A Systematic Literature Review of Data Science Techniques, Applications, and Tools: A Comprehensive Analysis

Apri Junaidi
Computer Science,
Faculty of Computing
Malaysia, Indonesia
junaidi20@graduate.utm.my

Faradhysa Camila
Computer Science,
Faculty of Computing
Malaysia, Indonesia
cfaradhysa@graduate.utm.my

Wafaa Zahira Binti Anas
Computer Science,
Faculty of Computing
Johor, Malaysia
wafaazahira@graduate.utm.my

Zainab Ali Albashah
Computer Science,
Faculty of Computing
Malaysia, Indonesia
zainabali@graduate.utm.my

***Abstract*— Data science has become an essential discipline, needed for enhancing competitiveness as well as improving organisations and industries operations with the help of advanced analytical tools. This paper provides a systematic review of the current available tools and their efficiency in data science in relation to key areas and methods of this modern and rapidly developing field. The review aims to address three key research questions: what kind of data science algorithms are used in practice and for what purposes, how data science is affecting the decision making and functioning of various industries, and what approaches are effective for an analytics project, as well as the assistance provided by existing tools. Hence, this review gathers knowledge from different peer-reviewed articles, obtained from Web of Science, Science Direct, Scopus, and IEEE by categorising literature based on the implementation of data science algorithms in industrial environments, measures the impact of such implementations on businesses and assesses the performance and enablers of the predictive analytics tools. The results contribute to the understanding of what data science techniques and tools are available, how they perform in practice, and the areas where further research and development might be useful.**

Keywords — Technique and algorithm, application, tools and libraries.

I. Introduction

Data science is a set of methods, techniques, tools, and processes that enables to obtain useful and valuable information from structured and unstructured data based on scientific methodologies [6]. Data science provides a significant role to current analytics with its approaches and methods affecting fields. Understanding a variety of data science techniques and their applicability is crucial because data volume and data variety will continuously grow in the future. The systematic literature review aims at providing an understanding of the current state of data science in terms of the techniques, applications, and tools. The review will address several key areas: in the domain of data science, what tools it uses, what technique and algorithms it involves, and a feasibility assessment of the different types of tools to reach an optimum result.

Data science techniques comprise a wide array of activities that are conducted to mine knowledge from data and this entail; Data gathering, Data cleaning, Data pre-processing, Data analysis and modelling, Data visualisation. In general, data science methods offer a flexible and effective approach for identifying insights, forecasting results, and improving operations across multiple industries, such as production, medicine, finance, and others [11].

In modern data science application of algorithms is expected to further augment the decision-making process and incorporate real time data for improved efficiency and accuracy. The use of algorithms in data science is expected to advance to other levels where they will substitute many human decisions hence enhancing automation in numerous sectors [6].

As the data science technology and algorithms are used in different fields like business intelligence, data analysis and big data analysis. These techniques entail such concepts such as data mining, statistical analysis, and other methods of machine learning to make information from large sets more favourable. Data science techniques help organisations make and inform decisions, forecast the future and find patterns in the data. Data sciences hold a significant part in the Business Intelligence applications for big data analysis wherein the features of classes, parameters, and observation help solve various problems [3].

Furthermore, the applicability of various tools and libraries proposed in the field of data science is an essential component of this analysis. Some of the popular tools that can be used for the same includes the Scikit-learn libraries in Python and for those working in R, there is the caret package. Data science tools are essential to support the capabilities of organisations in utilising data resources for analytical decision-making [9]. Therefore, this study seeks to outline a comprehensive and integrated understanding of the effectiveness of data science techniques, its applications and tools through a systematic review proposal on the existing literature with an intention of ascertaining the current trends and research gaps.

II. Background Study

A. Techniques and Algorithm

Data science methods and approaches are used in many disciplines to support using large datasets to make better decisions. Techniques and algorithms of data science include methods that are used in the collection, processing, analysis and interpretation of large amounts of data that have demonstrated the potential to enhance the process of decision-making [12].

Some of the commonly applied data science methods and approaches that are used in the area of rescheduling are; Machine Learning (ML) for pattern detection and decision support in rescheduling scenarios, Genetic Algorithms (GA) for enhanced meta heuristic search in scheduling, Tabu Search (TS) for optimization of rescheduling in production, and Simulated Annealing (SA) for solution of combinatorial problems in production [11]. The other technique and algorithm is deep-learning and it entails using algorithms referred to as deep neural networks to learn the specific information that will be used to decide or predict a certain outcome through the use of large data sets and without having to program the result [7].

One of the machine learning algorithms is; Naïve Bayes classifier which is probabilistic in nature used for classification which estimates the probability of an instance belonging to a class given the feature scores. It performs well with small training data and is used in text categorization, disease diagnosis, and spam detection. Although it uses the feature independence assumption, it can perform well against many complex models and is considered an effective tool in data science [12].

B. Application, Tools and Libraries

Data tools and libraries are crucial to enable organisations to harness the data resources to support analytical decision-making [9]. Data science tools and libraries are considered as components that are providing a systematic set of tools for effective data manipulation and analysis. By using these data science resources, data processing was made faster, the accuracy of analysis was enhanced and the results they obtained were more stable and consistent [1].

As for data modelling, Python is more preferred over R because Python consumes less memory overhead and is more efficient as well as having better documentation particularly when using the Jupyter notebook. The integration of using Python along with QlikView meant that there was a direct API plug-in to import and analyse Python outputs on the application. Despite Tableau or Power BI having slightly improved visualisation formats, QlikView was favoured by the organisation since it had existing cloud servers, and senior management is acquainted with the application's dashboard [9]. Python is applied in environments such as Colab, Visual Studio Code, Spyder based on the Python language. There were some libraries that were used and these included; Pandas to manage the data, Numpy for the numerical computations, Matplotlib for data visualisation and Sklearn for the purposes of machine learning. These tools enabled data cleaning, data manipulation, statistical analysis, graph construction, and dimensionality reduction, thus improving the study's analysis. With the help of tools which are used widely Pandas, Matplotlib, Numpy, Plotly, Seaborn, and Sklearn enhanced my skills of data management to a great extent [1].

III. Research Method

In this paper, the Systematic Literature Review (SLR) approach is used to systematically review and analyse different analytics methodologies in literature with an overall goal of assessing the tools currently available with regards to their performance, based on studies done so far. Articles that will be searched with relevant keywords will be searched in several reputable online research databases such as Web of Science, Science Direct, Scopus, and IEEE since these platforms contain vast volumes of peer-reviewed articles with considerable emphasis on data science. It entails identifying key search terms, searching the databases and using facets to screen articles and related information. The selected articles will be critically read and the data related to various analytics types, instruments used, their utilisation, and stated productivity will be determined. Research results will be discussed to reveal trends and generalities, the state of knowledge in the field, and potential areas for research and analysis of the effectiveness of data science analytics approaches and methods, as well as to propose ideas for targeted research.

A. Research Question

Table 3.1: Research Questions

No.	**Research Questions**
RQ1	What are the main techniques and algorithms used in data science?
RQ2	What are the key applications of data science across various domains?
RQ3	What tools and libraries are most commonly used in data science projects?

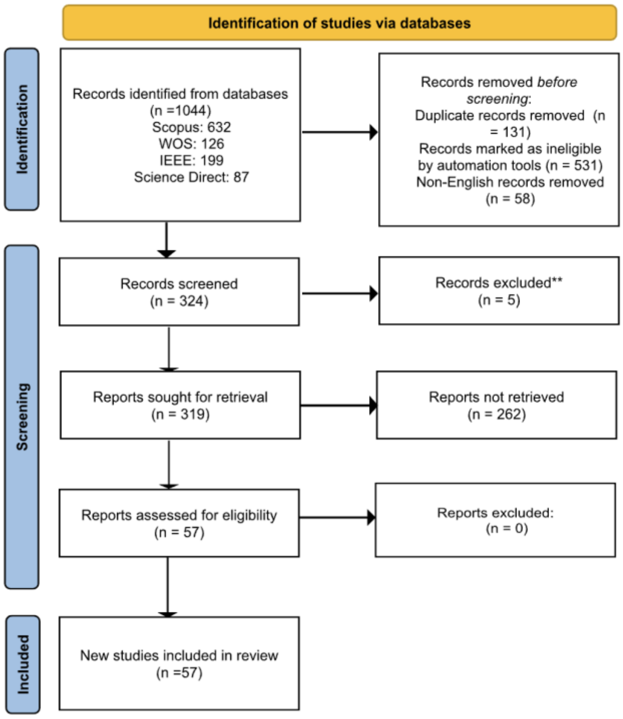
B. Quality Assessment

Table 3.2: Quality Assessment

No.	Quality Assessment Questions	Relevant to Research Question
QA1	How does the author(s) present the implementation of techniques and algorithms in the studies?	RQ1
QA2	How does the author(s) discuss the key applications of data science across various domains?	RQ2
QA3	How does the author(s) present and evaluate the tools and libraries used in data science projects?	RQ3

IV. Result

A. Study Fund



Identification of studies via databases
Identification
Records identified from databases
(n =1044)
Scopus: 632
WOS: 126
IEEE: 199
Science Direct: 87
Records removed *before screening*:
Duplicate records removed (n = 131)
Records marked as ineligible by automation tools (n = 531)
Non-English records removed (n = 58)
Screening
Records screened
(n = 324)
Records excluded**
(n = 5)
Reports sought for retrieval
(n = 319)
Reports not retrieved
(n = 262)
Reports assessed for eligibility
(n = 57)
Reports excluded:
(n = 0)
Included
New studies included in review
(n =57)

Figure 4.1: PRISMA Flow

In this paper, articles are searched in different databases using keywords used; ("data science") AND ("techniques" OR "algorithm") AND ("application") AND ("tools" OR "libraries"). In the Identification phase, a total of 1,044 records are identified from four databases including Scopus (632 records), Web of Science (126 records), IEEE (199 records) and ScienceDirect (87 records). Next, the records are further screened to remove 131 duplicate records, 58 papers with non-English languages and 531 records that are marked ineligible by automation tool resulting in 324 unique records screened.

B. Inclusion and Exclusion Criteria

The article papers are selected based on the following inclusion criteria:

- Articles are published in the years of 2019 until 2023.
- Articles contain and state clearly in the title and abstract about data science algorithms, techniques, tools and libraries that are applied and used in different studies of fields and industries.

The article papers are excluded from the research based on the following exclusion criteria:

- Articles that do not leverage any data science technologies of algorithms, techniques, tools and libraries.
- Articles that are reviews or surveys.

C. Background Analysis

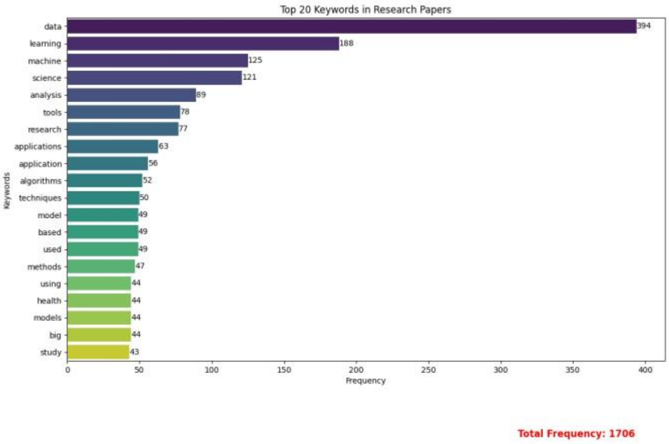


Figure 4.2: Top 20 Keywords in Research Papers Bar Chart

The bar chart 4.2 gives a view of 20 most used keywords in the context of the research papers as per the frequency of occurrence. The keyword "data" has been found the most frequently used followed by the keywords "learning," "machine," "science," and "analysis." The bar chart on the right quantitatively supports the importance of these terms in the existing academic literature, where the length of each bar depicts the frequency of the particular term used. This view can be considered as complementary to the word cloud as it offers the specific quantitative data on the frequency of terms and again emphasises the importance of data related themes and methodologies in the research area.

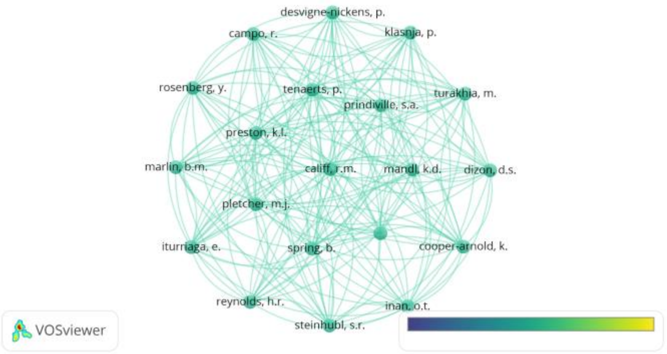


Figure 4.3: Co-Authorship Network Graph

The figure 4.3 represents the co-authorship network graph obtained from VOSviewer, which demonstrates the collaborative ties between authors in the analysed studies. Here, each node corresponds to an author, and the size of the node translates to the number of papers published by the author, while the lines connecting the nodes signify co-authorship relations, with thicker lines denoting more frequent collaborations. Central authors such as Preston, K. L. Califf, R. M., and Rosenberg, Y. also have large nodes and many connections indicating that they played a rather active part in the collaborations. The high density indicates that authors are frequently connected and closely interacting with each other, while the authors at the periphery of the network are Desvigne-Nickens, P., and Iturriaga, E. and have comparatively low collaboration levels. In general, the network can be characterised as a cohesive research community that supports significant levels of interaction and research progress within the domain.

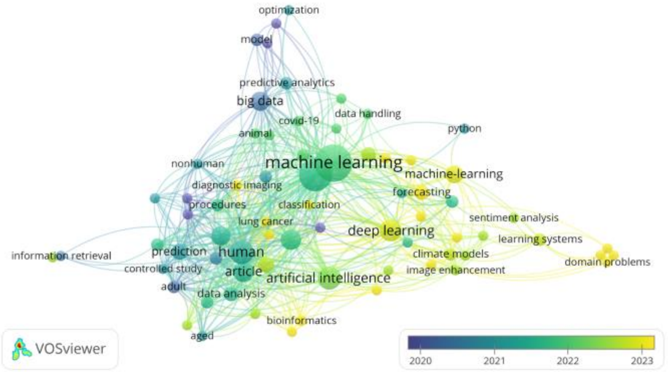


Figure 4.4: Keywords Network Visualization Graph

The network visualisation graph obtained from the VOSviewer based on 57 primary studies illustrates the terms obtained from the titles and abstracts of the papers in terms of their occurrence and co-occurrence patterns. The graph reveals three main clusters: While the green cluster is related to intrusion detection systems with keywords like ‘dataset’, ‘feature’, ‘technique’, ‘algorithm’, and ‘deep learning’, the blue cluster is concerned with the validation phase which includes the terms like ‘experimental result’; the yellow cluster deals with IoT security concerns with terms like ‘internet’, ‘thing’, ‘device’, ‘IoT device’, ‘attack’, ‘detection’, and ‘botnet’. From this analysis, it can be concluded that the studies are methodologically oriented towards specific areas such as IoT security and IDS as they are based on a solid experimental foundation, which supports the conclusion that this review is valid.

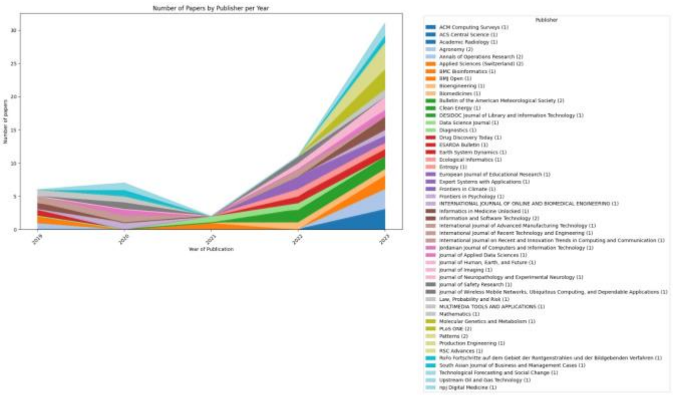


Figure 4.5: Number of Papers by Publisher per Year Area Chart

The chart in figure 4.5, shows the total records of research papers published by various publishers from the year 2019 to 2023. The thickness of the bar illustrates the number of papers being published, and the coloured bar is the representation of the number of publishers. The chart shows a general trend of publications rising and peaking in 2021 to 2023, and the author believes that this indicates expansion in the research field. The different colours and segments simply depict the myriad of publishers that are involved in producing literature for academics hence showing that there are many players in the market with a huge and diverse interest in the subject matters.

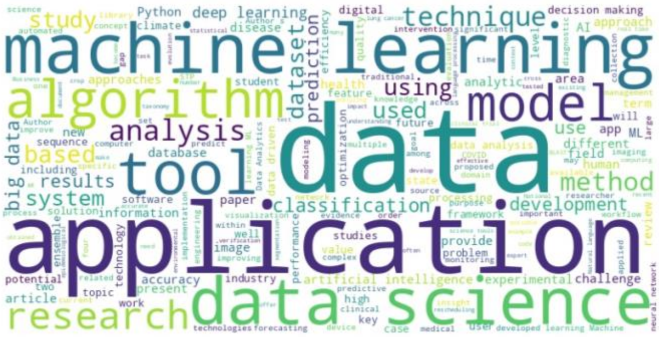


Figure 4.6: Data Science Word Cloud

The word cloud in figure 4.6 depicts the general frequency of the keywords in the research papers. The size of the word and its contrast to the background means that the frequently used terms are depicted in larger and bolder fonts. Popular words such as “machine learning”, “data”, “algorithm”, “application”, and “science” indicate that these are the key topics in the field. This visualisation allows one to easily recognize the main topics and trends in the field of research and the focus on data science, machine learning, and algorithms.

D. Machine Learning Algorithm: Supervised Learning

Supervised learning is a method in learning in which a machine is trained together with labelled examples about the features of data. In this case, the input dataset consists of examples and the corresponding labels or targets which characterise important aspects of the associated data. The algorithms used in supervised learning are typically classifiers and regressors which differ based on the type of output variable they predict: the classification algorithms output result in categorisation or classification of classes, while regression algorithms help in predicting a real value or a probability [13].

The paper presents information about supervised learning with reference to radiomics as an integration of image analysis facilitated by computers accompanied by machine learning. In particular, it explains that the use of quantitative analysis targets image descriptors concerning tumour characteristics and these descriptors can be associated with clinical end points by training statistical or machine learning techniques on the information. This is called supervised machine learning because it deals with labelled data to fit models that can predict on unseen, unlabeled data [4].

In [12], the researchers have used the Gaussian Naïve Bayes (GaussianNB) as a machine-learning approach to predict the probability of the case falling under a particular class which is either guilty or not guilty based on the case score of features.

E. Machine Learning Algorithm: Unsupervised Learning

Also known as unstructured learning, unsupervised learning is one of the subcategories of machine learning where it is very difficult to have a labelled data set or data sets with output data. Unlike supervised, there is no training for the algorithm on how to read the data since it is an unsupervised process. The main purpose of the unsupervised learning is to discover more concealed data relations and structures that are intrinsic. It is widely implemented in exploratory data analysis to recommend structures, groups, and clusters in the data. Popular techniques under unsupervised learning include clustering where methods like K-means clustering, hierarchical clustering, DBSCAN, and Apriori algorithm are encountered. Such types of learning are particularly beneficial in cases when it is impossible to label the data with some valuable information by hand when the amount of data is rather significant. An important advantage of unsupervised learning is that the results that have been obtained can reveal the discovered characteristics and relations that are not apparent initially. Contrary to supervised learning the latter does not involve labelled output data for training the net. Some of the applications of unsupervised learning are for instance in the case of detecting outliers or identifying new patterns and also in cases where it is needed to reduce the dimensions of the data [2].

Clustering and dimensionality reduction are used in this work to identify the properties of Science and Technology Parks (STPs). Unsupervised learning is used for patterns and relationships to be learned without output labels allowing the study of the inherent structures of the data. STPs can be clustered according to some traits or characteristics they possess that make it easy for the researchers to suggest similarities between different parks. The use of unsupervised learning in the research helps the researchers to gain profound understanding of the general state of Spanish STPs, confirm various sorts of STP, and receive quantitative data regarding the certain characteristics. Thus, the use of descriptive and machine learning, such as unsupervised learning, enhances the understanding of the analysed STPs and helps to make more accurate strategic decisions [14].

F. Ensemble Methods

Ensemble is a technique which is used in order to increase prediction accuracy of the classifier. It is indeed a strong meta classification method that improves the weak learners by incorporating strong learners into them. Here in this paper the author has used the Ensemble technique to improve the Various algorithms used for the prediction of heart disease. The reason why different classifiers are combined is to get a higher accuracy as compared to the accuracy given by a single classifier [10].

Ensemble methods like using multiple classifiers to improve the reliability of and decision making as seen with the classification of lung cancer to different histological subtypes. The ensemble methods prove to be more effective in refining the models by increasing the level of correct classification in cases where traditional approaches can be inefficient. Ensemble methods are extremely useful in increasing the stability and the ability of machine learning models to generalise to unseen data, which explains why they are essential in lung cancer subtyping given CT image features [5].

G. Deep Learning Algorithm

Deep Learning (DL) is a Machine Learning subset which encompasses a learning algorithm based on neural networks that has many artificial neurons known as perceptrons. In deep learning, artificial neurons can have multiple inputs and work mathematically like the basic biological neuron producing an output based on a set of mathematical computations. The most rudimentary of its kinds in deep learning is a neural network with at least one processor, one input, and one output, known as perceptron. Three main types of deep learning algorithms are commonly used: MLP, MLP with more than one hidden layer, Convolutional Neural Network or CNN, and Recurrent Neural Network or RNN. During training from the data, deep learning systems imbibe pattern detection and decision-making capabilities that allow the systems to self-decide. Still, DL relies on the mechanisms of the human brain by functioning based on tasks, such as pattern matching, visual recognition, customer segmentation, and discovering pinpoints of a business process inefficiency. Due to such advantages of deep learning models, deep learning can be used across different fields because of its high accuracy, specificity, and malleability whereby it is significantly incorporated in environment science and engineering disciplines. Recent DL structures like Boundary Regulated Network (BR-Net) have been proposed for the tasks such as image feature extraction and classification indicating the effectiveness of DL for processing high resolution remote satellite image data. Some of the methods of DL have been applied in studies that used convolutional neural network (CNN) models for analysing front-mounted camera data on experience of safe driving to develop DL models for actual use [8].

For whole lung and lung lesions evaluation on CT images of NHPs exposed to SARS-CoV-2, deep-learning-based quantification method was employed. A new strategy for the deep-learning-based automated segmentation of the whole lung and lung lesions was developed due to the discrepancy concerning the ground truth. There are several models which were developed by using the CNN on one part of the training data whereas not training the whole training dataset on a single model. To make sure that the network can recognize objects of many different sizes, feature pyramid network (FPN), a kind of CNN that predicts at pyramid levels, was used [7].

H. Tools and Libraries

Tools are defined as software programs or environments that perform or support data analysis, visualisation and modelling tasks, and models deployment. It is also important to note that common data science tools nowadays are equipped with GUI, and their purpose is to provide the means for optimising the common steps in data science processes. Libraries can be defined as sets of pre-built scripts which the users integrate into their projects to accomplish one or another function. Various libraries offer user capabilities for data manipulation, statistics, machine learning, deep learning, and many others, meaning that the developers do not have to create many solutions from scratch.

Python language's relieving and understandable nature ensures that it is able to be used effectively in the implementation of complex algorithms with the BLAST-like algorithm. Indeed, the application of python in npysearch makes it easier for the integration and development of the bioinformatics tools. As for the npysearch, it works with C++ by using pybind11 common for Python and PySide6 which is part of Python's powerful and numerous libraries and tools. This enables good dealing between the Python and C++ modules, improving on the performance and integration of the researched algorithm [15].

XCast, a Python climate forecasting toolkit described in the given project on SourceForge, uses Scikit-Learn for machine learning, which allows the user to apply different algorithms for climate prediction. Through the integration of Scikit-Learn with XCast, real-time deterministic and probabilistic statistical forecasts can be developed, which is explained in a case on South Asian Summer Monsoon Rainfall. This they can achieve using Scikit-Learn where XCast users are able to form a multitude of models using the Machine Learning methodologies, thereby making the climate prediction more precise and accurate. Some of the benefits attained through the application of Scikit-Learn in XCast include enhanced capacity in forecasting especially deterministic and probabilistic forecasts for other hard to predict climate factors like the South Asian Summer Monsoon Rainfall [16].

V. Conclusion

This systematic literature review has reviewed and examined the broad category of data science algorithms, techniques, applications, and tools along with libraries. The analysis of the literature concludes that data science is a critical component of various domains such as health care, environment, and industries. The most common supervised learning models in today's performance include decision trees and neural networks as they work best in tasks with accurate predictions and classifications. Exploratory data analysis and working with 'big data' require algorithms for clustering and for decreasing the dimensions of data. Ensemble methods have demonstrated great potential for the improvement of both the accuracy and the stability of the developed predictive models with the help of the use of several algorithms at a time. The next level deep learning involving CNN and RNN maintains to transform sectors that involve intricate data analysis such as images and voice. The review also includes the description of tools and libraries, which help to implement the discussed algorithms, like XCast Python and Jupyter Notebooks. In addition to facilitating the data science process, these resources also introduce complex approaches to a new level of users. Thus, the field of data science is constantly growing due to the constant appearance of new algorithms and technologies. In this way, utilising these strategies, practitioners can potentially expand the possibilities for analysis and influence within different domains. On that, this review will be considered as a starting point in the current state analysis and a beginning of the search for improvements.

Acknowledgement

This work is supported by the Faculty of Computing, Universiti Teknologi Malaysia.

References

[1] D. V. Rodríguez-Almonacid, J. G. Ramírez-Gil, O. L. Higuera, F. Hernández, and E. Díaz-Almanza, "A Comprehensive Step-by-Step Guide to Using Data Science Tools in the Gestion of Epidemiological and Climatological Data in Rice Production Systems," *Agronomy*, vol. 13, no. 11, 2023, doi: 10.3390/agronomy13112844.

[2] S. Al-Rbeawi, "A Review of Modern Approaches of Digitalization in Oil and Gas Industry," *Upstream Oil Gas Technol.*, vol. 11, 2023, doi: 10.1016/j.upstre.2023.100098.

[3] V. V. Kolisetty and D. S. Rajput, "A review on the significance of machine learning for data analysis in big data," *Jordan. J. Computers Inf. Tech.*, vol. 6, no. 1, pp. 41–57, 2020, doi: 10.5455/jjcit.71-1564729835.

[4] B. Feuerecker *et al.*, "Artificial Intelligence in Oncological Hybrid Imaging," *RoFo Fortschr. Geb. Rontgenstrahlen Bildgeb. Verfahr.*, vol. 195, no. 2, pp. 105–114, 2023, doi: 10.1055/a-1909-7013.

[5] B. Dunn, M. Pierobon, and Q. Wei, "Automated Classification of Lung Cancer Subtypes Using Deep Learning and CT-Scan Based Radiomic Analysis," *Bioeng.*, vol. 10, no. 6, 2023, doi: 10.3390/bioengineering10060690.

[6] V. Steinwandter, D. Borchert, and C. Herwig, "Data science tools and applications on the way to Pharma 4.0," *Drug Discov. Today*, vol. 24, no. 9, pp. 1795–1805, 2019, doi: 10.1016/j.drudis.2019.06.005.

[7] S. M. S. Reza *et al.*, "Deep-Learning-Based Whole-Lung and Lung-Lesion Quantification Despite Inconsistent Ground Truth: Application to Computerized Tomography in SARS-CoV-2 Nonhuman Primate Models," *Acad. Radiol.*, vol. 30, no. 9, pp. 2037–2045, 2023, doi: 10.1016/j.acra.2023.02.027.

[8] J. Valente, J. António, C. Mora, and S. Jardim, "Developments in Image Processing Using Deep

Learning and Reinforcement Learning," *J. Imaging*, vol. 9, no. 10, 2023, doi: 10.3390/jimaging9100207.

[9] Y. Karulkar and S. Jain, "Forecasting Business Persistency at HDFC Life: Smart Insights Powered by Data Analytics," *South Asian J. Bus. Manag. Econ.*, vol. 9, no. 3, pp. 343–358, 2020, doi: 10.1177/2277977920958573.

[10] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Inform. Med. Unlocked*, vol. 16, 2019, doi: 10.1016/j.imu.2019.100203.

[11] Y. Li, S. Carabelli, E. Fadda, D. Manerba, R. Tadei, and O. Terzo, "Machine learning and optimization for production rescheduling in Industry 4.0," *Int J Adv Manuf Technol*, vol. 110, no. 9–10, pp. 2445–2463, 2020, doi: 10.1007/s00170-020-05850-5.

[12] J. Mitchell, S. Mitchell, and C. Mitchell, "Machine learning for determining accurate outcomes in criminal trials," *Law Probab. Risk*, vol. 19, no. 1, pp. 43–65, 2020, doi: 10.1093/lpr/mgaa003.

[13] E. Sajno, S. Bartolotta, C. Tuena, P. Cipresso, E. Pedroli, and G. Riva, "Machine learning in biosignals processing for mental health: A narrative review," *Front. Psychol.*, vol. 13, 2023, doi: 10.3389/fpsyg.2022.1066317.

[14] O. Francés, J. Abreu-Salas, J. Fernández, Y. Gutiérrez, and M. Palomar, "Multidimensional Data Analysis for Enhancing In-Depth Knowledge on the Characteristics of Science and Technology Parks," *Appl. Sci.*, vol. 13, no. 23, 2023, doi: 10.3390/app132312595.

[15] S. Schmid, A. Jeevannavar, T. R. Julian, and M. Tamminen, "Portable BLAST-like algorithm library and its implementations for command line, Python, and R," *PLoS ONE*, vol. 18, no. 11 November, 2023, doi: 10.1371/journal.pone.0289693.

[16] K. J. C. Hall and N. Acharya, "XCast: A python climate forecasting toolkit," *Front. Clim.*, vol. 4, 2022, doi: 10.3389/fclim.2022.953262.