Competition and innovation: Evidence from financial services. *Journal of Banking & Finance, 37*(5), 1590–1601.
Bustos, O, & Pomares-Quimbaya, A (2020). Stock market movement forecast: A systematic review. *Expert Syst Appl, 156*, Article 113464.
Campagner, A., & Ciucci, D. (2018). Three-way and semi-supervised decision tree learning based on orthopartitions. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 748–759). Springer, Cham.
Cavalcante, R. C., Brasileiro, R. C., Souza, V. L., Nobrega, J. P., & Oliveira, A. L. (2016). Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications, 55*, 194–211.
Chau, K. W., Cao, Y., Anson, M., & Zhang, J. (2003). Application of data warehouse and decision support system in construction management. *Automation in construction, 12*(2), 213–224.
Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management, 59*(2), Article 102798.
Choi, M., & Kim, H. (2013). Social relation extraction from texts using a support-vector-machine-based dependency trigram kernel. *Information processing & management, 49*(1), 303–311.
Chu, Y. M., Shankaralingappa, B. M., Gireesha, B. J., Alzahrani, F., Khan, M. I., & Khan, S. U. (2022). Combined impact of Cattaneo-Christov double diffusion and radiative heat flux on bio-convective flow of Maxwell liquid configured by a stretched nano-material surface. *Applied Mathematics and Computation, 419*, Article 126883.
Dener, C., & Min, S. Y. S. (2013). Financial management information systems and open budget data: do governments report on where the money goes?. World Bank Publications.
Ebadi, M. J., Hosseini, A., & Hosseini, M. M. (2017). A projection type steepest descent neural network for solving a class of nonsmooth optimization problems. *Neurocomputing, 235*, 164–181.
Elsalamony, H. A. (2014). Bank direct marketing analysis of data mining techniques. *International Journal of Computer Applications, 85*(7), 12–22.
Gour, V., Sarangdevot, S. S., Tanwar, G. S., & Sharma, A. (2010). Improve performance of extract, transform and load (ETL) in data warehouse. *International Journal on Computer Science and Engineering, 2*(3), 786–789.
Hamoud, A. K., Abd Ulkareem, M., Hussain, H. N., Mohammed, Z. A., & Salih, G. M. (2020). Improve HR decision-making based on data mart and OLAP. In *, 1530. Journal of Physics: Conference Series*, Article 012058. IOP Publishing.
Jafari, H., Malinowski, M. T., & Ebadi, M. J. (2021). Fuzzy stochastic differential equations driven by fractional Brownian motion. *Advances in Difference Equations, 2021*(1), 1–17.
Jaleel, R. A., & Abbas, T. M. (2020). Design and Implementation of Efficient Decision Support System Using Data Mart Architecture. In. In *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)* (pp. 1–6). ppIEEE.
Liu, H., Huang, S., Wang, P., & Li, Z. (2021). A review of data mining methods in financial markets. *Data Science in Finance and Economics, 1*(4), 362–392.
Malekshah, S., Alhelou, H. H., & Siano, P. (2021). An optimal probabilistic spinning reserve quantification scheme considering frequency dynamic response in smart power environment. *International Transactions on Electrical Energy Systems, 31*(11), e13052.
Malekshah, S., Rasouli, A., Malekshah, Y., Ramezani, A., & Malekshah, A. (2022). Reliability-driven distribution power network dynamic reconfiguration in presence of distributed generation by the deep reinforcement learning method. *Alexandria Engineering Journal, 61*(8), 6541–6556.
Malekshah, S., Banihashemi, F., Daryabad, H., Yavarishad, N., & Cuzner, R. (2022). A zonal optimization solution to reliability security constraint unit commitment with wind uncertainty. *Computers and Electrical Engineering, 99*, Article 107750.
Mehregan, E., Asadollahi, M. (2022). Innovative Sourcing and Supply Chain. Rose Publication PTY LTD, v01, Melbourne, Australia.
MJ Al Taleb, T., Hasan, S., & Yousif Mahdi, Y. (2020). Indexed View Technique to Speed-up Data Warehouse Query Processing. *The ISC International Journal of Information Security, 12*(3), 81–85.
Mohammadi, H., Kazemi, R., Maghsoudloo, H., Mehregan, E., & Mashayekhi, A. (2010). System dynamic approach for analyzing cyclic mechanism in land market and their effect on house market fluctuations. In *, 25. Proceedings of the 29th International Conference of the System Dynamics Society, July* (p. 29).
Montebruno, P., Bennett, R. J., Smith, H., & Van Lieshout, C. (2020). Machine learning classification of entrepreneurs in British historical census data. *Information Processing & Management, 57*(3), Article 102210.
Nassirtoussi, AK, Aghabozorgi, S, Wah, TY, et al. (2014). Text mining for market prediction: A systematic review. *Expert Syst Appl, 41*, 7653–7670.
Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental Journal of Computer Science and Technology, 8*(1), 13–19.
Nithya, N., & Kiruthika, R. (2021). Impact of Business Intelligence Adoption on performance of banks: a conceptual framework. *Journal of Ambient Intelligence and Humanized Computing, 12*(2), 3139–3150.
Pandya, R., & Pandya, J. (2015). C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications, 117*(16), 18–21.
Pilarczyk, K. (2016). Importance of management information system in banking sector. *Annales Universitatis Mariae Curie-Skłodowska. Sectio H. Oeconomia, 50*(2), 69–80.
Sadoghi, M., Bhattacherjee, S., Bhattacharjee, B., & Canim, M. (2016). L-store: A real-time OLTP and OLAP system. *arXiv preprint. arXiv:1601.04084*.
Seliya, N., Abdollah Zadeh, A., & Khoshgoftaar, T. M (2021). A literature review on one-class classification and its potential applications in big data. *Journal of Big Data, 8*(1), 1–31.
Shahid, M. B., Sheikh, U., Raza, B., Shah, M. A., Kamran, A., Anjum, A., & Javaid, Q. (2016). Application of data warehouse in real life: State-of-the-art survey from user preferences' perspective. *International Journal of Advanced Computer Science and Applications, 7*(4), 415–426.
Soni, S (2011). Applications of anns in stock market prediction: A survey. *Int J Comput Sci Eng Technol, 2*, 71–83.
Wang, Y., Jia, Y., Tian, Y., & Xiao, J. (2022). Deep reinforcement learning with the confusion-matrix-based dynamic reward function for customer credit scoring. *Expert Systems with Applications, 200*, Article 117013.
Warners, H. L. H. S., & Randriatoamanana, R. (2016). Datawarehouser: A Data Warehouse artist who have ability to understand data warehouse schema pictures. In *2016 IEEE Region 10 Conference (TENCON)* (pp. 2205–2208). IEEE.
Xing, FZ, Cambria, E, & Welsch, RE (2017). Natural language based financial forecasting: a survey. *Artif Intell Rev, 50*, 49–73.
Xu, F., Pan, Z., & Xia, R. (2020). E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. *Information Processing & Management, 57*(5), Article 102221.
Zanaty, E. A. (2012). Support vector machines (SVMs) versus multilayer perception (MLP) in data classification. *Egyptian Informatics Journal, 13*(3), 177–183.
Zhao, T. H., Khan, M. I., & Chu, Y. M. (2021). Artificial neural networking (ANN) analysis for heat and entropy generation in flow of non-Newtonian fluid between two rotating disks. *Mathematical Methods in the Applied Sciences*.

Expert Systems With Applications 237 (2024) 121555

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Research on learning behavior patterns from the perspective of educational data mining: Evaluation, prediction and visualization

Guiyun Feng, Muwei Fan *

School of Management, Guizhou University, Guiyang, 550025, China

ARTICLE INFO

Keywords:
Educational data mining
Learning behavior patterns
Evaluation methodologies
Classification algorithms

ABSTRACT

The rapid growth of educational data creates the requirement to mine useful information from learning behavior patterns. The development of data mining technology makes educational data mining possible. The paper intends to use a public educational data set to study learning behavior patterns from the perspective of educational data mining, so as to promote the innovation of educational management. Firstly, in order to reduce the dimension of data analysis that facilitates the improvement in efficiency, principal component analysis is carried out to reduce the number of attributes in the data set. The significant attributes in the rotating principal component matrix rather than principal components which are not closely related to learning behavior patterns are extracted as the research variables. Then, a pseudo statistic is proposed to determine the number of clusters and the preprocessed data set is clustered according to the extracted attributes. The clustering results are applied to add class labels to the data, which is convenient for the later data training. Finally, six classification algorithms J48, K-Nearest Neighbor, Bayes Net, Random Forest, Support Vector Machine and Logit Boost are used to train the data with labels and build prediction models. At the same time, the performance and applicable conditions of six classifiers in terms of accuracy, efficiency, error, and so on are discussed and compared. It is found that the performance of the integrated algorithm is better than that of a single classifier. In the integrated algorithm, compared with Random Forest, the running time of Logit Boost is shorter.

1. Introduction

Data mining is capable of finding the useful information hidden behind simple data and processing large data sets. Educational data mining (EDM) is the application of data mining in the field of education, which was proposed by Romero and Ventura (2007) . EDM bridges the gap between two disciplines: education on the one hand, computing sciences on the other, where both data mining and machine learning as subfields of computing sciences are the focus (Bakhshinategh, Zaiane, ElAtia, & Ipperciel, 2018). The goal of EDM is to better understand how students learn and identify the settings in which they learn to improve educational outcomes and to gain insights into and explain educational phenomena (Romero & Ventura, 2013). It is concerned with developing methods for exploring the unique types of data that come from educational environments (Bakhshinategh et al., 2018). In short, EDM aims to utilize data mining technology to find the information hidden behind massive educational data, and use relevant information to promote the progress and development of education.

There are many reasons to study students' learning behavior patterns. For example,

- if students are ranked according to the traditional ranking system, it is very likely that many students are concentrated in one rank, resulting in the phenomenon that there is no one in other ranks. This is not an objective method of evaluating student performance. The reason for this phenomenon is that a certain course is difficult. Many students' scores are concentrated between 60 and 80, and no one gets 90, so no one gets "excellent". Clustering, as an important method in EDM, can solve this problem well by learning the law from real data and ranking them objectively.
- if we can find students with academic warning risk at the initial stage of the course, we could improve their performance opportunely (Riestra-Gonzalez, Paule-Ruiz, & Ortin, 2021). If we want to find students at risk and help them in time, we can use historical data to predict students' future performance. The classification algorithm in data mining just has such a function which can not only predict future performance, but also provide prediction metrics such as accuracy.
- a large amount of data related to students' behavior is useless in front of anyone without being processed. By using data mining to process data and present the results to stakeholders in a visual

* Corresponding author.
E-mail addresses: 1051184826@qq.com (G. Feng), mwfan@gzu.edu.cn (M. Fan).

https://doi.org/10.1016/j.eswa.2023.121555
Received 30 March 2023; Received in revised form 28 August 2023; Accepted 9 September 2023
Available online 15 September 2023

form, the data can be transformed into useful information which can be made full use of to improve teaching and learning.

Thus, it is very appropriate to apply EDM to study learning behavior patterns in terms of technology and research content which corresponds to the title of the paper. By using the methods which are closely related to EDM to study the learning behavior patterns of students, we can analyze, evaluate and predict the learning behavior. Therefore, this paper intends to study the learning behavior patterns that can best reflect the connotation of educational data from the perspective of educational data mining so as to provide reference for stakeholders. What is more, we can also innovate educational management by using visual research results. With the purpose of achieving research objectives, three research questions of this paper are proposed.

1. The rapid growth of educational data makes traditional processing methods unsuitable for studying learning behavior patterns, so data mining comes into play. Data mining technology has been applied in the field of education, and formed an important research field – education data mining. Therefore, how to study learning behavior patterns from the perspective of educational data mining is an urgent problem to be solved.
2. Learning behavior pattern is a relatively abstract phrase. To implement it into specific data mining technology, learning behavior pattern must be concrete. Thus, what aspects of learning behavior patterns can be studied?
3. The ultimate purpose of our research is to promote the development of education and innovate the management of education. How to promote the development of education and innovate the management of education by studying the learning behavior pattern from the perspective of educational data mining?

The remaining sections of this paper are organized as follows. Section 2 introduces the related work of the research topic in this paper, including analysis, evaluation and prediction of learning behavior. Section 3 analyzes the research methods used in this paper and describes the research results briefly. Section 4 answers the research questions according to the process and results of the research. The last section concludes the full text and makes an outlook to the research directions in the future.

2. Literature review

2.1. Analysis and evaluation of learning behavior

Many researchers use or improve the existing data mining technology to analyze and evaluate learning behavior. Manoharan, Ganesh, Felciah, and Banu (2014) introduce a deterministic model based on the clustering algorithm to analyze and monitor students' performance. Busalim, Masrom, and Wan (2019) study and analyze the effects of social software addiction and self-esteem on academic performance. Crivei, Czibula, Ciubotariu, and Dindelegan (2020) explore the usefulness of unsupervised machine learning methods, principal component analysis and association rule mining in analyzing students' academic achievement data in order to develop a supervised learning model for students' achievement prediction. Delgado, Morán, José, and Burgos (2021) adopt a new unsupervised clustering technology based on self-organizing mapping (SOM) artificial neural network (ANN) model to analyze online learning records. Mai, Bezbradica, and Crane (2022) propose a new method to deal with the problems of noise and trend effect in data to analyze students' learning behavior, and achieve success in detecting students with similar learning behavior and results.

Dynamic evaluation method has been proved to be a tool to find students' learning potential. In a learning environment where learning is currently mediated through technology, dynamic evaluation methods have a significant impact on students' academic performance (Zhang, Lai, Cheng, & Chen, 2017). Varela, Montero, Vásquez, Guiliany, Mercado, et al. (2019) use clustering technology as a useful management strategy tool to divide the population into homogeneous groups according to students' characteristics and skills to evaluate learning behavior. The evaluation of students' academic performance can be regarded as a clustering problem, and the hybrid clustering method is applied to evaluate academic performance in the educational environment (Yadav, 2020). Karthikeyan, Thangaraj, and Karthik (2020) propose a hybrid educational data mining (HEDM) model to analyze students' academic performance, and combine Naive Bayes and J48 classifiers to classify students' performance. Kumar, Balamurugan, and Sasikala (2021) establish a multi-tier student performance evaluation model (MTSPEM) using a single classifier and an integrated classifier to evaluate student records, and conduct a comparative evaluation to prove the effectiveness of the proposed model. The analysis and evaluation of students' performance will not only help colleges and universities improve the quality of education, but also help enhance the overall performance and identify students at risk, which aim to optimize the management of educational resources (Mallik, Roy, Maheshwari, Pandey, & Rautray, 2019).

2.2. Prediction of learning behavior and dropout rate

In order to improve the accuracy of learning behavior prediction, some researchers utilize a variety of algorithms or models for comparison. Huang and Ning (2013) construct four types of mathematical models to predict students' academic performance, including multiple linear regression, multi-layer perceptron network, radial basis function network and support vector machine. Agrawal, Nigam, and Sahu (2018) apply two clustering algorithms to predict students' academic execution. In order to provide reliable admission standards for colleges and universities, Mengash (2020) uses ANN, decision tree, support vector machine and Naive Bayes to predict students' behavior. The results demonstrate that the accuracy of ANN is the highest. Turabieh et al. (2021) propose an improved Harris Hawks optimization algorithm to search the most valuable features in the student achievement prediction problem, and evaluate the whole prediction system using k-nearest neighbor, multilayer recurrent neural network, Naive Bayes and ANN. The results indicate that the combination accuracy of the optimization algorithm and multilayer recurrent neural network is the highest. Lee and Recker (2022) examine how student and instructor participation in online discussions impacts students' course performance. Multilevel modeling results show that online listening behaviors significantly predict students' course performance.

Some researchers also link other aspects related to learning to predict learning behavior. Przepiorka, Blachnio, Cudo, and Kot (2021) analyze the relationship between social anxiety and social skills by studying the use of smart phones to predict physical symptoms and academic performance. Zaffar, Hashmani, Habib, Quraishi, Irfan, et al. (2022) design a hybrid feature selection framework to identify important features and relevant features to predict students' performance. In order to reveal the problem of the internal relationship between questions and skills, Gao, Zhao, Li, Zhao, and Zeng (2022) propose a deep cognitive diagnosis framework, which enhances the traditional cognitive diagnosis methods to predict learning behavior through deep learning.

Some scholars have studied the prediction of dropout rate alone. Heredia, Amaya, and Barrientos (2015) apply C4.5 and ID3 to predict the possibility of dropping out, and compare the results of the two algorithms. In order to solve the problem that it is difficult to ensure the accuracy of manually extracted features, Lin, Liu, and Yi (2018) propose an integrated framework with feature selection to predict the dropout rate in massive online open courses (MOOCs), including feature generation, feature selection and dropout rate prediction. Two aspects of students' performance have been concerned, which are predicting students' academic performance and combining typical progress

with prediction results. By focusing on a few courses with particularly good or poor performance, teachers can provide timely warning and support for students with poor performance, and provide advice and opportunities for students with good performance (Asif, Merceron, Ali, & Haider, 2017)

Most of the existing studies focus on the analysis and prediction of students' academic performance, and researchers usually predict learning behavior from the aspects which are relevant to academic performance. The purpose is to monitor students' academic performance in advance and find the space for students' progress, aiming to provide decision-making reference for the development of colleges and universities and to provide reference for innovating educational management. Existing studies either utilize clustering to analyze and evaluate academic performance, or apply a variety of classification algorithms to predict academic performance. On the basis of previous studies, this paper intends to comprehensively employ supervised and unsupervised learning technology to evaluate and predict learning behavior patterns. And different from the methods of extracting attributes previously, we select the significant attributes in the rotated component matrix as research variables by using principal component analysis in order to reduce dimension. We also compare the prediction accuracy, error and operation efficiency of different algorithms. At the same time, we use the results of data visualization to discuss and compare the application conditions of each classification algorithm and prediction model in the process of research.

3. Material and methods

3.1. Data

3.1.1. Data source

This study chooses a public educational data set from UCI (https://archive.ics.uci.edu/ml/datasets/Higher+Education+Students+Performance+Evaluation+Dataset). The data set is suitable for both the size of the data and the matching degree of the research content, and a public data set can be used for the readers to reproduce the experiment. There are 145 instances and 33 attributes in the data set. Except for the student number and course number, the remaining 31 attributes are divided into three parts: personal information, family information and information related to learning. Personal information has 10 attributes, family information has 6 attributes, and information related to learning has 15 attributes.

3.1.2. Data standardization

The student attribute matrix (Fan & Frederick, 2018) is used to quantify the student attribute for further analysis. There are no missing values in the data set, and the attribute characteristic is integer. Different attributes have different scales, so the data are preprocessed by dimensionless standardization.

3.2. Experiment and results

3.2.1. Attribute extraction

Feature selection (FS), also known as attribute extraction, is the most commonly used method in dimension reduction. It selects the proper subset of the feature from all the features of the original data in order to reduce the redundant information, noise, and the dimension of data analysis. It applies the proper subset of the feature to train the model constructed by learning algorithm, which aims to improve the learning performance of the algorithm. Attribute selection measures include information gain (e.g., ID3, C4.5), Gini index (e.g., SLIQ, SPRINT), and G-statistics. Considering the nonlinearity of data and the number of attributes and instances, the attributes are extracted with the help of principal component analysis (PCA). Before feature selection using PCA, we need to perform KMO and Bartlett's test in order to judge the data set whether to be suitable to use PCA. The result of the test is demonstrated in Table 1. The KMO and significance in the Bartlett test of sphericity (KMO > 0.5 and $P < 0.05$) indicate that the variables are highly correlated, which is enough to provide a reasonable theoretical basis for PCA.

Table 1
KMO and Bartlett's test.

Kaiser–Meyer–Olkin measure of sampling adequacy		0.568
Bartlett's test of sphericity	Approx. Chi-Square	1002.700
	df	465
	Sig.	0.000

According to the previous study, it is generally believed that the principal components with eigenvalues greater than 1 or cumulative contribution rate greater than 85% are representative. In the paper, we calculate both cumulative contribution rate and eigenvalue. We find that if we choose the principal components in which a cumulative contribution rate is greater than 85%, the final result almost covers all the principal components. In this situation, we cannot realize the goal of dimension reduction. Therefore, we select the principal components with eigenvalues greater than 1. Moreover, corresponding cumulative contribution rate reaches 63.660% that is acceptable to some extent. Table 2 shows 11 principal components with eigenvalues greater than 1 and cumulative contribution rate obtained by PCA.

The principal components are not the remaining variables after the original variables have been filtered, but the "comprehensive variables" after recombining the original variables. In the previous step, we reduce dimension by selecting the representative principal components calculated by PCA. These components are comprehensive variables which cannot replace original attributes that are direct illustration of learning behavior and will be used for the analysis of learning behavior pattern. Therefore, we make full use of rotated component matrix which is calculated from the previous step. We select the attributes with large positive load or small negative load in the rotation component matrix (Table 3) as the variables for subsequent research, which is functioned as extracting the attributes. The specific implementation condition is that if the absolute value of positive load or negative load in the row of each attribute is greater than 0.7, the attribute is extracted as the research variable. Table 3 shows the rotation component matrix composed of 11 principal components and 31 attributes. According to the definition of the matrix, the greater the absolute value of positive or negative load, the greater the correlation between the attributes of the row where the load resides and the principal component. In order to accurately study the learning behavior pattern and dimension reduction, we need to extract the highly correlated attributes. And considering the dimension of data analysis, the absolute value of the load is chosen as 0.7. Therefore, Partner, Study_hours, Read_frequency1, Read_frequency2, Preparation1, Preparation2, Take_notes, Listen, GPA and Expected_GPA are extracted as research variables. From the results of attribute extraction, nine attributes are attributes relevant to learning, and only one attribute is personal information. The use of personal statistics has no significant impact on the prediction accuracy (Tomasevic, Gvozdenovic, & Vranes, 2020), so the research variables extracted in this study are reasonable to certain extent.

3.2.2. Using a pseudo F statistic to determine the number of clusters

Clustering belongs to unsupervised learning. Clustering is not to categorize the data according to the existing rules, but to categorize the data by learning the law in the data. The result of clustering is closer to reality, which makes this method widely applied in various areas. In terms of the problem we solve, fast clustering, as a typical prototype clustering algorithm, is simple and fast, and has fast convergence of the objective function. These make the algorithm highly efficient in processing data sets. One of the disadvantages of fast clustering, also known as classic K-means clustering, is that the determination of cluster number is manually subjective. Thus, we add a step where a pseudo *F*

Table 2
Eigenvalues and contribution rate of principal components.

Principal components	Extract sums of squares loading		
	Eigenvalues	Variance contribution rate/%	Cumulative contribution rate/%
1	3.083	9.945	9.945
2	2.559	8.254	18.199
3	2.337	7.538	25.737
4	2.090	6.742	32.479
5	1.854	5.982	38.461
6	1.569	5.060	43.521
7	1.398	4.510	48.031
8	1.328	4.282	52.314
9	1.256	4.052	56.366
10	1.182	3.812	60.178
11	1.079	3.482	63.660

Table 3
Rotated component matrix.

	1	2	3	4	5	6	7	8	9	10	11
Age	−0.131	0.268	0.548	0.041	−0.020	−0.007	−0.217	0.328	−0.267	−0.299	0.192
Sex	−0.026	0.279	0.553	−0.048	−0.103	−0.002	0.079	0.114	0.307	0.420	−0.104
High_school	−0.076	0.043	0.189	0.004	−0.111	−0.078	0.004	0.690	0.209	−0.182	0.066
Scholarship	−0.022	0.371	−0.691	0.084	−0.125	−0.002	0.092	−0.065	0.085	0.012	0.093
Work	0.068	0.026	0.000	−0.179	−0.357	−0.206	0.019	−0.247	0.159	0.558	0.136
Activity	−0.256	−0.006	0.239	−0.548	−0.094	−0.113	0.061	0.200	−0.185	0.295	−0.113
Partner	−0.139	−0.038	−0.170	0.107	0.018	−0.277	−0.070	−0.047	−0.094	0.102	0.749
Salary	0.031	0.028	0.151	−0.086	0.007	0.501	−0.201	−0.147	−0.442	−0.025	0.043
Transportation	0.626	0.095	−0.018	−0.204	0.138	−0.311	−0.154	0.176	−0.084	0.026	0.126
Accommodate	0.637	−0.063	−0.112	−0.121	0.117	−0.018	0.071	−0.090	0.125	−0.407	−0.152
M_education	0.575	−0.068	0.195	0.156	−0.117	0.275	0.000	−0.255	0.039	0.139	0.120
F_education	0.447	−0.282	0.406	0.083	0.011	0.021	0.102	−0.342	0.114	0.126	0.231
Number_sb	−0.677	0.085	0.115	−0.063	0.202	−0.059	0.072	−0.116	0.024	−0.163	−0.061
Parent_status	−0.071	0.265	0.085	0.218	0.406	0.189	−0.466	0.002	−0.080	0.041	−0.183
M_occupation	0.676	0.050	−0.118	−0.043	−0.044	0.179	0.101	−0.009	−0.045	−0.005	−0.215
F_occupation	−0.084	−0.028	−0.086	0.162	0.052	−0.232	−0.146	−0.205	−0.025	0.079	−0.684
Study_hours	−0.068	−0.088	−0.073	0.242	0.716	−0.301	0.105	−0.143	0.031	0.096	−0.060
R_frequency1	−0.064	−0.014	0.123	0.769	0.111	−0.124	−0.005	−0.042	0.091	0.041	−0.007
R_frequency2	−0.072	0.070	−0.063	0.724	0.047	0.082	0.076	0.229	−0.117	0.029	−0.071
Attendance	0.152	−0.146	−0.552	0.024	0.189	−0.018	0.308	0.215	−0.212	0.008	0.032
Impact	0.096	−0.141	−0.178	0.113	−0.002	0.213	0.037	0.526	−0.170	0.102	0.112
Attend_class	−0.088	−0.129	0.108	0.103	−0.261	−0.080	−0.030	−0.112	−0.601	0.106	0.060
Preparation1	0.158	−0.102	−0.082	0.023	0.077	0.761	−0.048	0.119	0.147	−0.025	−0.065
Preparation2	−0.084	0.199	0.004	0.002	0.722	0.400	0.063	−0.026	0.103	0.005	0.090
Take_notes	0.093	0.082	−0.094	0.071	0.135	−0.146	0.760	0.140	0.118	0.037	−0.008
Listen	−0.073	0.018	0.031	−0.048	−0.200	−0.043	−0.002	0.035	0.109	−0.749	0.027
Discussion	−0.144	0.299	−0.060	−0.023	0.027	0.180	0.550	−0.310	0.037	0.006	0.052
Flip_class	−0.178	0.177	0.507	0.140	−0.020	−0.029	0.455	−0.021	−0.151	−0.023	−0.033
GPA	−0.048	0.824	0.073	−0.096	0.031	−0.034	0.095	−0.074	0.178	−0.033	0.041
Expect_GPA	0.013	0.860	−0.031	0.140	0.090	−0.058	0.077	0.037	0.101	0.054	−0.059
Grade	−0.086	0.244	0.178	0.167	−0.153	0.061	0.035	−0.150	0.680	0.041	0.022

statistic is proposed to determine the cluster number objectively in the algorithm, and the calculation equation is shown in (1).

$$F = \frac{\left(T - P_k\right) / (k-1)}{P_k / (n-k)} \tag{1}$$

T represents the sum of squares of total deviations. P_k is the sum of squares of intra class deviations when data are clustered into k classes. n represents the size of the sample. The pseudo F statistic is used to evaluate the effect of clustering into k classes. If clustering effect is good, the sum of squares of deviations between classes is larger than the sum of squares of deviations within classes, so the clustering level with large pseudo F statistics and small number of clusters should be taken.

Taking 10 research variables as inputs for fast clustering, the output intra-class distance and the number of known instances can be utilized to calculate the pseudo statistic corresponding to each cluster number. The number of clusters k starts from 2 as input. The general principle is $k_{max} \leq \sqrt{n}$ (Ramze Rezaee, Lelieveldt, & Reiber, 1988), and we take $k_{max} = \sqrt{n}$. n is equal to 145, so, the maximum value k is 12. However, there are only 10 research variables, and it is generally considered that the number of clusters does not exceed the number of variables. Therefore, the range of cluster number k in this study is $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. We calculate the pseudo F statistics for the number of clusters within this range, and the results are shown in Fig. 1. Based on the principle that the clustering level with large pseudo F statistics and small number of clusters should be taken, it can be found from Fig. 1 that when the number of clusters is 5, the number of clusters is small and the pseudo F statistic is large. Therefore, it is more reasonable for the data to be clustered into 5 classes. At this time, the number of iterations is 7 and the sum of squares of intra-class deviations is 84.76. 145 instances are clustered into five classes. The detailed clustering results show that there are 56 instances in the first class, 26 instances in the second class, 10 instances in the third class, 14 instances in the 4th class and 39 instances in the 5th class.

3.2.3. *Classifiers*

Classification belongs to supervised learning. Supervised learning trains labeled data, realizes mapping from input to output, and then applies this mapping relationship to unknown data to achieve classification and prediction. The biggest difference between supervised learning and unsupervised learning is whether the data is labeled or not. Therefore, if the supervised learning algorithm is to be trained

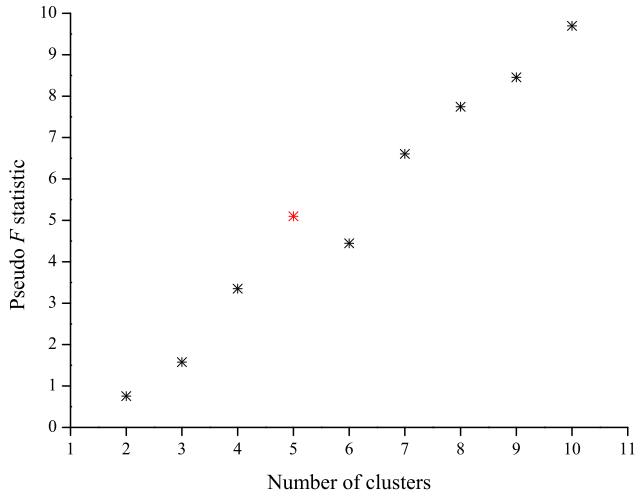


Fig. 1. Pseudo F statistics corresponding to different cluster numbers.

the model, the original data must be labeled or labeled in some way. Supervised learning falls into two main categories: Classification and regression. For discrete data, classification is suitable; For continuous data, regression is suitable. The data in this paper is discrete, so only the classification algorithm is applied. Due to the need for a labeled data set to interpret the analyzed data samples, supervised machine learning technology is considered to be more suitable for prediction tasks in educational data mining than unsupervised machine learning technology (Tomasevic et al., 2020).

Classifier, as the name suggests, is the implementation of classification tasks. There is a simple example to illustrate how to achieve prediction task by applying a classifier. Students who pass are classified as the first category, while students who fail are classified as the second category. In order to focus on those who fail, it is necessary to classify students based on historical performance data. The classification criteria of the classifier are that a student whose score is greater than or equal to 60 (in a percentage system) is the first category, and a student whose score is less than 60 is the second category. The result of classification is a prediction of students' future academic performance. If one is divided into the second category, the student needs to be paid more attention to. At this time, the classifier plays a role of early warning.

There is no class label on the metadata. The previous clustering results are applied to add labels to the preprocessed data (Feng, Fan and Chen, 2022), so that the data can be trained with the classification algorithm. The classification models are constructed based on six algorithms: J48, K-nearest neighbor (KNN), Bayes Net, Random Forest, Support vector machine (SVM) and Logit Boost. Among them, Random Forest and Logit Boost are integrated algorithms. WEKA integrates a large number of machine learning algorithms that can undertake data mining tasks. The experiments are based on WEKA, and the parameters of the six classification algorithms are default. Several classification algorithms will be briefly introduced below.

- **J48**: J48 is an improved algorithm of ID3. Improvements are made in the following four aspects: (1) The information gain rate is used to select attributes. (2) Pruning is performed during tree construction. (3) The discrete processing of continuous attributes can be completed. (4) The incomplete data can be processed. When constructing a tree, the algorithm requires multiple sequential scanning and sorting of the data set, resulting in low efficiency. However, due to its easy to understand classification rules, the accuracy is high.
- **KNN**: Cover and Hart (1953) proposed a KNN classification algorithm based on distance. KNN algorithm, also described as reference sample plot method, determines the distance between the samples to be classified and each training sample, and then chooses the k samples which are closest to the samples to be classified as the k nearest neighbors of the samples to be classified. If most of individuals among k similar samples in the feature space belong to one class, the sample also belongs to this class (Goldberger, Roweis, Hinton, & Salakhutdinov, 2005). As a typical non-parametric method, calculations can be performed by simply searching for similar units. Even if the system is linearly indivisible, this method can still be applied.
- **Bayes Net**: When the nodes, the states of nodes and the connections between nodes increase, the calculation of simple Bayes Net is very complex and the calculation of probability propagation becomes very heavy, which limit the application of Bayes Net in practice. Until (Pearl, 1986) proposed the message passing algorithm (polytree algorithm), and Lauritzen and Spiegelhalter (1988) further proposed the junction tree algorithm using the concept of message passing, it provided an effective algorithm for the probability propagation of Bayes Net and laid the foundation for practical application. The drawbacks of this classification method are how to effectively calculate probability when the model becomes complex, and how to handle continuous variables, which are research directions in recent years.
- **Random Forest**: Random Forest integrates the characteristics of Bagging series algorithms and random selection, and introduces random attribute selection for prediction in the training process of decision tree, which was proposed by Breiman (2001) . The algorithm adopts the put-back sampling strategy to extract samples from the original data set, and uses the non-put-back sampling strategy to extract different features as input variables. It constructs a decision tree on each new data set, and synthesizes the prediction results of multiple decision trees as the prediction results of the whole Random Forest. Random Forest has a good tolerance for outlier and noise, and is not easy to overfit. It is widely used in medicine, bio-informatics, management and other fields.
- **SVM**: Cortes and Vapnik (1995) formally proposed support vector machine. SVM is a supervised binary classifier based on VC (Vapnik–Chervonenkis) dimension theory of statistics and the principle of structural risk minimization. SVM can automatically find those support vectors that have better discrimination ability for classification by training. The classifier constructed from above theory can maximize the interval between classes, so it has better adaptability and higher discrimination ability. SVM has a deep theoretical foundation, which can ensure that the extremal solution is the global optimal solution rather than the local optimal solution. Thus, SVM has good generalization ability for unknown samples.
- **Logit Boost**: Boosting was first proposed by Schapire (1989) . Because it required priori knowledge of the performance of weak learnability, its application was limited. Freund and Schapire proposed an improved method: adaptive boosting, called AdaBoost, which did not require prior knowledge of weak learnability (Freund & Schapire, 1997). Friedman, Hastie, and Tibshirani (2000) lately improved AdaBoost: logit adaptive boosting, called Logit Boost. AdaBoost algorithm adopts exponential loss function, while Logit Boost adopts negative log likelihood loss function. Logit Boost constructs a basic weak classifier on the existing sample set, repeatedly calls the weak classifiers, and gives more weights to the samples with wrong classification each time. These weak classifiers are weighted and synthesized to obtain a strong classifier. The superposition model is fitted through maximum likelihood estimation in order to obtain a classification model with higher precision in a strong classifier. The main idea is that

Table 4
The confusion matrix.

Actual category	Prediction category	
	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

the strong classifier is made up of an army of weak classifiers (Demirer, Pierdzioch, & Zhang, 2017; Goessling, 2017; Kanamori & Takenouchi, 2013).

3.2.4. Evaluation metrics of classifying

In order to effectively evaluate the classification results, Kappa coefficient, accuracy, precision, recall, F-Measure, Youden index, receiver operator characteristic (ROC) curve and precision–recall (P–R) curve are used to evaluate the results from the performance of the classifier.

10-fold cross validation is applied in the training process with the purpose of avoiding the influence of over-fitting on the results and check the classification effect. The advantage of 10-fold cross validation is that it allows all data to participate in the training and testing process, fully reflecting the characteristics of "crossover". The basic idea of 10-fold cross validation is to randomly divide the data into 10 parts and conduct 10 experiments. Each time, 9 parts are used as training data, and the remaining part is used as test data. The experiment will repeat 10 times until all the data is used as the training set for validation. The average value of each metric in the 10 experiments is used as the final result. After training data, the classification results are compared with the class label. The evaluation metrics of the six classifiers are calculated, and corresponding metrics are compared. The confusion matrix based on the binary classification problem is applied to calculate each metric. The confusion matrix is shown in Table 4. True positive (TP) represents the number of positive samples that are judged as positive samples. True negative (TN) means the number of negative samples that are judged as negative samples. False positive (FP) denotes the number of negative samples that are judged to be positive samples. False negative (FN) indicates the number of positive samples that are judged to be negative samples.

In classification problems, the most common evaluation metric is accuracy, which can directly reflect the correct proportion. However, in practice, the sample size of each category is often unbalanced. Without adjustments on such unbalanced data sets, the model is prone to bias toward larger categories and abandon smaller ones. At this point, a metric that penalizes the model's "bias" is necessary. Kappa coefficient is utilized to evaluate the difference between the classification results of the classifier and the random classification. The value range is [−1,1]. Kappa value is positively correlated with the accuracy of the classifier. The closer the value is to 1, the more accurate the algorithm is. The calculation equation of Kappa coefficient based on confusion matrix is shown in (2). p is calculated as shown in Eq. (3) which denotes that the numerator is the sum of the diagonal elements, and the denominator is the sum of all the elements (in fact, p is the accuracy). q is calculated as shown in Eq. (4) which means that the numerator represents the sum of elements in column r multiplied by the sum of elements in row r and then all products are summed up, and the denominator represents the square of the sum of all elements, which is the sum of the "product of the actual quantity and the predicted quantity" corresponding to all the categories respectively, and divided by the "square of the total number of samples". According to the calculation of Kappa, the more unbalanced the confusion matrix, the higher the q, the lower the Kappa, which can achieve the goal of punishing the model with strong "bias".

$$Kappa = \frac{p-q}{1-q} \tag{2}$$

$$p = \frac{\sum_{i=j} f_{ij}}{\sum_i \sum_j f_{ij}} = \frac{TP+TN}{TP+FN+FP+TN} \tag{3}$$

$$q = \frac{\sum_r \left(\sum_i f_{ir} \cdot \sum_j f_{rj}\right)}{\left(\sum_i \sum_j f_{ij}\right)^2} = \frac{(TP+FN)(TP+FP)+(FP+TN)(FN+TN)}{(TP+FN+FP+TN)^2} \tag{4}$$

The accuracy, precision, recall (also called sensitivity), and F-Measure range from 0 to 1, and the calculation equations are shown in (5), (6), (7) and (8), respectively. With regard to F-Measure in the paper, α is taken as 1. Accuracy is a direct evaluation metric, which means the ratio of correctly classified samples to all samples. This is an easy to understand metric, but there is an obvious problem in case of imbalanced samples. For example, when people predict whether an earthquake will occur in a certain area on a certain day, the occurrence of an earthquake is classified as 0 and the non occurrence is classified as 1. If a classifier classifies all test cases into 1, the accuracy is high. But once an earthquake occurs, the loss caused by the classification results will be difficult to estimate. This indicates that accuracy is not a comprehensive and scientific metric. Compared with the calculation equations of precision and recall, it can be seen that in some situations these two metrics contradict each other. The F-Measure, as a combination of these two metrics, is the harmonic average of precision and recall when α is taken as 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} = Sensitivity \tag{7}$$

$$F-Measure = \frac{(\alpha^2+1) \times Recall \times Precision}{\alpha^2(Recall+Precision)} \tag{8}$$

True positive rate (TPR) is the ratio of positive samples determined as positive samples by the classifier, also known as sensitivity. In effect, it is the recall of the 'positive category'. True negative rate (TNR) is the ratio of negative samples determined as negative samples by the classifier, also known as specificity. Actually, it is the recall of the 'negative category'. False negative rate (FNR) is the ratio of positive samples determined as negative samples by the classifier. Some researchers also call it false reject rate (FNR). False positive rate (FPR) is the ratio of negative samples determined as positive samples by the classifier. Some researchers also call it false acceptance rate (FAR). It was originally negative, but it is recognized as positive. The calculation equations of TPR, TNR, FNR and FPR are shown in (9), (10), (11) and (12), respectively.

$$TPR = \frac{TP}{TP+FN} = Sensitivity \tag{9}$$

$$TNR = \frac{TN}{FP+TN} = Specificity \tag{10}$$

$$FNR = \frac{FN}{TP+FN} = 1 - Sensitivity \tag{11}$$

$$FPR = \frac{FP}{FP+TN} = 1 - Specificity \tag{12}$$

TPR and FPR which are two indicators of ROC curve do not depend on the specific distribution of the class. The curve takes the positive samples and negative samples into account, and will not change significantly with the change of the proportion of positive and negative samples. It is an effective tool for more balanced research on classification problems. TPR and FPR can be represented by sensitivity and (1-specificity), so two coordinate axes of the curve can also be composed of sensitivity and (1-specificity). The closer the ROC curve is to the upper left, the better the performance of the classifier is. When two ROC curves intersect, it is impossible for us to directly see which has better performance. At this time, it is necessary to compare the

Table 5
The specific values of seven metrics.

	Kappa	Accuracy	Precision	Recall	F-Measure	AUC	Youden index
J48	0.7291	0.800	0.813	0.800	0.804	0.910	0.737
KNN	0.7723	0.834	0.842	0.834	0.834	0.887	0.775
Bayes Net	0.8475	0.890	0.889	0.890	0.887	0.983	0.850
Random Forest	0.8566	0.897	0.898	0.897	0.894	0.990	0.854
SVM	0.8653	0.903	0.908	0.903	0.895	0.968	0.865
Logit Boost	0.8655	0.903	0.903	0.903	0.901	0.986	0.864

size of area under curve (AUC). The calculation equation of AUC is shown in (13). AUC is the area enclosed by the coordinate axis under the ROC curve. The larger the AUC, the better the performance of the classifier. When AUC = 0.5, the classifier is the same as random prediction, which is similar to coin tossing and the probability of both sides is 50%. When 0 < AUC < 0.5, the use of the classifier is worse than random prediction. In this situation, only reverse prediction is better than random prediction. When AUC>0.5, the classifier is better than random prediction. When AUC = 1, the classifier at this time is a perfect classifier. Generally, such a classifier cannot exist.

$$AUC = \frac{TPR}{FPR} = \frac{Sensitivity}{1 - Specificity} \tag{13}$$

The two indicators related to the Youden index are sensitivity and specificity. The sensitivity and specificity can be represented by TPR and FPR, so the Youden index can also be calculated by TPR and FPR. The range of the index is [−1,1]. The larger the index value, the better the performance of the classifier in predicting positive examples. When the index value is negative, it has no application value. The Youden index can be used in conjunction with the ROC curve. The geometric representation of the Youden index maximizes the vertical distance from the point on the ROC curve to the x-axis to ensure that the TPR is large while the FPR is as small as possible. The equation for calculating the Youden index is shown in (14). Although it is technically possible to obtain a value less than 0 from this equation, a value less than 0 only implies that the positive and negative labels have been switched.

$$\text{Youden index} = Sensitivity + Specificity - 1 = TPR - FPR \tag{14}$$

Kappa, accuracy, precision, recall, F-Measure, AUC and Youden index of the classifiers constructed by the six algorithms are shown in Table 5. From the Kappa value alone, except for J48 and KNN, other classifiers perform well, among which Logit Boost performs best. In order to clearly compare the performance of each classifier in terms of each metric, seven metrics are shown in Fig. 2. The polygonal line reflects the performance of the six classifiers in predicting positive examples, and value corresponding to the point on the polygonal line is the Youden index of each classifier. Histograms reveal six metrics of six classifiers. The larger the seven evaluation metrics in Fig. 2, the better the performance of the classifier. J48 and KNN perform poorly in terms of Kappa, accuracy, precision, recall and F-Measure. In terms of AUC, J48, Bayes Net, Random Forest, SVM and Logit Boost exceed 0.9. Bayes Net, Random Forest, SVM and Logit Boost exceed 0.95, among which Random Forest performs best. The Youden index of the six classifiers is between 0.7 and 0.9, and the Youden index of SVM is the largest. Although the AUC of SVM is smaller than that of Bayes Net, Random Forest and Logit Boost, it has the best performance in predicting positive examples. Therefore, if we pay more attention to the rate of the correction in the problem of predicting academic risk, we should not choose J48 and KNN.

The ROC curves of the classifiers constructed by the six algorithms under five classes are shown in Fig. 3. Compared with (a), (b) and (e), the ROC curves of Bayes Net and Logit Boost are relatively smooth and the change range of each class is small in (c) and (f). The area enclosed by the coordinate axis under the ROC curves of Random Forest and Logit Boost in (d) and (f) are larger, indicating that the performance of

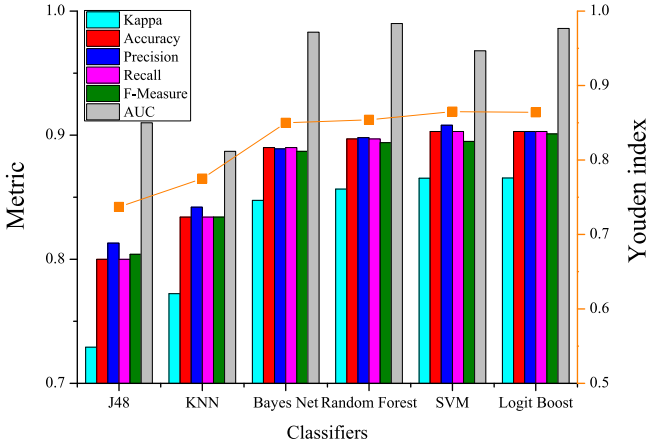


Fig. 2. Seven evaluation metrics of six classifiers.

these two classifiers is better. Therefore, when we consider the positive samples and negative samples at the same time, we had better utilize Random Forest and Logit Boost to construct classifiers to predict results.

When the classification result is unbalanced, due to the strong robustness of the ROC curve, the ROC curve may not change significantly, but it has a great impact on the model that attaches importance to accuracy. Precision and recall, which are two indicators of P–R curve, focus on positive samples. When positive samples are more important in class imbalance, P–R curve is better than ROC curve. If the P–R curve of a classifier is more convex to the upper right, it means that the performance of the classifier is better. If the P–R curve of one classifier is completely covered by the P–R curve of another classifier, the latter has better performance. If there is an intersection between the two curves, the area under the curve and the coordinate axis or the F-Measure can be used for comparison. The larger the area or the larger the F-measure value, the better the performance of the classifier. In addition to comparing the performance through calculation, one of the most important advantages of P–R curve is that the performance of classifier can be seen intuitively. Break-even point (BEP) is the value when precision is equal to recall. The larger the value corresponding to the intersection of P–R curve and the line that precision is equal to recall, the better the performance of the classifier.

The P–R curves formed by the six classifiers in the training process are shown in Fig. 4. The intersection values of the six curves and the line composed of BEP from small to large are J48, KNN, Bayes Net, SVM, Random Forest and Logit Boost respectively. The intersection values of Random Forest and Logit Boost coincide. The classifiers constructed by Random Forest and Logit Boost algorithm have the best performance. Therefore, if positive samples are more important in imbalanced problems, we could apply Random Forest and Logit Boost in the problem of predicting academic risk.

Integrated algorithms are divided into two categories according to whether there are dependencies between base classifiers: Bagging series algorithms without dependencies between base classifiers and Boosting series algorithms with dependencies between base classifiers. Random

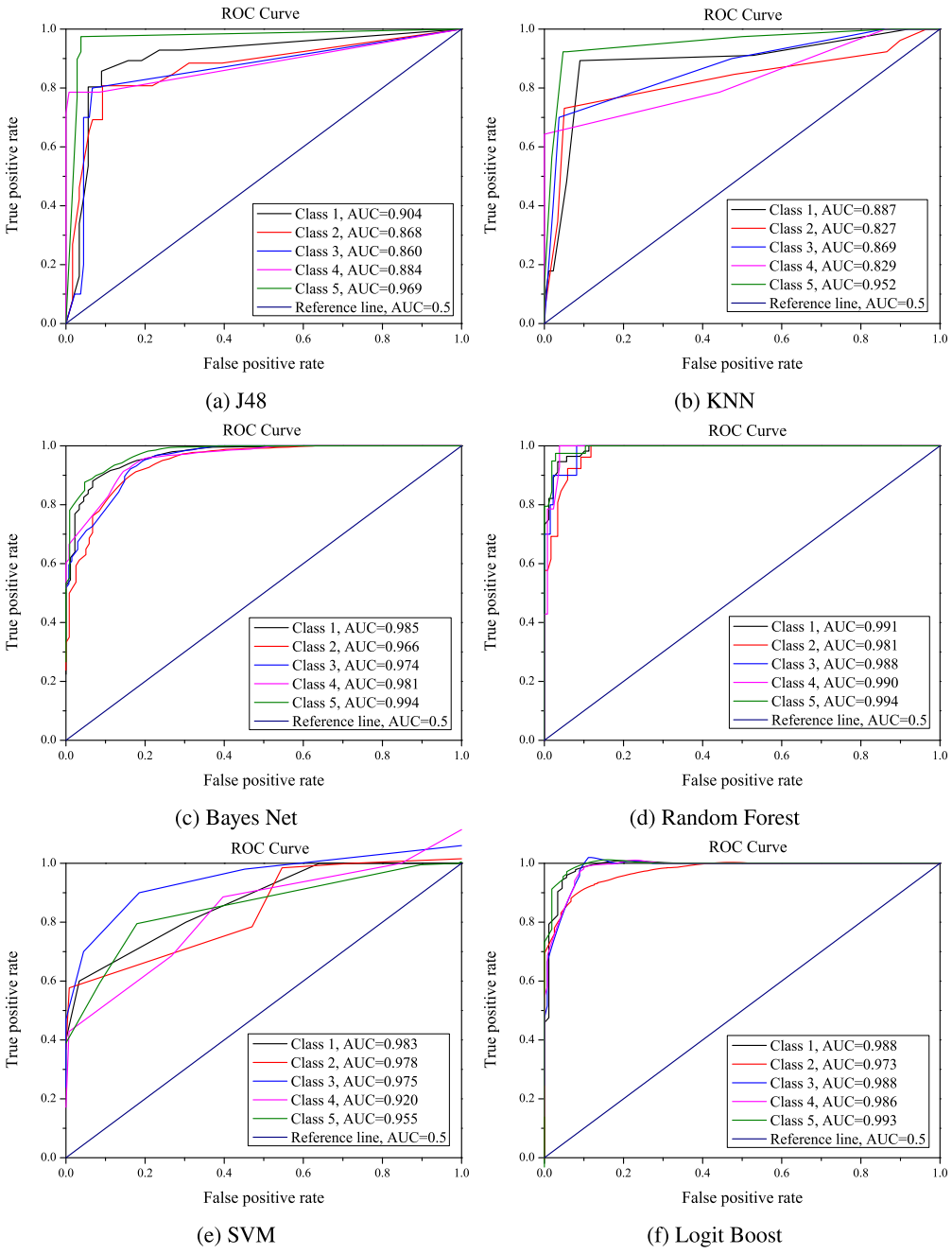


Fig. 3. ROC curves of six classifiers under five classes.

Forest is one of Bagging series algorithms, and Logit Boost is one of Boosting series algorithms. From the above results, it can be concluded that the performance of two integrated algorithms is better than that of a single classifier, but in terms of error and running time, the Logit Boost algorithm with dependency between base classifiers performs better. We also summarize the applicable conditions of each classifier in the actual prediction of academic performance according to the different expression forms of metrics.

4. Discussion

4.1. How to study learning behavior patterns from the perspective of educational data mining?

EDM is dominated by data mining, machine learning, statistics and other methods. With the development of large open online data sets and data mining technology, the field of EDM has attracted more and more attention. In this paper, we choose a public educational data set from UCI which is famous for data mining. In order to extract the features with great significance, we utilize PCA, and eigenvalues and contribution rate of principal components are shown in Table 2. However, we extract research variables from rotated component matrix in Table 3 instead of directly adopting the principal components as research variables. A pseudo statistic is proposed to determine cluster number objectively in Fig. 1 that is applied to add labels to metadata. The classification algorithm is applied to construct the prediction model with the extracted research variables. Seven evaluation metrics have a sharp contrast in Fig. 2. The intersections between BEP line and P–R curves which pay more attention to positive samples are shown in Fig. 4. According to different metrics, we summarize the application conditions which are convenient for the education manager to select the appropriate classifier to mine and analyze the educational data set

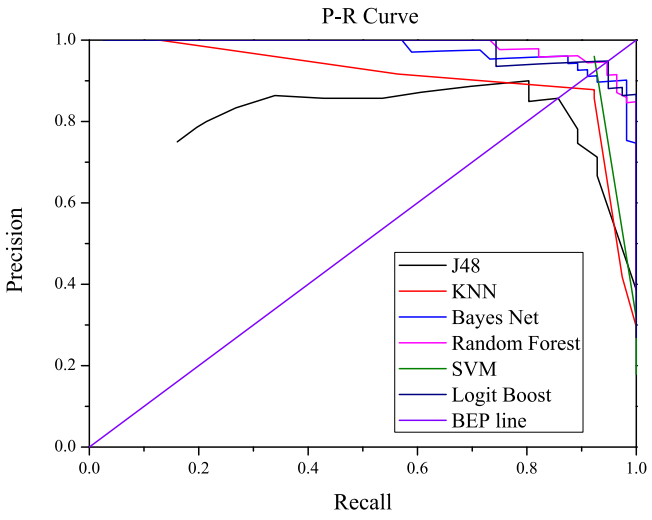


Fig. 4. P–R curves of six classifiers.

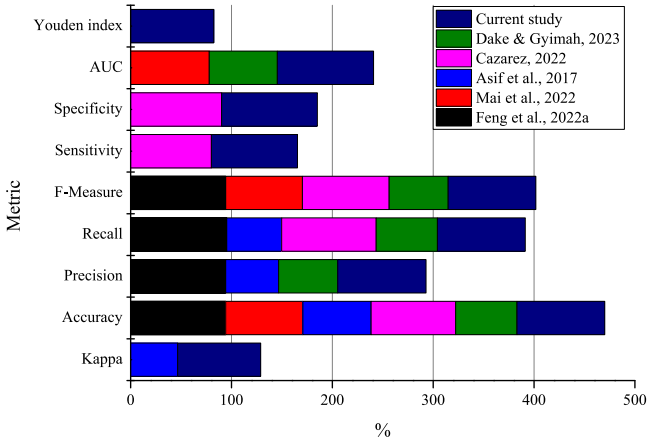


Fig. 5. Comparison between 5 previous works and current study.

in terms of the size of the educational data set, time and required accuracy in the actual situation. We compare the results of five works with the results of this paper, as shown in Fig. 5. Dake and Gyimah (2023), Feng, Fan and Ao (2022) and current study all employed accuracy, precision, recall, and F-Measure. It is found that accuracy, precision, recall, and F-Measure are used more frequently than Kappa, sensitivity, specificity, and AUC. Youden index is only used in this paper. In terms of Kappa, current study is obviously superior to the result from Asif et al. (2017) . With regard to sensitivity and specificity, current study is slightly better than (Cazarez, 2022). Compared with Dake and Gyimah (2023) and Mai et al. (2022), the performance of current study has a great advantage about AUC. Although not every metric is optimal, current study outperforms others in most metrics.

4.2. What aspects are included in the research on learning behavior patterns?

The public data set selected in this study contains not only information involved in learning, but also students' personal information and family information. Although personal information and family information are not directly related to learning behavior patterns, they will affect learning behavior patterns to some degree. In the results of attribute extraction in Table 3, only one attribute related to personal information is significant. The above is the preparation for studying the learning behavior patterns.

- As the name of the data set indicates, the 1st aspect of the research on learning behavior patterns is to evaluate learning behavior patterns. By visualizing the results of cluster number in Fig. 1, we can objectively determine students' learning groups in the data set. We can give students corresponding ranks according to the actual examination situation, instead of getting "excellent" only when they score more than 90 in the hundred mark system. We can also discover students' learning group preferences by analyzing the data in the same cluster.
- The 2nd aspect of research on learning behavior patterns in this paper is to predict the academic performance in the future through the analysis of the existing learning behavior. As can be seen from the title of the paper, one of the purposes of our research is to predict the students' academic performance based on their historical performance. 'Prediction' can be realized by classifying. In fact, PCA, clustering, and cross validation are all served for classifying. Firstly, PCA is applied to reduce dimension to extract more relevant attributes. Secondly, these extracted attributes as research variables are input aiming at clustering. Clustering results are used as the label of data. Then, labeled data can be classified in order to realize the goal of predicting. Finally, cross validation is utilized to obtain some metrics with the purpose of evaluating the effect of classifying. By using machine learning algorithms to build prediction classifiers, the evaluation metrics of different classifiers are displayed in Figs. 2, 3, and 4. The applicable conditions of different classifiers are compared and concluded according to the above evaluation metrics.
- The 3rd aspect of the research runs through the whole research process. We can instantly see the changing trend of the number of clusters and a pseudo-statistic proposed in this paper in Fig. 1, which aims to determine the number of clusters. Figs. 2, 3 and 4 clearly show the comparison of various performance metrics of each classifier. The teaching manager can choose the corresponding algorithm for prediction according to a certain aspect that should be taken into account when actually predicting academic performance.

4.3. How to innovate educational management by studying learning behavior patterns?

Simple statistical analysis methods are no longer suitable for analyzing the growing educational data sets. Therefore, in order to optimize educational management, we must innovate the analysis method radically. Data mining can discover the information hidden behind massive data, so educational data mining, which is a growing field, came into being. The research on learning behavior patterns is the most direct embodiment of educational data mining. Table 3 illustrates that personal information does not significantly affect the academic performance of students, so stakeholders had better attach more importance to attributes which are relevant to learning. According to quantitative standards in Eq. (1) rather than subjective score segmentation criteria, we determine the cluster number. When teachers pay more attention to the evaluation metrics of classifiers performance, they should avoid using J48 and KNN in Fig. 2. From Fig. 3, we can find that AUC of Random Forest and Logit Boost is larger, which indicates that the models constructed by these two classifiers perform better. If the classification result is unbalanced and positive samples are more important in class imbalance, we recommend to use Random Forest and Logit Boost, as can be seen in Fig. 4. By using data mining methods to study learning behavior patterns, we can extract the indirect information behind educational data sets. It is not only helpful for teaching managers to evaluate students' learning behavior patterns objectively, but also conducive to timely helping students who may be at risk in the future and encouraging the progress and development of excellent students.

5. Summary and prospect

With the help of data mining technology, this paper studies the learning behavior patterns from three aspects: evaluation, prediction and visualization. Now we summarize four places different from previous studies in the paper. First, in order to add labels to the data, clustering is performed and a pseudo statistic is proposed to determine the number of clusters objectively which avoids the arbitrariness of subjectivity. Second, although PCA is applied to extract attributes, the final selection of research variables is not the comprehensive variables extracted by PCA, but the attributes with greater significance in the rotating component matrix, which reduces the dimension of data analysis to certain extent and can improve the efficiency of later prediction because of the reduction of the number of attributes. Third, when using different classification algorithms to construct prediction classifiers, it is found that the accuracy, efficiency and error of the integrated classifier with dependent base classifier are better than that of a single classifier. The visualization of the analysis results makes the ‘‘calm’’ data set show intuitive and clear information, which is not only beneficial to the innovation of educational management, but also promote the development of students. Fourth, in the application of classification algorithms, we summarize the applicable conditions of different algorithms. When calculating the evaluation metrics of the classifier, we also summarize the applicable situation according to different visual forms.

The existing research and this paper all analyze, evaluate and predict learning behavior with different functional modules. In the future, we can consider designing an integrated framework to integrate data preprocessing technology, unsupervised learning technology, supervised learning technology and data visualization technology into EDM. As long as education managers input data, they can get the output of each functional module. Using the output of different functional modules can realize the idea of EDM proposed at the beginning, which is to serve education with technology and to develop education with innovation.

CRediT authorship contribution statement

Guiyun Feng: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Muwei Fan:** Writing – review & editing, Supervision, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Agrawal, S., Nigam, S., & Sahu, K. (2018). Prediction of students academic execution using K-Means and K-Medoids clustering technique. In *2018 2nd international conference on trends in electronics and informatics (ICOEI)* (pp. 1308–1315). USA: IEEE, http://dx.doi.org/10.1109/ICOEI.2018.8553747.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177–194. http://dx.doi.org/10.1016/j.compedu.2017.05.007.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, *23*(1), 537–553. http://dx.doi.org/10.1007/s10639-017-9616-z.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Busalim, A. H., Masrom, M., & Wan, N. (2019). The impact of facebook addiction and self-esteem on students' academic performance: A multi-group analysis. *Computers & Education*, *142*, Article 103651. http://dx.doi.org/10.1016/j.compedu.2019.103651.

Cazarez, R. L. U. (2022). Accuracy comparison between statistical and computational classifiers applied for predicting student performance in online higher education. *Education and Information Technologies*, *27*(8), 11565–11590. http://dx.doi.org/10.1007/s10639-022-11106-4.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. http://dx.doi.org/10.1007/BF00994018.

Cover, T. M., & Hart, P. E. (1953). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. http://dx.doi.org/10.1109/TIT.1967.1053964.

Crivei, L. M., Czibula, G., Ciubotariu, G., & Dindelegan, M. (2020). Unsupervised learning based mining of academic data sets for students' performance analysis. In *2020 IEEE 14th international symposium on applied computational intelligence and informatics (SACI)* (pp. 11–16). USA: IEEE, http://dx.doi.org/10.1109/SACI49304.2020.9118835.

Dake, D. K., & Gyimah, E. (2023). Using sentiment analysis to evaluate qualitative students' responses. *Education and Information Technologies*, *28*(4), 4629–4647. http://dx.doi.org/10.1007/s10639-022-11349-1.

Delgado, S., Morán, F., José, J. C. S., & Burgos, D. (2021). Analysis of students' behavior through user clustering in online learning settings, based on self organizing maps neural networks. *IEEE Access*, *9*, 132592–132608. http://dx.doi.org/10.1109/ACCESS.2021.3115024.

Demirer, R., Pierdzioch, C., & Zhang, H. (2017). On the short-term predictability of stock returns: A quantile boosting approach. *Finance Research Letters*, *22*(3), 35–41. http://dx.doi.org/10.1016/j.frl.2016.12.032.

Fan, Y., & Frederick, W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education*, *123*, 97–108. http://dx.doi.org/10.1016/j.compedu.2018.04.006.

Feng, G., Fan, M., & Ao, C. (2022). Exploration and visualization of learning behavior patterns from the perspective of educational process mining. *IEEE Access*, *10*, 65271–65283. http://dx.doi.org/10.1109/ACCESS.2022.3184111.

Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, *10*, 19558–19571. http://dx.doi.org/10.1109/ACCESS.2022.3151652.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. http://dx.doi.org/10.1006/jcss.1997.1504.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, *38*(2), 337–374.

Gao, L., Zhao, Z., Li, C., Zhao, J., & Zeng, Q. (2022). Deep cognitive diagnosis model for predicting students' performance. *Future Generation Computer Systems*, *126*, 252–262. http://dx.doi.org/10.1016/j.future.2021.08.019.

Goessling, M. (2017). LogitBoost autoregressive networks. *Computational Statistics & Data Analysis*, *112*(3), 88–98. http://dx.doi.org/10.1016/j.csda.2017.03.010.

Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. (2005). Neighbourhood components analysis. *Advances in Neural Information Processing Systems*, *17*, 513–520.

Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student dropout predictive model using data mining techniques. *IEEE Latin America Transactions*, *13*(9), 3127–3134. http://dx.doi.org/10.1109/TLA.2015.7350068.

Huang, S., & Ning, F. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, *61*, 133–145. http://dx.doi.org/10.1016/j.compedu.2012.08.015.

Kanamori, T., & Takenouchi, T. (2013). Improving Logitboost with prior knowledge. *Information Fusion*, *14*(2), 208–219. http://dx.doi.org/10.1016/j.inffus.2011.11.004.

Karthikeyan, V. G., Thangaraj, P., & Karthik, S. (2020). Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation. *Soft Computing*, *24*(24), 18477–18487. http://dx.doi.org/10.1007/s00500-020-05075-4.

Kumar, E., Balamurugan, S., & Sasikala, S. (2021). Multi-tier student performance evaluation model (MTSPEM) with integrated classification techniques for educational decision making. *International Journal of Computational Intelligence Systems*, *14*(1), 1796–1808. http://dx.doi.org/10.2991/ijcis.d.210609.001.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *50*(2), 157–224.

Lee, J.-E., & Recker, M. (2022). Predicting student performance by modeling participation in asynchronous discussions in university online introductory mathematical courses. *Educational Technology Research and Development*, *70*(6), 1993–2015. http://dx.doi.org/10.1007/s11423-022-10153-5.

Lin, Q., Liu, Y., & Yi, L. (2018). An integrated framework with feature selection for dropout prediction in Massive Open Online Courses. *IEEE Access*, *6*, 71414–71484. http://dx.doi.org/10.1109/ACCESS.2018.2881275.

Mai, T. T., Bezbradica, M., & Crane, M. (2022). Learning behaviours data in programming education: Community analysis and outcome prediction with cleaned data. *Future Generation Computer Systems*, *127*, 42–55. http://dx.doi.org/10.1016/j.future.2021.08.026.

Mallik, P., Roy, C., Maheshwari, E., Pandey, M., & Rautray, S. (2019). Analyzing student performance using data mining. In Y.-C. Hu, S. Tiwari, K. Mishra, & M. Trivedi (Eds.), *Ambient communications and computer systems, Vol. 904* (pp. 307–318). Singapore: Springer, http://dx.doi.org/10.1007/978-981-13-5934-7_28.

Manoharan, J. J., Ganesh, S. H., Felciah, M. L. P., & Banu, A. K. S. (2014). Discovering students' academic performance based on GPA using K-Means clustering algorithm. In *World congress on computing and communication technologies* (pp. 200–202). USA: IEEE, http://dx.doi.org/10.1109/WCCCT.2014.75.

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, *8*, 55462–55470. http://dx.doi.org/10.1109/ACCESS.2020.2981905.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, *29*(3), 241–288. http://dx.doi.org/10.1016/0004-3702(86)90072-X.

Przepiorka, A., Blachnio, A., Cudo, A., & Kot, P. (2021). Social anxiety and social skills via problematic smartphone use for predicting somatic symptoms and academic performance at primary school. *Computers & Education*, *173*, Article 104286. http://dx.doi.org/10.1016/j.compedu.2021.104286.

Ramze Rezaee, M., Lelieveldt, B. P. F., & Reiber, J. H. C. (1988). A new cluster validity index for the fuzzy *c*-mean. *Pattern Recognition Letters*, *19*(3–4), 239–246. http://dx.doi.org/10.1016/S0167-8655(97)00168-2.

Riestra-Gonzalez, M., Paule-Ruiz, M., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, *163*, Article 104108. http://dx.doi.org/10.1016/j.compedu.2020.104108.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. http://dx.doi.org/10.1016/j.eswa.2006.04.005.

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12–27. http://dx.doi.org/10.1002/widm.1075.

Schapire, R. E. (1989). The strength of weak learnability. In R. Rivest, D. Haussler, & M. K. Warmuth (Eds.), *Proceedings of the 2nd annual workshop on computational learning theory* (pp. 197–227). San Francisco (CA): Morgan Kaufmann, http://dx.doi.org/10.1016/B978-0-08-094829-4.50030-1.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*, *143*, Article 103676. http://dx.doi.org/10.1016/j.compedu.2019.103676.

Turabieh, H., Al Azwari, S., Rokaya, M., Alosaimi, W., Alharbi, A., Alhakami, W., et al. (2021). Enhanced Harris Hawks optimization as a feature selection for the prediction of student performance. *Computing*, *103*(7), 1417–1438. http://dx.doi.org/10.1007/s00607-020-00894-7.

Varela, N., Montero, E. S., Vásquez, C., Guiliany, J. G., Mercado, C. V., et al. (2019). Student performance assessment using clustering techniques. In Y. Tan, Y. Shi (Eds.), *Data mining and big data* (pp. 179–188). Singapore: Springer, http://dx.doi.org/10.1007/978-981-32-9563-6_19.

Yadav, R. S. (2020). Application of hybrid clustering methods for student performance evaluation. *International Journal of Information Technology*, *12*(3), 749–756. http://dx.doi.org/10.1007/s41870-018-0192-2.

Zaffar, M., Hashmani, M. A., Habib, R., Quraishi, K. S., Irfan, M., et al. (2022). A hybrid feature selection framework for predicting students performance. *Computers, Materials & Continua*, *70*(1), 1893–1920. http://dx.doi.org/10.32604/cmc.2022.018295.

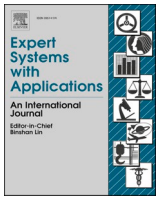
Zhang, R. C., Lai, H. M., Cheng, P. W., & Chen, C. P. (2017). Longitudinal effect of a computer-based graduated prompting assessment on students' academic performance. *Computers & Education*, *110*, 181–194. http://dx.doi.org/10.1016/j.compedu.2017.03.016.

Expert Systems With Applications 237 (2024) 121470

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Breast cancer diagnosis based on hybrid SqueezeNet and improved chef-based optimizer

Qirui Huang [a,*], Huan Ding [a], Mehdi Effatparvar [b,*]

[a] School of Information Engineering, Nanyang Institute of Technology, Nanyang, He Nan Sheng 473004, China
[b] Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran

ARTICLE INFO

Keywords:
Diagnosis
Breast Cancer
Mammogram
SqueezeNet
Improved Version of Chef-based Optimization Algorithm

ABSTRACT

The most frequent disease in women and the one that accounts for the majority of cancer-related fatalities in females is breast cancer. An important milestone in breast cancer CAD systems is the automatic recognition and delineation of masses in mammograms. We have developed a novel technique in this study to identify potential mass candidates in mammograms automatically. The suggested method is a pipeline method with three stages. First, a noise reduction method based on median filtering is performed to eliminate the noises. Then, the region of interest is threshold out of the pre-processed images using the Kapur technique. Afterward, to eliminate useless features, optimum feature extraction and selection are designed. It is established by an Improved version of the Chef-Based Optimization (ICBO) algorithm. Lastly, a classification is implemented based on an optimal SqueezeNet model for the final diagnosis. This classifier is also optimized based on the ICBO algorithm to maximize efficiency. Simulations of the method are applied to the MIAS database, and its results are compared with different state-of-the-art methods to show the method's advantage.

1. Introduction

A breast tumor is a kind of cancer in which the unchecked proliferation of aberrant cells results in a lump in the breast tissue. The World Health Organization (WHO) reported that this tumor affects 2.1 million women annually and accounts for the majority of female cancer-related fatalities. 627,000 women suffered breast cancer deaths in 2018, according to studies (Cai, 2021). In actuality, this malignancy is responsible for 15% of all cancer-related fatalities in women. The third leading cause of mortality for women is the most prevalent cancer. As a result, human mortality may drop if a mechanism could be developed to identify this illness.

One of the most popular techniques used by radiologists to identify malignant tumors and cysts to diagnose and test for tumors is X-ray mammography. The American NCC (National Cancer Center) reports that ten to thirty percent of breast tumors cannot be spotted using a radiologist because mammography pictures are challenging to evaluate (Navid Razmjooy & Ghadimi, 2018). One of the indications and symptoms of cancer shown in mammography pictures are masses and tiny calcium particles, which are extremely minute. It is quite challenging to identify these signs appropriately.

Generally speaking, masses are divided into two groups: benign and malignant, each of which has distinct visual features (Xu, 2020). Malignant cells have uneven and non-uniform shapes, unclear edges, angular edges, and sometimes boot-like or lobular shapes. At the same time, benign masses are oval, have obvious edges, and have no angles.

Very small calcareous particles are typically shown as noisy particles in mammography pictures, and their low brightness makes it challenging for radiologists and medical professionals to identify them. The difficulty of diagnosing breast tumor masses, which is primarily related to human mistakes in the precision of diagnosis, must be explained in some manner by scientists (Akbary, 2019; Bagheri, et al., 2018). It is because the timely and precise identification of cancerous tumors, along with their various sorts, is of key significance in the individuals' health in society.

To eliminate human mistakes, aid in early detection, and enhance the treatment of this condition, it is crucial to supply and develop an automatic approach utilizing the image processing and detection methods, as well as to improve and optimize current systems of diagnosing. One of the most fatal illnesses affecting women worldwide is breast cancer. It has been shown via a recent study that this illness is among the deadliest ones for female patients. Because of this, it is crucial

* Corresponding authors.
E-mail addresses: qirui@nyist.edu.cn (Q. Huang), effatparvarmehdi@gmail.com (M. Effatparvar).

https://doi.org/10.1016/j.eswa.2023.121470
Received 19 January 2023; Received in revised form 2 September 2023; Accepted 3 September 2023
Available online 5 September 2023

to establish it early. In recent years, an automated decision-making approach has been employed to identify this illness more swiftly (Cai, 2019; Dehghani, 2020; Ebrahimian, 2018). Computer science research in the area of breast cancer diagnostics could be a significant contribution to the field. In the last several years, image-analyzed methods and machine-learning algorithms have been put forth in the field of breast tumor detection using mammography pictures, which can reduce costs and increase the efficacy of the therapy (Eslami, et al., 2018; Fan, 2020; Firouz and Ghadimi, 2016; Gao, 2019). Neural networks have been one of the most used technologies in this subject in recent years.

The approaches used for extracting features and categorization determine how well the computer-aided recognition method operates. With the emergence of the deep learning machine learning model, which is essential for features extracted (Guo & Razmjooy, 2019). Deep learning offers a wealth of semantic data that makes it possible to learn picture structures. Among the studies on computerized intelligence systems' use in diagnosing breast cancer, it can be mentioned (Huang, 2019) that they presented a novel fruit fly optimization model that improved the Support Vector Machine for the high-level feature-based detection of breast cancer (Liu, 2020). The diagnosis of breast cancer is produced using an improved system for machine learning. For the initial time, breast cancer is diagnosed using the high-level characteristics that were extracted from the participants. It can effectively discriminate malignant from benign breast cancers and help the doctor make a medical assessment. The efficiency of the suggested strategy was examined using the 10-fold cross-validation approach. The experimental findings show that the suggested LFOA-SVM approach outperforms its competitors regarding several performance criteria. It can effectively discriminate malignant from benign breast cancers and help the doctor make a medical assessment.

Cai showed how technology, specifically Convolutional Neural Networks (CNN) and an advanced thermal exchange optimizer, can aid in the early detection of breast cancer (Cai, 2021). By using pattern recognition and image analysis techniques to automatically identify abnormal spots in mammographic images, this technique reduces human error and speeds up assessment. The first step in this process involved enhancing the quality of mammography pictures through image contrast enhancement and noise reduction techniques. As breast cancer was a frequent gynecological condition, early detection was key for successful treatment. Utilizing image-analyzed methods and a pipeline framework, mammographic pictures were examined to diagnose malignant tumors. The results of this study were applied to the MIAS mammogram database and showed a promising prediction performance of 93.79% for identifying cancer patients through the suggested technique. With the use of advanced technology, like CNN and image-analyzed techniques, healthcare providers can improve early detection rates and provide better treatment options for breast cancer patients.

Cao et al. (Cao, 2020) proposed NF-Net, an effective approach for training breast tumor classification models with noisy labels. The approach incorporated two softmax layers to prevent overfitting and a teacher-student module to distill knowledge from clean labels. Compared to existing works, the method achieved 73% accuracy, 69% precision, 80% recall, and 0.74 F1-score. The overcomes labeled shortage in training models using BI-RADS ratings. Despite noisy labels, NF-Net reduced their negative effect.

Zebari et al. (Zebari, 2021) proposed a new method for classifying benign or malignant breast cancer from mammogram images. The method used hybrid thresholding and machine learning to derive the ROI, which was divided into five blocks. The wavelet transform was applied to each block to remove noise, and an improved fractal dimension approach was used to extract multiple features from each block. The number of features was then reduced using a genetic algorithm, and the results were classified using five classifiers combined with ANN. The approach was tested on four benchmark mammogram datasets using single- and double-dataset evaluations. Results showed that the proposed method outperformed state-of-the-art models, with better results on different datasets.

Jabeen et al. (Jabeen, 2022) proposed a new framework for breast cancer classification from ultrasound images, which utilized deep learning and feature fusion. The framework consisted of five steps, including data augmentation, using a pre-trained DarkNet-53 model, extracting features, selecting the best features using improved optimization algorithms, and fusing the best features for classification using machine learning. The experiment was conducted on an augmented BUSI dataset and achieved an accuracy of 99.1%. The proposed framework outperformed recent techniques in breast cancer classification.

Maqsood et al. (Maqsood, Damaševičius, & Maskeliūnas, 2022) developed a deep learning system that efficiently identified breast cancer using an end-to-end training strategy, which used mammography images for computer-aided recognition. The approach included a modified contrast enhancement method that refined image detail and a TTCNN to enhance the classification performance. The performance of TTCNN was analyzed by extracting deep features from various CNN models and selecting the best features using the entropy-controlled firefly method. The proposed approach achieved an average accuracy of 97.49% on DDSM, INbreast, and MIAS datasets, outperforming previous methods. The findings showed that automatic deep learning algorithms offer the potential to improve clinical tools for more accurate mammogram screening.

To segment breast lesions, a quantization-assisted U-Net technique was suggested by Meraj et al. (Meraj, 2021). U-Net and quantization were the first and second segmentation steps, respectively. U-Net and quantization were the first and second segmentation steps, respectively. Quantization was a helpful tool when using U-Net-based segmentation to separate precise lesion regions from sonography pictures. The separated lesions were then employed in the Independent Component Analysis (ICA) approach to extract features, which were subsequently merged with deep automated features. For assessment and comparison, public ultrasonic-modality-based datasets like the Open Access Database of Raw Ultrasonic Signals (OASBUD) and the Breast Ultrasound Imaging Dataset (BUSI) were employed. The same characteristics were derived from the OASBUD data. However, classification was carried out after feature regularization utilizing the lasso technique. The data, which was collected, enabled stakeholders to suggest a Computer-Aided Design (CAD) method for identifying breast cancer utilizing ultrasound modalities.

Stephan et al. (Stephan, 2021) investigated a combined artificial bee colony with an optimal whale algorithm for better breast cancer detection. The drawbacks of traditional approaches could be solved by Computer-Aided Diagnosis (CAD), which enabled radiologists to make precise decisions. The accuracy of breast cancer detection could be increased by a computer-aided diagnosis approach based on ANN (Artificial Neural Networks) utilizing a population-based strategy. HAW integrated ABC's exploitative employee bee attacking phase with the whale optimization bubble net attacking technique. For WBCD, WDBC, WPBC, DDSM, MIAS, and INbreast, the AW-RP variation obtained better precision of 99.2%, 98.5%, 96.3%, and 99.1% with a low-complexity ANN system. Bees used humpback whales to locate better locations for food sources during the employee bee assaulting phase. Several breast cancer statistics data were used to analyze these combined variations.

Sha et al. (Sha, Hu, & Rouyendegh, 2020) recommended a best-practice deep learning approach for automated breast cancer screening. Breast cancer therapy could be made simpler and more successful with earlier detection. The complete approach to finding the malignant area in the mammography picture was proposed in this research. It used a grasshopper optimization technique, ideal picture classification, and image noise reduction. The suggested method outperformed Conventional approaches in terms of sensitivity (96%), specificity (93%), PPV (85%), NPV (97%), precision (92%), and effectiveness (92%).

Bourouis et al. (Bourouis, 2022) detected breast cancer using

ultrasound pictures and a neural network tailored to a *meta*-heuristic method. Among all cancers that affected women worldwide, breast cancer was the most dangerous. The only approach to enhance treatment choices, which therefore reduced mortality and raised survival rates, was through early detection. The Computer-Aided Diagnostic method used in this study to find abnormalities in breast ultrasound pictures was its key innovative feature. Compared to previous approaches like SOM-SVM (87.5%) and MBA-RF, the suggested GWO-WNN approach (98%) provided better precision (96.85). On 346 ultrasound pictures, the suggested scheme's classification performance was verified. According to a numerical study, the suggested work might produce greater prediction performance compared to the current approaches.

As is evident from the literature, there has been an increase in recent years in the use of deep learning and bio-inspired algorithms to address cancer detection issues. The major goal of this work is to present a new bio-inspired optimization technique for SqueezeNet's ideal placement in mammography pictures for the detection of breast tumor tissues. The Improved Chef-Based Optimization Algorithm, a novel bio-inspired optimizer algorithm, is used in the segmentation step to enhance SqueezeNet's performance. The graphical abstract of the proposed method is given in Fig. 1.

2. Image noise removal

The quality of medical images does become an essential part of up-to-date medicine and affects the doctors' diagnosis and treatment accuracy. In medical pictures, low contrast and resolution make the correct diagnosis a demanding task that affects the accuracy and speed of specialists' diagnoses. Because of the random physical feature of imaging systems, the presence of noise in the picture is inevitable. For example, the brightness and temperature of imaging sensors are among the most important factors affecting the amount of image noise. Also, since image sensors count the number of received photons, which is a random quantity, images generally have photon counting noise. In addition, for various reasons, while converting the image from one format to another, such as imaging, copying, scanning, digitizing, channel transmission, displaying, printing, or compressing the image, various types of noise are always added to the images (Guo, 2021). Thus, improvement of the medical image is necessary for reflecting disease information clearly and accurately.

A Mammogram image usually has noise and radiological artifacts. Therefore, first, in the pre-processing phase, noise removal should be established, and then the mammogram image is found. Image noise reduction defines a process for reducing or removing the noise from the input image.

Every recording tool has characteristics that subject it to noise. Random or white noise are two types of noise (a signal that has been distributed uniformly in whole frequencies as its power density function). Image deterioration in measurements of the environment (e.g., inadvertent disruption), exaggeration (e.g., distortion), geometric distortion, and unequal attenuation of the sensor noise sensor attenuation (e.g., camera movement or out-of-focus) are only a few of the causes of image noise.

To rebuild and return the deformed image to its original condition using ideal models, noise is removed from breast tumor images. The employment of the median filter is by far the most significant and well-known technique for noise removal. The median filter may be employed to reduce noise while preserving the edges of the image. The median value of a pixel's neighbors is used to substitute it in median filtering, i.e.,

$$y(i,j) = med[x(i,j) : (i,j) \in n] \tag{1}$$

Here, n describes the surrounding neighbor in the image position (i, j).

Fig. 2 indicates a sample noise reduction example for a breast cancer image using the median filter method.

The aforementioned outcome demonstrates that the median filter works well for the breast cancer noisy image.

3. Image segmentation

Image thresholding is used in this instance to segment breast cancer images. In other words, if image segmentation yields good results, good feature extraction will be made. The Otsu approach is used in this study's thresholding (Razmjooy, 2020). The following formulation is used in the procedure to find a threshold level that will reduce the image variance:

$$\sigma_{\delta}^2(t) = \delta_1(t)\sigma_1^2(t) + \delta_2(t)\sigma_2^2(t) \tag{2}$$

Where, δ_i and σ_i^2 represent the possibility for two discrete classes with t (a threshold amount) and the classes' variance.

This technique describes a type of minimizing problem for the variance amount that is class-like to maximize the variance amount that is class-in.

$$\sigma_{\omega}^2(t) = \sigma^2 - \sigma_{\delta}^2(t) = \delta_1(t)\delta_2(t)[\mu_1(t) + \mu_2(t)]^2 \tag{3}$$

Where, $\mu_i(t)$ defines the mean value, which is continually updated. Otsu's approach does initialize the amounts of $\rho_i(0)$ and $\mu_i(0)$ to the entire potential threshold values after analyzing the histogram of the intensity levels (Xu, Wang, & Razmjooy, 2022). Then, the best threshold in maximum is obtained for the thresholding procedure once the values of $\rho_i(t)$ and $\mu_i(t)$ are adjusted. After that, morphological procedures involving hole filling, opening, and shutting are used to remove the

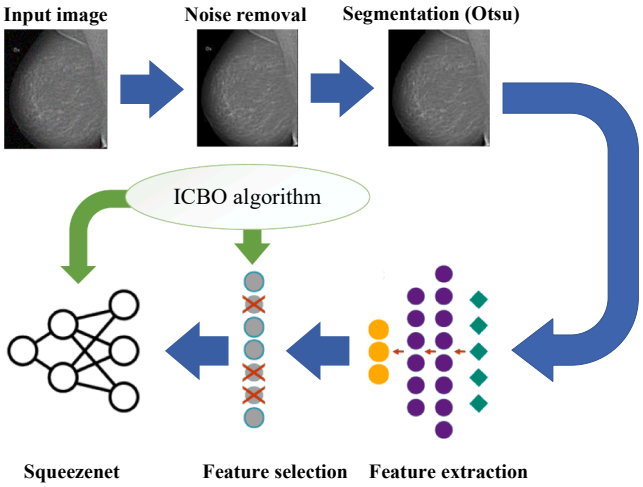


Fig. 1. Graphical abstract of the proposed method.

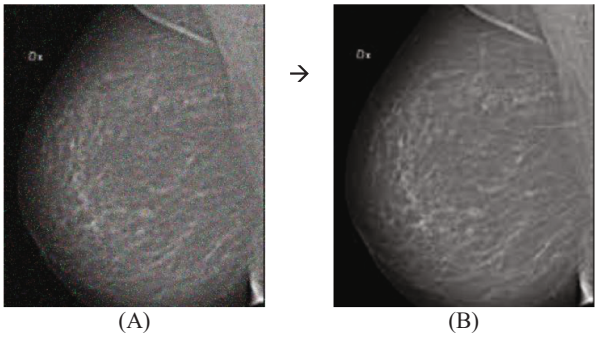


Fig. 2. Sample noise reduction example for breast cancer image: (A) noisy image and (B) image after noise reduction.

remaining portions of breast cancer. The following formulation is used to fill the image's empty spaces:

$$F_k = (F_{k-1} \ominus S) \bigcap A^c, k = 1, 2, \cdots \tag{4}$$

Here, S and A signify, in turn, the structuring factor and the area element. The structuring element is an identity matrix of size 5.

The brighter features are removed, while maintaining the other gray surfaces using morphological opening. The formula for this operation is as follows (Liu, 2020):

$$F \odot se = (F \Theta se) \oplus se \tag{5}$$

Then, the narrow portions of the generated image are connected via morphological close. The following equation gives the conclusion:

$$F \odot se = (F \oplus se) \Theta se \tag{6}$$

Fig. 3 indicates some examples of breast cancer segmentation using the suggested technique.

4. Improved chef-based optimization algorithm (ICBO)

4.1. Inspiration of the CBO algorithm

Those, who are interested in cooking and studying cooking, take part in training workshops for enhancing their expertise in cooking and become a chef. In this method, like other *meta*-heuristic algorithms, initialization is done for numerical individual solutions, and finally, the finest individual solution is specified as a solution to the problem through the implementation of algorithm iterations. Hence, the inspiration of the CBO Algorithm is the flow that turns a person into a chef during the training workshop.

A defined number of chef trainers are considered in a training workshop. A chef trainer is accountable for training in each class. When cooking techniques and skills are taught by chef trainers, they also try to enhance their abilities through practice and based on the training of the finest chefs. Cooking learners attempt to pick up and mimic the performance of the chef trainer. Also, practicing the skills learned by cooking learners will make them progress more in cooking. At the end of the workshop and as a result of the instructions given to the students, they can work as expert chefs. In the following sections, the previously mentioned concepts are simulated, and the CBO algorithm is designed.

4.2. Initialization of the algorithm

Two groupings of individuals, including cooking learners and cooking trainers, constitute the population of the suggested CBO algorithm. The constituent members of the CBO algorithm define the candidate solution and specify details concerning the variables of the problem. The collection of individuals of the algorithm is defined in the following matrix, where each individual is a vector:

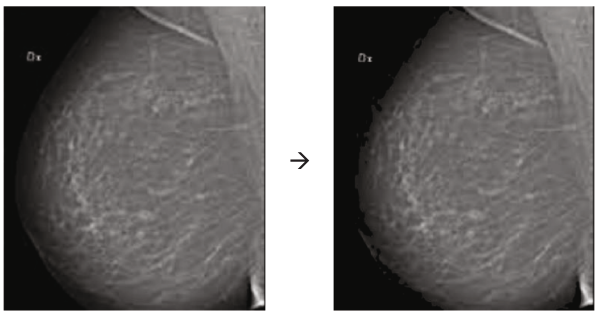


Fig. 3. Some examples of breast cancer diagnosis.

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_i \\ \vdots \\ Z_N \end{bmatrix}_{N \times m} = \begin{bmatrix} z_{1,1} & \cdots & z_{1,j} & \cdots & z_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{i,1} & \cdots & z_{i,j} & \cdots & z_{i,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{N,1} & \cdots & z_{N,j} & \cdots & z_{N,m} \end{bmatrix}_{N \times m} \tag{7}$$

Where, the population matrix of the CBO algorithm is denoted by Z, the i^{th} individual of the CBO algorithm is defined by $Z_i = (z_{i,1}, z_{i,2}, \cdots, z_{i,m})$, $z_{i,j}$ states its j^{th} coordinate, the size of individuals is indicated by N, and the number of problem variables of the fitness function is expressed by m. The location of the CBO approach individuals at the outset of its execution is set at random for $j = 1, 2, \cdots, m$ and $i = 1, 2, \cdots, N$ throughout the following formula:

$$z_{i,j} = lb_j + r.(ub_j - lb_j), \tag{8}$$

Where, r defines a random amount between 0 and 1, ub_j and lb_j indicate the higher and the lower restrictions of the j^{th} problem variable.

By devoting the proposed amounts of any individual of CBOA into the variables, a related fitness function value is assessed. Therefore, the cost function is gained for the number of individuals in the procedure (N). The values of the fitness function can be illustrated as follows:

$$F = \begin{bmatrix} F_1 \\ \vdots \\ F_i \\ \vdots \\ F_N \end{bmatrix}_{N \times 1} = \begin{bmatrix} F(Z_1) \\ \vdots \\ F(Z_i) \\ \vdots \\ F(Z_N) \end{bmatrix}_{N \times 1} \tag{9}$$

Where, F and F_i define the vector of the fitness function amounts and the fitness function amount computed for the i^{th} individual of the CBO algorithm.

To select the best solution from the CBO algorithm population, the fitness function's value is the criterion of action because it determines the quality of each candidate solution. The finest candidate solution is the individual, who has the finest value for the fitness function. Throughout the algorithm's operation, per epoch, the individuals of the CBO method have been renewed, and the related amounts of the fitness function have been obtained.

Hence, the finest individual in any epoch is renewed based on the fitness functions' value.

4.3. Mathematical modeling of the CBO algorithm

Each of the groups that make up the algorithm, which includes the group of chef trainers and the group of cooks in training, is updated in distinct ways during the epochs of the algorithm. According to analyzing the fitness functions' values, several CBOA individuals with more satisfactory values of the fitness function are set as the chef trainer. Consequently, by sorting the algorithm's swarm matrix rows in increasing order based on the value of the fitness function, the set of the first N_C individuals are chosen as the set of chef trainers, and the remains set of $N - N_C$ individuals are determined as the set of the cooking learners. The following equations represent the sorted population matrix of the CBO algorithm and the ordered fitness function vector.

$$ZS = \begin{bmatrix} ZS_1 \\ \vdots \\ ZS_{NC} \\ ZS_{NC+1} \\ \vdots \\ ZS_N \end{bmatrix}_{N\times m} = \begin{bmatrix} zs_{1,1} & \cdots & zs_{i,j} & \cdots & zs_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ zs_{Nc,1} & \cdots & zs_{Nc,j} & \cdots & zs_{Nc,m} \\ zs_{Nc+1,1} & \cdots & zs_{Nc+1,j} & \cdots & zs_{Nc+1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ zs_{N,1} & \cdots & zs_{N,j} & \cdots & zs_{N,m} \end{bmatrix}_{N\times m} \tag{10}$$

$$FS = \begin{bmatrix} FS_1 \\ \vdots \\ FS_{Nc} \\ FS_{Nc+1} \\ \vdots \\ FS_N \end{bmatrix}_{N\times m} \tag{11}$$

Where, N_C defines the number of chef trainer, ZS indicates the arrayed population matrix of the algorithm, and FS represents a vector that sorts the number of fitness functions in increasing order. In the ZS, individuals from ZS_1 to ZS_{Nc} define the set of chef trainers, and individuals from ZS_{Nc+1} to ZS_N define the set of cooking learners. The values of the fitness functions related to ZS_1 to ZS_N is represented by the vector FS_I.

Stage 1: The renewing approach for the set of chef trainers (renewing of ZS_1 to ZS_{NC})

In a cooking workshop, it is presumed that numerous chef trainers are accountable for instructing cooking abilities to learners. The methods of improving cooking ability that chef trainers choose, include the following two items. One of the methods is to follow the example of the finest chef trainer and learn from her skills. This procedure illustrates the CBO algorithm exploration and global search powers.

This procedure makes the top chefs increase their abilities according to the finest chef before the learners start learning. Likewise, this procedure inhibits the algorithm from getting trapped in the local optimum and leads to analyzing various regions of the solution space more efficiently and accurately. Based on this procedure, a novel location for any chef trainer is achieved when $i = 1, 2, \cdots, N_C$ and $j = 1, 2, \cdots, m$ by the next formula:

$$zs_{i,j}^{C/S1} = zs_{i,j} + r.(BC_j - I.zs_{i,j}) \tag{12}$$

Where, $zs_{i,j}^{C/S1}$ represents the novel computed location for the i^{th} arrayed individual of CBO algorithm on the first procedure ($C/S1$) of renewing the chef trainer in the j^{th} direction, BC defines the best chef trainer, which is determined by ZS_1 in matrix ZS, BC_j indicates the j^{th} orientation of the best chef trainer, r denotes a random amount between 0 and 1, and I denotes an amount that is chosen throughout the operation from the collection (Cai, 2021; Navid Razmjooy and Ghadimi, 2018) at random. This novel location is allowable to the CBO algorithm when it leads to improving the amount of the fitness function. The modeling of the stated situation is demonstrated below:

$$ZS_i = \begin{cases} zs_{i,j}^{C/S1}, FS_i^{C/S1} < F_i; \\ zs_i, else, \end{cases} \tag{13}$$

Where, $FS_i^{C/S1}$ defines the fitness function amount of the individual $zs_i^{C/S1}$.

In another procedure, the cooking ability of each chef trainer gets enhanced through their practice and activities. This procedure illustrates the CBO method's exploitation capability and local search. If any parameter in the problem is assumed ability of cooking, a chef trainer is likely to attempt to enhance entire of those abilities to obtain a more suitable value of fitness function.

This method makes each individual find better solutions in her/his neighborhood without considering the location of other individuals. It is possible to achieve better solutions by small alterations in the location of candidates and based on local exploitation and search. Therefore, nearby any chef trainer in the solution space, a location is created for $j = 1, 2, \cdots, m$ at random by the following equations:

$$lb_j^{local} = \frac{lb_j}{t}, \tag{14}$$

$$ub_j^{local} = \frac{ub_j}{t}, \tag{15}$$

Where, ub_j^{local} and lb_j^{local} represent the higher and lower boundaries of the j^{th} problem variable, and t indicates the counter of iteration.

$$zs_{i,j}^{\frac{C}{S2}} = zs_{i,j} + lb_j^{local} + r.\left(ub_j^{local} - lb_j^{local}\right),$$

$$i = 1, 2, \cdots, N_c, j = 1, 2, \cdots, m, \tag{16}$$

$$ZS_i = \begin{cases} ZS_{i,j}^{C/S2}, FS_i^{C/S2} < F_i; \\ ZS_i; else, \end{cases} \tag{17}$$

Where, $ZS_{i,j}^{C/S2}$ defines the novel computed location of the i^{th} CBO algorithm individual based on the second procedure ($C/S2$) of chef trainer renewing in the j^{th} orientation, and $FS_i^{C/S2}$ indicates its fitness function amount.

Stage 2: renewing approach of a cooking learner set (renew of ZS_{NC+1} to ZS_N)

Cooking learners participate in the workshop for learning cooking abilities and becoming a chef. In the proposed algorithm, three procedures are considered for this purpose. In the first procedure, the selection of the cooking class for learning is carried out randomly by the learner. As a result of using this procedure, learners are provided with dissimilar trainers for training, and they can learn various abilities (individuals of the population move to other areas of the solution space). Whereas, if the population individuals only move towards the finest individual i.e. when the training is done only through the finest trainer, the exact global search in the solution area will be impossible. This procedure has been modeled in the CBO algorithm in a manner that for any cooking learner, a novel location is obtained at first that is gained using the instruction and admonition of the chef trainer, for $i = N_C + 1.N_C + 2, \cdots, N, j = 1, 2, \cdots, m$ as follows.

$$zs_{i,j}^{S/S1} = zs_{i,j} + r.\left(CI_{ki,j} - I.zs_{i,j}\right), \tag{18}$$

Where, $zs_{i,j}^{S/S1}$ defines the novel computed location for the i^{th} arrayed individual of CBO algorithm based on the first procedure ($S/S1$) of renewing the cooking learner in the j^{th} orientation, and $CI_{ki,j}$ indicates the chosen chef trainer by the i^{th} cooking learner, in which k_1 is determined by the collection [1,2,…, NC] at random.

This novel location relocates the former location for any CBO algorithm individual if it leads to improving the fitness function's value. The modeling of the stated conception is as follows:

$$ZS_i = \begin{cases} ZS_{i,j}^{S/S1}, FS_i^{S/S1} < F_i; \\ ZS_i, else \end{cases} \tag{19}$$

Where, $FS_i^{S/S1}$ denotes the value of the fitness function of $ZS_{i,j}^{S/S1}$.

In this procedure, any parameter in the CBO algorithm has presumed

an ability of cooking; any cook learner aims to instruct the whole chef trainer one ability and thoroughly emulate the trainer. The exploration ability and global search of the proposed algorithm are increased through this procedure. Updating one variable instead of updating all individual solution variables is one of the advantages of using this procedure. It may not be essential to renew the entire individual location orientation to obtain better solutions.

The mentioned ability in the proposed algorithm illustrates a special element of a cooking abilities vector of a chosen chef trainer $CI_K(K\in\{1,2,\cdots,N_C\})$. Therefore, the mathematical simulation of the second procedure is described below:

For any cooking learner ZS_i (individuals of CBO algorithm with $i = N_{C+1}, N_{C+2}, \cdots, N$), one chief trainer, which is indicated by the vector $CI_{ki} = (CI_{ki+1}, \cdots, CI_{ki,m})$, is chosen at random (an individual in the CBO algorithm with the index k_i, which is chosen at random from the collection $[1, \cdots, N_C]$), then it is chosen by its $\updownarrow^{th}$ orientation at random (therefore, a numeral $\updownarrow$ from the collection $[1, \ldots m]$, which denotes the chosen trainer's "ability'') and using this amount $CI_{ki,\updownarrow}$, the $\updownarrow^{th}$ orientation of the i^{th} cooking learner vector ZS_i is substituted (hence,$zs_{i,\updownarrow}$).

Based on the noted conception, a novel location is obtained for any CBO algorithm cooking learner utilizing the next equation:

$$zs_{i,j}^{S/S2} = \begin{cases} CI_{ki,j}, & j = l; \\ zs_{i,j}, & else, \end{cases} \tag{20}$$

Where $\updownarrow$ is determined at random from the collection $[1,2,\cdots,m]$, $i = N_{C+1}, N_{C+2}, \cdots, N$, $j = 1, 2, \cdots, m$. Therefore, it is substituted with the former location according to the following equation, if it enhances the aim value of the fitness function.

$$ZS_i = \begin{cases} ZS_i^{S/S2}, & FS_i^{S/S2} < F_I; \\ ZS_i, & else, \end{cases} \tag{21}$$

Where, $ZS_{i,j}^{S/S2}$ defines the novel computed location for the i^{th} arrayed individual of CBO algorithm based on the second procedure $(S/S2)$ of renewing cooking learners on the j^{th}, $FS_i^{S/S2}$ denotes its fitness function's value.

In the third procedure, any cooking learner wants to enhance her cooking abilities based on her activities and practices. This procedure displays the CBOA's exploitation capability and local search. In this procedure, by increasing the exploitation and potency of local search in the algorithm, it is possible to obtain better solutions close to the determined solutions. In this procedure, like the chef trainers' local search technique, cook learners want to touch suitable outcomes with tiny and detailed stages. When any variable in the problem is assumed an ability of cooking, a cook learner is likely to enhance entire abilities for obtaining a finer fitness function amount.

On the basis of the described conception, near any cooking learner in the solution area, a location is randomly created using formula 9 and formula 8, so a novel location is computed as follows:

$$zs_{i,j}^{S/S3} = \begin{cases} zs_{i,j} + lb_j^{local} + r.\left(ub_j^{local} - lb_j^{local}\right), & j = q; \\ zs_{i,j}, & j \neq q, \end{cases} \tag{22}$$

Where, $zs_{i,j}^{S/S3}$ defines the novel computed location for the i^{th} arrayed individual of CBO algorithm based on the third procedure $(S/S3)$ of renewing cooking learners in the j^{th} orientation, and q indicates a random number chosen from the collection $[1,2,\cdots,m]$, $i = N_{C+1},N_{C+2},\cdots,$ and N, $j = 1, 2, \cdots, m$. When this novel random location leads to improvement of the fitness function value, it is allowable for renewing ZS_i, which is expressed below:

$$ZS_i = \begin{cases} ZS_i^{S/S3}, & FS_i^{S/S3} < F_i; \\ ZS_i, & else, \end{cases} \tag{23}$$

Where, $FS_i^{S/S3}$ denotes the value of the fitness function of $ZS_i^{S/S3}$.

4.4. Iteration process of the CBO algorithm

The iteration of the proposed algorithm is executed by renewing whole individuals of the population. The algorithm executes the subsequent epoch with these novel locations, and the set of chef trainers and cooking learners is determined again. Likewise, the population individuals are renewed during the execution of the CBO algorithm stages on the basis of the stated equations until the ultimate epoch of the method. When the maximum amount of the CBOA repetition parameter is achieved, the finest individual result achieved throughout the execution is offered as the outcome of the problem.

4.5. Improved chef-based optimization (ICBO) algorithm

Several optimization problems may be solved using the original chef-based optimization method, a recently published metaheuristic algorithm. On the other hand, the chef-based optimization technique may face some issues when it comes to solving optimization issues. As a consequence, we developed an improved version of the chef-based optimization algorithm to deal with these problems. To improve the results for the diagnosis of breast cancer, two changes were made to the research.

First, the chaos map method is applied (Ramezani, Bahmanyar, & Razmjooy, 2021). A known approach as a chaos map makes use of unanticipated chaotic variables rather than random variables. Chaos sequences, which may be seen in dynamic and nonlinear systems, are restricted, non-periodic, and non-convergent. They do straightforward explorations quicker as a consequence compared to searches that are possibility-based random ones (Navid Razmjooy & Ghadimi, 2018). Using chaotic variables rather than random variables in metaheuristic approaches allows for effective exploration of the solution space due to the dynamic nature of the turbulence sequence.

Alternative chaos maps were used in optimization approaches to easily build different sequences by changing the conditions at the beginning of the series. In order to improve the CBOA's convergence speed and strike a balance between the two stages of exploitation and exploration, the sinusoidal chaotic map function is utilized in this study. By doing so, the solution space may be explored more effectively, and the local optimum can be avoided. To modify the CBOA by the use of this chaos map, the random quantities of the Chaos function are used in place of the r random value. The sinusoidal map's formula is as follows:

$$r^* = P.r^2\sin(\pi.r) \tag{24}$$

Where, r describes a chaotic random amount that is produced in the current epoch, $P = 2.3$ defines a tunable variable and r is set to 0.5.

Like other iterative optimization techniques, the CBOA algorithm starts the optimization by producing the primary swarm at random. A regulator variable is, therefore, required to specify the size of the population. A fundamental problem is that estimating the population number in a problem is a complex and time-consuming operation. The self-adaptive population alters the population size with each repetition. The population amount is automatically changed on every iteration, which means the user does not have to bother about it. The following equation determines the initial population size before the primary loop of the algorithm:

$$PopSize = 10 \times D \tag{25}$$

Where, D describes the problem dimensions.

Next, the following steps are taken to reach the updated new population size:

$$PopSize_{new} = round(PopSize + \delta^*PopSize) \tag{26}$$

Where, δ refers to an arbitrary number that falls in the interval [-0.5, 0.5]. Up to 50% of the current population size can be changed, either by increasing or decreasing the population. It can increase or decrease the

population by as much as 50% of the current level. Whole individuals of the present population are kept, and the new population is generated via elitism if the population size gained for the future epoch is greater than the population size acquired for the preceding epoch ($PopSizenew > PopSize$).

4.6. Algorithm authentication

In this section, three well-known metaheuristic optimization techniques, including the Multi-Verse Optimizer (MVO) (Mirjalili, Mirjalili, & Hatamlou, 2016), Owl Search Algorithm (OSA) (Jain, 2018), Pigeon-Inspired Optimization Algorithm (PIO) (Cui, 2019), and the original CBOA (Trojovská & Dehghani, 2022) have been compared with the improved Chef-Based Optimization algorithm. Six benchmark functions are utilized for validation of the proposed method compared with others (Razmjooy, xxxx). The utilized functions are illustrated in Table 1.

Table 2 exemplifies the parameter adjustment for the studied algorithms.

To produce a fair comparison, the algorithms' population size and maximum iteration are set to 45 and 200 for all algorithms, respectively. Similar to this, every method is run 35 times on each of the six test functions to provide a reliable result. The mean value (Avg) and standard deviation (StD) values for each algorithm are employed to test the case suite, and the findings are represented in the following (Table 3).

As can be observed in Table 3, the proposed ICBO algorithm, which is more efficient than the original version, delivers the lowest average value for all standard test functions, followed by the suggested Original CBOA method. Additionally, the prominence of the proposed ICBO algorithm relative to the others is highlighted compared to others. Along with being accurate, the suggested approach with the lowest standard deviation value also demonstrates superior dependability over several runs.

5. Feature extraction

The reduction of the feature space dimension is one of the crucial elements in the accuracy and effectiveness of classifiers. The decrease in feature space size is mostly two main factors: decreased computing cost and increased classification accuracy. Feature extraction and feature selection are typically the two steps involved in decreasing the dimension of a feature space. Selecting a feature allows the creation of an object with less information using feature extraction. To individually characterize each collection of traits, these attributes must have certain qualities. The two samples cannot be separated in the classification section by any category if the collection of these traits is the same for the two samples. The segmented brain tumor pictures used in this investigation were extracted using 19 common image attributes. Then, by deleting comparable features and lowering the overfitting value, the suggested ISOA has been used to accomplish optimal feature selection to enhance the system's speed and performance. The following is the formulation for the features under consideration.

$$Area = \sum_{i=1}^{M}\sum_{j=1}^{N} p(i,j) \tag{27}$$

Table 2
Parameters adjustment for the studied algorithms.

Multi-verse optimizer (MVO) (Mirjalili et al., 2016)	WEP_{min}	0.1
	WEP_{max}	0.9
	$Coefficient(P)$	5
Owl Search Algorithm (OSA) (Jain, 2018)	T_{dead}	16
	$\|P\|$	8
	Acc_{low}	0.1
	Acc_{high}	1
Pigeon-inspired Optimization Algorithm (PIO) (Cui, 2019)	Number of Pigeons	200
	Space dimension	15
	Map and compass factor	0.1
	Map and compass operation limit	165
	Landmark operation limit	200
	Inertia factor (w)	1
	Self-confidence factor (c_1)	1.1
	Swarm confidence factor (c_2)	1.1

Table 3
Comparison outcomes of the investigated algorithms.

Function	Algorithms					
	MVO (Mirjalili et al., 2016)		OSA (Jain, 2018)		PIO (Cui, 2019)	
	AVG	SD	AVG	SD	AVG	SD
F_1	5.63 E-7	7.15E-7	3.36 E-8	9.29–8	5.59 E-9	3.93E-10
F_2	8.59E-5	10.36E-5	7.41E-6	9.81E-6	5.55 E-7	8.86E-7
F_3	0.83	0.59	1.088	1.0063	1.036	1.0031
F_4	0.077	0.055	0.043	0.035	0.039	0.029
F_5	−198.68	39.35	−213.34	30.28	−292.34	28.16
F_6	4.28E-8	8.33–8	5.75E-9	9.92E-9	6.36E-10	10.92E-10

Function	Algorithm			
	CBOA		ICBO	
	AVG	SD	AVG	SD
F_1	7.10 E-11	5.32E-12	8.95 E-12	7.27E-13
F_2	8.51 E-8	10.39E-8	5.38E-9	7.86E-9
F_3	1.011	0.99	0.87	0.52
F_4	0.0058	0.0026	0.0048	0.002
F_5	−305.89	20.09	−386.81	12
F_6	5.62 E-11	9.48E-11	5.49 E-12	7.67E-12

Table 1
Utilized benchmark functions for validation of the proposed method.

Function Name	Function equations	Dim	Range	F_{min}
Sphere	$F_1(x) = \sum_{i=1}^{n} x_i^2$	30	[−100,100]	0
Schwefel2.22	$F_2(x) = \sum_{i=1}^{n} \|X_i\| + \prod_{i=1}^{n} \|x_i\|$	30	[−10,10]	0
Rosenbrock's	$F_3(x) = \sum_{i=1}^{n-1} [10^2 (x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$	30	[−30,30]	0
Quartic	$F_4(x) = \sum_{i=1}^{n} ix_i^4 + random[0,1)$	30	[−128,128]	0
Schwefe	$F_5(x) = \sum_{i=1}^{n} -x_i \sin(\sqrt{\|x_i\|})$	30	[−500,500]	−418.9829
Ackley	$F_6(x) = -20\exp(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}) - \exp(\frac{1}{n}\sum_{i=1}^{n} \cos(2\pi x_i)) + 20 + e$	30	[−32,32]	0

$$Contrast = \sum_{i=1}^{M}\sum_{j=1}^{N} p(i,j)(i-j)^2 \tag{28}$$

$$Rectangularity = \frac{Area}{a \times b} \tag{29}$$

$$Entropy = -\sum_{i=1}^{M}\sum_{j=1}^{N} p(i,j)\log p(i,j) \tag{30}$$

$$Perimeter = \sum_{i=1}^{M}\sum_{j=1}^{N} b_p(i,j) \tag{31}$$

$$Homogeneity = \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{p(i,j)}{1+|i-j|} \tag{32}$$

$$Elongation = \frac{2\sqrt{Area}}{a\sqrt{\pi}} \tag{33}$$

$$Mean = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} p(i,j) \tag{34}$$

$$Variance = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} (p(i,j)-\mu) \tag{35}$$

$$StD = Variance^{\frac{1}{2}} \tag{36}$$

$$Irregularityindex = 4\pi \times \frac{Area}{Perimeter^2} \tag{37}$$

$$Eccentricity = 2a^{-1}\left(a^2-b^2\right)^{0.5} \tag{38}$$

$$Formfactor = \frac{Area}{a^2} \tag{39}$$

$$Energy = \sum_{i=1}^{M}\sum_{j=1}^{N} p^2(i,j) \tag{40}$$

$$Correlation = \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{p(i,j)-\mu_r\mu_c}{\sigma_r\sigma_c} \tag{41}$$

$$\begin{aligned}
\varphi_1 &= \eta_{20}+\eta_{02}\\
\varphi_2 &= (\eta_{20}-\eta_{02})^2+4\eta_{11}^2\\
\varphi_3 &= (\eta_{30}-3\eta_{12})^2+(3\eta_{21}-\mu_{03})^2\\
\varphi_4 &= (\eta_{30}+3\eta_{12})^2+(3\eta_{21}+\mu_{03})^2
\end{aligned} \tag{42}$$

Here, *bp* signifies the length of an exterior side of the border pixel, a and b are, in turn, the major and minor axes, and μ and σ define, in turn, the mean value and the standard deviation value. Additionally, $p(i,j)$ defines the intensity value of the pixel at point (i,j), and *MN* indicates the image size.

6. Features selection

One method for boosting the effectiveness and speed of digital image processing is feature selection. The feature selection approach selects feature from a whole set of characteristics that are helpful in categorization. Because there are numerous characteristics in these tasks, most of which are either irrelevant or have a low information burden, feature selection is crucial in many applications (such as classification). Although removing these aspects does not result in any information issues, it does make the targeted application's computing workload heavier and saves a lot of worthless data along with valuable data.

Some characteristics are more crucial than others regarding cancer diagnosis, while others may be overlooked. The Matthew correlation coefficient is used as the cost function in this work to provide an optimum feature selection approach based on the suggested ICBO algorithm. The Matthew correlation coefficient determines the relationship between the TP rate and FP rate, i.e.,

$$Costfunction = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TN+FP)\times(TP+FP)\times(TP+FN)\times(TN+FN))}} \tag{43}$$

where, *TP*, *FP*, *TN*, and *FN* describe, in turn, True Positive, False positive, True Negative, and False Negative.

The Matthew correlation coefficient is used as the cost function in this work to optimize the feature selection approach using the suggested ICBO algorithm. The coefficient measures the relationship between the true positive (TP) rate and the false positive (FP) rate.

Initially, all 16 features, including Area, Contrast, Rectangularity, Entropy, Perimeter, Homogeneity, Elongation, Mean, Variance, Standard Deviation (StD), Irregularity Index, Eccentricity, Form Factor, Energy Correlation, and *Phi* features $(\varphi_1, \varphi_2, \varphi_3, \varphi_4)$ were considered for selection. Then, by applying the ICBO algorithm, the features were ranked based on their fitness scores (Matthew correlation coefficient). The top five features that demonstrated the highest relevance for the task were selected for further analysis. Table 4 presents the selected features along with their corresponding fitness scores.

These selected features will be utilized for subsequent image processing tasks, focusing on improving the accuracy and computing the efficiency.

7. ICBO/SqueezeNet structure

This study presented ICBO/SqueezeNet, a useful method for diagnosing breast cancer. In the beginning, input data from a particular dataset is taken into consideration and subjected to pre-processing. Following that, the optimal feature selection approach is used to choose the optimal features. Finally, SqueezeNet, which was trained using the ICBO optimization technique, is used to diagnose breast cancer.

7.1. SqueezeNet structure

Today, high efficiency is typical in many applications. Convolutional Neural Networks were developed for the artificial processing and forward identification of multidimensional data, such as photographs. To better mimic the operation of the human visual system, its structure has been changed.

As with other artificial neural networks, convolutional neural networks are composed of neurons that hold weights and biases, and they incorporate a decision-making process in the upper layers (Alferaidi, 2022). The size of each hidden layer will be quite huge, and its updating process will take a very long time if fully linked layers and standard neural networks are utilized for picture categorization.

The high memory usage caused by convolutional neural networks' many parameters is a serious issue. SqueezeNet is an architecture with 50 times fewer parameters that is identical to the AlexNet design and is ideal for gadgets like mobile phones (Tian, 2021). In reality, a convolution layer can get combined with the pooling layer based on the activation functions.

In SqueezeNet, a technique known as global average pooling is employed in place of their customary fully linked layer. In contrast to networks that use completely connected layers for classification, this technique creates a feature map for each class after the final layer rather

Table 4
Selected features along with their corresponding fitness scores.

Rank	Feature	Fitness Score
1	Area	0.88
2	Contrast	0.81
3	Rectangularity	0.76
4	Entropy	0.71
5	Perimeter	0.69

than adding a whole layer, as is done in networks that utilize fully connected layers.

They are connected on top of feature maps, they profit from feature maps, and Softmax receives the feature vector results directly. Due to the lack of optimization parameters, one advantage of this layer is that it only causes more problems than it solves (Ranjbarzadeh, 2022). On the other hand, this layer is more resilient to local changes and more in line with the idea of convolutional networks.

The SqueezeNet is composed of various fire modules, each of which has a squeeze convolution layer and an expansion layer. The Fire Module of a SqueezeNet is seen in Fig. 4.

The squeeze convolution layer's output is sent to the following expand layer in the fire modules, as seen in Fig. 3. The SqueezeNet also starts with a single convolution layer and progresses through eight fire modules before arriving at the final convolution layer. Additionally, SqueezeNet performs the maximum pooling operation in two steps, as shown in Fig. 5.

7.2. Optimized SqueezeNet based on the ICBO algorithm

After choosing the best attributes, it is time to divide the photos into two categories: malignant instances and healthy cases. An improved deep neural network-based approach is used to carry out this operation. In this work, ICBO is used to create the best classifier possible using SqueezeNet.

The maximum number of iterations and the population size are both set at 100. The Mean Square Error (MSE) between the goal value and the output of the neural network is the cost function for reducing SqueezeNet.

$$Ft = \frac{1}{M}\sum_{j=1}^{M}\left(D_j - Y_j\right)^2 \tag{44}$$

Where, M represents the total number of samples, and D_j and Y_j describe the desired and the output values by the SqueezeNet.

The main purpose is to minimize the cost function using the suggested ICBO algorithm, i.e.,

$$\min Ft \tag{45}$$

8. Dataset

The proposed approach is examined in this study using the Mammographic Image Analysis Society (MIAS) database (Suckling, 2019), which is collected by a team of UK academics donated this database to be used in other mammography research projects, including those involving medical imaging. This database includes 322 digital mammography pictures that were gathered by specialists and marked with all abnormal areas to evaluate the effectiveness of the approaches. To speed up processing and simplify the programming, the MIAS dataset's original size, which is 1024 after filtering, was cut in half for the presented research (Fig. 6).

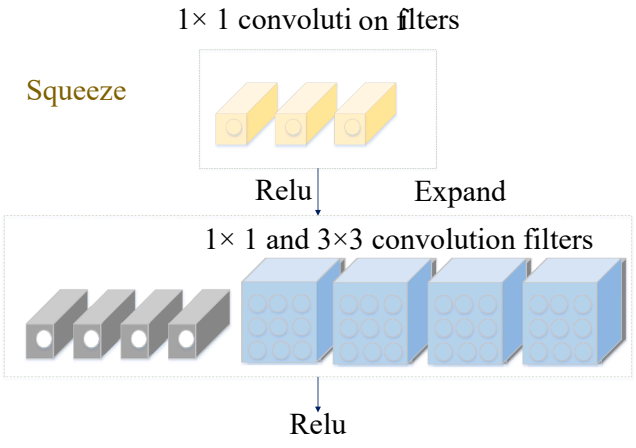


Fig. 4. The fire module in SqueezeNet.

9. Simulations and result

As was already explained, there are three basic phases to the process. Noise reduction is done initially. In this stage, the noise in the original input mammography pictures has been removed using median filtering. The pre-processed pictures were then threshold to separate the suspicious regions based on the Kapur approach. The superfluous portions of the segmented picture are also removed using mathematical morphology, which includes mathematical closure, opening, and filling. The segmented image's features are extracted in the second stage. To deliver just useful characteristics, the feature selection approach is used.

The Improved Chef-Based Optimization (ICBO) algorithm, a recently discovered metaheuristic technique, is used to do this. The classifying process is the third. In this stage, an optimal SqueezeNet model has been designed. Then, the efficiency of SqueezeNet was increased using a new, enhanced version based on the ICBO algorithm. The MIAS database has been used to simulate the process. To implement the code, a 64-bit MATLAB R2017b was employed. The system configuration is shown in Table 5.

In the classification step, a total of 322 images (before augmentation) and 622 images (after augmentation) were used, while 257 images (before augmentation) and 497 images (after augmentation) are selected to train their model, representing 80% of the available data. This training set was used to learn the model to recognize patterns and features in the images that correspond to different types of breast tissue, including benign and malignant tumors. The remaining 20% of the available data, or 65 images (before augmentation) and 125 images (after augmentation), were used as a test set to evaluate the performance of the trained model. This method of dividing the available data into training and test sets is a common practice in machine learning and is known as cross-validation. By using a separate set of images for testing, the authors were able to assess the accuracy of their model in identifying breast tissue types that were not included in the training set. It is important because it allows the researchers to determine if the model is capable of being generalized to new and unseen data. It should be noted that the specific ratio of training to test data used in this study may have been chosen based on factors, such as the size of the available dataset, the complexity of the classification task, and the computational resources available for training the model. However, the use of a separate test set is an essential step in the machine-learning pipeline and is crucial for evaluating the effectiveness of a trained model. For the initial analysis, precision, specificity, accuracy, sensitivity, MCC (Matthew's Correlation Coefficient), and F1 score are analyzed. The F1 score is a statistical measure commonly used to evaluate the similarity between two sets of data or two binary variables. The advantages of the F1 score include calculation simplicity, robustness, high sensitivity, versatility, and popularity. Overall, the F1-score is a useful and versatile similarity measure that provides a simple and effective way to evaluate the similarity between two sets of data. The explained indicators' mathematical formulas are given in the following:

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{46}$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \tag{47}$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{48}$$

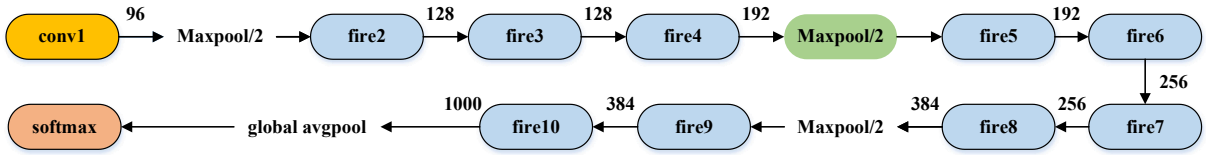


Fig. 5. The SqueezeNet.

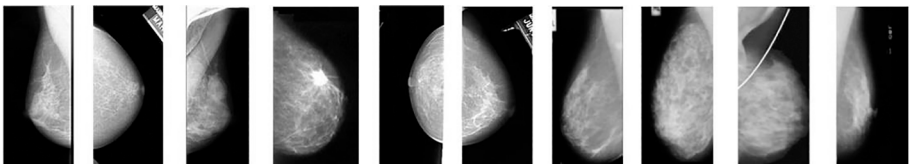


Fig. 6. Some samples of the MIAS database.

Table 5
System configuration.

Name	Setting
Hardware	Xeon Processors (1 core, 2 threads)
CPU	2.3 GHz
Disk space	100 GB
RAM	13 GB
GPU memory	16 GB
GPU	Nvidia K80 GPU
Cache	46 MB

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{49}$$

$$F1 - score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \times 100 \tag{50}$$

$$MCC = \frac{TP \times TN - TP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \times 100 \tag{51}$$

Here, the *TP*, *FP*, *FN*, and *TN* describe, in turn, True Positive, False Positive, False Negative, and True Negative.

Table 6 illustrates the simulation results of the suggested strategy for the indicators in contrast to the simple SqueezeNet, SqueezeNet/CBOA; moreover, the suggested SqueezeNet/ICBO demonstrate its effectiveness.

Fig. 7 illustrates the classification examination for the studied methods based on a bar plot for more clarification.

As can be observed from the results, the proposed Modified SqueezeNet/CBOA method performs better than the conventional SqueezeNet and the SqueezeNet based on original BOA in terms of accuracy (95%), precision (97%), and specificity (97.99%). Based on the information provided, the suggested solution achieves high precision, accuracy, and specificity, which are important measures of the performance of a classification system. Precision is the fraction of the true positive predictions among all the positive predictions made by the model, and it measures how well the model identifies positive samples. A high precision indicates that the model has a low false positive rate. Accuracy is the fraction of the correct predictions made by the model, and it measures how well the model classifies samples in general. A high accuracy indicates that the model can classify samples correctly. Specificity is the fraction of the true negative predictions among all the negative predictions made by the model, and it measures how well the model identifies negative samples. A high specificity indicates that the model has a low false negative rate. Therefore, based on the reported precision, accuracy, and specificity values, it can be concluded that the Modified SqueezeNet/CBOA method has a high performance in classifying samples, with a good balance between true positive and true negative predictions. However, it is important to note that the performance of the proposed method should be evaluated in comparison with other state-of-the-art methods to determine its true effectiveness.

Six alternative models, including Multi-Image Modalities (MIM) (Mahmood, 2020), Convolutional Neural Network (CNN) (Chougrad, Zouaki, & Alheyane, 2018), Deep Belief Networks (DBN) (Abdel-Zaher & Eldeib, 2016), MR Imaging (MRI) (Orel & Schnall, 2001), Multiple Instance Learning (MIL) (Sudharshan, 2019), Interval-Based (IB) (Liu, 2020) have been used for analysis and choosing the optimal kernel type. the *TP*; *FP*, *FN*, and *TN* describe, in turn, True Positive, False Positive, False Negative, and True Negative. Equations (52) and (53) provide the mathematical formula for these indicators:

$$PPV = \frac{TP}{TP + FP} \tag{52}$$

$$NPV = \frac{TN}{TN + FN} \tag{53}$$

The comparison outcomes of several breast cancer diagnostic algorithms applied to the MIAS dataset are shown in Table 7.

The results presented in Table 6 suggest that the proposed Modified SqueezeNet/CBOA method is more effective for detecting breast cancer from mammograms than conventional methods. The reported accuracy of 92% is the highest among the compared methods, indicating that the proposed method is better at correctly classifying mammograms as either positive or negative for breast cancer. The Negative Predictive Value (NPV) and Positive Predictive Value (PPV) of the proposed method are also reported to be 80% and 77%, respectively.

These values indicate that the proposed method is more accurate in identifying negative and positive findings, respectively, and can better distinguish between true negative and true positive results.

Table 6
Comparison results of breast cancer images' classification based on the proposed SqueezeNet/ICBO toward two other techniques.

Procedure	Precision	Specificity	Sensitivity	MCC	Accuracy	F1-score
SqueezeNet	92.99	93.00	78.00	54.00	92.00	85.00
SqueezeNet/ICBO	96.00	96.00	84.99	68.99	93.99	87.99
SqueezeNet/ICBO	97.00	97.99	93.00	94.00	95.00	90.00

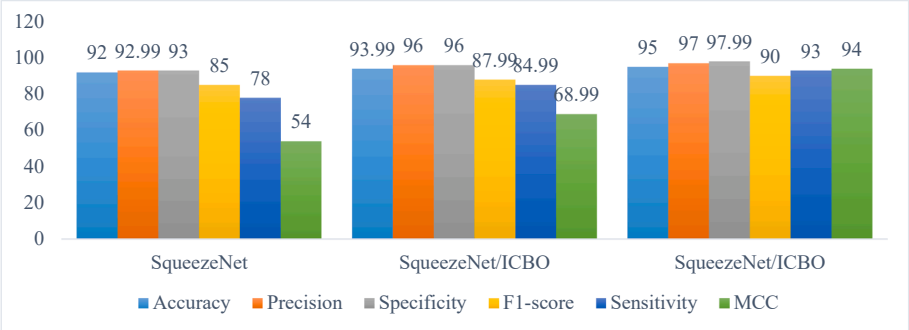


Fig. 7. Results of comparing the classification of breast cancer images using the proposed SqueezeNet/ICBO approach in contrast to two other techniques.

Table 7
Comparison analysis of different diagnosis techniques applied to the MIAS dataset.

Method	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)
SqueezeNet/ ICBO	62.00	86.00	77.00	80.00	95.00
MIM (Mahmood, 2020)	48.00	73.00	48.00	67.00	74.00
CNN (Chougrad et al., 2018)	56.00	66.00	60.00	75.00	87.00
DBN (Abdel-Zaher & Eldeib, 2016)	70.00	48.00	57.00	74.00	84.00
MRI (Orel & Schnall, 2001)	72.00	49.00	59.00	76.00	75.00
MIL (Sudharshan, 2019)	76.00	73.00	62.00	73.00	80.00
IB (Liu, 2020)	48.00	50.00	44.00	58.00	76.00

Furthermore, the proposed method is reported to have higher sensitivity and specificity values compared to the other methods. Sensitivity is the proportion of true positive samples that are correctly identified by the model, while specificity is the proportion of true negative samples that are correctly identified by the model. The higher sensitivity and specificity values of the proposed method indicate that it is more successful in correctly identifying both positive and negative samples, respectively.

Overall, the results suggest that the proposed Modified SqueezeNet/CBOA method is a more efficient and accurate approach for breast cancer detection from mammograms. However, it is important to note that these results may depend on the specific dataset and experimental setup used, and further evaluation on larger and more diverse datasets may be needed to establish the generalizability of the proposed method.

10. Evaluating algorithms using a 5-fold cross-validation technique

5-fold cross-validation is a commonly used method that can help to estimate the performance of a model on new data. By using cross-validation, we can assess how well the model generalizes to new data and also detect any overfitting issues. Reporting the averaged results from cross-validation can provide a more accurate estimate of the model's performance than simply reporting the results from a single train-test split. Therefore, in the present study, we also utilized the 5-fold cross-validation method to evaluate the performance of various classification algorithms for breast cancer classification using the Wisconsin Breast Cancer dataset. The average performance metric was calculated by taking the mean of the performance scores obtained from each five-fold. The results obtained were tabulated in Table 8.

Table 8
Average evaluation criteria using 5-fold cross-validation.

Method	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)	AUC (%)
SqueezeNet/ICBO	80.00	84.00	81.00	80.00	90.00
MIM (Mahmood, 2020)	76.00	76.00	69.00	67.00	86.00
CNN (Chougrad et al., 2018)	70.00	69.00	73.00	75.00	88.00
DBN (Abdel-Zaher & Eldeib, 2016)	68.00	51.00	75.00	74.00	87.00
MRI (Orel & Schnall, 2001)	67.00	52.00	76.00	76.00	87.00
MIL (Sudharshan, 2019)	68.00	75.00	75.00	73.00	85.00
IB (Liu, 2020)	77.00	62.00	57.00	58.00	76.00

According to the results shown in Table 7, the SqueezeNet/ICBO method performed better than the other algorithms in terms of various performance metrics. The proposed algorithm achieved high Recall (80%), Precision (84%), F1-Score (81%), and Accuracy (80%) values, indicating its effectiveness in correctly identifying positive cases of breast cancer. Furthermore, SqueezeNet/ICBO also showed the highest Area Under the Curve (AUC) score of 90%, which is a widely used metric for evaluating the overall performance of a classification model. It suggests that the proposed method is capable of achieving high true positive rates while keeping the false positive rates relatively low. Based on these experimental findings, it can be confidently concluded that SqueezeNet/ICBO is the most accurate algorithm for classifying breast cancer using the Wisconsin Breast Cancer dataset. These results are significant as breast cancer is a serious health issue affecting millions of women worldwide, and accurate diagnosis is critical for effective treatment and improved patient outcomes.

The comment raises a valid concern about the potential impact of data augmentation by rotating/flipping entire volumes in CT examinations. The concern is that this approach may cause the patient's body to be localized in non-standard positions, which could affect the accuracy and reliability of the classification results. To address this concern, the applicability of data augmentation by rotating/flipping entire volumes should be validated through appropriate experiments and comparisons with other methods. One such method could be augmentation by rotating small image patches, which may be less likely to cause distortions in the patient's body position.

By comparing the performance of these two approaches, it would be possible to determine which method is more effective and reliable for the specific task and dataset at hand. It is important to note that the choice of data augmentation method should be based on the characteristics of the dataset, the complexity of the task, and the limitations of the available computational resources.

Because data augmentation by rotating/flipping the entire volume

Table 9
Average evaluation criteria using 5-fold cross-validation for the proposed method before and after data augmentation.

Method	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)	AUC (%)
Without augmentation	80.00	84.00	81.00	80.00	90.00
With augmentation	86.50	90.00	88.00	92.00	98.00

causes the patient's body to be localized in non-standard positions in the CT examination, its applicability should be validated in the following. This validation has been assessed by adding 300 more augmented images. Table 9 indicates the average performance metric five-fold validation based on the proposed method before and after augmentation.

The statement refers to the use of data augmentation techniques in classification tasks, which can improve the performance of the model in various ways. Data augmentation involves creating new training data by applying different transformations to the original data, such as rotating and flipping images. Using data augmentation, the model is exposed to more diverse examples of the same class, which can help it learn more robust and invariant features relevant to classification. It can make the model less sensitive to variations in the input data, such as changes in lighting, orientation, or background, which can improve its accuracy and generalization ability. In addition, data augmentation can increase the size of the training set, which can prevent overfitting and improve the model's ability to generalize to new, unseen data. It is particularly important when the available training data is limited or imbalanced, as it can help to mitigate the effects of bias and variance in the model. Overall, using augmentation in classification can lead to a higher performance in terms of accuracy, robustness, generalization, and efficiency by leveraging the power of data diversity and variability.

11. Conclusions

One of the most prevalent causes of death in females is breast cancer, which is caused by the abnormal growth of breast cells. Imaging is the most frequent technique in the diagnosis of this kind of cancer. Imaging can be done in different ways, the most common of which is mammography. The main purpose of this study was to propose a method on the basis of a metaheuristic algorithm and deep learning for the accurate diagnosis of breast cancer mammogram images. The proposed approach was divided into three fundamental steps. Noise reduction is the first part. By the use of median filtering, the noise in the original input mammography images is eliminated at this step. The suspect regions were then threshold out of the pre-processed images using the Kapur technique. Then, to eliminate pointless features, an optimum feature selection was used. It was accomplished using an Improved version of the Chef-Based Optimization (ICBO) algorithm. Finally, classification was performed using the best SqueezeNet model. Then, a new, improved version of SqueezeNet was designed on the ICBO algorithm to maximize its effectiveness. The technique was eventually used on the MIAS database, and its outcomes were compared with those of SqueezeNet/CBOA, proposed SqueezeNet/ICBO, and various other cutting-edge techniques, including Multi-Image Modalities (MIM), Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), MR Imaging (MRI), Multiple Instance Learning (MIL), and Interval-Based (IB) demonstrate the method's superiority. One main shortcoming of the proposed method due to its high complexity is speed. Therefore, using a technique to improve the speed of the proposed system along with its accuracy can be a good idea for future research.

CRediT authorship contribution statement

Qirui Huang: Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Huan Ding:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Mehdi Mehdi Effatparvar:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. *Expert Systems with Applications, 46*, 139–144.

Akbary, P., et al. (2019). Extracting appropriate nodal marginal prices for all types of committed reserve. *Computational Economics, 53*(1), 1–26.

Alferaidi, A., et al. (2022). Distributed Deep CNN-LSTM Model for Intrusion Detection Method in IoT-Based Vehicles. *Mathematical Problems in Engineering, 2022*.

Bagheri, M., et al. *A novel wind power forecasting based feature selection and hybrid forecast engine bundled with honey bee mating optimization.* in *2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)*. 2018. IEEE.

Bourouis, S., et al. (2022). Meta-heuristic algorithm-tuned neural network for breast cancer diagnosis using ultrasound images. *Frontiers in Oncology, 12*, Article 834028.

Cai, W., et al. (2019). Optimal bidding and offering strategies of compressed air energy storage: A hybrid robust-stochastic approach. *Renewable Energy, 143*, 1–8.

Cai, X., et al. (2021). Breast Cancer Diagnosis by Convolutional Neural Network and Advanced Thermal Exchange Optimization Algorithm. *Computational and Mathematical Methods in Medicine, 2021*.

Cao, Z., et al. (2020). Breast tumor classification through learning from noisy labeled ultrasound images. *Medical Physics, 47*(3), 1048–1057.

Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep convolutional neural networks for breast cancer screening. *Computer methods and programs in biomedicine, 157*, 19–30.

Cui, Z., et al., *A pigeon-inspired optimization algorithm for many-objective optimization problems.* Sci. China Inf. Sci., 2019. 62(7): p. 70212:1-70212:3.

Dehghani, M., et al. (2020). Blockchain-based securing of data exchange in a power transmission system considering congestion management and social welfare. *Sustainability, 13*(1), 1.

Ebrahimian, H., et al. (2018). The price prediction for the energy market based on a new method. *Economic research-Ekonomska istraživanja, 31*(1), 313–337.

Eslami, M., et al., *A New Formulation to Reduce the Number of Variables and Constraints to Expedite SCUC in Bulky Power Systems.* Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, 2018: p. 1-11.

Fan, X., et al. (2020). High voltage gain DC/DC converter using coupled inductor and VM techniques. *IEEE Access, 8*, 131975–131987.

Firouz, M. H., & Ghadimi, N. (2016). Concordant controllers based on FACTS and FPSS for solving wide-area in multi-machine power system. *Journal of Intelligent & Fuzzy Systems, 30*(2), 845–859.

Gao, W., et al. (2019). Different states of multi-block based forecast engine for price and load prediction. *International Journal of Electrical Power & Energy Systems, 104*, 423–435.

Guo, Z., et al. (2021). Novel computer-aided lung cancer detection based on convolutional neural network-based and feature-based classifiers using metaheuristics. *International Journal of Imaging Systems and Technology*.

Guo, G., & Razmjooy, N. (2019). A new interval differential equation for edge detection and determining breast cancer regions in mammography images. *Systems Science & Control Engineering, 7*(1), 346–356.

Huang, H., et al. (2019). A new fruit fly optimization algorithm enhanced support vector machine for diagnosis of breast cancer based on high-level features. *BMC Bioinformatics, 20*(8), 1–14.

Jabeen, K., et al. (2022). Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. *Sensors, 22*(3), 807.

Jain, M., et al. (2018). Owl search algorithm: A novel nature-inspired heuristic paradigm for global optimization. *Journal of Intelligent & Fuzzy Systems, 34*(3), 1573–1582.

Liu, Q., et al. (2020). Computer-aided breast cancer diagnosis based on image segmentation and interval analysis. *Automatika, 61*(3), 496–506.

Mahmood, T., et al. (2020). A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities. *IEEE Access, 8*, 165779–165809.

Maqsood, S., Damaševičius, R., & Maskeliūnas, R. (2022). TTCNN: A breast cancer detection and classification towards computer-aided diagnosis using digital mammography in early stages. *Applied Sciences, 12*(7), 3273.

Meraj, T., et al. (2021). A quantization assisted U-Net study with ICA and deep features fusion for breast cancer identification using ultrasonic data. *PeerJ Computer Science, 7*, e805.

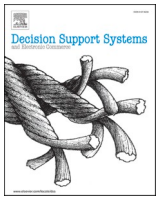
Mirjalili, S., Mirjalili, S. M., & Hatamlou, A. (2016). Multi-verse optimizer: A nature-inspired algorithm for global optimization. *Neural Computing and Applications, 27*(2), 495–513.
Navid Razmjooy, F. R. S., & Ghadimi, N. (2018). A Hybrid Neural Network – World Cup Optimization Algorithm for Melanoma Detection. *Open Medicine, 13*, 9–16.
Orel, S. G., & Schnall, M. D. (2001). MR imaging of the breast for the detection, diagnosis, and staging of breast cancer. *Radiology, 220*(1), 13–30.
Ramezani, M., Bahmanyar, D., & Razmjooy, N. (2021). A New Improved Model of Marine Predator Algorithm for Optimization Problems. *Arabian Journal for Science and Engineering*, 1–24.
Ranjbarzadeh, R., et al. (2022). Nerve optic segmentation in CT images using a deep learning model and a texture descriptor. *Complex & Intelligent Systems*, 1–15.
Razmjooy, N., et al. (2020). Computer-aided diagnosis of skin cancer: A review. *Current medical imaging, 16*(7), 781–793.
Razmjooy, N., M. Ashourian, and Z. Foroozandeh, *Metaheuristics and Optimization in Computer and Electrical Engineering*. Springer.
Sha, Z., Hu, L., & Rouyendegh, B. D. (2020). Deep learning and optimization algorithms for automatic breast cancer detection. *International Journal of Imaging Systems and Technology, 30*(2), 495–506.
Stephan, P., et al. (2021). A hybrid artificial bee colony with whale optimization algorithm for improved breast cancer diagnosis. *Neural Computing and Applications, 33*(20), 13667–13691.
Suckling, J. *The mini-MIAS database of mammograms*. 2019; Available from: http://peipa.essex.ac.uk/info/mias.html.
Sudharshan, P., et al. (2019). Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications, 117*, 103–111.
Tian, Q., et al. (2021). A New optimized sequential method for lung tumor diagnosis based on deep learning and converged search and rescue algorithm. *Biomedical Signal Processing and Control, 68*, Article 102761.
Trojovská, E., & Dehghani, M. (2022). A new human-based metahurestic optimization method based on mimicking cooking training. *Scientific Reports, 12*(1), 1–24.
Xu, Z., et al. (2020). Computer-aided diagnosis of skin cancer based on soft computing techniques. *Open Medicine, 15*(1), 860–871.
Xu, Y., Wang, Y., & Razmjooy, N. (2022). Lung cancer diagnosis in CT images based on Alexnet optimized by modified Bowerbird optimization algorithm. *Biomedical Signal Processing and Control, 77*, Article 103791.
Zebari, D. A., et al. (2021). Breast cancer detection using mammogram images with improved multi-fractal dimension approach and feature fusion. *Applied Sciences, 11*(24), 12122.

Decision Support Systems 175 (2023) 114039

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



All eyes on me: Predicting consumer intentions on social commerce platforms using eye-tracking data and ensemble learning

Patrick Mikalef [a,b,*], Kshitij Sharma [a,c], Sheshadri Chatterjee [d], Ranjan Chaudhuri [e], Vinit Parida [f,g,h], Shivam Gupta [i]

[a] Department of Computer Science, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Norway
[b] Department of Technology Management, SINTEF Digital, Trondheim, Norway
[c] University of Science and Technology, Trondheim, Norway
[d] Department of Computer Science & Engineering, Indian Institute of Technology Kharagpur, India
[e] Indian Institute of Management Ranchi, Jharkhand, India
[f] Department of Social Sciences, Technology and Arts, Luleå University of Technology, Sweden
[g] Entrepreneurship and Innovation, Luleå University of Technology, Sweden
[h] School of Management, University of Vaasa, Vaasa, Finland
[i] Department of Information Systems, Supply Chain Management & Decision Support, NEOMA Business School, Reims, France

ARTICLE INFO

Keywords:
Eye-tracking
Social commerce
Ensemble learning
Prediction
Machine learning

ABSTRACT

Understanding what information is important for consumers when making a purchase-related decision has been a key question for researchers and practitioners ever since the advent of empirical research in commerce. Nevertheless, our knowledge of what information is important has been formed primarily through post-purchase conscious capturing approaches, such as surveys and questionnaires. To overcome these limitations, we ground this research on an exploratory study that captures eye-tracking data during a decision-making task of product selection. Grounded on the dynamic attention theory, we utilize different information types and formats present on a popular social commerce platform, to identify elements which are important when deciding about online product purchase decision. Specifically, we employ a series of prediction algorithms and use an ensemble learning setup to predict the aspects that contribute to product selection by consumers. Our analysis highlights the most important informational cues to accurately predict product selection among alternatives. In addition, the results showcase how such elements shift in importance during the temporal sequence of comparing different product alternatives. Our results provide insight into how we can understand the journey of decision-making for social commerce customers when navigating through information to select a product. In addition, it opens the discussion about the shifts that eye-tracking in combination with machine learning can create for researchers and marketers.

1. Introduction

Ever since the emergence of online commerce, a primary question for research and practice has been to understand what informational cues trigger consumers to select certain products during their online purchases [32,74,75]. A key aspect in this quest has been to present consumers with sufficient information about a product to make ease the decision-making process [16,43]. Social commerce platforms which have become increasingly prevalent over the last decade have added to the complexity of this task, as they typically incorporate both marketer- and consumer-generated information about a product [17,44,80]. In addition, such information is often presented in different formats, which facilitates consumers to select from a variety of informational cues that will ease the selection of products [6]. Yet, while there has been a significant advancement in the front of presenting information to consumers, we still have limited knowledge on how consumers utilize such information, and more importantly what type of information is critical for them during their decision-making process [59]. Furthermore, we have limited knowledge on how the importance of different informational cues evolves during the product selection process [49].

* Corresponding author.
E-mail addresses: patrick.mikalef@ntnu.no (P. Mikalef), kshitij.sharma@ntnu.no (K. Sharma), ranjan@nitie.ac.in (R. Chaudhuri), vinit.parida@ltu.se (V. Parida), shivam.gupta@neoma-bs.fr (S. Gupta).

https://doi.org/10.1016/j.dss.2023.114039
Received 2 December 2022; Received in revised form 30 May 2023; Accepted 14 June 2023
Available online 17 June 2023

A growing stream of research has sought to understand more about how users of social commerce platforms process different formats of information by capturing physiological data through eye-tracking approaches [7,10,23,71]. Eye movements and attributes of the pupil can help us understand the connection between eye movements and cognitive processes, and specifically how visual stimuli affect humans in decision making processes [4]. Specifically, the eye-mind hypothesis argues that the direction of our eye movements helps to assess what an individual is thinking and where the attention of the individual is focused [35]. The resulting data gathered from eye tracking methods can be analyzed statistically and graphically to provide evidence towards specific visual patterns and the sequence of gazing [11]. By examining the saccades along with other patterns of eye movements, researchers can ascertain the attractiveness of a given product, as well as the interaction of the user with different types and formats of information that are provided [45]. Such analyses provide valuable insight into what informational cues have the highest impact and which ones are ignored by users during their decision-making process [9]. Furthermore, such approaches can enable researchers and practitioners to understand how the attention of individuals dynamically changes over time during the decision-making process [58,76].

Prior studies have provided us with rich insight on the importance of customer satisfaction during the decision-making process of product selection [78] and its significance on prompting purchase intentions [33,36,37,40]. An important part in realizing customer satisfaction has to do with the quality information that is provided to users during their decision-making process [17,60,78,80]. Delving into this topic through eye-tracking methods, recent studies have found that users tend to complement different types and formats of information when deciding on which products to purchase [71]. Fei et al. [23] find that the type of information presented on online commerce platforms and the format or presentation have differentiating effects on consumer intentions. Building on an e-commerce platform using an eye-tracking approach, Brand and Reith [10] show that that both the type and representation of information, have an impact on how users make decisions on credibility evaluations. Thus, there is a growing consensus that the type and representation format of information that users are provided with on online platforms, has an effect on their decision-making process [2,5,66,81].

While there is a growing understanding concerning the role of information cues on online platforms, we still have limited knowledge on how the utilization of the different types and formats of information influences a user's decision-making process during product selection. In addition, there is still a limited understanding concerning which informational cues are most important during the different phases of decision-making for users of social commerce platforms. From a methodological perspective, eye-tracking approaches provide several strengths in capturing detailed data on the gaze patterns of users, however, a challenge is in making inferences from such data. In this study we build on machine learning approaches to identify aspects of the product selection process that can accurately predict which product will be selected by consumers [25]. Doing so enables us to understand how users utilize different types of information present on social commerce platforms, and how the importance of such information changes during the decision-making process. We, therefore, seek to understand not only what type of insight eye-tracking can enable during the decision-making process of consumers, but also how such insight can be used to optimize the product selection experience. To address this research gap, the aim of this study is to address the following research questions (RQs).

RQ1. What informational cues are important for predicting consumer choices during the decision-making process on social commerce platforms?

RQ2. How does the importance of informational cues for consumers change over time during the decision-making process on social commerce platforms?

To answer these questions, we conducted two separate studies with participants in a lab setting. During these studies, participants were provided with several sets of products to select from on a popular social commerce platform. Using an eye-tracking device, we captured aspects of their visual processing patterns during product selection, as well as the types of information they used on the social commerce platform to reach a final decision. Using this data, we employed several machine learning and prediction algorithms to understand which features were the best predictors of product selection, as well as how these features shifted in importance over the duration of product selection. Our results highlight key elements that play an important part in decision-making for product selection. They also highlight the strengths of combining eye-tracking with advanced ways of analyzing data through machine learning. Based on these results, we draw several key research and practical implications.

2. Theoretical background

2.1. Dynamic attention theory

The Dynamic Human-Centered Communication System theory (referred to as Dynamic Attention theory) [41,73] has been used in this study to understand how the attention of a user in an online platform such as social commerce leads to certain outcomes, and how attention shifts over time. In this theory, attention is considered as a process involving interaction between humans as well as environment [41]. Dynamic attention theory helps to assess the intention of the consumers for the consumption of information and supports interpreting how the consumers make a decision of which product to select. Dynamic attention theory is founded on human-computer interaction (HCI) research which presumes that an active observer is supposed to use stable elements available in the environment which is also known as distributed cognition to facilitate decision-making [56].

In terms of dynamic attention theory, recording of eye tracking highlights that humans tend to acquire information from the available environment by using low-efforts gazing strategies compared to using high-efforts based on a strategy of memorization [42]. This human behavioral attribute has been the driver of design for many social commerce platforms and is the main driver by which user of such online platforms make a decision towards purchasing a product or a service. Understanding and predicting within the context of dynamic attention theory can be studied through the use of eye tracking approaches [70]. By tracking attention through an eye-tracking method, movements of eye occur in the direction of the object which receives more attention in terms of the dynamic attention theory, but "the resulting fixation and fixation durations may not be long enough for memory encoding" [73]. Hence, eye tracking methods can provide a more accurate and nuanced assessment of the implications of this theory.

Through the lens of dynamic attention theory, eye tracking is considered as a specific type of communication between the consumers and the product at hand, which is a complex dynamic system comprising of a human, a message, a congenial medium, and a location [41]. Thus, dynamic attention theory provides enables us to understand how the attention of a user in relation to a product evolves over time while a decision has been made. In the context of this study, dynamic attention theory is used to gain insight into the consumers' decision-making process and to try to understand the key elements that lead to the selection of a specific product. To gain such insight, the use of eye-tracking data is deemed as the most suitable as it is possible to record gaze patterns, saccades, and attention over a product decision-making activity.

2.2. Eye-tracking and decision-making

Applying eye tracking in studies of human behavior aids in capturing real-time information on individuals' fixations and visualization patterns as consumers; in turn granting researchers the ability to study the

behavioral-environmental processes more effectively behind transaction decisions [19,67]. Eye tracking is considered as a measurement of the movements of eye for determining the gaze location and to ascertain where the attraction is mainly focused [26]. Research using eye-trackers has built on a number of different measures, as for example fixation duration, an indicator of concentration and focus on a particular object or area [68]. There is not a particular study where the authors have provided a universally accepted definition of fixation in the context of eye tracking research because such definition differs between on-screen system of eye tracking and mobile eye tracking system [28]. Fixation may be interpreted as the duration during which the pupils of a person remained still on an object even the object is in movement [29]. Several studies have shown that there are mainly two types of systems that drive human visual attention which include stimuli-related attention known as bottom-up system, and observer related attention which is called top-down system [9].

In the context of eye tracking research, a seminal study by Ellis and Smith [22] highlights that the process of eye movement is either stratified random or completely random or may be statistically dependent stochastic. A study conducted in a physical store by Wästlund et al. [69] showed that when entering a shop, customers tend to typically start browsing the central areas of store shelves, and in sequence proceed to the peripheral areas on the horizontal axis. In this context, Chandon et al. [15] found that products which are placed on the horizontal central line as well as placed in the upper areas are found to have received better attention by the consumers in the context of on-screen self-test approach.

When considering online commerce, several studies have shown that the location of information that is depicted plays a critical role in purchase decision-making, where most tend to first observe the screen from the center and move towards the right side [64]. Researchers have demonstrated that since the central portion is considered as the optimal location containing maximum amount of preliminary information, people are used to have their attention centered to this part of the screen first [3]. Some research has also demonstrated that objects which are placed in the central area of the field of vision are attended by the people in a faster way for fostering initial saccades [13].

Nevertheless, the position of informational cues on the screen is not the only aspect that drives user attention and focus when making purchase decisions on online platforms. Several studies have shown that the type and format of information also play an important role in user attention [23,25,71]. Yang [77] found that positive and negative framing, or peripheral cues, can have an effect on purchase intentions by increasing attention on the cue message. Furthermore, on par with Yang's findings, Zhang and Benyoucef [80] found that there is a positive relationship between peripheral cues and purchase intention. During the decision phase, a consumer's mind is more heavily impacted by negative product reviews than positive ones [63]. As a result, product reviews may directly impact a consumer's decision on whether to purchase a product or service, depending on how they are framed and represented. The influence of information types is also noted in a study by Guerreiro et al. [27] who show that hedonic products present higher levels of fixation than utilitarian products.

While these studies provide valuable insight into the eye-tracking of individuals in relation to decision-making processes of consumers, they come under the limitation that they do not build on appropriate approaches of capturing the dynamic cognitive process that underpin such eye signals. In this study we leverage a series of prediction algorithmics and an ensemble learning setup to uncover how important eye-tracking information can provide us with real-time information about consumers preferences and decision-making processes. From a theoretical point of view, this approach facilitates a better understanding of how visual cues and information consumption influence decision making. From a methodological perspective, the combinations of methods and techniques allows to understand in real-time the types of informational sources that are important for consumers when making a purchase-related decision and to predict with high accuracy the products that will be selected. These insights can help us understand how users of online social commerce platforms utilize different types and formats of information present on them, as well as uncover the temporal sequence of use of such information.

3. Research methodology

3.1. Participants and procedure

We conducted two data collection studies to understand the relation between the customers' decision-making processes and their gaze-patterns. The choice of two studies was done on the basis of minimizing any potential bias due to the types and products that participants were offered to choose from as well as to differentiate the number of products that were needed to be compared to reach a decision. Thus, we followed relevant guidelines to mitigate the potential of bias in our tests by increasing the variance in a controlled manner [14,31]. The main idea for having two different studies (both in the terms of types of products and types of eye-trackers) was to have more variance in the model [8], which could lead to higher generalizability of the outcomes. Most of the measurements that we get from the two different eye-trackers is the same, for example, fixation duration, first fixations in the areas of interest, pupil diameter, saccade length and saccade velocity. Therefore, the two datasets were similar in nature. Moreover, having one head-mounted eye-tracker and one screen-based eye-tracker also allows us to study the consumer behavior in two different but ecologically valid settings. With a single study, the model would have been biased towards a specific setting for data collections and types of products. This would hinder the generalizability of our outcomes. On the other hand, we did not choose two different social commerce platforms to control the variability of the datasets.

In the first study, we recruited 30 participants (9 females, 22 males) with an average age of 24.84 years (SD = 6.58 years). The participants were shown 12 products from different categories, and they chose six out of 12 products. These products appeared in pairs of two from which they needed to select one each time, and spanned different categories in order to ensure that there was no bias introduced due a specific type of product. For the second study we recruited 23 participants (10 females, 13 males) with an average age of 27.5 years (SD = 7.15 years). The participants were shown three products and they chose one product out of three. We used a slight manipulation in the two studies on the number of products that the participants needed to compare among in order to reach a decision, the first being between two and the second being between three [31].

Participants for both studies were recruited through an open call with a brief description of the step and requirements of the study. We offered each participant a financial compensation equal to approximately $20. The participants in both studies were familiar with Amazon as a company and had average experience with shopping on Amazon having purchased on the website more than once. During the studies, they participants signed an informed consent form that provided information concerning the purpose of the study and the way their data would be treated. They were notified that all data would remain anonymous, and the purpose of the study was solely for research purposes. Once they were briefed about the objective of the study, they were also informed that they needed to just select the product of their choice after assessing the relevant information, without the need to purchase it. We did however instruct them that they should treat the decision-making task as an actual product they would buy. Moreover, all the participants had 20 × 20 vision (with or without correction). A threshold was defined and set for the calibration of eye-tracking, which is crucial to determine whether or not to re-calibrate and conduct the study. This threshold was set for the validation accuracy and was selected to be allowed a maximum of 2.50 degrees. The achieved validation accuracy was below 1.0 for all respondents, with a high of 0.98 and a low of 0.30.

Hence, no results were discarded due to an inability to produce a sufficient validation accuracy at this phase.

The eye-tracker's validation was conducted by repeating the same 5-point calibration scheme. Five points appear on screen, one after the other, and the participants are told to look at them for a few seconds. Then the in-built algorithm of the eye-tracker was used to validate the calibration of the eye-tracker. The validation threshold of 2.5 degrees corresponds to <10 pixels on the screen from a viewing distance of 50–70 cm. This is an acceptable error margin in eye-tracking studies [30].

3.2. Data collection

During the data collection participants were given as much time as needed to select each product among the pairs they were presented with. Throughout the entirety of the study, respondents were free to ask questions to the researcher regarding the study, however, they were encouraged to complete the study with minimal disruption and assistance. In order to prevent the end result of the research from being undermined by sub-optimal research scenarios, measures to improve the reliability of the eye tracking research were emphasized. Given that most eye tracking devices utilize infrared light reflecting from the pupil, alongside complex algorithms to track eye movements, the lighting in the study's environment must be stable. In other words, it is important to mitigate the amount of fluctuating infrared light and maintain consistent lighting levels to ensure high accuracy. Thus, we selected a controlled laboratory setting where incoming light and appropriate settings could be carefully calibrated. After participants signed a consent form, we collected non-sensitive demographic data, such as age and gender of the respondents through a quick demographic survey. In sequence, a brief introduction to the eye tracking system is given, in addition to a quick explanation of what will be tested during the ensuing study. Prior to the actual data collection, a calibration exercise was done in order to ensure the correct syncing of every individual participant's retinal movement to the recording equipment, which is an important part for the validity of measurements and results.

We utilized two different types of eye-trackers to control for possible movement of users during the studies and variations in sampling accuracy [20]. First with a stationary eye-tracker, we used Tobii Pro X3–120 and collected data at 120 HZ. The screen was a full HD 24-in. monitor with a display resolution of 1920 × 1080. Second, we used mobile eye-tracking glasses, where the participants were wearing the eye-tracking glasses and they were provided a laptop to watch the products' webpages. We used SMI ETG (eye-tracking glasses) and Tobii Eye-tracking glasses to collected data at 60 Hz. The laptop screen was a full HD 17-in. with as display resolution of 1680 × 1050. For both the setups, we ensured that the laboratory was appropriately lit, mitigating effects from external elements of light and annoyances, and elements of obstruction and distraction were minimized before conducting the actual study. The use of two sampling methods was done to minimize the potential noise or error points in data collection, as well as to control for any differences between the two groups of users [14]. During the assignment of participants to tasks, there was a random allocation of the devices used. Our analysis showed that there were no significant differences between the participants that used the mobile and stable eye-trackers [31]. Moreover, the data exported from both types of eye-trackers has the same structure and we made sure the calibration and validation errors were similar in both cases.

3.3. Variables and measurements

3.3.1. Dependent variable

In this contribution, we attempt to predict whether the given product was chosen to be bought by the participants or not, using the eye-tracking data. This is a binary variable with two non-ordinal categories: selected and rejected. This variable was measured through a question which asked respondents to add their selected product in the purchase basket.

3.3.2. Independent variables

Tobii's default algorithm (i.e., in-build function in the Tobii software for gaze data processing) was used to identify fixations and saccades (for details please see Olsen [47]). A filter (i.e., in-build function in the Tobii software) was used to remove the raw gaze points that were classified as blinks. Eye movement data provide the mean, variance, minimum, maximum and median of several parameters, such as pupil diameters, fixation details, saccade details, blink details, and event statistics. Table 1 provides an overview of the extracted features as well as the respective reference from the literature. All the Areas of interests (AOIs) were calculated as a proportion of the time students spend looking at the different areas of the screen.

Amazon has developed and implemented social commerce tools that have become familiar concepts to most of its users. Such tools include customer reviews and the product rating system, which function as our primary components comprising the *social popularity*, also known as peer influence. Customer reviews are peer reviews posted by other individuals who have purchased and used, or have experience with, the product or service in question. Amazon also possesses the ability to display time left of sale and the remaining quantity currently available for products when applicable. The latter comprises *scarcity*, present for the product if it is sufficiently low in stock, demonstrated by Amazon with red text at the right-hand side of the product display. The third category of AOIs is *product information*, which is readily available for all products on Amazon, based on what the vendor has provided as available description and information, e.g. product description, technical specifications, and product summary. Finally, the remainder of the available information on Amazon's web page for the product display is referred to as *distractions*. This comprises all visible elements that are not directly related to the product itself, but rather to Amazon's related product advertisements, based on their recommendation system. Examples of this include related products, recommended products, frequently bought together, and any other recommendation system appearances. The AOIs are selected to obtain substantial data on the relevant variables while attempting to balance sensitivity and selectivity for the targeted areas (Fig. 1). They are also chosen to represent and distinguish between the variables (product information, social popularity, scarcity, and distractions) as well as several additional unrelated informational sections presented by Amazon. Table 3 summarizes the AOIs for each product. It is important to point out that because the areas of the different AOI vary a lot in the webpage used, there was a normalization scheme used while computing the time spent on each individual AOI. The time on the different AOIs was normalized for the

Table 1
Features extracted from the eye-tracking data.

Eye-tracking parameters	Features extracted
Diameter	Pupil [53] (mean, median, min, max, SD)
Fixation	Fixation duration [61] (mean, median, min, max, SD)
	Fixation dispersion [34] (mean, median, min, max, SD)
	Skewness of fixation duration histogram [55]
Saccade	Ratio of forward saccades to total saccades [39] (scanpath velocity)
	Ratio of global and local saccades [79] (threshold on sac. vel.)
	Skewness of saccade velocity histogram [52]
	Saccade velocity [57] (mean, median, min, max, SD)
	Saccade length [38] (mean, median, min, max, SD)
	Saccade amplitude [50] (mean, median, min, max, SD)
	Saccade duration [65] (mean, median, min, max, SD)
Events	Num. Fixations, Num Saccades, Fixation to saccade ratio
Others	Time spent on Areas Of Interest (AOIs) (see the Table 2 for the details about AOIs)
	Cognitive load (mean, SD, skewness) [21]
	Index of information Processing (mean, SD, skewness) [46]

Table 2
Description of AOI categories.

AOI category	AOI names
Product Information	Additional Details, Other Technical Details, Price, Product Description, Product Images Small, Product Main Image, Product Summary, Technical Details, Title, Zoomed Image
Social Popularity	4 Stars and Above, Customer Questions and Answers, Detailed 3 Star Reviews, Detailed All Reviews, Detailed Negative Reviews, Review Summary, Summary Reviews, Top Critical Review, Top Positive Review, Top Reviews
Scarcity	Quantity
Distractions	Compare Similar Products, Customers Also Viewed, Frequently Bought Together, Inspired By, Recommended Products, Related Products, Sponsored Products

Table 3
prediction results from using the full data length.

Metric	Training	Testing
Precision	88.12	83.33
Recall	87.25	86.96
F1-Score	87.68	85.11
Accuracy	87.56	85.42

area of the AOI by dividing the time spent on the AOI by the square root of the area of the given AOI.

3.4. Prediction algorithms

3.4.1. Support vector machines

SVM maps an input x onto a multidimensional space using kernel functions (linear, radial or polynomial), and then any kind of regression can be used to model the input data in the new feature space. The quality of estimation is measured by the ε-intensive loss function given by Cortes and Vapnik [18].

3.4.2. Gaussian process models

GPM is similar to SVM; the only difference is that the mapping from the original space to a multidimensional space is governed by Gaussian latent variables that are parametrized using different kernel functions [12]. In this study, we set the kernel functions to take linear, radial, and polynomial forms.

3.4.3. Random forest

Random Forests are ensembles of decision trees mostly used for classification and/or regression purposes. The training algorithm for RF applies the general technique of bagging repeatedly selects a random sample with replacement of the training set, fits trees to these samples, and uses these replicates as new testing sets. The random forest is able to permute the given feature set and compute the feature importance for each feature in a given dataset, by optimizing one of the modelling parameters, e.g., root mean squared error, proportion of variance explained; or in the case of classifications, precision and/or recall. Using the individual feature importance from RFs, one can put a threshold either on the number of features (10 in our case) or on the importance values of the features to select the required number of features.

3.5. Ensemble learning setup

One way of using the results from multiple models is to use a weighted average from all the prediction algorithms. The weights for individual prediction are considered based on their accuracy during the validation phase. There are 3 major advantages of these methods [24,48,54]: 1) We can compare the performance of the ensemble methods to the diversification of our models predicting cognitive performance. It is advised to keep a diverse set of models to reduce the variability in the prediction and hence, to minimize the error rate. Similarly, the ensemble of models will yield better performance on the test case scenarios (unseen data), as compared to the individual models in most of the cases. 2) The aggregate result of multiple models always involves less noise than the individual models. This leads to model stability and robustness. 3) Ensemble models can be used to capture the linear, as well as the non-linear relationships in the data. This can be accomplished by using two different models and forming an ensemble of both.

The main reason for selecting the models was to introduce the three categories of prediction models was to make sure that we are using the benefits of the models and maximize the prediction performance. For example, SVM is a preferred model when the dataset is relatively small and have high dimensional feature set. Moreover, SVM is known to reduce the risk of over fitting [51]. One drawback of SVM is that it might be difficult to choose the kernel function (Amari & [1]) and therefore, we chose to use all three of them in this contribution. Considering the Gaussian process models, they also have similar advantages as SVMs with one difference. The major difference in GPM and SVM is the use of gaussian latent variables to map the original feature space onto the higher dimensional space. Once again, there is no clear indication, from the data, about which method of creating the higher dimensional space from the original space. Therefore, we decided to use both methods to optimize performance. Finally, we included Random Forest in our ensemble because it is known to be efficient while handling multidimensional data when the different dimensions have varied distributions and ranges. This is the case in our dataset as well, for example, the underlying distribution of the five major categories of features (diameter, fixation, saccades, events, and other) are different from each other.

3.6. Training validation and testing

We perform out-of-sampling testing (i.e., leave-one-participant-out), dividing all 3 first datasets into 3 subsets: (1) training, (2) validation, and (3) testing. We keep the testing set aside (15%). The datasets are split based on participant identifiers. All the models are trained and validated using the training and validation sets with a cross validation. The cross-validation is performed using leave-one-participant-out. We used the following metrics to evaluate the performance of the ensemble classifier:

1. Precision = TP / (TP + FP);
2. Recall = TP / (TP + FN);
3. Accuracy = (TP + TN) / (TP + TN + FP + FN);
4. F1 score = 2TP / (2TP + FP + FN).

Where,
TP = true positive;
FP = false positive;
TN = true negative;
FN = false negative.

For evaluating the prediction quality the “chosen” class is the “positive” class. For the baseline prediction, we selected the “random prediction baseline”, due to the balanced nature of our dataset.

The validation in the prediction pipeline was done using leave-one-participant out scheme, this is the most suited method of validation for datasets with smaller sample size [72] as opposed to K-fold validation scheme that is suited for larger datasets, for example, as it is the case with Aribag & Schwartz [2]. In this process the training phase is repeated while leaving one participant out for validation every time. This process is repeated until all the participants in the training data have been left out once. Once the process is complete the validation accuracy is calculated as the mean of all the leave-one-participant-out iterations.

Fig. 1. AOI example from one product.

3.7. Prediction with partial temporal data

We also predicted the dependent variable using partial data, to test "*how quickly can we predict whether the customer will buy the product?*". For this, we took 75% of the data (approx. 6–7 min of eye-tracking data) from each participant and used the methods described above to predict the dependent variable. Then, we keep removing 10% data based on the time up to 15% of the data (approx. 1–2 min of eye-tracking data) (Fig. 2). For each partial dataset we evaluate the prediction performance and the set of most important features. This set is chosen based on the features' importance computed from the random forest classifier and has a value of >75 (out of 100).

4. Results

In this section, we will present the results from the two prediction setups. The first set of results is about the using the whole data duration for training, validation and testing along with the variable importance. Second, we will present the prediction results from using the partial data duration along with the changing variable importance for each of the most important predictors.

4.1. Basic prediction results

First, we predicted whether the product was "chosen" based on the complete data and we obtain a good prediction quality as it can be seen in the Fig. 3 and in the Table 3. The most important variables are shown

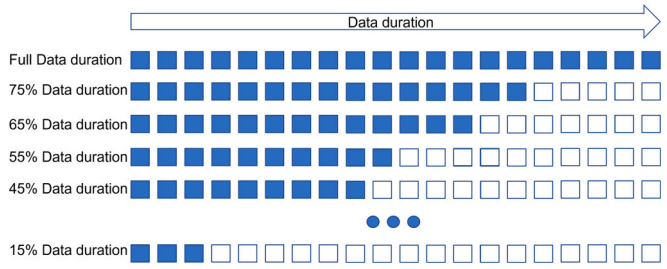


Fig. 2. Schematic representation of the different data duration for the prediction of whether the customer selected or rejected a given product. Each block represents 5% of the data. Blue blocks show the data used for prediction and the white boxes show the remaining unused data in each iteration. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

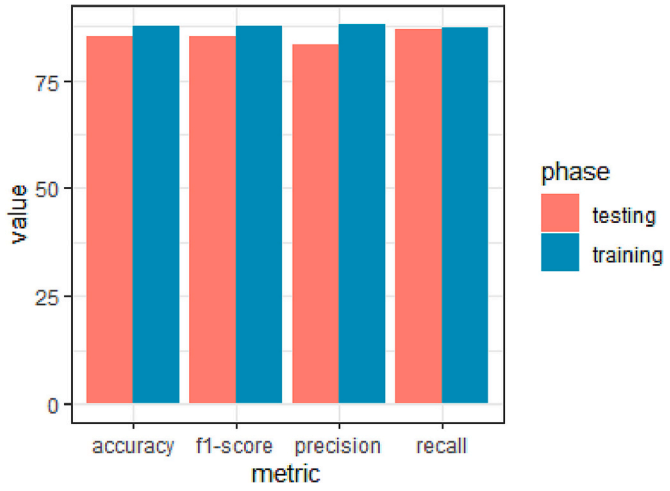


Fig. 3. Visual comparison of the prediction quality metrics from using the full data duration.

in the Fig. 4. We can observe that the precision (88.12) and recall (87.25) from the training phase are similar to those in the testing phase (precision = 83.33; recall = 86.96)., this shows that we successfully avoided over-fitting of the data, because the testing was done using an out of sample setup. In both these phases, we clearly improve the random assignment baseline (precision = 0.50, recall = 0.50, f1-score = 0.50, accuracy = 0.50). Moreover, the most important features for predicting the dependent variables are cognitive load, information processing index, time on AOIs (price, image, recommendations, quantity, review summary, description), fixation duration, scanpath velocity, saccade velocity mean, saccade velocity skewness. The list of most important variables in the predict model show that there are clear patterns from the gaze data that can help us in classifying whether the customer would have selected the product or not. For example, we observe that apart from the AOI-based features, many other variables appear as the most important variable. This indicates that the importance of the visual information processing behavior is as important as the behavior induced by the information provided by the e-commerce platform. We performed single feature predictions to show how the consumer choice would be predicted with the features concerning the time on the different AOIs. We performed Single feature predictions (Appendix A, Table A3) to compare the most important feature with other AOI-based features that are in the top 12 most important features.

Moreover, we also showcase the effectiveness of the ensemble methods used in this contribution by two different methods. First, we compare the proposed method by replacing one of the methods (at a time) with logistic regression and artificial neural networks. The comparison results are shown in Appendix A (Tables A4 and A5). None of the two methods provide better prediction performance than the one used in the paper. Second, we also compare the ensembling of the three models using feature fusion with the ensembling using the weighted average. In the Appendix A (Table A6), we show the precision and recall when one of the five major categories of the features (Diameter, fixation, saccade, Events, others) are missing from one of the three major categories of the prediction algorithms groups (SVM, GPM, RF). We observe that none of the ensembles are better than the weighted average method of ensebmling the models. One clear explanation for such results is the lack of information in each individual model which leads to the poor performance. Finally, we compare the ensembling of the three models using another scheme feature fusion with the ensembling using the weighted average. In the Appendix A (Table A7), we show the precision and recall when two of the five major categories of the features (Diameter, fixation, saccade, Events, others) are missing from one of the three major categories of the prediction algorithms groups (SVM, GPM, RF). We observe that none of the ensembles are better than the weighted average method of ensembling the models. One clear explanation for such results, is the lack of information in each model resulting in poor performance.

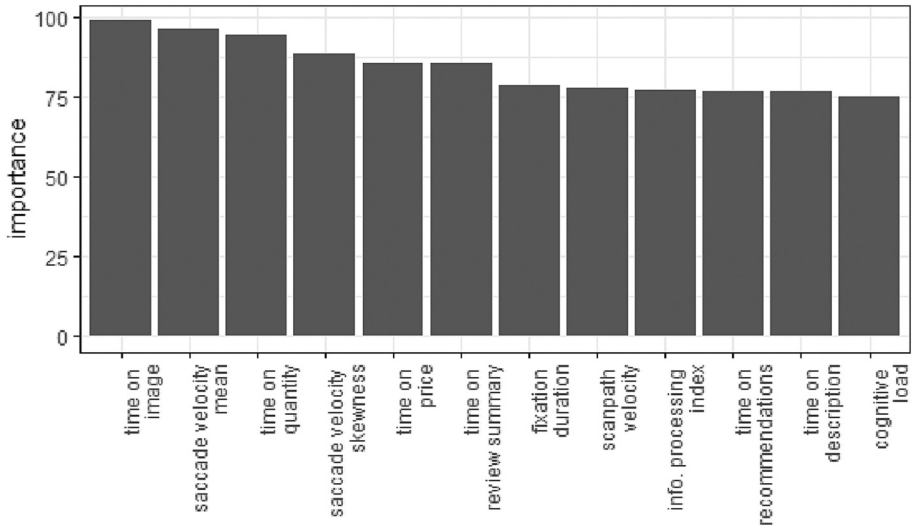


Fig. 4. The most important variables from the ensemble learning pipelines.

Table 4
prediction quality from the partial data prediction.

Data	Training phase				Testing phase			
	Recall	Precision	F1-score	Accuracy	Recall	Precision	F1-score	Accuracy
75%	84.80	82.78	84.60	83.92	80.62	80.47	80.82	80.91
65%	83.34	82.44	83.25	82.30	80.01	80.06	80.60	78.72
55%	83.22	81.95	82.85	81.13	77.92	79.83	78.82	76.58
45%	82.32	79.61	80.13	79.20	77.79	78.18	78.18	76.31
35%	82.09	79.53	79.90	78.63	76.45	77.83	77.73	76.26
25%	78.56	79.45	79.71	78.58	76.41	76.33	76.16	76.01
15%	< 25	< 25	< 25	< 25	< 25	< 25	< 25	< 25

Table 5
most important features from the partial data prediction.

Feature	Data partitions					
	75%	65%	55%	45%	35%	25%
Cognitive load	80.46	92.84	79.65	93.85	94.27	91.48
Info. processing index	75.76	81.83	87.78	81.24	89.04	83.99
Time on price	90.47	95.88	80.79	75.84	79.70	90.08
Time on image	91.96	86.74	93.44	88.17	91.77	82.94
Time on recommendations	98.29	92.79	95.89	90.88	75.17	97.18
Time on quantity	97.18	90.77	80.62	93.99	79.48	84.26
Time on review summary	80.16	76.11	89.20	95.85	77.33	82.43
Time on description	75.16	82.03	88.19	80.75	92.24	93.71
Fixation duration	97.43	92.89	92.20	95.97	94.65	84.31
Scanpath velocity	96.68	76.03	97.42	82.50	84.83	78.94
Saccade velocity mean	80.36	93.10	98.06	78.46	75.53	81.14
Saccade velocity Skewness	81.08	81.48	91.87	79.84	75.97	95.20

4.2. Partial prediction results

Next, we present the results from predicting the dependent variable using shorter periods of the data. Tables 4 and 5 present the prediction quality and the variable importance for the different data lengths, respectively. We observe from the Table 5 that the set of the most important features remains the same as it was for the full data prediction, with changes in the importance order. We can also observe that the prediction quality does not deteriorate until the 35% data partition and at the 25% data partition the recall is the only metric that changes by a considerable difference. This shows that by using a little over two minutes of eye-tracking data, we can predict whether the customer is going to buy a given product. Form Table 5, we also observe that the most important feature set remains consistent across the different data slices (how much data we are using in the terms of duration). However, there are slight differences in the ranking of these most important variables but there is nothing that bursts out of order in the terms of feature importance. This is indicative of the fact that there is a considerable amount of information in the gaze data from the participants while they process e-commerce websites, and this information is also consistent over time.

We compare the proposed method with two other methods, which have similar data collection settings as in our studies. The first (Appendix A, Table A1) utilizes Hidden Markov Models (HMM, [76]) and the second (Appendix A, Table A2) utilizes ANOVA-based prediction [25].

5. Discussion and conclusions

In this research we have sought to understand how consumers utilize different types of informational cues when purchasing online, and specifically, when attempting to make a choice between alternative products. Online commerce platforms nowadays incorporate different types and formats for presenting information to consumers, both marketer- and consumer-generated. Nevertheless, we still know very little about how consumers interact with such content and which types of information they utilize during the process of making a purchase-related decision. Furthermore, there is a lack of understanding concerning how users during their decision-making process utilize different informational cues to reach a decision. Using an eye-tracking approach and building on ensemble learning methods of prediction, we uncover what aspects of information on social commerce sites enhance decision-making.

In relation to our first research question (RQ1), our analysis revealed that there are certain informational cues that decision-makers placed more focus on and where more important in explaining a purchase decision. Specifically, the most important cues included the time a user spent on the image, the saccade velocity, the time on quantity, saccade velocity skewness, time on price, and time on review summary. These outcomes demonstrate that certain informational cues are of high significance for decision-makers in the context of social commerce, since the process of deciding what product to purchase is a result of carefully examining the visual characteristics of the product at hand, as well as the expected cost, scarcity, and experiences of others. The saccades, which correspond to the eye movement from one point of fixation to another, also indicate that when there is increased speed and dynamics of fixation that is a strong indicator of consumer choices during the decision-making process. Taken together, these findings indicate that both different informational cues presented on screen as well as patterns of gazing behavior can capture consumer decision-making. Specifically, in regard to the informational cues, the results show that consumers are particularly drawn to the visual stimuli from product images, as well as the trade-off between price and scarcity, in relation to other consumer experiences, as presented by the relevant information.

When exploring the second research question (RQ2), our analysis reveals that there are indeed fluctuations in the importance of different aspects in relation to decision-making on social commerce platforms. More precisely, we find that by using different segments of activity the informational cues play an important role on decision-making shift. In fact, time on recommendations appears to be a very important determinant during the first quarter of the allocated time, while it later shifts to the price information, and in sequence to the quantity of the product. This temporal sequence of importance for the different information ques. allows us to gain a better understanding of the decision-making process that users of social commerce platforms go through when deciding about what product to purchase. Within the context of social commerce product selection, it highlights that consumers initially rely on experiences provided by others, which may indicate that they perceive this type of information as more credible than that of marketers. In sequence, once trust has been established towards a product, informational cues that revolve around the assessments of the trade-off between price and its scarcity receive more importance. Finally, details about the product such as the visual appearance and a re-confirmation of opinions from other users conclude decision-making. Such insight provides us with a more nuanced understanding of decision-making as a process that is dynamic and changing, and where the importance of informational cues shifts over time.

5.1. Research implications

This study has built on a combination of data collection approaches and methods to analyze eye-tracking data which is gaining momentum in the domain of consumer decision-making. Our work contributes by extending the knowledge in this domain concerning how different informational cues influence consumer decision-making during product selection of social commerce websites, where there is a plethora of different information types and formats. In addition, we explore the dynamic nature of decision-making identifying the temporal sequence of user preferences when interacting with such information. In addition, the approaches used to analyze such types of data can be transferred to other application areas where there is a presence of such complex phenomena. Specifically, this study contributes to ongoing research in the following ways.

When it comes to understanding how consumers interact with information presented on online commerce platforms, research to date has identified several important aspects and information types that trigger consumer intentions [23,69,71]. To date, most studies have attempted to examine the value or importance of certain types of information (user- vs marketer-generated) when selecting product online, or on identifying for users perceive certain types of prompts when those are presented to them. Nevertheless, contemporary social commerce platforms manage to contain diverse types of information generated by marketers and consumers, as well as different formats of presenting such information. In our study we have defined several different areas of interest (AOIs) which correspond to some of the information users are exposed to during a purchase decision. Our findings indicate that users make use of certain types of information over others, and that the value of information presented on such websites does not have equal weight when it comes to ensuring that consumers make an informed decision. In our analysis we have included predictors of the layout, as well as on the gaze patterns of consumers. Specifically, the analysis pinpoints to the fact that certain types of information such as the time spent on the image, price, and on reviews have a strong effect on decision-making prediction. Thus, this finding indicates that we can infer more detailed information when can capture, or control for both.

Nevertheless, another key finding is that the importance of informational sources and gaze patterns dynamically changes during the decision-making process. As users interact with online interfaces the significance of certain types of information either increases or decreases in importance. This finding shows that there is a temporal significance of key aspects for users as they are making decisions. Several of the identified informational cues shift in importance for users of social commerce platforms as they decide which product to select. This finding provides some context to our first point, that the importance of certain types of information and how they are represented should be considered in the context of when exactly during the decision-making process they are utilized. This finding indicates that informational sources may have an ephemeral value in satisfying consumer requirements. Extending on such a reasoning entail that we need to develop a more nuanced understanding of how consumers of information utilize such sources over time, rather than in a snapshot in time. Thus, a fruitful domain for future studies is to adopt a more dynamic perspective to understanding how users interact with online commerce platforms, and conceptualize decision-making as an active process of interaction, between the user and the available information.

Adding to the above, our approach opens future research avenues for understanding designing dynamic interfaces that can satisfy user requirements. With the prevalence of smart glasses and more advanced eye-trackers in devices that are not intrusive, there is a renewed interest in such dynamic interfaces, and presenting users with the right type and format of information that is needed at the different stages of the decision-making process. Such interfaces that can capture data in real-time from eye-tracking inferences and dynamically alter the information that is presented to users or consumers is likely to be an important area of research in the years to come. To date, we have been accustomed to static interfaces that only change after they have been prompted by user action. The same applies also in the case of social commerce websites where consumers need to initiate an action to receive information that may interest them, such as clicking on an option to get additional information. By capturing real-time data from eye-movements, there is a renewed interest in developing interfaces that can present important information automatically.

A key aspect in being able to analyze data such as eye movements and make accurate assessments of user intentions is utilizing advanced methods of prediction. In this study, we showcase how different prediction algorithms can be used on eye-tracking data in order to provide a more detailed understanding of aspects that contribute to product selection. While this study is not the first to combine such data and analysis approaches [4,23,62], it does showcase how it can yield interesting findings within the area of social commerce by focusing on key types of information presented on such online platforms. The combination of rich data from physiological data with advanced methods of analyzing such data can enable the examination of complex phenomena in emerging digital technologies (e.g., smart glasses, smart windshields etc.). We therefore argue that such approaches will likely be very useful in future research as a more in-depth lens of understanding user behavior when interacting with digital technologies.

5.2. Practical implications

Apart from several important research implications, the study also provides some interesting insights for practitioners, both on how to develop interfaces for social commerce platforms as well as for future advancements in the domain. In terms of designing and developing interfaces on social commerce platforms, our results provide some useful information to UX designers on aspects that are of increased importance to consumers when making a product selection. They also highlight the need to design based on principles of efficiency and use from the end-user's perspective. As marketers want to improve the efficiency of product selection and browsing by providing the right type of information to consumers, it is important that such input is utilized in the design process. Prior studies have shown that information overload can negatively affect product selection on such platforms, so a key take for designers is not to include as much information as possible on social commerce websites but rather to focus on the ones that are of importance to consumers.

In addition, our results highlight that there are certain aspects from the analysis that users find particularly important in their decision-making process. These can be leveraged from practitioners in the design of online platforms for optimizing the decision-making process of users. Furthermore, the temporal order of importance can provide insight into the structure of design and how to present important information in a way that eases the user's decision-making. For instance, time on recommendations was found to be an important aspect of information for the users of social commerce platforms during product selection which highlights the focus that should be placed on presenting such information in a clear and distinct way, without proximity to other aspects that might distract the user's attention and using appropriate font size and style to facilitate easy readability. Similarly, the time on image features as one of the most important aspects of prediction for purchase intention, which indicates that is a need for practitioners to design interfaces that can capture the intricacies of the product that is visualized. In other words, images of products on social commerce platforms must be of high quality and featured in a way that can make them easily accessible and interactive to users.

Furthermore, an interesting practical implication has to do with emergence of novel wearable devices and how they can be leveraged to create dynamic interfaces for information presentation. A growing number of users are now using smart glasses, or even laptops that have enhanced abilities of understanding face gestures and track eye

movements. This trend denotes a progression in terms of information that can be utilized by marketers to provide more personalized information to consumers and enhance their product selection when purchasing online. For instance, it has become common not to utilize location data to provide accurate advertisements, or third-party browsing history for product suggestions. With novel devices that can track eye-movement, we are likely to see new design paradigms where information that is presented to consumers is adjusted along the product-browsing journey. In addition, much of what is presented to users may be dynamically adjusted based on several factors such as individual preferences, type of product, or even based on real-time data like where the consumer is focusing their attention or if there is a detection of drowsiness, lack of attention, or cognitive overload. Such new waves of designing interfaces with the integration of real-time physiological data are likely to herald a new era for online purchasing.

5.3. Limitations and future research

Although in this study we have attempted to minimize the presence of any bias and to provide results that are generalizable, the outcomes do not come without limitations. First, we have based our analysis on a product selection study which was conducted in a lab and with a preselection of products for participants. While such a set-up approximates reality, it does not accurately capture the process of product selection of individuals. A more realistic approach would have been to allow individuals to search and find different products on their own, and possibly select between multiple different options. Nevertheless, we did not use this approach as we wanted to be able to compare similar conditions for all participants. A second limitation is that during the participant selection and data collection, we used individuals who live in Scandinavia. Such individuals are well accustomed to using social commerce platforms such as Amazon, so their use and consumption of information are likely to differ from users that use this platform for the first time or have limited experience. In addition, we did not perform a separate analysis to check how younger vs older users would reach a conclusion about which product to select, and what types of information or aspects were more important in determining their selection. Finally, each platform that is used for social commerce presents different types of information and in different formats. For instance, platforms such as Instagram which are primarily used to promote products and secondarily to enable purchase of the products via third-party websites will likely result in different aspects being important for users. Thus, it is interesting to identify how the different types of social media platforms trigger different types of information requirements from their users.

CRediT authorship contribution statement

Patrick Mikalef: Conceptualization, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Kshitij Sharma:** Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. **Sheshadri Chatterjee:** Investigation, Writing – review & editing. **Ranjan Chaudhuri:** Investigation, Writing – review & editing. **Vinit Parida:** Supervision, Resources, Project administration. **Shivam Gupta:** Supervision, Resources, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dss.2023.114039.

References

[1] S.-I. Amari, S. Wu, Improving support vector machine classifiers by modifying kernel functions, Neural Netw. 12 (6) (1999) 783–789.
[2] A. Aribarg, E.M. Schwartz, Native advertising in online news: trade-offs among clicks, brand recognition, and website trustworthiness, J. Mark. Res. 57 (1) (2020) 20–34.
[3] A.S. Atalay, H.O. Bodur, D. Rasolofoarison, Shining in the center: central gaze cascade effect on product choice, J. Consum. Res. 39 (4) (2012) 848–866.
[4] V. Bachurina, S. Sushchinskaya, M. Sharaev, E. Burnaev, M. Arsalidou, A machine learning investigation of factors that contribute to predicting cognitive performance: difficulty level, reaction time and eye-movements, Decis. Support. Syst. 155 (2022), 113713.
[5] D. Bačić, R. Henry, Advancing our understanding and assessment of cognitive effort in the cognitive fit theory and data visualization context: eye tracking-based approach, Decis. Support. Syst. 163 (2022), 113862.
[6] S. Banerjee, S. Bhattacharyya, I. Bose, Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business, Decis. Support. Syst. 96 (2017) 17–26.
[7] R. Batista Duarte, D. Silva da Silveira, V. de Albuquerque Brito, C.S. Lopes, A systematic literature review on the usage of eye-tracking in understanding process models, Bus. Process. Manag. J. 27 (1) (2021) 346–367.
[8] M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, Proc. Natl. Acad. Sci. 116 (32) (2019) 15849–15854.
[9] R. Boardman, H. Mccormick, Attention and behaviour on fashion retail websites: an eye-tracking study, Inf. Technol. People 35 (7) (2022) 2219–2240.
[10] B.M. Brand, R. Reith, Cultural differences in the perception of credible online reviews–the influence of presentation format, Decis. Support. Syst. 154 (2022), 113710.
[11] T.T. Brunyé, A.L. Gardony, Eye tracking measures of uncertainty during perceptual decision making, Int. J. Psychophysiol. 120 (2017) 60–68.
[12] R. Calandra, J. Peters, C.E. Rasmussen, M.P. Deisenroth, Manifold Gaussian Processes for Regression. 2016 International Joint Conference on Neural Networks (IJCNN), 2016.
[13] D. Camors, Y. Trotter, P. Pouget, S. Gilardeau, J.-B. Durand, Visual straight-ahead preference in saccadic eye movements, Sci. Rep. 6 (1) (2016) 1–9.
[14] B.T. Carter, S.G. Luke, Best practices in eye tracking research, Int. J. Psychophysiol. 155 (2020) 49–62.
[15] P. Chandon, J.W. Hutchinson, E.T. Bradlow, S.H. Young, Does in-store marketing work? Effects of the number and position of shelf facings on brand attention and evaluation at the point of purchase, J. Mark. 73 (6) (2009) 1–17.
[16] N. Chaudhuri, G. Gupta, V. Vamsi, I. Bose, On the platform but will they buy? Predicting customers' purchase behavior using deep learning, Decis. Support. Syst. 149 (2021), 113622.
[17] J.V. Chen, B.-c. Su, A.E. Widjaja, Facebook C2C social commerce: a study of online impulse buying, Decis. Support. Syst. 83 (2016) 57–69.
[18] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[19] R.-F. Day, Examining the validity of the Needleman–Wunsch algorithm in identifying decision strategy with eye-movement data, Decis. Support. Syst. 49 (4) (2010) 396–403.
[20] S. Dowiasch, P. Wolf, F. Bremmer, Quantitative comparison of a mobile and a stationary video-based eye-tracker, Behav. Res. Methods 52 (2020) 667–680.
[21] A.T. Duchowski, K. Krejtz, I. Krejtz, C. Biele, A. Niedzielska, P. Kiefer, M. Raubal, I. Giannopoulos, The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018.
[22] S.R. Ellis, J.D. Smith, Patterns of statistical dependency in visual scanning, in: Eye Movements and Human Information Processing, 1985, pp. 221–238.
[23] M. Fei, H. Tan, X. Peng, Q. Wang, L. Wang, Promoting or attenuating? An eye-tracking study on the role of social cues in e-commerce livestreaming, Decis. Support. Syst. 142 (2021), 113466.
[24] V.V. Gavrishchaka, M.E. Koepke, O.N. Ulyanova, Ensemble learning frameworks for the discovery of multi-component quantitative models in biomedical applications, in: 2010 Second International Conference on Computer Modeling and Simulation, 2010.
[25] S. Goyal, K.P. Miyapuram, U. Lahiri, Predicting consumer's behavior using eye tracking data, in: 2015 Second International Conference on Soft Computing and Machine Intelligence (ISCMI), 2015.
[26] D.J. Graham, J.L. Orquin, V.H. Visschers, Eye tracking and nutrition label use: a review of the literature and recommendations for label enhancement, Food Policy 37 (4) (2012) 378–382.
[27] J. Guerreiro, P. Rita, D. Trigueiros, Attention, emotions and cause-related marketing effectiveness, Eur. J. Mark. 49 (11/12) (2015) 1728–1750.
[28] J. Gwizdka, Exploring eye-tracking data for detection of mind-wandering on web tasks, in: Information Systems and Neuroscience: NeuroIS Retreat 2018, 2019.
[29] R.S. Hessels, D.C. Niehorster, M. Nyström, R. Andersson, I.T. Hooge, Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers, R. Soc. Open Sci. 5 (8) (2018), 180502.

[30] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, J. Van de Weijer, Eye Tracking: A Comprehensive Guide to Methods and Measures, OUP Oxford, 2011.
[31] K. Holmqvist, S.L. Örbom, I.T. Hooge, D.C. Niehorster, R.G. Alexander, R. Andersson, J.S. Benjamins, P. Blignaut, A.-M. Brouwer, L.L. Chuang, Eye tracking: empirical foundations for a minimal reporting guideline, Behav. Res. Methods 55 (1) (2023) 364–416.
[32] R.E. Hostler, V.Y. Yoon, Z. Guo, T. Guimaraes, G. Forgionne, Assessing the impact of recommender agents on on-line consumer unplanned purchase behavior, Inf. Manag. 48 (8) (2011) 336–343.
[33] C.-L. Hsu, K.-C. Chang, M.-C. Chen, The impact of website quality on customer satisfaction and purchase intention: perceived playfulness and perceived flow as mediators, IseB 10 (2012) 549–570.
[34] T. Jaarsma, H. Jarodzka, M. Nap, J.J. van Merrienboer, H.P. Boshuizen, Expertise under the microscope: processing histopathological slides, Med. Educ. 48 (3) (2014) 292–300.
[35] L. Jenke, K. Bansak, J. Hainmueller, D. Hangartner, Using eye-tracking to understand decision-making in conjoint experiments, Polit. Anal. 29 (1) (2021) 75–101.
[36] Y. Jiang, B.W. Ritchie, M.L. Verreynne, Building tourism organizational resilience to crises and disasters: a dynamic capabilities view, Int. J. Tour. Res. 21 (6) (2019) 882–900.
[37] Z. Jiang, J. Chan, B.C. Tan, W.S. Chua, Effects of interactivity on website involvement and purchase intention, J. Assoc. Inf. Syst. 11 (1) (2010) 34–59.
[38] E.M. Kok, H. Jarodzka, Before your very eyes: the value and limitations of eye tracking in medical education, Med. Educ. 51 (1) (2017) 114–122.
[39] C. Krischer, W.H. Zangemeister, Scanpaths in reading and picture viewing: computer-assisted optimization of display conditions, Comput. Biol. Med. 37 (7) (2007) 947–956.
[40] Y.-F. Kuo, C.-M. Wu, W.-J. Deng, The relationships among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services, Comput. Hum. Behav. 25 (4) (2009) 887–896.
[41] A. Lang, Dynamic human-centered communication systems theory, Inf. Soc. 30 (1) (2014) 60–70.
[42] A. Lang, R.L. Bailey, Understanding information selection and encoding from a dynamic, energy saving, evolved, embodied, embedded perspective, Hum. Commun. Res. 41 (1) (2015) 1–20.
[43] T. Mavlanova, R. Benbunan-Fich, G. Lang, The role of external and internal signals in E-commerce, Decis. Support. Syst. 87 (2016) 59–68.
[44] N. Meilatinova, Social commerce: factors affecting customer repurchase and word-of-mouth intentions, Int. J. Inf. Manag. 57 (2021), 102300.
[45] R.V. Menon, V. Sigurdsson, N.M. Larsen, A. Fagerstrøm, G.R. Foxall, Consumer attention to price in social commerce: eye tracking patterns in retail clothing, J. Bus. Res. 69 (11) (2016) 5008–5013.
[46] P. Mikalef, K. Sharma, I.O. Pappas, M. Giannakos, Seeking information on social commerce: an examination of the impact of user-and marketer-generated content through an eye-tracking study, Inf. Syst. Front. 23 (2021) 1273–1286.
[47] A. Olsen, The Tobii I-VT fixation filter, Tobii Technol. 21 (2012) 4–19.
[48] M. Papouskova, P. Hajek, Two-stage consumer credit risk modelling using heterogeneous ensemble learning, Decis. Support. Syst. 118 (2019) 33–45.
[49] L. Peng, W. Zhang, X. Wang, S. Liang, Moderating effects of time pressure on the relationship between perceived value and purchase intention in social E-commerce sales promotion: considering the impact of product involvement, Inf. Manag. 56 (2) (2019) 317–328.
[50] M.H. Phillips, J.A. Edelman, The dependence of visual scanning performance on saccade, fixation, and perceptual metrics, Vis. Res. 48 (7) (2008) 926–936.
[51] D.A. Pisner, D.M. Schnyer, Support vector machine, in: Machine Learning, Elsevier, 2020, pp. 101–121.
[52] S. Prasad, S.L. Galetta, Eye movement abnormalities in multiple sclerosis, Neurol. Clin. 28 (3) (2010) 641–655.
[53] L.P. Prieto, K. Sharma, Ł. Kidziński, M.J. Rodríguez-Triana, P. Dillenbourg, Multimodal teaching analytics: automated extraction of orchestration graphs from wearable sensor data, J. Comput. Assist. Learn. 34 (2) (2018) 193–203.
[54] X. Qiu, L. Zhang, Y. Ren, P.N. Suganthan, G. Amaratunga, Ensemble deep learning for regression and time series forecasting, in: 2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL), 2014.
[55] E.M. Reingold, E.D. Reichle, M.G. Glaholt, H. Sheridan, Direct lexical control of eye movements in reading: evidence from a survival analysis of fixation durations, Cogn. Psychol. 65 (2) (2012) 177–206.
[56] Y. Rogers, H. Sharp, J. Preece, Interaction Design: Beyond Human-Computer Interaction, John Wiley & Sons, 2023.
[57] M. Russo, M. Thomas, D. Thorne, H. Sing, D. Redmond, L. Rowland, D. Johnson, S. Hall, J. Krichmar, T. Balkin, Oculomotor impairment during chronic partial sleep deprivation, Clin. Neurophysiol. 114 (4) (2003) 723–736.
[58] M. Ryan, N. Krucien, F. Hermens, The eyes have it: using eye tracking to inform information processing strategies in multi-attributes choices, Health Econ. 27 (4) (2018) 709–721.
[59] M. Salehan, D.J. Kim, Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics, Decis. Support. Syst. 81 (2016) 30–40.
[60] A. Savoy, G. Salvendy, Factors for customer information satisfaction: user approved and empirically evaluated, Int. J. Human–Comput. Interact. 32 (9) (2016) 695–707.
[61] S. Schroeder, J. Hyönä, S.P. Liversedge, Developmental eye-tracking research in reading: introduction to the special issue, J. Cogn. Psychol. 27 (5) (2015) 500–510.
[62] M. Shojaeizadeh, S. Djamasbi, R.C. Paffenroth, A.C. Trapp, Detecting task demand via an eye tracking machine learning system, Decis. Support. Syst. 116 (2019) 91–101.
[63] K. Tzafilkou, N. Protogeros, Diagnosing user perception and acceptance using eye tracking in web-based end-user development, Comput. Hum. Behav. 72 (2017) 23–37.
[64] B.T. Vincent, R. Baddeley, A. Correani, T. Troscianko, U. Leonards, Do we look at lights? Using mixture modelling to distinguish between low-and high-level factors in natural image viewing, Vis. Cogn. 17 (6–7) (2009) 856–879.
[65] T. Vuori, M. Olkkonen, M. Pölönen, A. Siren, J. Häkkinen, Can eye movements be quantitatively applied to image quality studies?, in: Proceedings of the Third Nordic Conference on Human-Computer Interaction, 2004.
[66] D. Wang, X.R. Luo, Y. Hua, J. Benitez, Big arena, small potatoes: a mixed-methods investigation of atmospheric cues in live-streaming e-commerce, Decis. Support. Syst. 158 (2022), 113801.
[67] Q. Wang, S. Yang, M. Liu, Z. Cao, Q. Ma, An eye-tracking study of website complexity from cognitive load perspective, Decis. Support. Syst. 62 (2014) 1–10.
[68] S.V. Wass, T.J. Smith, M.H. Johnson, Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults, Behav. Res. Methods 45 (2013) 229–250.
[69] E. Wästlund, P. Shams, M. Löfgren, L. Witell, A. Gustafsson, Consumer perception at point of purchase: evaluating proposed package designs in an eye-tracking lab, J. Bus. Retail Manag. Res. 5 (1) (2010) 42–51.
[70] M. Wedel, R. Pieters, A review of eye-tracking research in marketing, Rev. Mark. Res. (2017) 123–147.
[71] J. Willems, C.J. Waldner, J.C. Ronquillo, Reputation star society: are star ratings consulted as substitute or complementary information? Decis. Support. Syst. 124 (2019), 113080.
[72] T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recogn. 48 (9) (2015) 2839–2846.
[73] B. Wooley, S. Bellman, N. Hartnett, A. Rask, D. Varan, Influence of dynamic content on visual attention during video advertisements, Eur. J. Mark. 56 (13) (2022) 137–166.
[74] Y. Wu, E.W. Ngai, P. Wu, C. Wu, Fake online reviews: literature review, synthesis, and directions for future research, Decis. Support. Syst. 132 (2020), 113280.
[75] J. Yang, R. Sarathy, J. Lee, The effect of product review balance and volume on online Shoppers' risk perception and purchase intention, Decis. Support. Syst. 89 (2016) 66–76.
[76] L. Yang, O. Toubia, M.G. De Jong, A bounded rationality model of information search and choice in preference measurement, J. Mark. Res. 52 (2) (2015) 166–183.
[77] S.-F. Yang, An eye-tracking study of the elaboration likelihood model in online shopping, Electron. Commer. Res. Appl. 14 (4) (2015) 233–240.
[78] V.Y. Yoon, R.E. Hostler, Z. Guo, T. Guimaraes, Assessing the moderating effect of consumer product knowledge and online shopping experience on using recommendation agents for customer loyalty, Decis. Support. Syst. 55 (4) (2013) 883–893.
[79] W.H. Zangemeister, T. Liman, Foveal versus parafoveal scanpaths of visual imagery in virtual hemianopic subjects, Comput. Biol. Med. 37 (7) (2007) 975–982.
[80] K.Z. Zhang, M. Benyoucef, Consumer behavior in social commerce: a literature review, Decis. Support. Syst. 86 (2016) 95–108.
[81] K. Zhao, P. Zhang, H.-M. Lee, Understanding the impacts of user-and marketer-generated content on free digital content consumption, Decis. Support. Syst. 154 (2022), 113684.

Patrick Mikalef is a Professor in Data Science and Information Systems at the Department of Computer Science. In the past, he has been a Marie Skłodowska-Curie post-doctoral research fellow working on the research project "Competitive Advantage for the Data-driven Enterprise" (CADENT). He received his B.Sc. in Informatics from the Ionian University, his M.Sc. in Business Informatics for Utrecht University, and his Ph.D. in IT Strategy from the Ionian University. His research interests focus the on strategic use of information systems and IT-business value in turbulent environments. He has published his work in over 150 international conferences and peer-reviewed journals including the Journal of the Association for Information Systems, European Journal of Information Systems, Journal of Business Research, British Journal of Management, Information and Management, Industrial Management & Data Systems, and Information Systems and e-Business Management. He serves as a senior editor of the European Journal of Information Systems and as editor of Information of Management. Patrick Mikalef is also a distinguished member of the Association for Information Systems (AIS).

Kshitij Sharma is an Associate Professor in Human-Computer Interaction and Collaborative/cooperative learning. His doctoral work was in the area of using multimodal data (EEG, eye-tracking, facial expressions, audio, dialogues, blood pressure, skin conductance, heart rate) to explain the differences between and predict, experts and novice groups; good and poor students; functional and non-functional groups. The main context for the application of his research has been education. His research interests are primarily in the area of Applied Machine Learning, Artificial Intelligence, and Human-Computer Interaction (HCI) with a heavy emphasis on groups' behavior and physiological data such as eye-tracking, EEG, facial expressions (theoretical and practical methods in digital interaction). He seek to understand relations between users' data (EEG, eye-tracking, system log data, users' actions) and the profile of the user (expertise, motivation, strategy, performance) based on empirical experimentation (controlled experiments) and mixed methods analysis (utilizing a multitude of digital technologies). The knowledge gained from these studies is then used to provide feedback to the group or adapt to the needs of the group in a proactive manner. For this effort, in his studies, he has combined eye-tracking and users' actions to

provide more comprehensive results through data science, statistics, and machine learning practices.

Sheshadri Chatterjee is a post-doctoral research scholar at the Indian Institute of Technology Kharagpur, India. He has completed a PhD from the Indian Institute of Technology Delhi, India. He is having work experience in different multinational organizations such as Microsoft Corporation, Hewlett Packard Company, IBM, and so on. Sheshadri has published research articles in several reputed journals such as Government Information Quarterly, Information Technology & People, Journal of Digital Policy, Regulation and Governance, and so on. Sheshadri is also a certified project management professional, PMP from Project Management Institute (PMI), USA, and completed PRINCE2, OGC, UK, and ITIL v3 UK.

Ranjan Chaudhuri was a Fulbright Fellow to the University of Alabama in Huntsville in 2012. He is an Associate Professor in Marketing Sciences at the National Institute of Industrial Engineering, Mumbai. In the recent past, Dr. Chaudhuri also served as a Faculty at the Indian Institute of Technology Kharagpur and the Indian Institute of Technology Delhi. Dr. Chaudhuri has over twenty years of industrial, teaching and research experience. Dr. Chaudhuri holds an MBA in Marketing and a PhD in Management Sciences. Dr. Chaudhuri authored/coauthored >120 publications in referred National and International Journals and Conference Proceedings and contributed chapters in seven books and authored one monograph published by a leading press in European Union.

Vinit Parida is a Chaired Professor for Entrepreneurship and Innovation, Associate Editor for Journal of Business Research, Member of Swedish ministry's high-level group on digital transformation of Swedish industry, Scientific Leader for NorrlandsNavet- A Kamprad Family Foundation Center for SMEs Growth and Innovation in Northern Sweden, and Board member for RE:Source Strategic Innovation Program (SIP) Vinnova. His research results have been published in 200+ leading international peer-reviewed journals, conferences, book chapters and industry/popular publications. Such as Academy of Management Journal, Journal of Management, Strategic Management Journal, Journal of Management Studies, Entrepreneurship Theory and Practice, Journal of Product Innovation Management, MIT Sloan Management Review, California Management Review, Long Range Planning, Industrial Marketing Management, Journal of Business Research, International Journal of Production Economics, Production and Operation Management, International Journal of Operations & Production Management, Strategic Entrepreneurship Journal, Entrepreneurship and Regional Development, Journal of Small Business Management, and Journal of Cleaner Production. He is active within different academic communities and has presented research results in well-known international conferences, such as Babson College Enterprise Research Conference, Research in Entrepreneurship and Small Business (RENT), The Annual ICSB World Conference, International Conference on Management of Technology (IAMOT), International Product Development Management Conference, and CRIP IPSS conference.

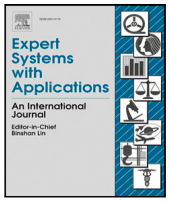
Shivam Gupta is a Professor at NEOMA Business School, France with a demonstrated history of working in the higher education industry. Skilled in Statistics, Cloud Computing, Big Data Analytics, Artificial Intelligence and Sustainability. Strong education professional with a Doctor of Philosophy (PhD) focussed in Cloud Computing and Operations Management from Indian Institute of Technology (IIT) Kanpur. Followed by PhD, postdoctoral research was pursued at Freie Universität Berlin and SUSTech, China. He has completed HDR from University of Montpellier, France. He has published several research papers in reputed journals and has been the recipient of the International Young Scientist Award by the National Natural Science Foundation of China (NSFC) in 2017 and winner of the 2017 Emerald South Asia LIS award.

Expert Systems With Applications 237 (2024) 121599

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



A cyber risk prediction model using common vulnerabilities and exposures

Arash Negahdari Kia, Finbarr Murphy, Barry Sheehan *, Darren Shannon

University of Limerick, Limerick, Ireland

ARTICLE INFO

Keywords:
Cyber risk prediction
Topic extraction
Supervised learners
Random forest
Time series

ABSTRACT

The cyber risk from malicious external attackers is a significant socio-economic problem. Cyber risk prediction is particularly difficult, given the constantly changing attack vectors. This study presents a model that automatically predicts cyber risks. The model is only based on common vulnerabilities and exposures (CVE) data and supervised prediction algorithms. This approach eliminates expert opinion bias in cyber risk prediction. Our supervised data-driven model, $CyRiPred$, CVE data into cyber risk groups by mapping the textual description field of the database into relevant Wikipedia article titles. Then $CyRiPred$ aggregates the occurrence and severity of extracted topics for the desired time unit and produces a time series fed to supervised regressors for prediction. The risks are calculated using predicted occurrence and impact. Finally, the cyber risks are ranked by their score, and the top ten risks are presented. The proposed model is evaluated, and the results are discussed.

1. Introduction

Cyberrisk targets software or hardware systems, whether standalone, distributed, or network architecture. Identifying and assessing cyber risks is vital for information system users, IT-based organisations, cybersecurity experts, and cyber risk insurers. Assessment and Identification of cyber risk require developing special techniques and procedures (Refsdal et al., 2015; Sheehan et al., 2021). To underwrite the risks, the insurer needs to understand the risks and the impact of a cyber-attack. A risk impact score depends on the impact and likelihood of risk occurrence (Alali et al., 2018). There is limited recorded information for academic research, and most predictions depend on experts' knowledge and subjective understanding of current and future risk scenarios.

When predicting future cyber risk, the insurance industry and cybersecurity companies produce cyber risk prediction and trends reports. However, the methodology used to create these reports is based on expert opinion, and annual reports are not sufficiently frequent. IBM, Norton, and many other companies provide these reports (Rawlings, 2015). While companies and research groups work on this area with different methodologies, there is still a gap between a reliable, comprehensive, real-time, data-driven method and their research.

There are many different approaches to looking at the problem in the cyber risk prediction literature. These approaches can be distinguished into two general categories: The first category includes those researches that try to predict cyber incidents and risks inside a specific individual system or network (Khodabakhsh et al., 2020; Sentuna et al., 2021). These researches are used to improve the efficiency of Intrusion Detection Systems (Buczak & Guven, 2016) by predicting a specific type of failure before it happens (Salfner et al., 2010). The second category is those the researches that are environment-independent and investigate the concept of cyber risk regardless of focusing on any particular application, system, or network ecosystem (Subroto & Apriyana, 2019). Our model is similar in the second category. However, to our knowledge, there is no peer-reviewed comprehensive study that is both environment-independent and application-independent and is entirely data-driven on the Common Vulnerabilities and Exposures dataset.

The ability to forecast cyber attacks will significantly limit the socio-economic impacts of such events. In our model, which we call $CyRiPred$ (Cyber Risk Prediction Model), whenever a new vulnerability comes in the National Vulnerabilities Database (NVD), our system immediately associates that risk to a related cybersecurity topic using our novel $WikiTopic$ sub-module, which uses a keyword extraction algorithm and a consecutive Wikipedia search. We further assign a severity score to each topic using aggregated impact scores in the CVE database. The CVE database gathers publicly known cybersecurity vulnerabilities and exposures (Cremer et al., 2022). Finally, we apply the aggregated cyber risk topic occurrences and impact scores to historical data to predict future emerging risk topics.

In general, research that identifies cyber risks lacks a method to generalise the CVEs into more meaningful groups. For example, many vulnerabilities are related to Microsoft Windows with different CVE codes. There were more than 140 K CVEs up to the year 2020. This

* Corresponding author.
E-mail addresses: arash.kia@ul.ie (A.N. Kia), finbarr.murphy@ul.ie (F. Murphy), barry.sheehan@ul.ie (B. Sheehan), darren.shannon@ul.ie (D. Shannon).

https://doi.org/10.1016/j.eswa.2023.121599
Received 3 May 2023; Received in revised form 1 September 2023; Accepted 10 September 2023
Available online 16 September 2023

highly granular data inhibits an analysis leading to a more general understanding of cyber vulnerabilities' underlying structure. A famous effort to categorise CVEs into more general groups is Common Weakness Enumeration (Christey et al., 2013). However, besides helping define a categorisation platform, Common Weakness Enumeration (CWE) has been criticised as being inaccurate, incomplete, inconsistent, or ambiguous (Black et al., 2015). There were more than 600 CWEs up to the year 2020.

To our knowledge, none of the models built to predict cyber risk are comprehensive. Each model pays attention to one dimension of cyber risk or a specific type. Chen et al. (2018) tried to forecast if a new CVE would be rejected by using its records as input features and Random Forest (RF), SVM, and Naïve Bayes (NB) models as binary forecasters. Bilge et al. (2017) used binary files data inside a network to discover new cyber risks. Schultz et al. (2020) built a system to forecast damages on network assets using probabilistic models. Application-based cyber risk prediction models for IoT, Connected and Autonomous Vehicles (CAV), etc., have also been mentioned.

In risk prediction and system protection, leveraging a 'first principles' methodology ensures that foundational truths and basic concepts are at the heart of the analysis. One can derive more reliable and insightful conclusions by dissecting complex systems into their elementary components and understanding their inherent properties and interrelations. A notable application of this approach is seen in the work of Chai et al. (2011), where they employ the first principles combined with social network theory to prioritise protection in the interconnected oil and gas industries within a networked critical infrastructure system. Their study breaks down the intricate web of industry interactions into elemental relationships, emphasising the fundamental dynamics of the network. Drawing inspiration from such a methodology, our model, $CyRiPred$, seeks to predict cyber risks using a data-driven approach rooted in the foundational data of Common Vulnerabilities and Exposures (CVE), circumventing the biases often inherent in expert-driven analyses.

Our model identifies a cyber risk topic in real-time, meaning that the model extracts the related cybersecurity topic whenever a new vulnerability comes into the National Vulnerabilities Database (NVD). This feature in our model differentiates it from the CWE, whose shortcomings are overcome by our novel $WikiTopic$ topic extraction algorithm that uses CVE categorisation in cyber risk groups. Using an open-source knowledge repository like Wikipedia also has the advantage of achieving a higher resolution for cyber risks compared to CWE, which has just over 600 enumerations. Wikipedia had been used as a topic repository in non-explicit topic extractors (Yun et al., 2011) and in topic mapping to video data (Roy et al., 2011). Wikipedia is the most extensive knowledge repository and the most comprehensive ontology resource at hand (Ensslin, 2011; Miz et al., 2017; Pilkauskas, 2010; Sinanc & Yavanoglu, 2013). The authors know that Wikipedia is not an academic peer-reviewed encyclopedia of cybersecurity topics. However, it is the best open-source knowledge repository at hand that is used in many intelligent NLP-based based tools in different expert domains (del Valle et al., 2018; Lämmel et al., 2013; Noraset et al., 2021; Schultz et al., 2000). After retrieving the topics, the likelihood of topic communities' occurrences and their severity from CVSS are extracted for each time unit. Then, the extracted time series of cyber risk topics occurrences and impacts are used to calculate the time series's risk score. Finally, the computed series are fed to ML prediction models. The model can update its prediction with every instant update in the NVD.

Our study provides a model for categorising the types of cyber risks as well as predicting the future of these risks based on an automated data-driven method. The model only requires CVE data, and then it identifies and predicts cyber risks automatically with novel Natural Language Processing (NLP) and Machine Learning (ML) algorithms. Cyber-attacks constantly change, and there is a societal and market demand for forward-looking early-warning systems (Kopp et al., 2017).

In the next section, a general overview of the model is presented. The $WikiTopic$ algorithm is explained in detail, describing the risk calculation and forecast models. Section 3 discusses the CVE dataset and offers some of its statistics. In Section 4, the results of the $CyRiPred$ are presented. Section 5 evaluates the model and all its modules and $WikiTopic$ algorithm. In Section 6, we discuss the results produced by our novel $CyRiPred$ model before a final concluding section.

2. Methodology

2.1. General overview of $CyRiPred$

$CyRiPred$ stands for cyber risk recognition and prediction model. The presented model is data-driven, taking historical data on cybersecurity vulnerabilities and exposures as input and predicting future cyber risks. CVEs in the NVD dataset are used as input. In our proposed methodology, for each CVE in NVD, we focus on three fields: a text field that describes the vulnerability or exposure, a numerical field called base score and records the severity of the vulnerability or exposure, and a date field that shows the first time the vulnerability or exposure was discovered. A comprehensive explanation of the dataset used in the proposed method is presented in Section 3.

Fig. 1 presents the general steps of our proposed methodology. In this section, each step – each box in Fig. 1 – is explained in detail. Initially, each description field in each CVE will be entered as input to a topic extraction algorithm to be labelled with a cybersecurity topic. In this way, significant groups of cybersecurity problems are recognised. The next step is to calculate the impact and likelihood of the occurrence of cybersecurity topics for each month and year (or any other time unit). We need to calculate the likelihood and impact of each cybersecurity topic to calculate the risk score. Risk is calculated as in Eq. (1).

$$Risk = Likelihood \times Impact \tag{1}$$

The likelihood of the occurrence of a cybersecurity topic is the total number of specific cybersecurity topics per time unit. The CVSS base score can be considered as an indicator of the impact of a CVE (Fruhwirth & Mannisto, 2009; Houmb et al., 2010). We calculate the average of base Scores for each cybersecurity topic as the impact in risk calculation. Finally, we will have each cybersecurity topic's time series of occurrence, impact, and risk score. By entering these time series as the input of a prediction model, we can produce a prediction for the next time unit (e.g., month or year).

Our model identifies known and new risks from an updating knowledge repository (in our case, Wikipedia) by assigning risk labels (entry topics) to new vulnerabilities and exposures from an updating dataset of vulnerabilities (in our case, CVE dataset) with the help of using a novel text-mining algorithm. Our model can attach to two updating knowledge repositories (One with risk labels and one with vulnerabilities and exposures) and identify and rank known and future cyber risks. The model predicts future risks based on the discovered patterns of occurrences and impacts of identified risks from the knowledge repository (Wikipedia in our case). The overview of $CyRiPred$ implementation is also presented graphically in Fig. 2. Here we should mention an important issue: In prediction research, there are two types of future risks: One that has been seen in the past or now or we have a trace of it in our data and those that have not been seen before. The nature of Machine Learning models is data-driven. They assume that the future can be predicted based on past historical data. The model cannot predict if a pattern has been seen in the historical data. Machine Learning methods are designed to forecast the future based on past data (they are data-driven). They cannot foresight a risk that has not been seen before, and there is no trace of it in the historical data. This can be seen as a limitation of data-driven machine-learning models. However, our novel model uses updating knowledge repositories for risk labelling and vulnerabilities and exposures. This helps our model capture new risks with an entry/label in our knowledge repository (Wikipedia).

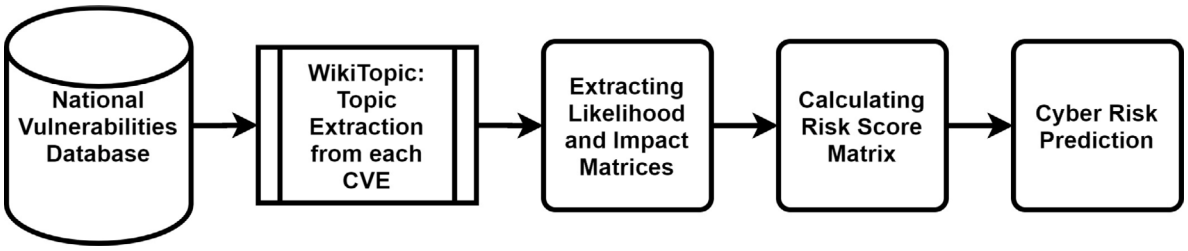


Fig. 1. General steps of the $CyRiPred$ methodology.

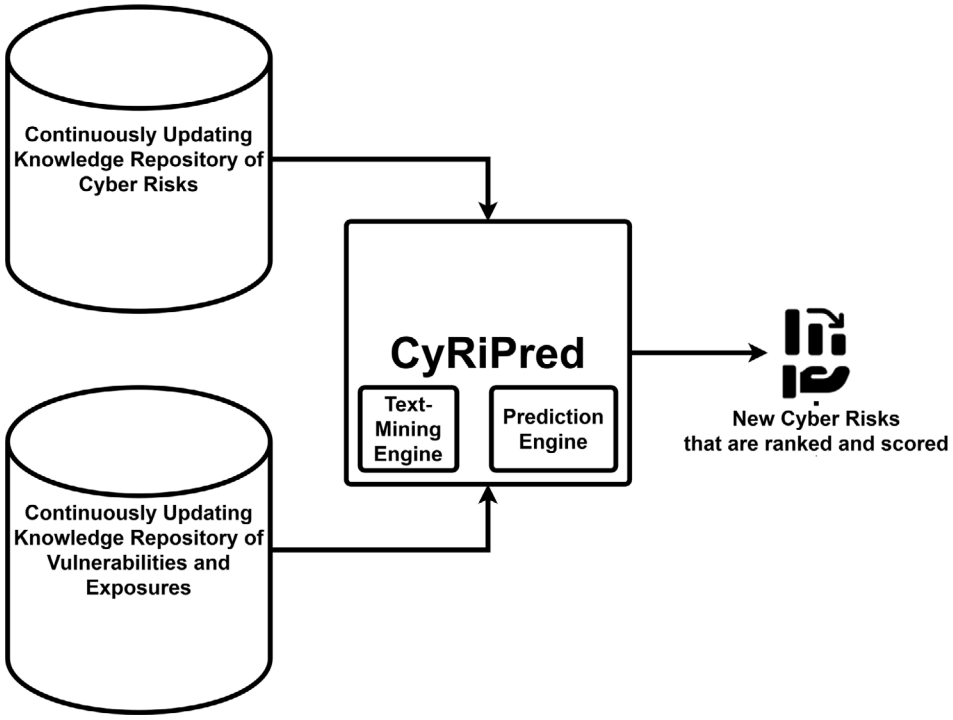


Fig. 2. An overview of how $CyRiPred$ is implemented.

2.2. $WikiTopic$: A novel explicit topic extraction algorithm

In the first step, a descriptive textual field of each CVE is entered into a topic extraction algorithm that explicitly assigns each CVE to a cybersecurity topic. All CVEs with the same cybersecurity topic are put into the same category. These categories can be considered clusters in unsupervised learning methods.

Most existing topic modelling algorithms provide multiple topics as output, where each topic contains various words, and each word in the topic has a probability score showing its importance in the extracted topic (Alghamdi & Alfalqi, 2015; Daud et al., 2010; Jelodar et al., 2019; Padmaja et al., 2018). These algorithms have statistical assumptions, for example, about the distribution of words and topics. In these algorithms, the algorithm user must set the number of issues. In our study, we do not assume prior knowledge about the number of cybersecurity topics that can be extracted from the NDV database. The most common topic modelling algorithms are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet allocation (LDA) (Blei & Lafferty, 2006; Blei et al., 2003; Dumais, 2004). These models require extensive development and modification to work with short texts (Qiang et al., 2020). Some other algorithms developed for short texts have the same drawbacks of too many parameters to adjust and default statistical assumptions about word frequency distribution in the corpus. These issues with the conventional topic modelling algorithms and the requirement for an algorithm that explicitly outputs one topic per CVE led us to develop our explicit topic extraction method called $WikiTopic$.

In the $WikiTopic$ algorithm, Wikipedia is used as a comprehensive open-source human-knowledge repository of topics. As a data source, Wikipedia offers advantages like a balanced perspective, high reliability due to constant editing, and comprehensive coverage of topics. Meanwhile, social media can sometimes introduce biases due to user sentiment or localised events. . Wikipedia topics have been used in research to extend LDA, but the problems mentioned before with LDA, such as assigning the number of topics and distribution assumption, remain (Hansen et al., 2013). Our approach is simple; we want to assign the topic of the most similar Wikipedia article to the input text. The key here is the definition of similarity. To find the most similar Wikipedia article to a text, we used the search API provided by Wikipedia and keywords extracted from the text by a keyword extraction algorithm. For keyword extraction, we use the Yet Another Keyword Extraction (YAKE) algorithm that performs better than other keyword extraction algorithms (Campos et al., 2018a, 2018b, 2020). YAKE is a statistical unsupervised keyword extraction algorithm that uses five different features for each term in text and calculates a keyword candidate score. Then, it outputs keywords by the computed scores. For further details of the YAKE, the interested reader can read (Campos et al., 2020).

Fig. 3 shows the flowchart of the proposed $WikiTopic$ algorithm. At first, the input text should be prepared by lemmatising and removing the stop-words. Lemmatising is the process of extracting each word's

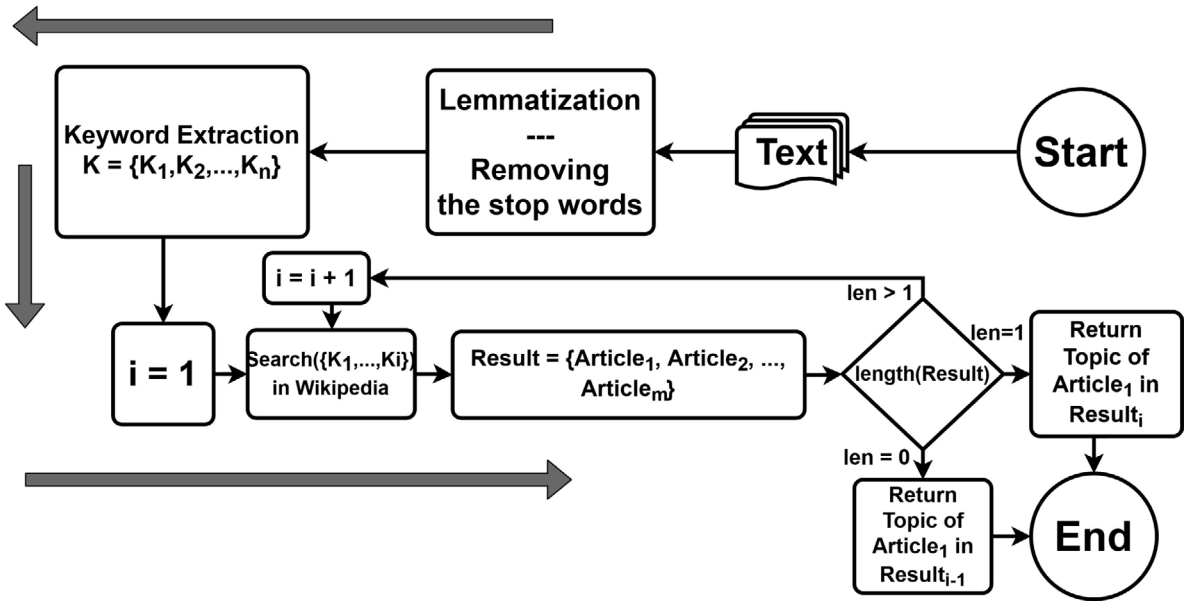


Fig. 3. Flowchart of *WikiTopic* explicit topic extraction algorithm.

root in the text and replacing it with the actual word. Stop words are the most common words used in a language. Examples of stop-words include "a, an, the, of" and other common words. The lemma, which is also called the non-infected or dictionary form of the word, is also a better input for many NLP applications and can increase the performance of keyword extraction algorithm and Wikipedia search API and reduce confusion that may happen with different inflected forms of a word (Müller et al., 2015).

The data preparation (Lemmatization and stop-word elimination) produces a list of keywords for the input text. Assume $K = \{K_1, K_2, \ldots, K_n\}$ are n different keywords extracted from the text where their importance as a keyword sorts them in decreasing order. The keyword importance score comes from the YAKE algorithm, which takes Term Frequency – Inverse Document Frequency (TF-IDF) and other variables into account (Campos et al., 2020). This means that K_1 is the most important keyword in the text, and K_n is the least important one. The *WikiTopic* algorithm uses iteration to search Wikipedia with an incremental list of $K' = \{K_1, K_2, \ldots, K_i\}$. In the first iteration of *WikiTopic*, the Wikipedia articles are searched with the K_1 keyword. Then, the Wikipedia API returns a list of articles as the search output with this keyword. In the second iteration of the *WikiTopic*, the Wikipedia is searched with K_1 and K_2. The iteration continues until the search result list becomes empty or only one article. Adding keywords to the search input narrows the search until we only have one article or no articles. If the search result's length is one, then the article's topic returned from the search will be the topic of the input text. Suppose the search result becomes empty in an iteration. In that case, the first search result of the previous iteration will be considered the closest article to the input text, and its topic will be returned as the output of *WikiTopic*. The *WikiTopic* is available to readers as an open-source Python package in PyPi and GitHub under GNU General Public License version v3.0.[1]

After running the *WikiTopic* on all CVEs, we will have a new field (feature) in our dataset, the cybersecurity topic for each CVE. Eq. (2) shows the list of CVEs and the list of topics as W_1 *to* W_p. Assume that we will have q different cybersecurity topics in the Topics list. Each of these cybersecurity topics will correspond to a column in Occurrence, baseScore, and Risk matrices in the next section.

$$\begin{aligned} &CVEs = \{CVE_1, CVE_2, \ldots, CVE_p\} \\ &Cybersecurity\ Topics = WikiTopic(CVEs) = \{W_1, W_2, \ldots, W_p\} \\ &Distinctive\ Cybersecurity\ Topics = T_1, T_2, \ldots, T_q \end{aligned} \tag{2}$$

2.3. Likelihood, impact, and risk calculation

Given q specific cybersecurity topics as shown in Eq. (2), we make two matrices with $t \times q$ dimension where t is the total number of data in each column (each time series of topic occurrence and baseScore for each cybersecurity topic). The time unit of each time series (each column in matrices of Eq. (3)) can be day, month, or year, depending on the aggregation method used on the date feature of the primary CVE dataset. There will be a comprehensive explanation of the CVE dataset in Section 3, where we will talk about the dataset used as the input of our *CyRiPred* model.

$$\begin{aligned} Topic\ Occurrence\ Matrix &= \begin{bmatrix} T_{1,1} & \cdots & T_{1,q} \\ \vdots & \ddots & \vdots \\ T_{t,1} & \cdots & T_{t,q} \end{bmatrix} \\ baseScore\ Matrix &= \begin{bmatrix} B_{1,1} & \cdots & B_{1,q} \\ \vdots & \ddots & \vdots \\ B_{t,1} & \cdots & B_{t,q} \end{bmatrix} \end{aligned} \tag{3}$$

From Eq. (3) and Topic Occurrence Matrix, $T_{i,j}$ is the number of occurrences for cybersecurity topic j in time i. In the *baseScore* Matrix, $B_{i,j}$ is the average base Score for cybersecurity topic j in time i. The risk equation will be written as Eq. (4), where the inner product of two matrices of Occurrence and *baseScore* will result in the risk score matrix.

$$RiskScore = Topic\ Occurrence \bullet baseScore \quad (innerproduct) \tag{4}$$

The risk score matrix is presented in Eq. (5) where $R_{i,j}$ is the risk score for cyber risk related to cybersecurity topic j in time i.

$$Risk\ Score\ Matrix = \begin{bmatrix} T_{1,1}B_{1,1} & \cdots & T_{1,q}B_{1,q} \\ \vdots & \ddots & \vdots \\ T_{t,1}Bt,1 & \cdots & T_{t,q}B_{t,q} \end{bmatrix} = \begin{bmatrix} R_{1,1} & \cdots & R_{1,q} \\ \vdots & \ddots & \vdots \\ R_{t,1} & \cdots & R_{t,q} \end{bmatrix} \tag{5}$$

In Fig. 4, a glimpse of what happens in actual data and results is presented to view better the significance of Eq. (5). The column

[1] *WikiTopic* Python Source Code and Package Information: https://github.com/ConKruG/WikiTopic

Occurrence (Frequency) × **baseScore (Impact)** = **Risk Score**

Date	Transport Layer Security	Cross-site scripting	...	SQL injection
2000	12	2	...	0
2001	18	12	...	7
⋮	⋮	⋮	...	⋮
2010	273	166	...	212
⋮	⋮	⋮	...	⋮
2019	226	155	...	207
2020	140	105	...	150
2021	Prediction	Prediction	...	Prediction

×

Date	Transport Layer Security	Cross-site scripting	...	SQL injection
2000	5.94	8.75	...	No Data
2001	6.31	6.53	...	6.67
⋮	⋮	⋮	...	⋮
2010	6.59	4.68	...	7.02
⋮	⋮	⋮	...	⋮
2019	5.18	4.48	...	5.27
2020	5.23	4.55	...	5.27
2021	Prediction	Prediction	...	Prediction

=

Date	Transport Layer Security	Cross-site scripting	...	SQL injection
2000	71.30	17.50	...	0.00
2001	113.60	78.40	...	46.70
⋮	⋮	⋮	...	⋮
2010	1798.40	777.50	...	1487.30
⋮	⋮	⋮	...	⋮
2019	1170.90	694.90	...	1090.00
2020	732.67	477.65	...	790.54
2021	Prediction	Prediction	...	Prediction

Fig. 4. Cyber risk score calculation after topic extraction and likelihood (Frequency or Occurrence) and baseScore (Impact) calculation. Transport Layer Security, Cross-site scripting, and SQL injection are three samples of topics extracted from the CVE description field with the $WikiTopic$ Algorithm. The multiplication operator in the figure is an inner product (Multiplication element by element).

Risk Classification Matrix =

Likelihood \ Severity	1	2	3
1	Low	Low	Medium
2	Low	Medium	High
3	Medium	High	High

Fig. 5. Risk classification matrix.

headers are some of the topics extracted from CVEs description fields. These topics are common for many CVE descriptions and, therefore, make clusters of CVEs under the name of a cybersecurity topic. Let us consider each column as a time series for the likelihood (frequency), $baseScore$ (impact), and the cyber risk score of each cybersecurity topic. We can use a prediction model to predict the future values for these time series.

Before discussing the prediction phase, it is instructive to consider the possibility of classifying cyber risks under the $CyRiPred$ model. Dividing the intervals of continuous values for occurrence and $baseScore$, risk can be classified with a risk classification matrix such as Fig. 5. According to NIST (Information Technology Laboratory, 2020), in some applications, the $baseScore$ is discretised by dividing it into three or five intervals. For example, in the three-interval division, the $baseScore$ interval from 0 to 3.99 indicates low vulnerability or exposure severity, 3.99 to 7 indicates medium severity, and 7 to 10, high severity. The different discretisation of the continuous values of $baseScore$ is presented in NIST-NVD online documents (ibid). The likelihood of a cybersecurity topic in our work is the number of occurrences of CVEs with that topic in a period. We can scale all the likelihood data into any interval and then discretise it the same as the $baseScore$. After discretising $baseScore$ and Topic Occurrence, we can use a risk classification matrix to classify cybersecurity risks extracted under $WikiTopic$, CVE clustering, and $baseScore$, $TopicOccurrence$ aggregation for the desired period (desired time unit).

2.4. Cyber risk prediction

Each column of $Occurrence$, $baseScore$, and $CyberRisk$ matrices in Eqs. (3), and (5) and in Fig. 4 are a time series of a cybersecurity topics likelihood, impact, and risk. For example, the time series of cybersecurity risk i is presented in Eq. (6). The $PredictionModel$ in Eq. (6) can be any regressor with a tolerable error in the application domain. Our model's cyber risk score is computed by predicting the risk time series for each cyber risk topic, as explained in Eq. (6). The cyber risks are first calculated using known occurrence and impact data, and then future risks are predicted. The minimum acceptable performance is a function of business conditions and the maximum prediction performance that business experts can provide without using an intelligent system. It may also be possible that each time series (each column in one of the three matrices for likelihood, impact, and risk corresponding to a different cybersecurity topic) can be more accurately predicted with a different regressor.

$$\begin{aligned} S_i &= R_{1,i}, R_{2,i}, \dots, R_{t,i} \\ R_{t+1,i} &= PredictionModel\left(S_i\right) \end{aligned} \tag{6}$$

After predicting all the values for the next year or month for likelihood, impact, and cyber risk score for cybersecurity topics, we classify each cyber risk with the risk matrix classification method in Fig. 5 or sort the cybersecurity topics to find out top-ranking future cybersecurity issues. In Section 4, we find a random forest regressor had the best performance in terms of average accuracy for all the extracted cybersecurity topics. Random forest is an ensemble machine learning model that consists of a set of decision trees on random subsets of a dataset (Ho, 1998). Each decision tree is trained on a random subset of a delay-embedded time series in our study. It should be mentioned that the time series should be converted to a delay-embedded format before being fed to the random forest. In a delay-embedded format, each time series will convert to a table where each row will have n consecutive elements, and the last element will be used as the training label. Random forest randomly selects several rows and columns of the delay-embedded time series and creates random subsets as new datasets. Finally, a decision tree regressor is trained for each subset. Each decision tree regressor gives us a prediction. The average prediction of all decision trees on all random subsets of data is the final prediction of the random forest model. It is important to mention that there might be correlation between different cyber risks that may provide an opportunity to predict multi-variate time series to capture these interrelationships. However, due to lack of enough data to capture the interrelationship between all the extracted time series and simplicity of the methodology using uni-variate time series prediction, we sticked with the univariate approach. We acknowledging the potential for future work in multivariate approaches.

3. Data

NVD presents data in both Javascript Object Notation (JSON) and eXtensible Markup Language (XML) format.[2] There are many different fields in the database for each CVE. Among all the features, we used three fields: Description (to use in topic extraction algorithm and label a topic to each CVE), $baseScore$ (to calculate the severity or impact of each cybersecurity topic group), and Date (to use in averaging and summation aggregation and sequencing the data in time series used in prediction models). These fields and two other fields are shown in Table 1. For clarity, we added two fields, a unique CVE code and a qualitative discretised field called Severity that labels each CVE into three classes of Low, Medium, and High according to their $baseScore$. The last column of Table 1 is the topic extracted from the Description field after running the $WikiTopic$ algorithm on the dataset.

[2] [dataset] CVE database: lastaccessed:1thJan.2021

Table 1
A glimpse of the dataset used in $CyRiPred$ model. The results of the $WikiTopic$ algorithm are presented in the "Retracted Topic with WikiTopic" field. These results should be read as "Related to PHP" or "Related to DragonFly BSD". A "Related to" is better to put at the beginning of all the topic results from the $WikiTopic$ so that it makes sense.

Date	CVE code	Description	baseScore	Severity	Extracted topic with $WikiTopic$
10/1/1988	CVE-1999–0095	The debug command in Sendmail is enabled, allowing attackers to execute commands as root.	10	HIGH	DragonFly BSD
⋮	⋮	⋮	⋮	⋮	⋮
2/8/2012	CVE-2012–1023	Open redirect vulnerability in admin/index.php in 4images 1.7.10 allows remote attackers to redirect users to arbitrary websites and conduct phishing attacks via a URL in the redirect parameter.	5.8	MEDIUM	PHP
⋮	⋮	⋮	⋮	⋮	⋮
9/1/2020	CVE-2020–6137	SQL injection vulnerability exists in the password reset functionality of OS4Ed openSIS 7.3. The password_stf_email parameter in the password reset page /opensis/ResetUserInfo.php is vulnerable to SQL injection. An attacker can send an HTTP request to trigger this vulnerability.	7.5	HIGH	Password cracking
⋮	⋮	⋮	⋮	⋮	⋮

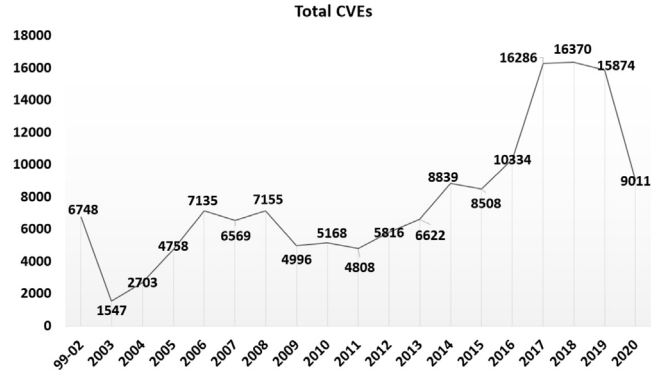


Fig. 6. Total number of CVEs for each year.

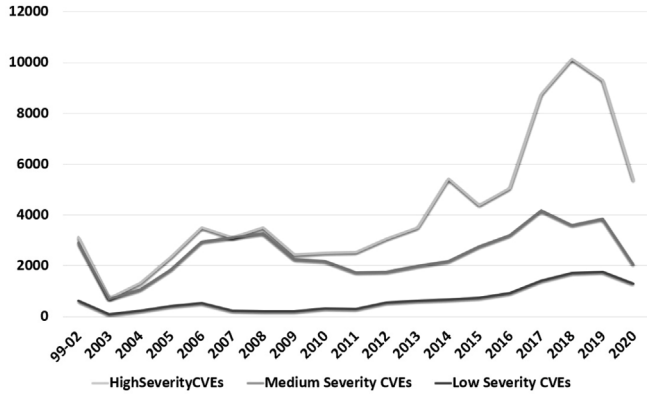


Fig. 7. Total number of severity classes recorded for CVEs per year.

There are 149,247 CVEs in our dataset. After eliminating the rejected/disputed data labelled in the NVD database, 140,213 CVEs remain. The first CVE was from 1988 and was recorded in NVD in 1999. The last CVE in our dataset is for 4th September 2020. The NVD is updated daily with new CVEs.

Fig. 6 shows graphs of total CVEs per year. The downward movement for 2020 is due to our dataset's not having data for the months after September. An increasing pattern is seen in the recorded CVEs over the years.

Fig. 7 shows the total number of CVEs with three different severity classes in NVD. The graph shows more increase in the high severity class than two other classes in recent years.

4. Results

In this section, the likelihood (number of occurrences of a cybersecurity topic group), the impact severity (average baseScore for a topic group), and the cybersecurity risk scores calculated for the extracted topic groups are ranked for each year and all years. There were 20,181 different cybersecurity topics extracted from the dataset, with 7875 topics with more than one occurrence. The top ten rankings for likelihood, impact, and cyber risk and the predictions for next year (2021) are then presented as an important part of the results made by $CyRiPred$. For some of the results (the first three cybersecurity topic groups in each table), we will briefly explain the topic itself. However, for further studies and more information on the technical details of computer networks and security, the reader can refer to Peterson and Davie (2007) and Gollmann (2010).

Table 2 presents the most frequent cybersecurity topic groups for the whole dataset. The results show that Transport Layer Security issues are on the top list of occurrences among all other cybersecurity problems following Cross-site scripting and SQL injection attacks. Transport Layer is the third layer of TCP/IP protocol (Transmission Control Protocol/ Internet Protocol). Cross-site scripting allows attackers to run client-side scripts from a website in the victim's client. In SQL injection attacks, malicious inputs can change the behaviour of an application that uses SQL in the back end to help the attacker gain unauthorised access or meddle with other security functions like authentication, confidentiality, anonymity, and privacy.

Table 3 presents the cybersecurity topic groups with the highest risk score among all the cybersecurity topics extracted with $CyRiPred$. Again, the first and third ranks are for Transport Layer Security issues and SQL injection attacks. But the second rank is for Adobe Flash Player

Table 2
Top ten most frequent cybersecurity topic groups among the CVEs in the NVD for all the years.

Rank	1	2	3	4	5	6	7	8	9	10
Most Frequent CAV Topics	Transport Layer Security	Cross-site scripting	SQL injection	List of TCP and UDP port numbers	Adobe Flash Player	IOS version history	Firefox version history	Cross-site request forgery	HTTP cookie	JavaScript

Table 3
Top ten riskiest cybersecurity topic groups among the CVEs in the NVD for all the years.

Rank	CAV topic	Impact (Average BaseScore)	Total	RiskScore
1	Transport Layer Security	6.19	3781	23 392.8
2	Adobe Flash Player	7.98	2862	22 832.3
3	SQL injection	6.62	3182	21051
4	List of TCP and UDP port numbers	5.89	3051	17 978.5
5	Cross-site scripting	4.81	3428	16 482.7
6	IOS version history	6.66	2356	15 686.4
7	Firefox version history	6.36	2107	13 397.6
8	Cross-site request forgery	5.4	1735	9371.9
9	Code injection	6.75	1070	7227.5
10	JavaScript	5.55	1282	7119.5

Table 4
Top ten most frequent cybersecurity topic groups for 2019, 2020 and prediction for 2021. In the evaluation section of the study, we predict 2019 and 2020 against actual observations.

Rank	2019	2020	2021 (level change from previous year)
1	List of TCP and UDP port numbers	List of TCP and UDP port numbers	List of TCP and UDP port numbers
2	IOS version history	**ZFS**	IOS version history (+3)
		(Not in the prediction for the next year)	
3	Adobe Flash Player	Windows Vista	Firefox (+1)
4	Firefox	Firefox	**Adobe Flash Player (New)**
5	Adobe ColdFusion	IOS version history	Transport Layer Security (+4)
6	Transport Layer Security	Android (operating system)	SQL injection (+1)
7	List of HTTP status codes	SQL injection	Windows Vista (−4)
8	SQL injection	**Cracking of wireless networks**	Adobe ColdFusion (+2)
		(Not in the prediction for the next year)	
9	Cross-site scripting	Transport Layer Security	Android (operating system) (−3)
10	Intel Active Management Technology	Adobe ColdFusion	**Cross-site scripting (New)**

vulnerabilities and exposures. We do not present a table for the top ten most severe cybersecurity topic groups because there are more than ten topic groups with the highest score of severity ($baseScore$) which is 10.

In Table 4, the most frequent cybersecurity topics are presented for the last two years and also a prediction for the next year, 2021. Expected changes in the ranking for the year 2021 compared to its previous year are also shown in parenthesis. If a cybersecurity topic goes out of the top ten list in the predicted year, it is mentioned in the parenthesis. The Zettabyte File System (ZFS) and "Cracking of wireless networks" are not in the predicted top ten frequent cybersecurity topics for 2021. The ZFS is a file system developed by Sun Microsystems, with the latest release in 2018. The top rank in 2019 and 2020 and the predicted top risk for 2021 is "List of TCP and User Datagram Protocol (UDP) port numbers". This is precisely the title of the most similar Wikipedia article to the CVEs in these years. TCP and UDP ports are abstract connections in layer 3 of TCP/IP. Each port can be assigned to a particular application that needs communication with another client or server. UDP is a sub-protocol in layer 3 of TCP/IP for more error-tolerant connections with less delay. It should also be mentioned that topics like "IOS version history" or "Firefox version history" are automatically extracted for CVEs that talk about iOS or Firefox security related issues in their description field.

Table 5 shows the top ten cyber risks for the last two years (2019–20), and predictions for 2021 are presented. Given our time series of risk scores for each cybersecurity risk for each month from 1988 to 2020, the random forest regressor predicts the cyber risk scores of these cybersecurity topics for 2021. The random forest regressor is chosen due to its highest accuracy on average for all the cybersecurity topic time series. The last column of Table 5 is the main output of $CyRiPred$. Windows Vista is on the list of top ten cybersecurity risks for 2021.

Legacy systems such as Windows Vista are those systems that are old and not updated and supported by the developer company anymore, but they are still in use. Adobe Flash Player and Adobe Cold Fusion (a software for rapid web development) are in the list of predicted top ten cyber risks of 2021 by $CyRiPred$. $CyRiPred$ also predicts that ZFS and wireless cracking will drop off the top ten risks in 2021.

Moreover, the first rank for the last two years and the predicted year will be for "List of TCP and UDP ports". This means that attacks on layer three ports will be the highest risk in the future year. The difference between "Transport Layer Security" and "List of TCP and UDP ports" is that the first one is more general. Attacks, vulnerabilities, and exposures are related to layer three (transport layer) but not necessarily ports.

5. Evaluation of the model

In the evaluation phase for $CyRiPred$, first, we tested two sub-modules of the $CyRiPred$ model: $WikiTopic$ and Random Forest regressor. Then, we evaluated the model by comparing its outputs to the results presented in different cybersecurity trend reports mentioned in the Discussion section. For $WikiTopic$, the topic extractor module, we chose 100 random articles from four other Wikipedia alternatives and their title as our test target labels.

The alternatives to Wikipedia (alternative online open-source encyclopedias to gather labelled test data) used in the evaluation were

Table 5
Top ten cyber risks for 2019, 2020, and prediction for 2021.

Rank	2019	2020	2021 (level change from previous year)
1	List of TCP and UDP port numbers	List of TCP and UDP port numbers	List of TCP and UDP port numbers
2	IOS version history	Windows Vista	IOS version history (+2)
3	Firefox version history	**ZFS (Not in the prediction for the next year)**	Firefox version history (+2)
4	Adobe ColdFusion	IOS version history	**Adobe Flash Player (New)**
5	Adobe Flash Player	Firefox version history	**Adobe ColdFusion (New)**
6	Transport Layer Security	SQL injection	Windows Vista (−4)
7	Intel Active Management Technology	List of Qualcomm Snapdragon systems-on-chip	Transport Layer Security (+3)
8	SQL injection	Android (operating system)	SQL injection (−2)
9	List of HTTP status codes	**Cracking of wireless networks** Not in the prediction for the next year)	Android (operating system) (−1)
10	Android (operating system)	Transport Layer Security	List of Qualcomm Snapdragon systems (−3)

Table 6
Parameters of the base models for prediction.

Model	Parameters
Random Forest	n_estimators = 150
MLP Regressor	(31-7-3-1) - Activation = Logistic_Sigmoid, Learning_Rate = 0.001, Momentum = 0.9
KNN Regressor	K = 7
ARIMA	Best fit of p, d, q for each time series in the validation set
Support Vector Regressor	Kernel = RBF, Regularisation = 1.0

Scholarpedia, Citizendium, Britannica, and Infoplease,[3]. We chose two random paragraphs from each article. Then, we used $WikiTopic$ to extract the topic of these articles (two paragraphs as a delegate for the whole article) and finally calculated the accuracy of the $WikiTopic$. We sampled the articles because, in our application, the description data can be considered a short text (up to one paragraph). We used two sample paragraphs from each article in our test because some parts of the articles, like the history part, are unrelated to the article's topic. From 100 test articles, 96 articles were labelled correctly. The topic from Wikipedia was the same as the one in the alternative encyclopedia. This indicated an accuracy of 97% for the $WikiTopic$ algorithm.

For the prediction module of $CyRiPred$, time series are partitioned into train, validation, and test sets for prediction evaluation. Monthly aggregated data (summation for topic occurrence and average for topic severity) starts from October 1988; the final data is for September 2020. Therefore, the time series will have 354 elements for each series for 354 months. Our prediction model tries to predict 15 months after that to have 12 months of the year 2021. We used 2017 and 2018 as a validation set for parameter selection and 2019 and 2020 for testing. After testing different popular supervised regressors like MLP (Multi-Layer Perceptron), KNN (K-Nearest Neighbour), SVR (Support Vector Regression), RF (Random Forest), and ARIMA (Auto-Regressive Integrated Moving Average), the best results in terms of MSE (Mean Square Error), MAE (Mean Absolute Error), and R squared were for Random Forest (in average for all the topic time series).

Our model's approach toward cyber risk prediction is unique; Therefore, we cannot compare our model to any existing model in the literature. However, we compare Random Forest as the prediction core of our model to other base supervised learners. The models' configurations are presented in Table 6. These configurations have been set using exhaustive grid searches in validation sets.

The MSE (Mean Squared Error), MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error) results are presented in Table 7 for comparison. Formulas for these evaluation criteria are shown in Eq. (7). The MSE and RMSE raise the magnitude of significant errors and lessen the importance of minor errors due to squaring in their formulas. At the same time, MAE treats significant or insignificant errors the same. The R^2 criterion shows how well the data fit in a regression model.

Table 7
Prediction results of different base models.

Model	MSE	MAE	RMSE	R2
Random Forest Regressor	7.868	0.457	2.805	0.981
MLP Regressor	11.923	0.781	3.452	0.970
KNN Regressor	17.123	0.923	4.137	0.965
ARIMA	24.167	0.998	4.916	0.962
Support Vector Regressor	10.188	0.556	3.192	0.979

$$\begin{aligned} MAE &= \frac{1}{n}\sum_{1}^{n}|y_i-\hat{y}_i| \\ MSE &= \frac{1}{n}\sum_{1}^{n}(y_i-\hat{y}_i)^2, \quad RMSE=\sqrt{MSE} \\ R^2 &= 1-\frac{(y_i-\hat{y}_i)^2}{(y_i-\bar{y}_i)^2} \\ \hat{y} &: PredictedValue \\ \bar{y} &: MeanValue \end{aligned} \tag{7}$$

Evaluation criteria for the RF model with 150 regressor trees (150 estimators) were MAE = 0.457, MSE = 7.869, R Squared = 0.982). Finally, after choosing RF as our best fit model, we trained the model again with all 354 data to predict the next 15 elements (3 months in 2020 and 12 months in 2021). Table 8 presents the evaluation results of two main components of CyRiPred.

Table 9 is presented to show how significantly different our Random Forest model results are compared to other models. We used the Diebold–Mariano test for statistical significance results. The Diebold–Mariano test is the best choice for comparing the RMSE, MSE, or MAE of different prediction models. In our case, we compared the RMSEs of other models to the Random Forest model, which showed the best results in the test set. In our test, the H_0 hypothesis meant no difference between RMSEs of Random Forest and any other prediction models. At the same time, H_1 implied a significant difference between the RMSEs. In our Diebold–Mariano test, we assumed that a $P-Value$ less than 0.05 meant that the difference between RMSEs was statistically significant.

[3] Scholarpedia: http://www.scholarpedia.org Citizendium: https://www.infoplease.com, Britannica: https://www.britannica.com, Infoplease: https://www.infoplease.com

Table 8
Test results for WikiTopic topic extraction algorithm and random forest as the best predictor among different regressors.

WikiTopic Topic Extractor			
WikiTopic	Test Size 100 Test Articles with known topic	Alternative Encyclopedias Accuracy = 96%	Wikipedia Accuracy = 97%
Predictor (Test results for 2019–2020)			
Random Forest Regressor (150 Estimators)	MSE = 7.869	MAE = 0.457	R-Squared = 0.982

Table 9
P-values of the Diebold–Mariano test between the random forest model and other models.

Model	p-value
MLP Regressor	< 0.01
KNN Regressor	< 0.01
ARIMA	< 0.01
Support Vector Regressor	$0.042 < 0.05$

One can read the work of Diebold (2015) for a detailed explanation of the Diebold–Mariano test.

For a more qualitative testing approach for the whole $CyRiPred$ model, we excluded the data's last two years, 2019, 2020 and tried to predict cyber risk scores for the cybersecurity topic groups. We saw that the ranks did not change for the top ten cyber risks, and there was also no error in rank prediction for high-occurrence cybersecurity topics.

6. Discussion

We first consider how much the predictions of $CyRiPred$ are similar to the predictions of cyber risk companies.

Many annual cyber risk reports include a prediction for the most critical cyber risks of the following year. Our model does this job entirely data-driven without referring to experts. This benefits the cybersecurity and insurance companies like all those cyber risk reports. The difference between the output results of our model and those reports is that our model's output is expert-bias-free, and the procedure is clear, comprehensive and complete. Cybersecurity and cyber insurance companies can rely on the results of our model to make their future policies better than before when they know what they should expect in the future more accurately.

A comprehensive study by GovTech[4] investigates emerging technologies and cybersecurity risks over 20 different cybersecurity trend prediction reports for 2020. GovTech included the reports of famous companies like Kaspersky, McAfee, AT&T, Gartner, Forbes, and BitDefender. Their reports and other reports like those for Cybriant[5] and Cybercrime Magazine[6] focus on 2021, and their predictions are a mixture of $CyRiPred$'s predictions and their risk perception.

We calculated each topic's average polar sentiment score (instead of $baseScore$, which measured severity in CVSS) to confirm. The sentiment score can indicate risk perception (Chang & Wang, 2018; Gaspar et al., 2016; Wu et al., 2014). Sentiment scores were derived with the Vader Sentiment algorithm (Gilbert & Hutto, 2014). The worst risk perceptions for 2020 and 2021 were Ransomware and Denial-of-service attacks (Rank first and second). The other cybersecurity topics within the top ten worst risk perceptions that were not in the top ten high-risk cybersecurity topics were VirtualBox, File inclusion vulnerability, Java, WordPress, and Heartbleed, an OpenSSL cryptography bug. Ransomware, which has the worst risk perception, were among half of cybersecurity trend predictions for 2020. Other cybersecurity high-risk topics predicted with $CyRiPred$ were among the trend prediction reports of cybersecurity companies. Our predictions are automatically generated and are data-driven, showing the risk perception versus reality issue. We emphasise that our model can predict cyber risk perception just by changing the CVSS baseScore to polar sentiment score.

Statista and IBM report that the most common cybersecurity attacks companies experience are phishing. Attacks like phishing need social engineering techniques where a user is persuaded to open a link. The NVD dataset records vulnerabilities and exposures of IT systems and software. Therefore, $CyRiPred$ cannot make a forecast in the areas where the vulnerability is in the user.

We investigate the completeness of the $CyRiPred$ model from two different aspects: First, we use a comprehensive risk assessment framework named CURF (Core Unified Risk Framework) (Wangen, 2017; Wangen et al., 2018). Then, we investigate the completeness of the vulnerabilities and exposures dataset and the knowledge repository we use as the input of our model. Regarding the CURF cyber risk completeness framework, our model fulfils all the primary activities and parts of the secondary activities of CURF. Here, the primary activities of CURF are mentioned with fulfilled secondary activities in parenthesis inside double quotes: Risk Identification (Assigning risks to vulnerabilities and exposures under "Vulnerability" secondary activity), Risk Estimation ("Probability estimation" using the data-driven occurrence calculation and "impact estimation" using CVSS dataset impact scores), and part of Risk Evaluation ("Risk prioritisation" by ranking risks). Our model mainly lacks the "Risk Treatment" secondary activity from the Risk Evaluation primary activity of the completeness evaluation framework. The "Risk Treatment" was outside the scope of our model, which was built mainly to be a machine learning-based data-driven prediction model.

Regarding the completeness of the datasets we used, it is worth mentioning that the CVE dataset is the most comprehensive repository of vulnerabilities and exposures, continuously updated by the National Cyber Security Division of the U.S. Department of Homeland Security. Wikipedia is chosen as the knowledge repository of cyber risk labelling. Wikipedia is also the largest continuously updating open-source encyclopedia in the world (Zhao & Strotmann, 2021). All these measures guarantee that the completeness of the model will be achieved to the highest standards possible in cyber risk recognition and prediction.

In our CyPiPred framework, the WikiTopic extraction algorithm operates as a sequential process designed meticulously to ensure thoroughness and accuracy in topic extraction. We acknowledge that in today's age of real-time processing, sequential operations may raise concerns about efficiency. However, it is imperative to understand the contextual application of this algorithm. The CyPiPred model is conceptualised as a real-time cyber risk prediction tool and a periodic reporting mechanism. Its primary purpose is to produce comprehensive cyber risk rankings that organisations can utilise to assess and adjust their cybersecurity postures. Given this objective, the model is intended to be run on a scheduled basis, such as monthly, aligning with typical enterprise risk reporting cycles. In such a scenario, the time taken by the WikiTopic algorithm becomes less critical as the emphasis shifts towards accuracy and depth of topic extraction. In essence, while the speed of data processing is undeniably valuable, the

[4] GovTech cybersecurity trends report: https://www.govtech.com/blogs/lohrmann-on-cybersecurity/the-top-20-security-predictions-for-2020.html Online report, [last accessed: 1th Jan. 2021]

[5] Cybriant: https://cybriant.com/plan-today-for-cybersecurity-trends-in-2021/ Online report, [last accessed: 1th Jan. 2021]

[6] Cybercrime Magazine: https://cybersecurityventures.com/top-5-cybersecurity-facts-figures-predictions-and-statistics-for-2019-to-2021/Online report, [last accessed : 1th Jan. 2021]

objective of the CyPiPred model emphasises the quality and reliability of the extracted topics over sheer processing speed. This approach, though seemingly counter-intuitive in real-time, aligns more closely with organisational cybersecurity management's strategic, longer-term decision-making processes.

We also highlight the problem of legacy systems that shows itself in the results of $CyRiPred$. The reader can mention cybersecurity topic groups for systems and applications that have been obsolete for years (legacy systems). This does not mean they do not create new security problems, but the fact that they are no longer supported can make them riskier. Windows Vista topic in recent years is an example of this in the results tables produced by $CyRiPred$.

The synergy between data-driven models and expert opinions is not overlooked in the quest for comprehensive cyber risk prediction. The recent work by Ji et al. (2020) sheds light on a bi-objective optimisation model to aggregate expert opinions, enhancing objectivity and reliability in group decision-making. Their approach revolves around representing experts' views as probability density functions and focusing on the consensus and confidence levels for more grounded predictions. Such a fusion of algorithmic predictions with expert insights, encapsulating objective and subjective perspectives, can be a powerful tool to apprehend unexpected risks. While our model emphasises the algorithmic paradigm, future iterations might benefit from integrating these expert opinion frameworks, offering a more holistic approach to cyber risk prediction.

7. Conclusion

This study provides a general model for identifying major cyber risk groups from the most comprehensive database of security vulnerabilities. In addition to identifying and classifying cyber risks, our model predicts future risks based on the extracted time series. Using out-of-sample techniques, we demonstrate the efficacy of our model and further confirm this with detailed robustness checks.

This research overcomes the inefficiency of previous models to provide a comprehensive method for recognising and predicting cyber risks. Our $CyRiPred$ model does not have pre-assumptions and is not limited to specific applications, environments, or architectures.

As a suggestion for future research, extracting more complex structures compared to a simple classification of cyber risks from CVEs can be considered. Multi-level taxonomies can be extracted, and their evolution studied. This can potentially uncover the process of the creation of new cyber risks. By availability of more data in future, using different deep-learning models can be another choice to evaluate higher-performance models. In our research, we tried to show "HOW" our method predicts the cyber risks. As another suggestion for future research, interested researchers can think of "WHY"s instead of "HOW"s, which means discussing why a cybersecurity topic rank fluctuates in time. In light of our findings and the ever-evolving landscape of cyber threats, a logical next step would development of robust policies tailored to mitigate the highlighted risks. Given our model's focus on predicting cyber risks through CVE data, policy development can leverage these predictions to formulate proactive strategies that pre-emptively address potential vulnerabilities. This would bridge the gap between predictive analytics and actionable measures, offering a more comprehensive approach to cyber risk management. Exploring this synergy between prediction and policy could usher in a new era of cyber resilience, further enhancing the security of digital infrastructures.

Abbreviation

Application Programming Interface (API), Auto-Regressive Integrated Moving Average (ARIMA), Core Unified Risk Framework (CURF) , Common Vulnerabilities and Exposures (CVE), Common Vulnerabilities Scoring System (CVSS), Common Weakness Enumeration (CWE), Cyber Risk Recognition and Prediction Model ($CyRiPred$), Decision Support System (DSS), JSON (JavaScript Object Notation), K-Nearest Neighbour (KNN), Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Machine Learning (ML), Mean Absolute Error (MAE), Multilayer Perceptron (MLP), Mean Squared Error (MSE), Naïve Bayes (NB), National Institute of Standards and Technology (NIST), Natural Language Processing (NLP), National Vulnerabilities Database (NVD), Probabilistic latent semantic analysis (PLSA), Support Vector Machine (SVM), Support Vector Regression (SVR), Transmission Control Protocol/Internet Protocol (TCP/IP), Term Frequency/Inverse Document Frequency (TF/IDF), User Datagram Protocol (UDP), eXtensible Markup Language (XML), Yet Another Keyword Extraction (YAKE), Zettabyte File System (ZFS)

CRediT authorship contribution statement

Arash Negahdari Kia: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Visualization. **Finbarr Murphy:** Conceptualization, Methodology, Investigation, Writing – review & editing. **Barry Sheehan:** Conceptualization, Methodology, Investigation, Writing – review & editing. **Darren Shannon:** Conceptualization, Methodology, Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

CVE data is publicly available.

Acknowledgement

This work was funded by the European Union's Horizon 2020 research and innovation program via the MALAGA Project under grant agreement No 844864 funded this work.

References

Alali, M., Almogren, A., Hassan, M., Rassan, I., & Bhuiyan, M. (2018). Improving risk assessment model of cyber security using fuzzy logic inference system. *Computers & Security*, *74*, 323–339. http://dx.doi.org/10.1016/j.cose.2017.09.011.

Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications(IJACSA)*, *6*(1), http://dx.doi.org/10.14569/IJACSA.2015.060121.

Bilge, L., Han, Y., & Dell'Amico, M. (2017). Riskteller: Predicting the risk of cyber incidents. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. http://dx.doi.org/10.1145/3133956.3134022.

Black, P., Bojanova, I., Yesha, Y., & Wu, Y. (2015). Towards a periodic table of bugs. In *15th High confidence software and systems conference*.

Blei, D., & Lafferty, J. (2006). Correlated topic models. In *Advances in neural information processing systems. Vol. 18* (p. 147).

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176. http://dx.doi.org/10.1109/COMST.2015.2494502.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2018a). A text feature based automatic keyword extraction method for single documents. In *European conference on information retrieval*. http://dx.doi.org/10.1007/978-3-319-76941-7_63.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2018b). YAKE! Collection-independent automatic keyword extractor. In *European conference on information retrieval*. http://dx.doi.org/10.1007/978-3-319-76941-7_80.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289. http://dx.doi.org/10.1016/j.ins.2019.09.013.

Chai, C.-L., Liu, X., Zhang, W., & Baber, Z. (2011). Application of social network theory to prioritizing Oil & Gas industries protection in a networked critical infrastructure system. *Journal of Loss Prevention in the Process Industries*, *24*(5), 688–694.

Chang, W.-L., & Wang, J.-Y. (2018). Mine is yours? Using sentiment analysis to explore the degree of risk in the sharing economy. *Electronic Commerce Research and Applications*, *28*, 141–158. http://dx.doi.org/10.1016/j.elerap.2018.01.014.

Chen, Q., Bao, L., Li, L., Xia, X., & Cai, L. (2018). Categorizing and predicting invalid vulnerabilities on common vulnerabilities and exposures. In *2018 25th Asia-Pacific software engineering conference*. http://dx.doi.org/10.1109/APSEC.2018.00049.

Christey, S., Kenderdine, J., Mazella, J., & Miles, B. (2013). *Common weakness enumeration*. Mitre Corporation.

Cremer, F., Sheehan, B., Fortmann, M., Kia, A. N., Mullins, M., Murphy, F., & Materne, S. (2022). Cyber risk and cybersecurity: A systematic review of data availability. In *The Geneva papers on risk and insurance-issues and practice* (pp. 1–39). Springer.

Daud, A., Li, J., Zhou, L., & Muhammad, F. (2010). Knowledge discovery through directed probabilistic topic models: A survey. *Frontiers of Computer Science in China*, *4*(2), 280–301. http://dx.doi.org/10.1007/s11704-009-0062-y.

del Valle, E. P. G., García, G. L., Santamaria, L. P., Zanin, M., Ruiz, E. M., & González, A. R. (2018). Evaluating Wikipedia as a source of information for disease understanding. In *2018 IEEE 31st international symposium on computer-based medical systems* (pp. 399–404). IEEE.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, *33*(1), 1. http://dx.doi.org/10.1080/07350015.2014.983236.

Dumais, S. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*(1), 188–230. http://dx.doi.org/10.1002/aris.1440380105.

Ensslin, A. (2011). "What an un-wiki way of doing things": Wikipedia's multilingual policy and metalinguistic practice. *Journal of Language and Politics*, *10*(4), 535–561.

Fruhwirth, C., & Mannisto, T. (2009). Improving CVSS-based vulnerability prioritization and response with context information. In *2009 3rd International symposium on empirical software engineering and measurement*. http://dx.doi.org/10.1109/ESEM.2009.5314230.

Gaspar, R., Pedro, C., Panagiotopoulos, P., & Seibt, B. (2016). Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers in Human Behavior*, *56*, 179–191. http://dx.doi.org/10.1016/j.chb.2015.11.040.

Gilbert, C., & Hutto, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international conference on weblogs and social media* (p. 14).

Gollmann, D. (2010). Computer security. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(5), 544–554. http://dx.doi.org/10.1002/wics.106.

Hansen, J., Ringger, E., & Seppi, K. (2013). Probabilistic explicit topic modeling using wikipedia. In *Language processing and knowledge in the web* (pp. 69–82). Springer: http://dx.doi.org/10.1007/978-3-642-40722-2_7.

Ho, T. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. http://dx.doi.org/10.1109/34.709601.

Houmb, S., Franqueira, V., & Engum, E. (2010). Quantifying security risk level from CVSS estimates of frequency and impact. *Journal of Systems and Software*, *83*(9), 1622–1634. http://dx.doi.org/10.1016/j.jss.2009.08.023.

Information Technology Laboratory (2020). *National vulnerability database*. https://nvd.nist.gov/vuln-metrics/cvss.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. http://dx.doi.org/10.1007/s11042-018-6894-4.

Ji, C., Lu, X., & Zhang, W. (2020). A biobjective optimization model for expert opinions aggregation and its application in group decision making. *IEEE Systems Journal*, *15*(2), 2834–2844.

Khodabakhsh, A., Yayilgan, S. Y., Abomhara, M., Istad, M., & Hurzuk, N. (2020). Cyber-risk identification for a digital substation. In *Proceedings of the 15th international conference on availability, reliability and security* (pp. 1–7). http://dx.doi.org/10.1145/3407023.3409227.

Kopp, E., Kaffenberger, L., & Jenkinson, N. (2017). Cyber risk, market failures, and financial stability: International Monetary Fund. http://dx.doi.org/10.5089/9781484313787.001.

Lämmel, R., Mosen, D., & Varanovich, A. (2013). Method and tool support for classifying software languages with wikipedia. In *International conference on software language engineering* (pp. 249–259). Springer.

Miz, V., Benzi, K., Ricaud, B., & Vandergheynst, P. (2017). Wikipedia graph mining: dynamic structure of collective memory. arXiv preprint arXiv:1710.00398.

Müller, T., Cotterell, R., Fraser, A., & Schütze, H. (2015). Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. http://dx.doi.org/10.18653/v1/D15-1272.

Noraset, T., Lowphansirikul, L., & Tuarob, S. (2021). Wabiqa: A wikipedia-based thai question-answering system. *Information Processing & Management*, *58*(1), Article 102431.

Padmaja, C., Narayana, S., & Divakar, C. (2018). Probabilistic topic modeling and its variants–a survey. *International Journal of Advanced Research in Computer Science*, *9*(3), http://dx.doi.org/10.26483/ijarcs.v9i3.6107.

Peterson, L., & Davie, B. (2007). Computer networks: a systems approach: Elsevier.

Pilkauskas, P. (2010). Expertise classification of recommenders in the wikipedia recommender system. IMM-M.Sc.-2010-77.

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*, http://dx.doi.org/10.1109/TKDE.2020.2992485.

Rawlings, P. (2015). Cyber risk: Insuring the digital age. *British Insurance Law Association Journal*, *128*.

Refsdal, A., Solhaug, B., & Stølen, K. (2015). Cyber-risk management. In *Cyber-risk management* (pp. 33–47). Springer: http://dx.doi.org/10.1007/978-3-319-23570-7.

Roy, S., Mak, M.-T., & Wan, K. (2011). Wikipedia based news video topic modeling for information extraction. In *International conference on multimedia modeling*. http://dx.doi.org/10.1007/978-3-642-17829-0_39.

Salfner, F., Lenk, M., & Malek, M. (2010). A survey of online failure prediction methods. *ACM Computing Surveys*, *42*(3), 10. http://dx.doi.org/10.1145/1670679.1670680.

Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P., & Bork, P. (2000). SMART: A web-based tool for the study of genetically mobile domains. *Nucleic Acids Research*, *28*(1), 231–234.

Schultz, C., Nitao, J., Starr, J., & Compton, J. (2020). Probabilistic model for cyber risk forecasting. Google Patents.

Sentuna, A., Alsadoon, A., Prasad, P., Saadeh, M., & Alsadoon, O. H. (2021). A novel enhanced naïve bayes posterior probability (ENBPP) using machine learning: Cyber threat analysis. *Neural Processing Letters*, *53*(1), 177–209. http://dx.doi.org/10.1007/s11063-020-10381-x.

Sheehan, B., Murphy, F., Kia, A. N., & Kiely, R. (2021). A quantitative bow-tie cyber risk classification and assessment framework. *Journal of Risk Research*, *24*(12), 1619–1638.

Sinanc, D., & Yavanoglu, U. (2013). A new approach to detecting content anomalies in wikipedia. In *2013 12th International conference on machine learning and applications. Vol. 2* (pp. 288–293).

Subroto, A., & Apriyana, A. (2019). Cyber risk prediction through social media big data analytics and statistical machine learning. *Journal of Big Data*, *6*(1), 1–19. http://dx.doi.org/10.1186/s40537-019-0216-1.

Wangen, G. (2017). Information security risk assessment: A method comparison. *Computer*, *50*(4), 52–61. http://dx.doi.org/10.1109/MC.2017.107.

Wangen, G., Hallstensen, C., & Snekkenes, E. (2018). A framework for estimating information security risk assessment method completeness. *International Journal of Information Security*, *17*(6), 681–699. http://dx.doi.org/10.1007/s10207-017-0382-0.

Wu, D., Zheng, L., & Olson, D. (2014). A decision support approach for online stock forum sentiment analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *44*(8), 1077–1087. http://dx.doi.org/10.1109/TSMC.2013.2295353.

Yun, J., Jing, L., Yu, J., Huang, H., & Zhang, Y. (2011). Document topic extraction based on wikipedia category. In *2011 Fourth international joint conference on computational sciences and optimization*. http://dx.doi.org/10.1109/CSO.2011.119.

Zhao, D., & Strotmann, A. (2021). Mapping knowledge domains on wikipedia: an author bibliographic coupling analysis of traditional Chinese medicine. *Journal of Documentation*, http://dx.doi.org/10.1108/JD-02-2021-0039.

Dr. **Arash N. Kia** is apostdoctoral researcher in AI in School of Computer Science and Statistics, Trinity College Dublin. He was a Marie Curie Research Fellow at the University of Limerick from 2019 to 2022. He is now a member of Lero Software Research Centre in Ireland, Emerging Risk Group in Kemmy Business School, and Knowledge Discovery and Data Mining Laboratory at the University of Tehran. Arash s research interests and experiences include time series prediction, machine learning, cybersecurity, and network science.

Prof. **Finbarr Murphy** is Executive Dean (Interim) of Kemmy Business School, Head of Department, Accounting and Finance, and Senior Lecturer in Quantitative Finance and Emerging Risk at theUniversity of Limerick. A computer engineering graduate, Finbarr worked for over ten years in investment banking before completing his PhD in quantitative finance. A former Fulbright Scholar and Erasmus Mundus Scholar, his research interests include quantitative finance and more recently, emerging technological risk. In particular, he is interested in applying machine learning techniques to determine risks of emerging threats such as cyber attacks. He is currently engaged in several EU H2020 projects and Irish Science Foundation Ireland (SFI) projects.

Dr. **Barry Sheehan** lectures risk management and insurance in the Kemmy Business School at the University of Limerick. He is a cybersecurity researcher with insurance industry and academic experience. With a professional background in actuarial science, his research uses machine-learning techniques to estimate the changing risk profile produced by developing loss exposures such as autonomous vehicles, nanotechnology and cyber risk. He is contributing member of Emerging Risk Group (ERG) at the University of Limerick which has long-established expertise in insurance and risk management and has a continued success within large research consortia including several EU H2020 and FP7 research projects.

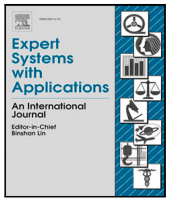
Dr. **Darren Shannon** is a Lecturer in the Accounting and Finance Department, University of Limerick, and a researcher in the Emerging Risk Group (ERG). Darren received his Ph.D. in Applied Statistics from the University of Limerick. His research centres on the insurance implications of connected and autonomous vehicles (CAVs); specifically, their cybersecurity vulnerabilities and the extent of their ability to reduce injury and collision rates.

Expert Systems With Applications 237 (2024) 121449

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



RFM model customer segmentation based on hierarchical approach using FCA

Chongkolnee Rungruang [a], Pakwan Riyapan [b], Arthit Intarasit [b], Khanchit Chuarkham [c], Jirapond Muangprathub [d,*]

[a] College of Digital Science, Prince of Songkla University, Songkla 90110, Thailand
[b] Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani 94000, Thailand
[c] Faculty of Commerce and Management, Prince of Songkla University, Trang Campus, Trang 92000, Thailand
[d] Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani 84000, Thailand

ARTICLE INFO

Dataset link: https://archive.ics.uci.edu

Keywords:
Customer segmentation
Formal concept analysis
Hierarchical concept
RFM model
Clustering

ABSTRACT

Nowadays, every business focuses on customer relationship management (CRM) to deliver their customers better services and to establish a competitive advantage over their competitors. Significantly, customer insights with solid customer relationships improve customer retention and satisfaction, thereby contributing to profit. Thus, customer segmentation based on cluster analysis is critical to customer identification in CRM. In addition, it can identify the potential customers and their needs to be matched with marketing strategies. However, unfortunately, this approach has led to a gap between the marketing persons who care about the business implications and clustering output with the data science complexity barrier. Moreover, most clustering methodologies give only groups or segments, such that customers of each group have similar features without customer data relevance. Thus, this work sought to address these concerns by using a hierarchical approach. This research proposes a new effective clustering algorithm by combining Recency, Frequency, and Monetary (RFM) model with formal concept analysis (FCA). This new methodology uses the advantages of FCA in building the knowledge representation; therefore, the obtained construction contains both implicit and explicit knowledge. Explicit knowledge shows information represented in the hierarchical structure model, while implicit knowledge is embedded in the structure with its implication properties. Thus, the knowledge structure from FCA reveals relationships among data points in an easily understood manner. The proposed model was evaluated and compared with K-means clustering and hierarchical clustering using the online retail II dataset from the UCI Machine Learning Repository. The proposed method provides enough and appropriate information for marketers to perceive the value of the clustering results for creating practical marketing strategies in real-world business by offering the marketers both customer segmentation and the relationships in customer data at the same time.

1. Introduction

Segmentation, Targeting, and Positioning (STP) marketing is a core marketing approach for creating superior customer value and supporting the development of products and services (Gupta, Justy, Kamboj, Kumar, & Kristoffersen, 2021; Munusamy & Murugesan, 2020). The first step mainly addresses determining important characteristics to differentiate each market segment, leading to market targeting and product positioning. In competitive markets, businesses need to understand their customers to generate suitable matched marketing strategies. However, it is difficult to understand a massive number of customers clearly (Chen, Zhang, Chu, & Yan, 2019; Deng & Gao, 2020). Because of this, before companies can apply marketing strategies to their customers, they often use customer segmentation (also called market segmentation) to categorize the customers. Customer segmentation divides customers into groups according to their similarities in needs, characteristics, or behaviors, to maintain customer relationships and increase profit. Moreover, customer segmentation is a tool of customer identification that is the primary process of customer

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: chongkolnee.r@psu.ac.th (C. Rungruang), pakwan.r@psu.ac.th (P. Riyapan), arthit.i@psu.ac.th (A. Intarasit), ckhanchit@hotmail.com (K. Chuarkham), jirapond.m@psu.ac.th (J. Muangprathub).

https://doi.org/10.1016/j.eswa.2023.121449
Received 2 April 2023; Received in revised form 2 September 2023; Accepted 2 September 2023
Available online 15 September 2023

relationship management (CRM) (Deng & Gao, 2020; Dolnicar, Grün, & Leisch, 2018a). Notably, customer segmentation supports businesses in enhancing customer retention and loyalty and helps identify the value of a customer (Ballestar, Grau-Carles, & Sainz, 2018; Nandapala & Jayasena, 2020; Wu & Liu, 2020).

There are several customer segmentation methods, but most of them are based on customers' behavioral, psychographic, geographic, and demographic information. However, the customers' behavioral information based on RFM analysis is emphasized because of using a small set of features to segment customers surveyed in Alves Gomes and Meisen (2023). These factors will be used to segment customers with many algorithms. Segmentation algorithms can be divided into four main groups. They include association algorithms (e.g., Apriori, FP growth, ECLAT partition), clustering (e.g., hierarchical clustering, K-means clustering, fuzzy C Means clustering, density-based clustering, affinity propagation clustering), classification algorithms (e.g., KNN, Naïve Bayes, SVM, C4.5, Decision Tree) and regression algorithms (e.g., Logistic regression) (Tsiptsis & Chorianopoulos, 2011). Clustering is an unsupervised learning approach and can be divided into two categories: hard clustering and soft clustering (Singh & Srivastava, 2020). In hard clustering, each object is assigned to only one cluster, while soft clustering can also be done in an overlapping manner in which an object can be part of more than one cluster. After we have generated a partition using a clustering algorithm, we need to evaluate the validity or quality of this partition. If several partitions are generated (e.g., with different numbers of clusters), we need ways to compare them before we conclude the clustering result. There are three broad categories for cluster validity: internal, external, and relative indices (Tsiptsis & Chorianopoulos, 2011). The main approaches to assess these include graphical representations and internal indices. The silhouette plot is an advanced graphical representation that provides visual information about the quality of the partition. Internal indices measure a partition's "intrinsic" quality (how well-separated the clusters are). We have seen that the mean silhouette value can be used as an internal index. There exist many other internal indices. The Calinski–Harabasz, the Davies–Bouldin, and the Dunn indices are the three that are most widely used (Gagolewski, Bartoszuk, & Cena, 2021). Fuzzy Clustering is an example of soft clustering, and cluster validity indices have also been defined for this approach, one such index being the fuzzy silhouette index.

The K-means clustering algorithm is the most popular and is commonly used (Fränti & Sieranoja, 2018; Khalili-Damghani, Abdi, & Abolmakarem, 2018). However, it is sensitive to cluster centroid initialization and outliers, both impacting the eventual results (Deng & Gao, 2020; Meng et al., 2020). Unlike K-means, hierarchical clustering can produce a robust result without assigning initial values. Furthermore, the hierarchical clustering method is the most intuitive way of grouping data (Dolnicar, Grün, & Leisch, 2018b). It allows a human to track the task of dividing a set of n observations (customers) into k clusters (segments) represented in a hierarchy graphically shown as a dendrogram. In contrast, the result of K-means is unstructured non-overlapping clusters; therefore, the hierarchical clustering result is more interpretable and informative for marketers.

Creating effective marketing strategies in real-world business requires business people to have both implicit and explicit knowledge. However, hierarchical clustering only provides customer groups or segments in a hierarchy, lacking some knowledge such as the relevance of customer data. To solve this problem, this study proposes a new clustering algorithm that uses the RFM model and FCA to build knowledge representation and segmentation. The resulting construction contains both implicit and explicit knowledge. Explicit knowledge is information represented in the hierarchical structure model, while implicit knowledge is embedded in the structure with its implication properties. The FCA knowledge structure reveals relationships among data points in an easily understood manner. Additionally, a customer data representation based on FCA can uncover latent issues in customer information. The main contribution of this study is the use of FCA to build a knowledge structure for customer segmentation. FCA is ideal for identifying groups of customers with specific common properties or features. The presented structure has advantages in discovering both explicit and implicit knowledge. The marketers need these types of knowledge to create practical marketing strategies in real-world business. Namely, the explicit knowledge shows a group of customers with the same behavior, while the implicit knowledge shows the behavior associated with using the service in their business. To address this problem, this current study proposes a new effective clustering algorithm using the advantages of the RFM model and FCA to build knowledge representation and segmentation. The obtained construction contains both implicit and explicit knowledge. Explicit knowledge shows information represented in the hierarchical structure model, while implicit knowledge is embedded in the structure with its implication properties. Thus, the knowledge structure from FCA reveals relationships among data points in an easily understood manner. Above all, it is emphasized that a customer data representation based on FCA can be used to discover latent issues in customer information. Practically, the presented knowledge structure provides a decision-maker's approach to exact both customer segmentation and discover their behavior relevance.

This new method's clustering was compared with the most popular method, K-means clustering, and hierarchical clustering, using the online retail II dataset from the UCI Machine Learning Repository. After the segmentation, business implications and interpretation are discussed. Finally, conclusions are shared.

2. Background

2.1. Customer segmentation

Customer segmentation is the process of dividing the customer group into subgroups according to similarities (Choi, Choi, Yoon, & Joung, 2020; Deng & Gao, 2020; Singh & Mittal, 2021; Zeybek, 2018). A simple customer segmentation approach includes geographic, demographic, psychographic, and behavioral segmentation. Behavioral segmentation, being the most commonly used method, includes the following main steps: (1) business understanding and design of the segmentation process; (2) data understanding, preparation, and enrichment; (3) identification of the segments with cluster modeling; (4) evaluation and profiling of the revealed segments; and (5) deployment of the segmentation solution, design, and delivery of the differentiated strategies (Chorianopoulos, 2016; Tsiptsis & Chorianopoulos, 2011).

Segmentation can use several alternative algorithms, including an association algorithm, clustering algorithm, classification algorithm, or regression algorithm. Among these, clustering is the most precise and effective method for customer segmentation (Tsiptsis & Chorianopoulos, 2011).

Customer segmentation with a clustering algorithm uses machine learning to classify customer data based on similarity. All clustering algorithms aim to minimize the distances within clusters and maximize the distances between them. Moreover, the different clustering algorithms and outcomes are related to the objective (cost) function to measure the quality of clustering, the underlying structure assumed, the similarity measure, and how to consider the number of clusters.

Above all, most clustering techniques give only groups or segments without customer data relevance, whereas the marketing persons need an effective customer segmentation that gives appropriate knowledge in an easily understood format.

2.2. RFM model

RFM model is one of the most prevalent behavioral segmentations (Alves Gomes & Meisen, 2023; Hosseini, Abdolvand, & Harandi, 2022; Khalili-Damghani et al., 2018; Peker, Kocyigit, & Eren, 2017; Wang, Tsai, & Ciou, 2020). This model groups existing customers with their recency, frequency, and monetary values and does not focus on attracting new customers but identifies the best customers to perform targeted marketing. Recency is the number of days since the last purchase. Frequency refers to the total number of purchases. Meanwhile, monetary is the total purchase value during a specific period (Dedi, Dzulhaq, Sari, Ramdhan, Tullah, & Sutarman, 2019; Sokol & Holỳ, 2021).

The RFM scoring process is carried out to demonstrate RFM analysis using the quintile method. The customers are split into quintiles (five equal groups), and each customer is given a score based on which quintile he belongs to Christy, Umamakeswari, Priyatharsini, and Neyaa (2021). Accordingly, the scores assigned in the first quintile with the highest values for frequency (F) is 5, while the other quintiles are scored with 4, 3, 2, and 1 in rank order. This process is also undertaken for monetary (M). In contrast, the last quintile with the smallest recency (R) values is coded as 5. Next, we sequentially give the other quintiles based on recency values the scores 4, 3, 2, and 1. Finally, all the customers are ranked using R, F, and M values. Accordingly, the best customer group is 5-5-5, and the worst is 1-1-1.

However, the other approaches to the RFM model use the actual values of each RFM factor as the segmentation variables. The original RFM values are determined using various clustering techniques with required data preprocessing. Especially, K-means is the most popular and most commonly used method (Ernawati, Baharin, & Kasmin, 2021). Unfortunately, it is sensitive to cluster centroid initialization and outliers, so these affect the clustering results (Christy et al., 2021). In addition, it gives only complete separation of customer groups without customer data relationships. To ensure the accuracy of the K-means clustering results, the RFM values undergo preprocessing to identify and eliminate outliers. These outliers are then treated separately from the rest of the data (Chen, Sain, & Guo, 2012). Therefore, we utilize the RFM scoring technique with the quintile method to avoid the need for valid outlier removal. This study sought to create a practical approach to address these problems by using the RFM model, whose advantage is using a very small number of variables (only three variables), with clustering based on FCA. However, we can incorporate new variables into the RFM model to improve the accuracy and gain more information. An example is the RFMT model introduced by Zhou, Wei, and Xu (2021), which includes interpurchase time (T) to enhance customer segmentation. Other models, such as LRFM (Chang & Tsay, 2004), RFMTC (Yeh, Yang, & Ting, 2009), and RFMD (Noori, 2015), also include additional variables.

2.3. Formal concept analysis

Formal concept analysis (FCA) is a mathematical theory of concept formation based on lattice theory, applied in classification and concept discovery to organize information and discover relationships (Ganter & Wille, 2012; Wille, 2009). Moreover, FCA follows a human-centered approach and supports exploration operations through the concept lattice to organize information and discover relationships embedded in the binary relations between a pair of sets (called objects and attributes, respectively). A node in the concept lattice is an objects/attributes pair, called a (formal) concept. A concept consists of the extent (all objects belonging to the concept) and the intent (attributes describing the concept).

The core knowledge of the FCA approach is based on a simple data representation: a binary table called a formal context that is transformed into a mathematical structure called a concept lattice (Castellanos, Cigarrán, & García-Serrano, 2017). $\mathbb{K} := (G, M, I)$ is a formal context in which G represents a set of objects, M represents a set of attributes and $I \subseteq G \times M$ represents a set of is-a or has-a relationships between G and M, defined by gIm, which is read as the object g has the attribute m. I is the incidence relation of the context (G, M, I).

From this formal context, a set of formal concepts can be generated. To define a formal concept, the following derivation operations are needed. For any subsets A and B, $A \subseteq G$ and $B \subseteq$ M:

$$A \mapsto A' := \{m \in M \mid gIm \quad \forall g \in A\}$$
$$B \mapsto B' := \{g \in G \mid gIm \quad \forall m \in B\}$$

Thus, a formal concept is a pair (A, B) where $A \subseteq G$ is a set of objects (the extent of the formal concept), and $B \subseteq M$ is a set of attributes (the intent of the formal concept), which has the following properties:

- If all objects a in A are tagged with an attribute b, then b must be included in B (i.e., $B = A'$ the intent of the formal concept includes all the attributes shared by the objects in the extent).
- Conversely, if an object a is tagged with all the attributes in B, then a must be included in A (i.e., $A = B'$: the extent of the formal concept includes all those objects filtered out by the intent).

Formal concepts can be ordered according to their extents by applying the partial order relationships (Benavent, Castellanos, de Ves, García-Serrano, & Cigarrán, 2019) in Eq. (1), where a formal concept (A, B) with extent A is considered a sub-concept of another formal concept (C, D) with an extent C when the objects in A are contained into the objects in C.

$$(A, B) \leq (C, D) \Leftrightarrow (A \subseteq C \Leftrightarrow D \subseteq B). \tag{1}$$

The organization that results from this order relationship can be proven to be a lattice, a concept lattice, associated with the formal context denoted by $\mathfrak{B}(G, M, I)$. Since concept lattices are ordered sets, they can be naturally displayed in Hasse diagrams.

Frequency (support) is one of the most popular measures in the theory of pattern mining (Kuznetsov & Makhalova, 2018). Frequency arises from the assumption that the most "interesting" concepts are frequent ones:

$$\text{supp}(A, B) = \frac{|A|}{|G|} \tag{2}$$

The support provides an efficient level-wise algorithm of semilattice computing:

$$B_1 \subset B_2 \rightarrow \text{supp}\left(B_1\right) \geq \text{supp}\left(B_2\right). \tag{3}$$

In this study, we say that a set of attributes is frequent if its support exceeds a certain threshold. Thus, the frequency of an attribute set means that it is frequent.

Practically, clustering based on the FCA method contains four steps (Zhang, Zhao, & Yan, 2018). First, the dataset is pretreated, and feature items are extracted from data elements. Then, a formal context is constructed by taking data elements as objects and feature items as attributes. Next, the concept of formal contexts is extracted, and the concept lattice of the formal contexts is constructed. Finally, a Hasse diagram as a highly informative visualization can be generated.

It is emphasized that the FCA provides a well-defined mathematical framework to discover implicit and explicit knowledge in an easily understood format by using formal context and a Hasse diagram that clearly represents the concepts' generalization/specialization relationships (Ganter & Wille, 2012). In addition, FCA can construct an informative concept hierarchy providing valuable information on various specific domains. Therefore, we propose performing RFM model customer segmentation based on FCA.

Table 1
Study of RFM-based customer segmentation.

Resource	Dataset	Variables	Clustering techniques
Anitha and Patil (2022)	Online retail (Chen et al., 2012) Real-time retail	RFM	K-means
Hosseini et al. (2022)	Internet Banking	RFM	K-means and DBSCAN
Christy et al. (2021)	Online retail (Chen et al., 2012)	RFM	K-means, RM K-means, and Fuzzy C-means
Frasquet, Ieva, and Ziliani (2021)	Offline and Online retail	RFM + Sex + Age	Latent Class Analysis (LCA)
Rahim, Mushafiq, Khan, and Arain (2021)	Online retail (Chen et al., 2012)	RFM	K-means
Zhou et al. (2021)	Online retail	RFMT	Hierarchical
Shokouhyar, Shokoohyar, and Safari (2020)	Retail	RFM	K-means
Monalisa, Nadya, and Novita (2019)	Distribution and retail	RFM	Fuzzy C-means
Nakano and Kondo (2018)	Multichannel Retail	RFM	Latent Class Analysis (LCA)

2.4. Association rule mining

An association rule is just a statement about the conditional sample probability (called confidence) of an event w.r.t. another one, together with the statement of the joint sample probability of the two events (called support), where both events are described in terms of attribute sets. In FCA terms, for two subsets of attributes (called itemsets in data mining) Y_1 and $Y_2 \subseteq M$, the association rule $Y_1 \rightarrow Y_2$ has support $\frac{\left|(Y_1 \cup Y_2)'\right|}{|G|}$ and confidence $\frac{|(Y_1 \cup Y_2)'|}{\left|Y_1'\right|}$. The rule $Y_1 \rightarrow Y_2$ is called frequent if $\text{supp}\left(Y_1 \rightarrow Y_2\right) \geq \text{minsupp}$ for some threshold minsupp. In FCA, association rules are known as partial implications (Kuznetsov & Poelmans, 2013). This work applies the obtained FCA knowledge structure to extract implication rules to discover knowledge relationships in customer segmentation.

3. Related work

RFM model-based customer segmentation has been demonstrated using various clustering techniques, briefly summarized in Table 1. Most of these studies have applied K-means clustering to explicit customer segmentation surveyed in Hiziroglu (2013) and Alves Gomes and Meisen (2023). K-means is popular in the application areas because it is simple to understand, interpret, and apply. For example, Chen et al. applied K-means clustering and decision tree induction to segment customers from the online retail dataset of customer transactions in the UCI repository (Chen et al., 2012). In the same way, many studies have deployed dataset segmentation by using the K-means algorithm (Anitha & Patil, 2022, 2022; Chen et al., 2012; Christy et al., 2021). These studies have shown that data preprocessing is a crucial and time-consuming process before segmentation, especially before K-means clustering. For this type of clustering, data preprocessing needs to remove outliers, scale the ranges of data, and solve any long tails problems by data transformation (Tavakoli et al., 2018). In addition, we found that the number of segments chosen was between 2 and 10 groups. The number of segments applied should not be too large because it will make it difficult for the marketing analyst to interpret and design marketing strategies for selected customer segments. The Silhouette index is the most widely used validation index. The online retail dataset of customer transactions introduced by Chen et al. (2012) and placed in the UCI repository is the most often used. Therefore, this work uses this dataset for benchmarking and comparing the proposed model with K-means clustering and hierarchical clustering.

Hierarchical clustering methods are the most intuitive way of grouping data because they mimic how a human would approach the task of dividing a set of observations (customers) into k groups (segments). Moreover, this is one of the most frequently used methods and can produce a robust result. Unlike other unsupervised methods, such as K-means, for hierarchical clustering, it is not necessary to assign any initial values (Zhou et al., 2021). Zhou et al. (2021) used hierarchical clustering for customer segmentation by web content mining. They used Calinski–Harabasz index and Davies–Doulbin index to determine the optimal cluster number. In addition, Chen, Zhang, and Zhao (2017) developed an approach to analyze customer behavior in mobile app usage and identify associations between functions by combining three data mining techniques: RFM analysis, link analysis based on graph theory, and association rule learning. Their proposed approach can be used to visualize and gain the advantages of these associations. Moreover, it provides insights into customer behaviors and function usage preferences. Remarkably, the prior work based on hierarchical clustering shows that segments and relationships are important and valuable knowledge for analyzing customer behaviors. Thus, this work proposes customer segmentation based on a hierarchical structure using FCA.

Poelmans, Ignatov, Kuznetsov, and Dedene (2013) utilized FCA's visualization abilities to discover the main research topics of papers on FCA that were published between 2003 and 2011. Their findings showed that FCA-based techniques were used by researchers for knowledge discovery and ontology engineering in various application domains, including software mining, web analytics, medicine, biology, and chemistry data. Moreover, it is worth noting that FCA possesses visualization capabilities. However, research on applications to CRM in customer segmentation is lacking, while some studies, such as the one by Marchetto (2005), have focused on software mining to extract concerns from web applications using MDSOC Hyperspaces definition and FCA. Another example is the work by Ravi, Ravi, and Prasad (2017), which proposed a hybrid model comprising fuzzy formal concept analysis and concept-level sentiment analysis (FFCA+SA). That research aimed to conduct opinion mining for CRM within the financial services sector. However, customer segmentation was not the main focus. Hence, there is still much to be explored in this area.

In Lei, Yan, Han, and Jiang (2018), association rules were extracted from concept lattices generated by FCA. This study proposed a feasible and effective method based on concept lattice and attribute analysis to reduce the number of association rules. Furthermore, they controlled the number of concepts by using rough attribute values and improved the concept's quality. Above all, this study emphasized that association rules from concept lattices can provide latent and valuable knowledge.

However, in practice, a concept lattice has many formal concepts that lead to the high complexity of conceptual clustering. Nguyen, Tran, Quan, Nguyen, and Le (2019) introduced a framework named MarCHGen (Malware Concept Hierarchy Generation) to generate a malware concept hierarchy. In this framework, they established frequent concepts on a concept lattice. A frequent concept has object sets larger than a defined threshold. Furthermore, only the frequent concepts are kept on the lattice. Therefore, a frequent lattice will be pruned to be less complex than the original lattice.

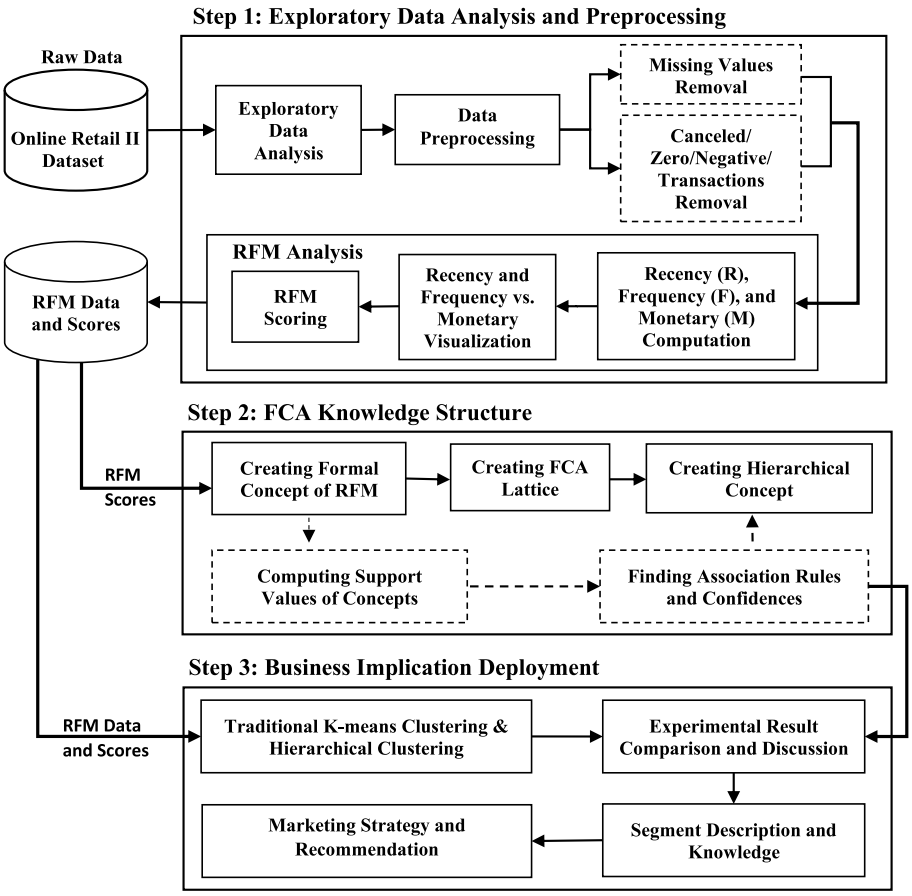


Fig. 1. Tasks and techniques of the proposed approach.

Above all, hierarchical clustering gives only groups or segments of customers in a hierarchy lacking some knowledge in the form of implications such as the relevance of customer data. To counter this problem, this research proposes a new effective clustering algorithm using the advantages of the RFM model and FCA to build knowledge representation and segmentation. Moreover, this appears to be the first time that the RFM model has been combined with the FCA.

4. Research methodology

The main contribution of this study is in applying FCA to build a knowledge structure for customer segmentation. FCA is ideal for identifying groups of customers with certain common properties (or features). The advantages of the presented structure are the discovery of explicit and implicit knowledge. The explicit knowledge is derived from identifying the co-appearance of attributes, while the implicit knowledge stems from the implication rules in relationships inside the concept lattice structure generated by the FCA. This section describes the process of building a hierarchical structure using FCA and its knowledge acquisition. Moreover, the traditional clustering approaches, K-means, and agglomerative hierarchical clustering are compared with the proposed approach.

The proposed methodology can be divided into three main steps, as shown in Fig. 1. The first step involves data exploration and preparation, which is critical for eventual accurate insights. Afterwards, the FCA knowledge structure provides explicit and implicit knowledge in customer segmentation and the implication relationships, respectively. The final step focuses on deploying the acquired knowledge in the customer segmentation and the discovered knowledge in each segment. The corresponding details are explained in the subsections that follow.

4.1. Exploratory data analysis and preprocessing

Exploratory data analysis (EDA) is the primary data exploration to extract and understand the data patterns by using statistics and graphical representations. In this research, we use the online retail II dataset used in Chen et al. (2012), Chen, Guo, and Li (2019), Christy et al. (2021), and Rahim et al. (2021) from the UCI Machine Learning Repository. This dataset contains all the transactions over two years, occurring for a UK-based and registered, non-store online retail, from 1 December 2009 to 9 December 2011. The company mainly sells unique all-occasion giftware. Many customers of the company are wholesalers. The customer transaction dataset contains 1,067,371 records with eight variables, as shown in Table 2.

Practically, we identify the unique customers, details of data, inconsistent data, incomplete data (or missing values), and noisy data. We prepare the dataset by cleaning steps, such as removing missing values and canceled/zero/negative transactions. Afterwards, the obtained dataset must be preprocessed for the required RFM model-based clustering analysis. The main steps and relevant tasks involved in the data preparation are as follows:

After this data preparation, there were 805,549 transactions remaining in the target dataset. Next, the RFM analysis concept, which is a popular and significant customer segmentation approach in the retail industry, will be used to reduce the data before generating a knowledge structure. The recency of transactions, frequency, and amount the customer spent are determined to create RFM values. Moreover, recency, frequency, and monetary values are visualized to gain knowledge and understanding of these data. Next, the day of the last purchase was decided as the reference date to calculate recency. At last, each customer's RFM score was created.

Table 2
Variables in the customer transaction dataset (1,067,371 instances).

Variable name	Data type	Description; typical values and meanings
Invoice	Nominal	Invoice number; A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
StockCode	Nominal	Product (item) code; A 5-digit integral number uniquely assigned to each distinct product.
Description	Nominal	Product (item) name;
Quantity	Numeric	The quantities of each product (item) per transaction.
InvoiceDate	Numeric	Invoice date and time; The day and time when a transaction was generated.
Price	Numeric	Unit price; Product price per unit in sterling (Â£).
Customer ID	Nominal	Customer number; A 5-digit integral number uniquely assigned to each customer.
Country	Nominal	Country name; The name of the country where a customer resides.

Table 3
Variables in the RFM scores and data (5878 instances).

Variable name	Data type	Description
CustomerID	Nominal	Corresponding to each distinct customer id
Recency	Numeric	Recency in days
Frequency	Numeric	Frequency of purchase per customer
Monetary	Numeric	Monetary or total amount spent per customer
R_score	Numeric	Recency scores [1, 5]
F_score	Numeric	Frequency scores [1, 5]
M_score	Numeric	Monetary scores [1, 5]

To calculate the RFM score, we can see that all the quantities calculated here (recency, frequency, and monetary) have different ranges. All the customers are ranked by considering their recency, frequency, and monetary values, and R-F-M codes represent them. Thus, we first convert these quantities to scores based on the quintiles in the target dataset. This step of the RFM scoring was computed as follows.

1. Create an aggregated variable named $Total$ by multiplying $Quantity$ with $Price$, which gives the total amount of money spent per product/item in each transaction.
2. Create three essential aggregated variables $Recency$, $Frequency$, and $Monetary$ to calculate the values of these variables for every customer:
   - Recency for variable $Recency$: the difference between the analysis date and the most recent date that the customer has shopped in the store. The analysis date has been taken as the maximum date available for the variable $InvoiceDate$.
   - Frequency for variable $Frequency$: The number of transactions performed by each customer.
   - Monetary for variable $Monetary$: Total money spent by every customer in the store.
3. Calculate RFM scores. All the values for frequency and monetary in the first quintile are given 1 for F_score and M_score, given 2 for the second quintile, and so on. For recency, a more recent customer will have less recency value than a customer who has not shopped in a while. Therefore, the recency values in the first quintile are given 5 for R_score, 4 for the second quintile values, and so on.

The result from data preparation with RFM analysis consists of 5878 instances summarized in Table 3. This prepared dataset is provided to build the knowledge structure in the next step.

4.2. FCA knowledge structure

This step uses RFM scores as the data to generate a formal context. We first transform the RFM scores (with three attributes and five levels in each) into a formal context containing binary data using 15 attributes to analyze and discover the latent concepts addressed in the customers and their relationships. An example of RFM score formal context is shown in Table 4. The first customer in the first row of this table has RFM scores of 2-2-5. Afterwards, we create the formal concepts of RFM scores and then create the FCA lattice following Eq. (1) of the theory in Section 2.3.

The hierarchical concept of this lattice will be built to provide a knowledge base of the proposed system. At the same time, the support values of the concepts will be computed to determine the partial implication association rules and their confidences. To construct the less complex hierarchical concept lattice with a high degree of visualization, we determine only the concepts whose number of objects is larger than a certain threshold.

From the hierarchical concept lattice, the customer segmentation is derived from the sublattice that considers the top view of this lattice. The relationship of customer behavior in each segmentation is derived from the implication rules following Eqs. (2)–(3) of the theory in Section 2.3. Those data will be prepared as described in the following subsection.

4.3. Business implication deployment

This subsection begins with experimenting with traditional types of clustering, namely K-means and agglomerative hierarchical clustering, by using RFM data and scores dataset.

As is well-known, the K-means clustering algorithm is very sensitive to outliers (data anomalies) or variables of incomparable scales or magnitudes. To ensure optimal parameters for the K-means model, we followed the guidelines in Tavakoli et al. (2018) by removing outliers using Interquartile Ranges (IQR). Following the removal of outliers, only 5633 customers remained after this screening. We then addressed the Long Tail problem by applying Log Transformation to recency, frequency, and monetary. After that, we used Max-Min scaling to scale the ranges of frequency and monetary. Finally, we determined the optimal number of segments (best K for K-means). To find the best parameters for the K-means model, we proceeded with the following steps:

1. Removing the outliers: Some customers show unusual behavior in almost all businesses. Therefore, we had to remove these data for high-quality data analysis and apply machine learning methods. The critical point is that we used IQR to remove the outliers. To calculate the IQR, the dataset is divided into quartiles or four rank-ordered even parts via linear interpolation. These thresholds separating these quartiles are denoted by Q1 (also called the lower quartile), Q2 (the median), and Q3 (also called the upper quartile). Thus, we used the formula (4) below to find the thresholds serving as upper and lower bounds of recency, frequency, and monetary features.

$$\begin{aligned} \mathrm{IQR} &= \mathrm{Q3} - \mathrm{Q1} \\ \text{lower bound} &= \mathrm{Q1} - 1.5 \times \mathrm{IQR} \\ \text{upper bound} &= \mathrm{Q3} + 1.5 \times \mathrm{IQR} \end{aligned} \tag{4}$$

   Accordingly, we removed the values more extreme than the lower or the upper bound in each of recency, frequency, and monetary. Finally, there were only 5633 customers left by this screening.

Table 4
Example of RFM score formal context.

R1	R2	R3	R4	R5	F1	F2	F3	F4	F5	M1	M2	M3	M4	M5
0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
0	0	1	0	0	0	0	1	0	0	0	0	0	1	0
0	0	0	0	1	0	0	0	1	0	0	0	0	0	1
0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
0	1	0	0	0	0	0	1	0	0	0	0	1	0	0

2. Solving the Long Tail problem in recency, frequency, and monetary: The histograms of RFM values can be used to support our clustering step, but it leads to problems. Thus, we applied the Log transformation to change the distribution of our data from a long-tailed shape in histograms of RFM values to a more Normal distribution.
3. Scaling the range of frequency and monetary to have appropriate data for clustering methods: Since the range of values in our frequency was utterly different from monetary values, we had to apply normalization to both frequency and monetary data. We used Max-Min Scaling, which helped us scale our data to an appropriate range for clustering.
4. Finding the best Number of Segments (best K for K-means)

The above K-means clustering process results will be compared with our work. Unlike K-means, hierarchical clustering can produce a result without assigning initial values. In this work, we also used agglomerative hierarchical clustering to compare with our approach. Usually, the hierarchical clustering method is the most intuitive way of grouping data (Dolnicar et al., 2018b). The hierarchical clustering applied Ward's method for similarity measures. However, the number of targeted clusters is a crucial parameter that should be determined before applying both the K-means and the hierarchical clustering algorithm. The Silhouette index and Davies–Boulbin index were chosen to assess the performances of clustering models. A higher Silhouette index or a lower Davies–Bouldin index is preferred for an optimal cluster number.

After the traditional clustering experimental results, the customer segmentation will be discussed by creating a descriptive profile for each customer segment with different characteristics based on the results of the proposed hierarchical concept and knowledge reported in the previous section. Finally, to design and deliver differentiated marketing strategies, the segmentation solution with customer relationship management and marketing implications are recommended for each customer group.

5. Results and discussion

This section presents the first to final steps and examines the three main steps mentioned above. These steps include (1) analyzing and preparing the data, (2) creating a knowledge structure using FCA, and (3) deploying and its business implications. Finally, we conclude with an overall discussion.

5.1. Findings from exploratory data analysis and preprocessing

After preparing the dataset, RFM values were visualized in the first step, shown in Fig. 2. This figure highlights the presence of outliers in the RFM values and indicates that the RFM values are not normally distributed. According to the data at hand, it seems that most of our clients have completed their latest transaction in the last 100 days. Furthermore, the majority of these individuals have made fewer than 100 orders overall, with most of them having spent less than £1,000.

In addition, the three variables are not on comparable scales, and the value ranges are pretty different: Recency [0, 738], Frequency [1, 12890], and Monetary [0.0, 608821.7]. For this reason, we used the Spearman correlation coefficients to evaluate the relationships between the variables in the RFM model, as this is appropriate for non-normally distributed data and is robust against outliers (Rebekić, Lončarić, Petrović, & Marić, 2015; Schober, Boer, & Schwarte, 2018). Fig. 3 shows the correlation matrices, while Fig. 4 displays the distribution of RFM values. These figures underscore the strong correlation between the variables Frequency and Monetary.

Moreover, Fig. 5 illustrates the distribution among the different RFM scores. After analyzing the data, customers with a recency and frequency value of 5 seem to have the highest monetary value (designated as Monetary value 5). This implies that individuals who have made frequent and recent purchases are more likely to be valuable customers. On the other hand, customers who have made purchases infrequently and a long time ago tend to spend the least money, with a monetary value of 1.

Finally, the results of RFM data and scores consisting of 5878 instances are provided for the next step.

Table 5
The first 11 formal concepts from all 208 concepts.

Concept#	Attributes	Support	Number of customers
1	{}	1.000000	5878
2	{M5}	0.200068	1176
3	{M4}	0.199898	1175
4	{M3}	0.200068	1176
5	{M2}	0.199898	1175
6	{M1}	0.200068	1176
7	{F5}	0.200749	1180
8	{F5, M5}	0.143756	845
9	{F5, M4}	0.043722	257
10	{F5, M3}	0.012759	75
11	{F5, M2}	0.000510	3

5.2. FCA knowledge structure results

The second step involves the FCA knowledge structure. We used the fcaR package (Cordero, Enciso, López-Rodríguez, & Mora, 2022) to perform FCA with R. The resulting FCA contains 208 concepts. The first 11 formal concepts are shown in Table 5, and some concepts forming a sublattice are shown in Fig. 6. This table and figure show a group of customers sharing the same attributes (behaviors).

In addition, we extract association rules from the concept lattice. The general form is "⟨N⟩P = [C] => ⟨N′⟩C′", where N is the number of objects satisfying the premise, P is a precondition, C is the confidence of association rule, N′ is the number of objects meeting the premise, and C′ is the conclusion. For example, the first implication in Table 6, i.e., ⟨521⟩ R5 M5 = [81%] => ⟨422⟩ F5, means that there are 521 customers with a recency score of 5, and the monetary score of 5, and 422 customers among them have frequency score of 5. Thus, the confidence of this implication is 81%.

However, the number of formal concepts, namely 208, is excessively large. The concept lattice is partially ordered according to the extents or objects by set inclusion. Therefore, we use the top-ranked concepts

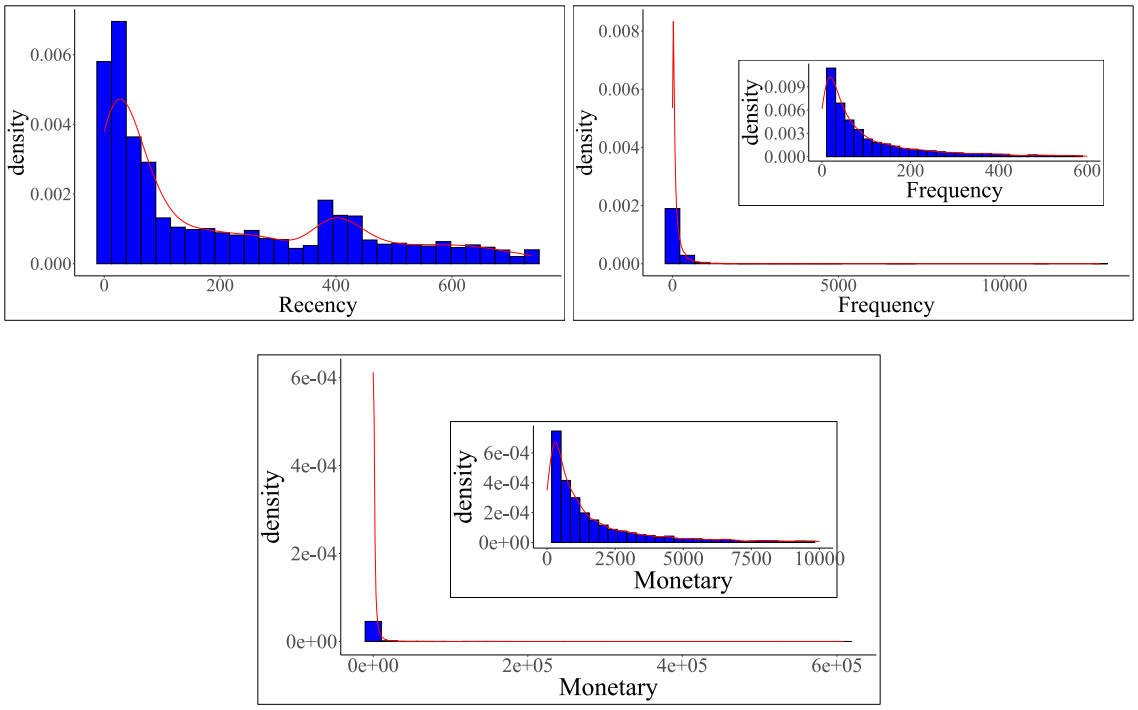


Fig. 2. Histograms of RFM values.

Table 6
List of association rules from concept lattice with confidence ≥ 60%.

No.	Rules	No.	Rules
1	<521 >R5 M5 = [81%] = ><422 >F5	10	<284 >R1 F1 = [68%] = ><192 >M1
2	<524 >R5 F5 = [81%] = ><422 >M5	11	<517 >R1 M1 = [67%]=><345 >F1
3	<450 >R1 F1 = [77%] = ><345 >M1	12	<170 >R3 M1 = [66%] = ><113 >F1
4	<329 >R4 M5 = [75%] = ><246 >F5	13	<1176 >M1 = [66%]=><775 >F1
5	<115 >R4 F1 = [73%] = ><84 >M1	14	<294 >R2 M1 = [65%] = ><192 >F1
6	<1176 >M5 = [72%]=><845 >F5	15	<131 >R4 M1 = [64%]=><84 >F1
7	<1180 >F5 = [72%]=><845 >M5	16	<64 >R5 M1 = [64%] = ><41 >F1
8	<350 >R4 F5 = [70%] = ><246 >M5	17	<197 >R3 F5 = [62%] = ><123 >M5
9	<1106 >F1 = [70%] = ><775 >M1	18	<188 >R3 F1 = [60%] = ><113 >M1

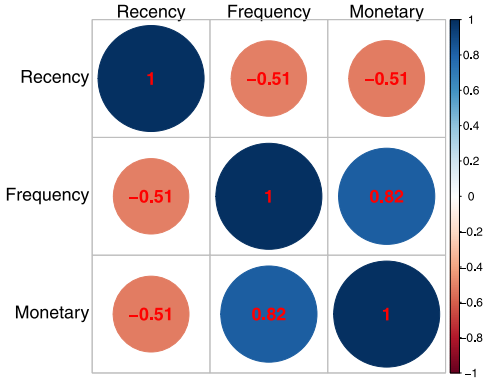


Fig. 3. Matrices of correlation with Spearman coefficients between the variables Recency, Frequency, and Monetary.

as the primary customer clusters. In this case, we include the concepts {M5}, {F5}, {M4}, {M3}, {M2}, and {M1}. Then, we considered the other concepts relating to these concepts.

The RFM scoring applies the quintile method by splitting data into five quintiles, and thus, in this case, we use only concepts whose support exceeds 0.04. We have a dataset of 5878 customers. A concept is considered if this concept consists of at least $0.04 \times 5878 = 235$ customers (20% of main clusters). Therefore, Fig. 7 depicts the customer hierarchical concept.

5.3. Business implications

For business implication deployment, the concept lattice in Fig. 6 and the hierarchical concept in Fig. 7 emphasize that the monetary score is the most significant variable in dividing customers into groups. Thus, we can simply split the customers with their monetary scores. Moreover, we also must monitor the customers having a frequency score of 5 in the highest quintile. The concept {F5} is associated with concept {F5, M5} and concept {F5, M4} belonging to the first and second quintile of monetary, respectively. These two concepts cover 19% of all customers. Accordingly, we know that the more frequently the customer purchases, the more profit the company gains. The most frequent spenders with recent purchases are customers with the highest spending amount. In contrast, customers who made their transactions long ago and less frequently contributed low monetary scores.

From Table 8, the total purchase value of concept {M5} is 77.26% of the overall purchase value. It conforms to the Pareto principle (Craft & Leake, 2002; Kim, Singh, & Winer, 2017) that "80% of sales comes from 20% of customers". Moreover, Table 8 shows that customers in concept {F5} spent 66.72% of company purchase value.

To analyze the clusters' characteristics, we refer to the information presented in Fig. 7 and Table 7.

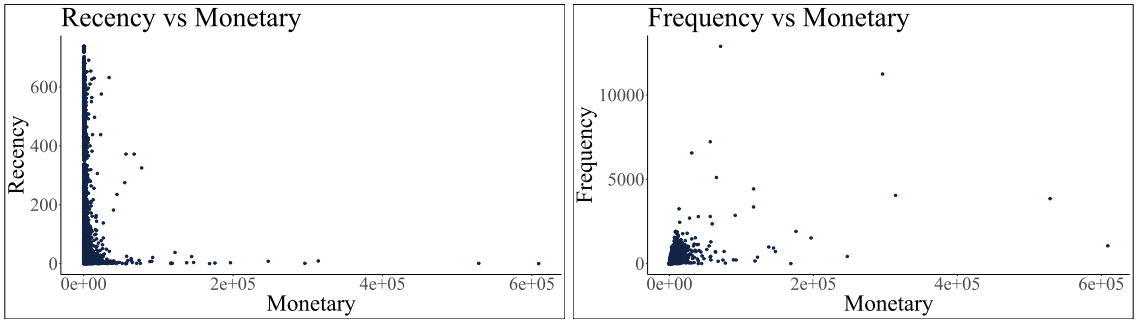


Fig. 4. Distribution of (a) Recency vs. Monetary and (b) Frequency vs. Monetary values.

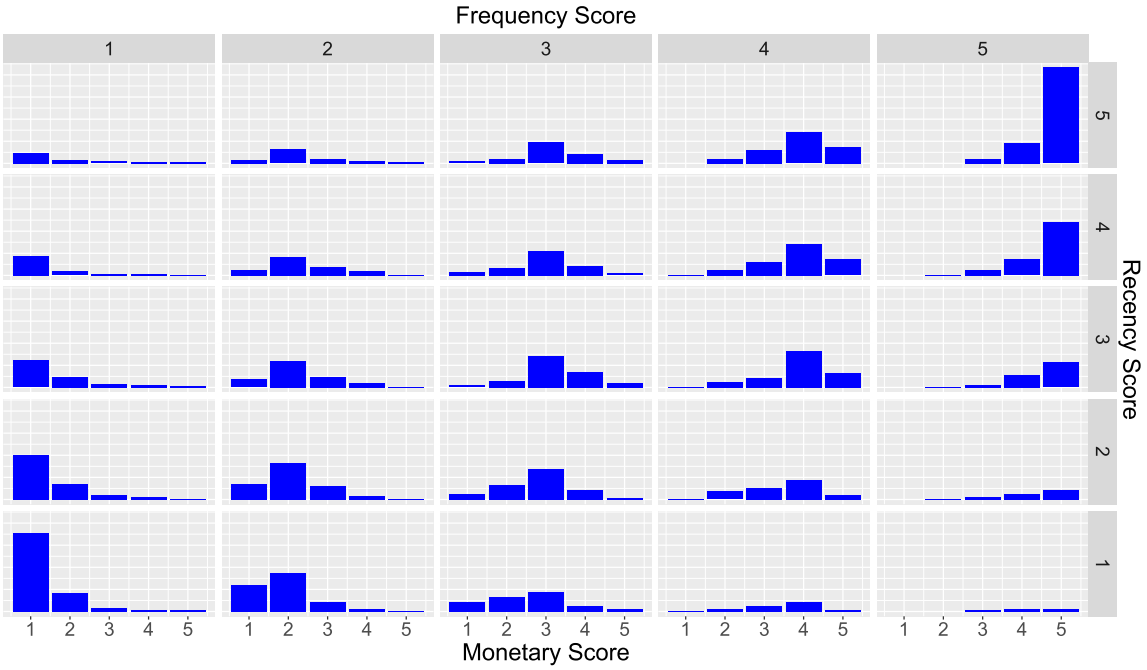


Fig. 5. RFM bar chart.

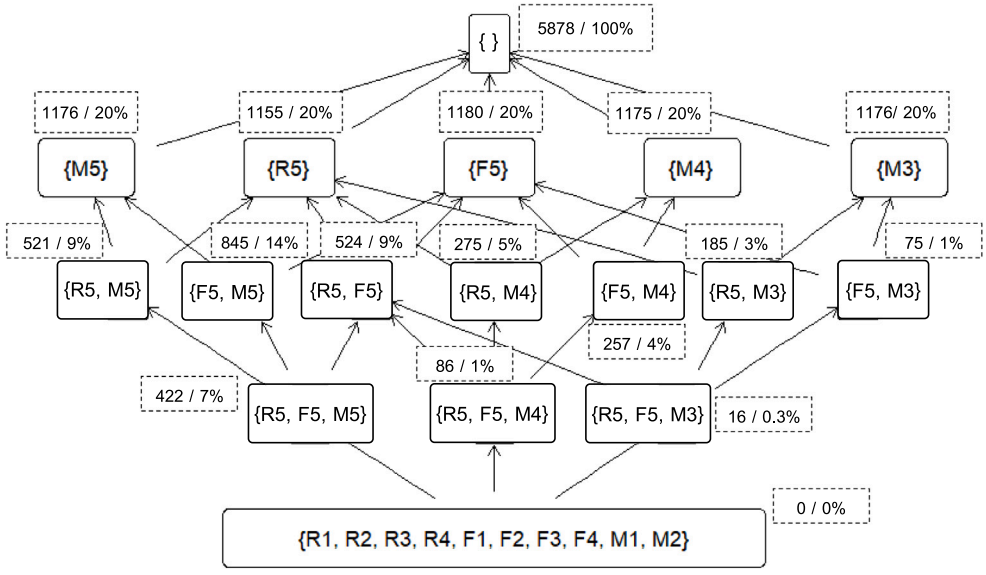


Fig. 6. Sublattice of concept 1, 2, 3, 4, 7, 8, 36, and 43 (non-specific object).

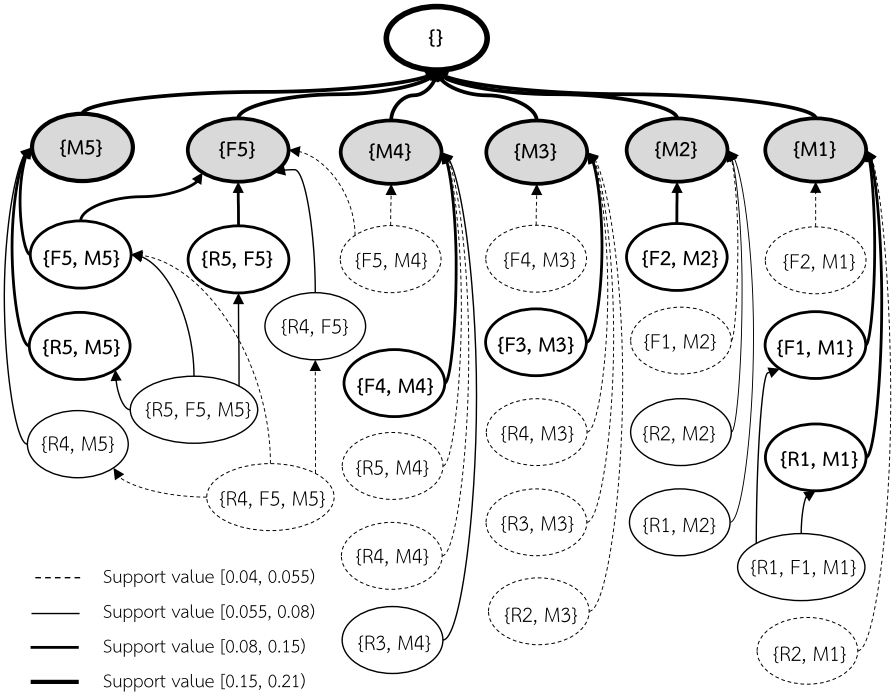


Fig. 7. RFM-based customer hierarchical concept.

Table 7
Description of the concepts related to Fig. 7.

Concepts	Support	Number of customers	Concepts	Support	Number of customers
{M5}	0.200068	1176	{R5, F5}	0.089146	524
{M4}	0.199898	1175	{R5, F5, M5}	0.071793	422
{M3}	0.200068	1176	{R4, M5}	0.055971	329
{M2}	0.199898	1175	{R4, M4}	0.049337	290
{M1}	0.200068	1176	{R4, M3}	0.043552	256
{F5}	0.200749	1180	{R4, F5}	0.059544	350
{F5, M5}	0.143756	845	{R4, F5, M5}	0.041851	246
{F5, M4}	0.043722	257	{R3, M4}	0.055291	325
{F4, M4}	0.099013	582	{R3, M3}	0.045253	266
{F4, M3}	0.042021	247	{R2, M3}	0.048826	287
{F3, M3}	0.098503	579	{R2, M2}	0.056992	335
{F2, M2}	0.105478	620	{R2, M1}	0.050017	294
{F2, M1}	0.048826	287	{R1, M2}	0.058013	341
{F1, M1}	0.131848	775	{R1, M1}	0.087955	517
{R5, M5}	0.088636	521	{R1, F1, M1}	0.058693	345
{R5, M4}	0.046785	275			

Cluster 1 ‘High Rollers’ ({M5}): This group comprises 1176 customers or 20% of the total customers. It includes customers who have spent considerable money, showcasing their ability to make substantial purchases and potentially contributing significantly to the business revenue. The 845 customers, or 72% of total customers in this cluster, have frequency scores in the top quintile (F5). Concept {F5, M5} is a sub-concept of this concept {M5}. Concept {R5, M5} is also a sub-concept of this concept {M5}. Consequently, it demonstrates that most of the customers in this group purchased products most recently. Regarding the recency scores, 521 customers, or 44% of cluster members, have recency scores in the top quintile (R5). Concept {R5, M5} is a sub-concept of this concept/cluster. Furthermore, 329 customers, or 28% of cluster members, are in the fourth quintile (R4) of recency values. For this reason, most customers in this cluster most recently bought products. Above all, the best customers with RFM scores 5-5-5, and 4-5-5 in sub-concept {R5, F5, M5} and {R4, F5, M5} include 668 customers or 11% of the whole population. These customers are exceptional in their spending habits and significantly impact the business’s success. They are regarded with prestige and recognition for their loyalty and high purchasing power. The implication is that they are highly valued customers.

Cluster 2 ‘Big Spenders’ ({M4}): This segment contains 1175 customers or 20% of all the customers. The customers in this group spent a significant amount of money but did not reach the same level as the High Rollers. Nonetheless, they have a significant impact on the business’s earnings. Moreover, the customers in this segment with frequency scores of F5 and F4 contribute 4% and 10% of the cluster members in sequence.

Cluster 3 ‘Moderate Spenders’ ({M3}): This cluster includes 1176 members or 20% of the total population. This group of customers spent an average amount of money, not as much as the High Rollers or Big Spenders, but still contributed significantly to the business’s revenue. In this particular group, 49% of customers possess frequency scores of 3 (F3) and monetary scores of 3 (M3).

Cluster 4 ‘Low Spenders’ ({M2}): This group includes 1175 members or 20% of the population. This group spent less than the previous groups, but still brought value to the business and should be given attention to maintain engagement and prevent them from leaving while encouraging spending. It is evident that the majority of customers in this particular group, accounting for 53% of cluster members, exhibit frequency scores of 2 (F2) and monetary scores of 2 (M2).

Table 8
Customer segmentation with FCA and total purchase value by cluster.

Cluster/Concept	Total purchase value	% of total purchase value	% of customer number
{M1}	193,389.6	1.09%	20%
{M2}	505,586.8	2.85%	20%
{M3}	1,063,264.1	5.99%	20%
{M4}	2,272,762.6	12.81%	20%
{M5}	13,708,426.1	77.26%	20%
{F5}	11,837,646.0	66.72%	20%

Cluster 5 'Lowest Spenders' ({M1}): This segment includes 1176 members or 20% of the population. This customer group makes the smallest financial contributions among the segments formed. Most customers in this group, with 66% of cluster members, have frequency scores of 1 (F1) and monetary scores of 1 (M1). They have the lowest potential to become loyal customers. In addition, 44% of cluster members have recency scores of 1 (R1) and monetary scores of 1 (M1). In the sub-concept of {R1, F1, M1}, 30% of all customers have RFM scores of 1-1-1, indicating that this group has the highest chance of being lost. These customers are considered the worst customers.

Cluster 6 'Frequent Buyers' ({F5}): Out of all the customers, 20% belong to this group, which has 1180 members. These customers make the most frequent purchases. Within this group, 72% of the members have frequency scores of 5 (F5) and monetary scores of 5 (M5). Additionally, 44% of these customers have recency scores of 5 (R5) and frequency scores of 5 (F5), indicating that they have made their most recent and frequent purchases of products.

We will showcase the effectiveness of our approach by presenting some examples. Fig. 7 is particularly useful for business people interested in customer retention. It highlights that customers in the groups {R3, M4}, {R3, M3}, and {R2, M3} spent a significant amount of money but were inactive for a while. These customers will likely get lost, resulting in a substantial financial loss for the business. As a result, it is necessary to take action affecting these customers. For another example, we have a considerable number of customers in the categories {F4, M4} and {F3, M3} who made multiple purchases involving significant amounts of money. By incentivizing them to make more frequent and higher-value purchases, we can potentially turn them into loyal advocates of our brand.

5.4. Discussion

To compare and discuss traditional K-means clustering and hierarchical clustering, the results of using the Silhouette index and Davies–Boulbin index to assess the performance of the traditional clustering model are shown in Fig. 8. We found that an appropriate choice is 4 – 6 clusters because they are meaningful and valuable in business implications. The number of customer segments should not be too large because it will be challenging to interpret and design marketing strategies. In addition, the choice in most of the prior works is 4 – 5 groups (Ernawati et al., 2021), and we need to compare the results with those studies. Therefore, we divided the customers into six main concepts based on FCA. The cluster profiles and results summarizing traditional K-means clustering and hierarchical clustering are provided in Figs. 9 and 10, and Tables 9 and 10.

After conducting experiments, it was discovered that outliers do not impact FCA. Therefore, it is unnecessary to eliminate outliers during data preprocessing. On the other hand, the K-means algorithm requires outlier management and data normalization as described in Section 4.3. The hierarchical clustering approach is also sensitive to outliers, which may result in imbalanced clustering. Our proposed technique leverages the RFM quintile method to generate RFM scores for all customers, which were then used to create a formal context. As a result, our approach simplifies the preprocessing stage compared to the other two methods.

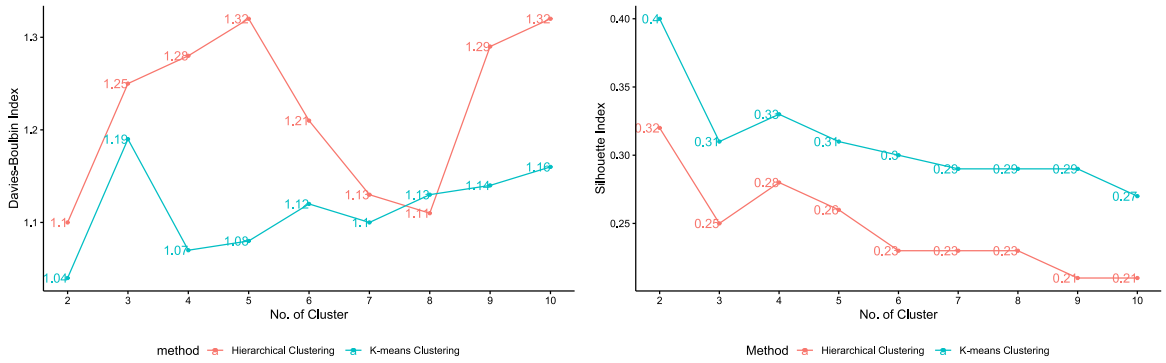
This new methodology combines RFM analysis and FCA; therefore, the relationships can be visualized in a Hasse diagram for the hierarchical concept. Moreover, this approach applies the obtained FCA knowledge structure to extract implication rules to discover knowledge relationships in customer segmentation. K-means clustering gives only completely separate customer groups without their relationships and implicit knowledge. For the hierarchical clustering technique, the results are a complete separation of customer groups with a hierarchically structured dendrogram. FCA also has a strong mathematical theory as support. Thus, the results of FCA can be assessed using mathematical theory. The results show the actual incidents that occur in the dataset. This new approach mainly creates hierarchical overlapping clusters. Therefore, a customer possibly (usually) belongs to more than one hierarchical concept or cluster.

A novel approach that combines RFM analysis and FCA has been introduced, enabling customer relationships to be visualized through a Hasse diagram for a hierarchical concept. This methodology utilizes the FCA knowledge structure to extract implication rules, facilitating the discovery of knowledge relationships in customer segmentation, in contrast to K-means clustering, which only provides distinct customer groups without their relationships and implicit knowledge. The hierarchical clustering technique also completely separates customer groups and is associated with a hierarchically structured dendrogram. This innovative methodology primarily generates hierarchical overlapping clusters, meaning a customer may belong to more than one hierarchical concept or cluster. Moreover, FCA is supported by a strong mathematical theory, allowing for rigorous evaluation of the results. The outcomes of our approach indicate the actual incidents that occur in the dataset. The results of our proposed approach are presented in a clear and understandable way with the help of different visual aids, such as the sublattice in Fig. 6 and the hierarchical concepts in Fig. 7.

From Section 5.3, the business implication deployment examples highlight how our new approach can be effectively implemented and visualized in the context of real incidents. This is in contrast to the limited representation of each customer group presented by K-means clustering and hierarchical clustering techniques, as shown in Table 9 and Table 10.

For this reason, business decision-makers have improved opportunities to deliver the customers an appropriately matched marketing strategy and increase customer retention and satisfaction. In addition, the FCA approach makes decision-makers able to clearly visualize the real incidence of customer purchase behavior. Above all, this new approach provides ease of processing and understanding of the results. The comparison summary is shown in Table 11.

To deploy in marketing strategy and recommendations, the total purchase value of the most valuable customers is almost 80 percent of total sales during the data accumulation period. On the other hand, the company gains only 1 percent of the total purchases from the worst customer group. The best customers are high-potential targets for new products. They will help the company to promote its products and brand. Therefore, the company should reward these customers and maintain close relationships to increase revenue. In addition, the company should understand why the worst customers or lost customers did not buy the products anymore. For customers with average frequency and good total sales, the company should offer promotions and recommend products for upselling. The company should provide special offers to increase customer visits, leading to more purchases.



(a) Davies-Boulbin index (b) Silhouette index

Fig. 8. Davies–Boulbin index and Silhouette index of hierarchical and K-means clustering at different cluster numbers.

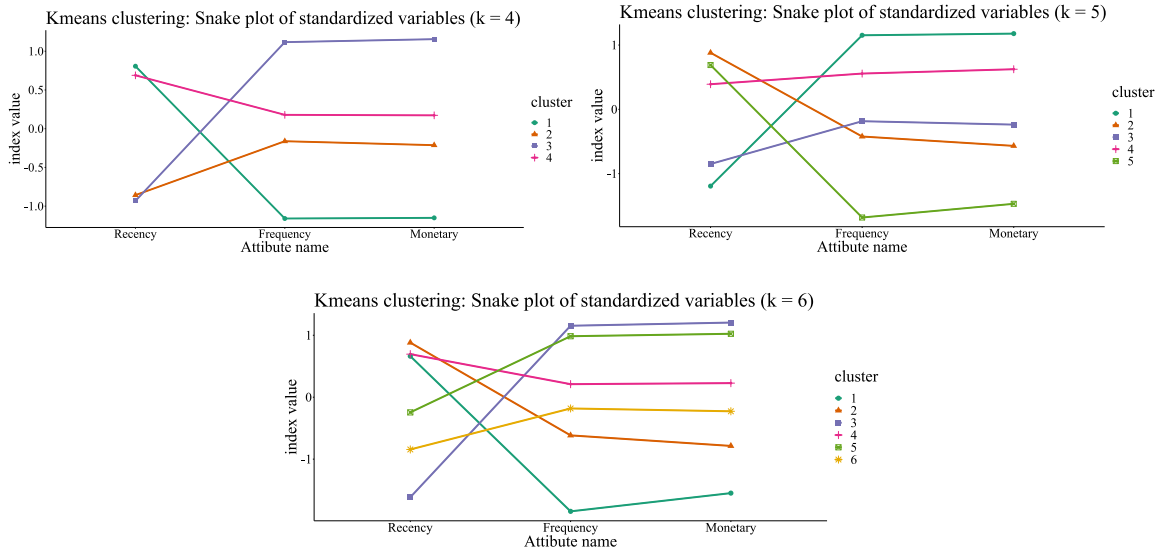


Fig. 9. Snake plots of standardized variables of K-means clustering output.

Table 9
K-means clustering output summary related to Fig. 9.

k = 4					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	High (394.26)	Low (13.66)	Low (248.20)	1472	Lost customers
2	Low (32.91)	Medium (48.10)	Medium (765.57)	1121	Potential loyalists
3	Low (37.75)	High (251.59)	High (4205.91)	1432	Champions
4	High (311.10)	Medium (75.13)	Medium (1299.57)	1608	Hibernating customers
k = 5					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	Low (20.95)	High (264.09)	High (4369.32)	1141	Champions
2	High (399.47)	Medium (30.88)	Medium Low (480.56)	1450	Hibernating customers
3	Low (32.73)	Medium (46.31)	Medium (737.69)	1061	Potential loyalists
4	Medium High (216.01)	Medium High (122.99)	Medium High (2173.25)	1263	At risk
5	High (363.34)	Low (6.31)	Low (176.06)	718	Lost customers
k = 6					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	High (360.09)	Low (5.00)	Low (168.14)	565	Lost customers
2	High (407.47)	Medium (23.20)	Medium Low (349.47)	1238	Hibernating customers
3	Low (8.91)	High (273.09)	High (4631.04)	683	Champions
4	High (308.96)	Medium High (73.98)	Medium High (1279.63)	1105	At risk
5	Medium (82.73)	High (208.63)	High (3479.89)	990	Loyal customers
6	Medium Low (32.58)	Medium (46.34)	Medium (747.53)	1052	Potential loyalists

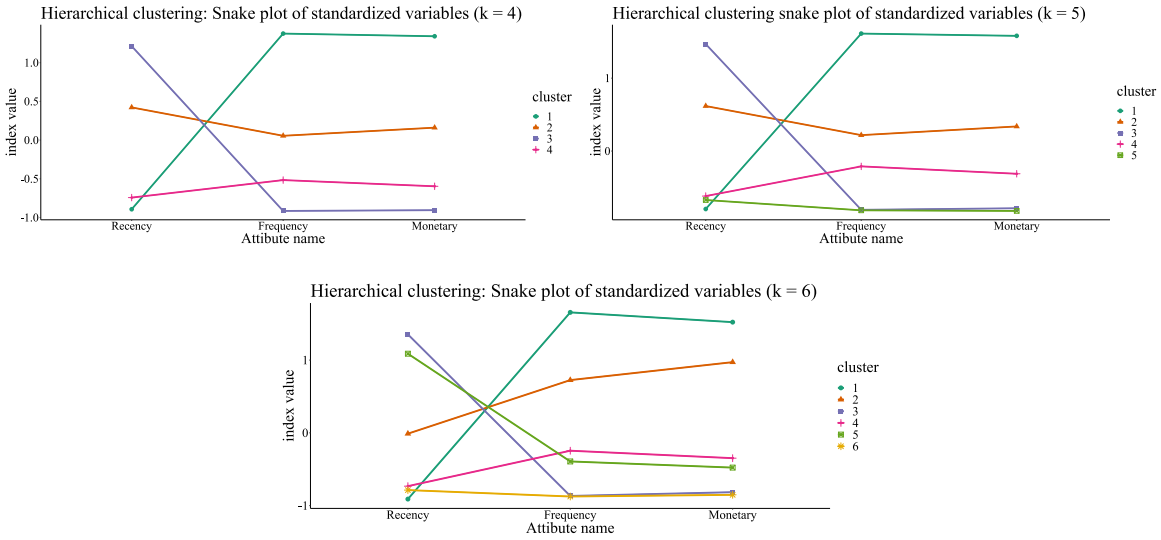


Fig. 10. Snake plots of standardized variables of hierarchical clustering output.

Table 10
Hierarchical clustering output summary related to Fig. 10.

k = 4					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	Low (14.01)	High (265.24)	High (4302.61)	913	Champions
2	Medium High (258.88)	Medium (122.63)	Medium (2223.53)	1594	Hibernating customers
3	High (406.21)	Low (17.78)	Low (346.06)	1688	Lost customers
4	Low (42.029)	Medium (60.81)	Medium (889.69)	1438	Potential loyalists
k = 5					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	Low (14.01)	High (265.24)	High (4302.61)	913	Champions
2	Medium High (258.88)	Medium (122.63)	Medium (2223.53)	1594	Hibernating customers
3	High (406.21)	Low (17.78)	Low (346.06)	1688	Lost customers
4	Low (44.75)	Medium Low (78.76)	Medium Low (1137.65)	1020	Potential loyalists
5	Low (35.39)	Low (17.03)	Low (284.63)	418	New customers
k = 6					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	Low (14.01)	High (265.24)	High (4302.61)	913	Champions
2	Medium (169.74)	Medium High (173.95)	Medium High (3371.74)	848	Loyal customers
3	High (406.21)	Low (17.78)	Low (346.06)	1688	Lost customers
4	Low (44.75)	Medium Low (78.76)	Medium Low (1137.65)	1020	Potential loyalists
5	High (360.20)	Medium Low (64.29)	Medium Low (918.33)	746	Hibernating customers
6	Low (35.39)	Low (17.03)	Low (284.63)	418	New customers

Table 11
FCA-based, K-means and hierarchical clustering comparison.

Issues	FCA	K-means	Hierarchical
Preprocessing	Less	More	More
Outlier	Not sensitive to outliers	Sensitive to outliers and outlier management needed	Sensitive to outliers leading to imbalanced clustering results
Data normalization	RFM quintile method used	Data normalization needed	Data normalization needed
Relationship information among segments	Relationship demonstration with Hasse diagram and hierarchical concept	Complete separation of customer groups	Complete separation of customer groups with hierarchically structured dendrogram
Mathematical theory support	Strong support	Cluster centroid initialization sensitivity	Imbalanced clustering results.
Implicit Knowledge	Included	Not included	Not included
Clustering type	Soft clustering	Hard clustering	Hard clustering
Clustering result	Hierarchical overlapping clusters	Unstructured non-overlapping clusters	Hierarchical structured non-overlapping clusters
Visualization	Easy and various	More difficult	More difficult

6. Conclusion

This research proposes a new effective clustering algorithm using the advantages of FCA to build a knowledge representation. This model combines the RFM model with FCA. Thus, the construction contains both implicit and explicit knowledge. Explicit knowledge shows cluster visualized information represented in the hierarchical structure model, while implicit knowledge is embedded in the structure with its implication properties. Thus, the knowledge structure from FCA reveals relationships among data points and easily understood results. Afterwards, the proposed model was compared with K-means clustering and hierarchical clustering using the online retail II dataset from the UCI Machine Learning Repository. In conclusion, the proposed method provides enough and appropriate information for marketers to perceive the value of the clustering results for creating practical marketing strategies in real-world business. This approach offers marketers both customer segmentation and relationships in customer data simultaneously. The advantage of the RFM model is the use of a very small number of variables (only three variables) to reduce the complexity of the model. However, we suggest adding new variables to the RFM model to increase the accuracy and gain more information in future studies. In addition, we will modify and improve this model by representing RFM values in a non-binary formal context in future studies using fuzzy sets. Then, FCA will be considered and compared with the results from this alternative approach. Therefore, we recommend employing it in the context of other industries in future research. This proposed method can also be applied to businesses other than online retail because the characteristics of the retail dataset are similar to our experiment. In addition, visualizations play a crucial role in achieving success in data-driven decision-making with this approach regarding information systems and application development. We suggest the application should provide interactive visualization options that enable users to add or choose features and adjust for better visualization effortlessly.

CRediT authorship contribution statement

Chongkolnee Rungruang: Methodology, Software, Writing – original draft, Visualization, Data curation, Validation. **Pakwan Riyapan:** Supervision, Validation, Writing – review & editing. **Arthit Intarasit:** Supervision, Validation, Writing – review & editing. **Khanchit Chuarkham:** Supervision, Validation, Writing – review & editing. **Jirapond Muangprathub:** Conceptualization, Methodology, Software, Supervision, Validation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We use dataset from UCI repository at: https://archive.ics.uci.edu.

Acknowledgments

The authors are deeply grateful to the Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Thailand. This research was financially supported by the Research and Development Office, Prince of Songkla University, Thailand, under grant No. SIT6502069S. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation. The authors also gratefully acknowledge the helpful check of the English language by Assoc. Prof. Dr. Seppo Karrila.

References

Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, http://dx.doi.org/10.1007/s10257-023-00640-4.

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785–1792. http://dx.doi.org/10.1016/j.jksuci.2019.12.011, URL https://www.sciencedirect.com/science/article/pii/S1319157819309802.

Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2018). Customer segmentation in e-commerce: Applications to the cashback business model. *Journal of Business Research*, *88*, 407–414. http://dx.doi.org/10.1016/j.jbusres.2017.11.047, URL https://www.sciencedirect.com/science/article/pii/S0148296317304939.

Benavent, X., Castellanos, A., de Ves, E., García-Serrano, A., & Cigarrán, J. (2019). FCA-based knowledge representation and local generalized linear models to address relevance and diversity in diverse social images. *Future Generation Computer Systems*, *100*, 250–265. http://dx.doi.org/10.1016/j.future.2019.05.029, URL https://www.sciencedirect.com/science/article/pii/S0167739X18307271.

Castellanos, A., Cigarrán, J., & García-Serrano, A. (2017). Formal concept analysis for topic detection: A clustering quality experimental analysis. *Information Systems*, *66*, 24–42. http://dx.doi.org/10.1016/j.is.2017.01.008, URL https://www.sciencedirect.com/science/article/pii/S030643791730087X.

Chang, H., & Tsay, S. (2004). Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation. *Journal of Information Management*, *11*(4), 161–203, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-79951576359&partnerID=40&md5=b5597c45dbdb9db7076022549b658424.

Chen, D., Guo, K., & Li, B. (2019). Predicting customer profitability dynamically over time: An experimental comparative study. In *Iberoamerican congress on pattern recognition* (pp. 174–183). Springer.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(3), 197–208. http://dx.doi.org/10.1057/dbm.2012.17.

Chen, H., Zhang, L., Chu, X., & Yan, B. (2019). Smartphone customer segmentation based on the usage pattern. *Advanced Engineering Informatics*, *42*, Article 101000. http://dx.doi.org/10.1016/j.aei.2019.101000, URL https://www.sciencedirect.com/science/article/pii/S1474034619305737.

Chen, Q., Zhang, M., & Zhao, X. (2017). Analysing customer behaviour in mobile app usage. *Industrial Management & Data Systems*.

Choi, H., Choi, E. K., Yoon, B., & Joung, H. W. (2020). Understanding food truck customers: Selection attributes and customer segmentation. *International Journal of Hospitality Management*, *90*, Article 102647. http://dx.doi.org/10.1016/j.ijhm.2020.102647, URL https://www.sciencedirect.com/science/article/pii/S0278431920301997.

Chorianopoulos, A. (2016). *Effective CRM using predictive analytics*. John Wiley & Sons.

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, *33*(10), 1251–1257. http://dx.doi.org/10.1016/j.jksuci.2018.09.004.

Cordero, P., Enciso, M., López-Rodríguez, D., & Mora, Á. (2022). fcaR, formal concept analysis with R. *R J.*, *14*, 341–361.

Craft, R. C., & Leake, C. (2002). The Pareto principle in organizational decision making. *Management Decision*, *40*(8), 729–733. http://dx.doi.org/10.1108/00251740210437699.

Dedi, Dzulhaq, M. I., Sari, K. W., Ramdhan, S., Tullah, R., & Sutarman (2019). Customer segmentation based on RFM value using K-means algorithm. In *2019 fourth international conference on informatics and computing* (pp. 1–7). http://dx.doi.org/10.1109/ICIC47613.2019.8985726.

Deng, Y., & Gao, Q. (2020). A study on e-commerce customer segmentation management based on improved K-means algorithm. *Information Systems and e-Business Management*, *18*(4), 497–510. http://dx.doi.org/10.1007/s10257-018-0381-3.

Dolnicar, S., Grün, B., & Leisch, F. (2018a). *Market segmentation analysis: Understanding it, doing it, and making it useful*. Springer Nature.

Dolnicar, S., Grün, B., & Leisch, F. (2018b). Step 5: Extracting segments. In *Market segmentation analysis* (pp. 75–181). Springer.

Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2021). A review of data mining methods in RFM-based customer segmentation. *Journal of Physics: Conference Series*, *1869*(1), http://dx.doi.org/10.1088/1742-6596/1869/1/012085.

Fränti, P., & Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. *Applied Intelligence*, *48*(12), 4743–4759. http://dx.doi.org/10.1007/s10489-018-1238-7.

Frasquet, M., Ieva, M., & Ziliani, C. (2021). Online channel adoption in supermarket retailing. *Journal of Retailing and Consumer Services*, *59*, Article 102374. http://dx.doi.org/10.1016/j.jretconser.2020.102374.

Gagolewski, M., Bartoszuk, M., & Cena, A. (2021). Are cluster validity measures (in) valid? *Information Sciences*, *581*, 620–636. http://dx.doi.org/10.1016/j.ins.2021.10.004, URL https://www.sciencedirect.com/science/article/pii/S0020025521010082.

Ganter, B., & Wille, R. (2012). *Formal concept analysis: mathematical foundations*. Springer Science & Business Media.

Gupta, S., Justy, T., Kamboj, S., Kumar, A., & Kristoffersen, E. (2021). Big data and firm marketing performance: Findings from knowledge-based view. *Technological Forecasting and Social Change*, *171*, Article 120986. http://dx.doi.org/10.1016/j.techfore.2021.120986, URL https://www.sciencedirect.com/science/article/pii/S0040162521004182.

Hiziroglu, A. (2013). Soft computing applications in customer segmentation: State-of-art review and critique. *Expert Systems with Applications*, *40*(16), 6491–6507. http://dx.doi.org/10.1016/j.eswa.2013.05.052, URL https://www.sciencedirect.com/science/article/pii/S0957417413003503.

Hosseini, M., Abdolvand, N., & Harandi, S. R. (2022). Two-dimensional analysis of customer behavior in traditional and electronic banking. *Digital Business*, *2*(2), Article 100030. http://dx.doi.org/10.1016/j.digbus.2022.100030, URL https://www.sciencedirect.com/science/article/pii/S2666954422000102.

Khalili-Damghani, K., Abdi, F., & Abolmakarem, S. (2018). Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Applied Soft Computing*, *73*, 816–828. http://dx.doi.org/10.1016/j.asoc.2018.09.001, URL https://www.sciencedirect.com/science/article/pii/S1568494618305052.

Kim, B. J., Singh, V., & Winer, R. S. (2017). The Pareto rule for frequently purchased packaged goods: an empirical generalization. *Marketing Letters*, *28*(4), 491–507. http://dx.doi.org/10.1007/s11002-017-9442-5.

Kuznetsov, S., & Makhalova, T. (2018). On interestingness measures of formal concepts. *Information Sciences*, *442–443*, 202–219. http://dx.doi.org/10.1016/j.ins.2018.02.032, URL https://www.sciencedirect.com/science/article/pii/S0020025516315791.

Kuznetsov, S. O., & Poelmans, J. (2013). Knowledge representation and processing with formal concept analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(3), 200–215.

Lei, Y., Yan, Y., Han, Y., & Jiang, F. (2018). The hierarchies of multivalued attribute domains and corresponding applications in data mining. *Wireless Communications and Mobile Computing*, *2018*.

Marchetto, A. (2005). A concerns-based metrics suite for web applications. *INFOCOMP Journal of Computer Science*, *4*(3), 11–22.

Meng, X., Liu, M., Wu, J., Zhou, H., Xu, F., & Wu, Q. (2020). Hierarchical clustering on metric lattice. *International Journal of Intelligent Information and Database Systems*, *13*(1), 1–16.

Monalisa, S., Nadya, P., & Novita, R. (2019). Analysis for customer lifetime value categorization with RFM model. *Procedia Computer Science*, *161*, 834–840. http://dx.doi.org/10.1016/j.procs.2019.11.190, URL https://www.sciencedirect.com/science/article/pii/S1877050919319015 The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.

Munusamy, S., & Murugesan, P. (2020). Modified dynamic fuzzy c-means clustering algorithm – Application in dynamic customer segmentation. *Applied Intelligence*, *50*(6), 1922–1942. http://dx.doi.org/10.1007/s10489-019-01626-x.

Nakano, S., & Kondo, F. N. (2018). Customer segmentation with purchase channels and media touchpoints using single source panel data. *Journal of Retailing and Consumer Services*, *41*, 142–152. http://dx.doi.org/10.1016/j.jretconser.2017.11.012.

Nandapala, E., & Jayasena, K. (2020). The practical approach in customers segmentation by using the K-means algorithm. In *2020 IEEE 15th international conference on industrial and information systems* (pp. 344–349). http://dx.doi.org/10.1109/ICIIS51140.2020.9342639.

Nguyen, T. B., Tran, C. D., Quan, T. T., Nguyen, M. H., & Le, T. A. (2019). MarCHGen: A framework for generating a malware concept hierarchy. *Expert Systems*, *36*(5), Article e12445.

Noori, B. (2015). An analysis of mobile banking user behavior using customer segmentation. *International Journal of Global Business*, *8*(2).

Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. *Marketing Intelligence & Planning*, *35*(4), 544–559. http://dx.doi.org/10.1108/MIP-11-2016-0210.

Poelmans, J., Ignatov, D. I., Kuznetsov, S. O., & Dedene, G. (2013). Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications*, *40*(16), 6538–6560. http://dx.doi.org/10.1016/j.eswa.2013.05.009.

Rahim, M. A., Mushafiq, M., Khan, S., & Arain, Z. A. (2021). RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services*, *61*, Article 102566. http://dx.doi.org/10.1016/j.jretconser.2021.102566, URL https://www.sciencedirect.com/science/article/pii/S0969698921001326.

Ravi, K., Ravi, V., & Prasad, P. S. R. K. (2017). Fuzzy formal concept analysis based opinion mining for CRM in financial services. *Applied Soft Computing*, *60*, 786–807. http://dx.doi.org/10.1016/j.asoc.2017.05.028, URL https://www.sciencedirect.com/science/article/pii/S1568494617302910.

Rebekić, A., Lončarić, Z., Petrović, S., & Marić, S. (2015). Pearson's or Spearman's correlation coefficient-which one to use? *Poljoprivreda*, *21*(2), 47–54.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763–1768. http://dx.doi.org/10.1213/ANE.0000000000002864.

Shokouhyar, S., Shokoohyar, S., & Safari, S. (2020). Research on the influence of after-sales service quality factors on customer satisfaction. *Journal of Retailing and Consumer Services*, *56*, Article 102139. http://dx.doi.org/10.1016/j.jretconser.2020.102139, URL https://www.sciencedirect.com/science/article/pii/S0969698919313311.

Singh, J., & Mittal, M. (2021). Customer's purchase prediction using customer segmentation approach for clustering of categorical data. *Management and Production Engineering Review*, *12*.

Singh, S., & Srivastava, S. (2020). Review of clustering techniques in control system: Review of clustering techniques in control system. *Procedia Computer Science*, *173*, 272–280. http://dx.doi.org/10.1016/j.procs.2020.06.032, URL https://www.sciencedirect.com/science/article/pii/S1877050920315362 International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020.

Sokol, O., & Holỳ, V. (2021). The role of shopping mission in retail customer segmentation. *International Journal of Market Research*, *63*(4), 454–470.

Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., & Rahmani, R. (2018). Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: A case study. In *2018 IEEE 15th international conference on E-business engineering* (pp. 119–126). http://dx.doi.org/10.1109/ICEBE.2018.00027.

Tsiptsis, K. K., & Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons.

Wang, S. C., Tsai, Y. T., & Ciou, Y. S. (2020). A hybrid big data analytical approach for analyzing customer patterns through an integrated supply chain network. *Journal of Industrial Information Integration*, *20*, Article 100177. http://dx.doi.org/10.1016/j.jii.2020.100177, URL https://www.sciencedirect.com/science/article/pii/S2452414X20300522.

Wille, R. (2009). Restructuring lattice theory: an approach based on hierarchies of concepts. In *International conference on formal concept analysis* (pp. 314–339). Springer.

Wu, T., & Liu, X. (2020). A dynamic interval type-2 fuzzy customer segmentation model and its application in E-commerce. *Applied Soft Computing*, *94*, Article 106366. http://dx.doi.org/10.1016/j.asoc.2020.106366, URL https://www.sciencedirect.com/science/article/pii/S1568494620303069.

Yeh, I.-C., Yang, K. J., & Ting, T. M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, *36*(3, Part 2), 5866–5871. http://dx.doi.org/10.1016/j.eswa.2008.07.018, URL https://www.sciencedirect.com/science/article/pii/S0957417408004508.

Zeybek, H. (2018). Customer segmentation strategy for rail freight market: The case of Turkish State Railways. *Research in Transportation Business & Management*, *28*, 45–53. http://dx.doi.org/10.1016/j.rtbm.2018.10.003, URL https://www.sciencedirect.com/science/article/pii/S2210539516301596.

Zhang, Z., Zhao, J., & Yan, X. (2018). A web page clustering method based on formal concept analysis. *Information*, *9*(9), http://dx.doi.org/10.3390/info9090228, URL https://www.mdpi.com/2078-2489/9/9/228.

Zhou, J., Wei, J., & Xu, B. (2021). Customer segmentation by web content mining. *Journal of Retailing and Consumer Services*, *61*, Article 102588. http://dx.doi.org/10.1016/j.jretconser.2021.102588, URL https://www.sciencedirect.com/science/article/pii/S0969698921001545.