Decision Analytics Journal 11 (2024) 100468

Contents lists available at ScienceDirect

Decision Analytics Journal

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

An intelligent recommender system using machine learning association rules and rough set for disease prediction from incomplete symptom set

Kamakhya Narain Singh [a,b,*], Jibendu Kumar Mantri [a]

[a] Department of Computer Application, Maharaja Sriram Chandra Bhanja Deo University, Baripada, India
[b] School of Computer Applications, KIIT Deemed to be University, Bhubaneswar, India

ARTICLE INFO

Keywords:
Machine learning
Association rules
Classifications
Patient care
Incomplete symptom set
Rough set

ABSTRACT

Digital devices are an integral component of the healthcare sector. With the advancement of modern technology with Artificial Intelligence (AI) and Machine Learning (ML), an automated diagnosis system with promising results is not a difficult task. This study aims to develop a recommender system (RS) for better diagnosis and improvement of patient care by hybridizing machine learning association rules (AR) and rough set theory (RST) to classify acute and life-threatening diseases. Initially data is preprocessed using binary, on-hot vector, and min–max scale to remove the noise. RST is used for feature selection to deal with incompleteness, inconsistency, and vagueness. We have designed an Associated Symptom Selection (ASS) algorithm to extract the mutually associated symptoms which need to be further matched in the existing database for prediction. ASS is especially helpful in detecting neurodevelopmental type diseases because the symptoms are usually not detectable by standard tests, and observations of behavioral expressions do general testing. The experiment is carried out using six popular ML classifiers such as AR, Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Linear Support Vector Machine (LSVM), and Naive Bayes (NB) on a publicly available datasets. Performance was compared among different classifiers regarding the accuracy, precision, recall, F1-score, and J-Score value. The experimental result shows that AR performs better on clinical data with an accuracy of 94.40%, precision of 90.73%, recall of 94.45%, F1-score of 92.55%, and J-score of 95.14% and on autism with 98.7% accuracy, 98% precision, 97.8% recall, 97.9% F1-score, and 97.12% J-score respectively.

1. Introduction

In the era of digital technology, digital devices have become an integral component of the healthcare sector. Clinicians prefer healthcare data in electronic formats. Digital technology has greatly improved clinical efficiency concerning standards of patient care and revolutionized healthcare practice. It also resulted in the centralized storage of patient medical records in electronic health records (EHR) [1]. The digital revolution has significantly enhanced the overall experience of both clinicians and patients [2]. With the advancement of modern technology with Artificial Intelligence (AI) and Machine Learning (ML), an automated diagnosis system with promising results is not a difficult task. But, there is a high possibility of getting incomplete data during online interaction. To deal with vague information, we propose a recommender system (RS) for better diagnosis and improvement of patient care. The novelty of this work lies in the hybridization of machine learning association rules and a rough set that needs only one symptom to identify the disease and generate case-specific advice.

RS is an advanced technology which is designed to help clinicians in making diagnostic decisions regarding individual patients. These systems are efficient intelligence technologies that produce particular instance recommendations based on multiple parts of health records [3]. RS is an active knowledge system, which uses incomplete set of patient data to generate case-specific advice. Our work is primarily based on disease detection with incomplete or insufficient symptom sets with a special focus on rural areas and the online interaction of patients using association rules machine learning techniques. The diagnosis of autism is especially difficult for a doctor in rural areas because the parents are not always able to express all the symptoms of their child. In addition, the more experienced specialists are also not very much accessible in rural areas [4]. So mostly, health assistants or less experienced doctors are available to deal with such patients. Also, there is a high possibility of getting inconsistent data during online interaction. Furthermore, It addresses the issue of a shortage of well qualified professionals and provides health care services to the large population.

There are many existing systems that have already been developed which have used complete symptom sets or partial symptom sets for the diagnosis and identification of Autism, but they often lead to an improper identification of Autism [5]. In order to address inconsistent

* Corresponding author at: Department of Computer Application, Maharaja Sriram Chandra Bhanja Deo University, Baripada, India.
E-mail addresses: kamakhya.vphcu@gmail.com (K.N. Singh), jkmantri@gmail.com (J.K. Mantri).

https://doi.org/10.1016/j.dajour.2024.100468
Received 29 December 2023; Received in revised form 8 April 2024; Accepted 17 April 2024
Available online 20 April 2024

Table 1
List of abbreviations.

Full form	Abbreviation	Full form	Abbreviation
Artificial Intelligence	AI	Linear Support Vector Machine	LSVM
Machine Learning	ML	Naive Bayes	NB
Recommender System	RS	Electronic Health Records	EHR
Association Rules	AR	Fuzzy Rough Set	FRS
Rough Set Theory	RST	Clinical Knowledge Base	CKB
Associated Symptom Selection	ASS	Random Forest	RF
Decision Tree	DT	K-Nearest Neighbors	KNN
Fuzzy Rough Set	FRS		

or incomplete data in clinical decisions aims to develop a framework for better diagnosis and improvement of patient care. The proposed framework uses an incomplete symptom set which is either only one or more sets of patient data to generate case-specific advice. Table 1 presents the list of abbreviations.

There is a need for a clinical decision support system framework to analyze patient-related data and enable professionals, patients, or other experts with advanced knowledge to strengthen patient care quality [6]. In this work, we proposed a framework based on the hybridization of rough set and association rules to classify the diseases. A rough set is used for feature selection and machine learning association rules are used for disease prediction from a preliminary symptom only. A variety of feature selection models have been proposed to deal with the problem of data uncertainty, inconsistency, and vagueness [7]. Pawlak et al. [8] developed the idea of rough sets, which has proven to be one of the most fruitful models to date. There are three basic zones of human knowledge: inside, outside, and boundary [9]. The notion of rough sets is founded on this philosophy. Rough sets are now used in a considerably wider range of industries than they were in the past, including computer networks, image processing, data mining, and even medicine. Mohamed, E. K. et al. [10] explored rough set theory from a topological perspective. Rough set and topology were then applied to a wide range of incomplete and ambiguous objects, as well as to more key structures, such as [9]. It has also been suggested that rough set models can be extended utilizing a variety of relationships such as tolerance and similarity, and dominance connections. Expansion (generalizations) of the positive zone is the primary goal of these extensions (generalizations). Deleting items in the negative region is a similar goal. Another essential term in rough set theory is the boundary region and accuracy measure, which are established using upper and lower approximations. Accuracy measures demonstrate how vast the boundary area of a set is, but they do not tell us anything about its structure. It is possible to gain some insight into the structure of the border region by using approximations, however, this knowledge is lacking. Four strategies for reducing the border region were presented in [11].

Association Rules play a major role in the health sector as they help to conduct intelligent diagnoses and extract valuable information to build knowledge bases automatically [12]. The Apriori algorithm is used to form association rules in [13] to diabetic patients database. In this research, the technique of identifying new, unexpected, and hidden patterns in clinical databases is considered. A clinical database maintains the chronic and acute disease symptoms of each patient. The association rule machine learning technique is used to extract the relevant symptoms from the EHR and match them with symptom set in the clinical database.

The contributions of this research are as follows:

- An efficient and new diagnosis framework is developed based on RST and Association Rules machine learning to diagnose disease in real-time health care applications.
- Supporting doctors with a new model to help them diagnose neurodevelopmental diseases using intelligent approaches.
- The proposed model is cost-effective and useful for various kinds of medical data in real life.
- The framework is useful in rural areas and for inexperienced or less experienced doctors.

The remaining parts are divided into different sections. The relevant work is presented in Section 2; the proposed framework and approach are presented in Section 3. Discuss implementation in Sections 4 and 5 concludes with a summary.

2. Related work

In this section, we will first describe knowledge representation approaches, which are utilized in various medical expert systems, and then we will discuss several medical applications where either rough set or association rules have been used to represent knowledge. After that, we do a comparison examination of the most prevalent knowledge representation methods.

The rough sets method [14] has been proposed as an interesting tool for discovering information from data. Using a rough set to mine data is a good option because the data is qualitative, making it difficult to examine using normal statistical methods. Learning from examples, extracting rules from a data set of interest, and discovering data regularities are all part of the rough set-based approach [15].

There are numerous methods for storing and distributing knowledge in the literature [16,17]. As a starting point, production rules are widely employed in the medical field [18,19]. Many medical expert systems, including MYCIN, PUFF, UMLS-based CDSS, CMDS, and others [20], have employed rule-based knowledge representation methodologies. Some probabilistic approaches have been used to deal with uncertainty in medical knowledge when implementing production standards, fuzzy logic and Bayesian networks are often used in clinical uncertainty resolution. When dealing with data that is ambiguous, incomplete, or unsure, fuzzy rough set (FRS) is a useful tool [21]. It has been utilized in a variety of domains, including decision making, feature selection, medical diagnosis, intrusion detection, and image retrieval, among others. With the introduction of a reflexive fuzzy b-neighborhood operator to deal with the problem of misclassifications and perturbations, B. Sun, W. Ma, et al. [22] suggested a covering-based variable precision fuzzy rough set. There are two approaches to inducing the granules from fuzzy complicators and implicatory: the first is to use fuzzy rough sets, and the second is to use granular variable precision rough sets [23]. FRS models based on implicatory and conjuncture were developed in [24] to allow for the investigation of more complex interactions and structures among items. Therefore, the theoretical background of this study is based on FRS, which is an effective method for learning knowledge of complex data.

The rough set theory was used by C. Zhao et al. [25] to develop a new classification approach based on a hierarchical granule structure. To identify the categorization rules quickly, the hierarchical granulation structure was used. Knowledge reduction was achieved by applying the upper/lower approximations of rough sets in the classification rules. To demonstrate the method's efficacy, a simulation was run on the WBC data set. In the simulation, it was shown that the proposed classification approach produced minimal classification rules and made the examination of information systems simple. The researchers, on the other hand, failed to explore the issue of attribute reduction. A

bacteremia diagnosis program developed by Y Vang et al. [26] uses production rules for knowledge representation. Using this knowledge representation technique, each rule is composed of two parts: a premise (which is a Boolean combination of predicate functions) and an action component. Each rule is also assigned a confidence factor, which indicates how strong the rule is, to reflect its strength. B Sun et al. [27] introduced a medical expert system, ONCOCIN, to facilitate the analysis of the temporal trends of various symptoms related to cancer patients. MYCIN's knowledge representation technique allows only backward-chaining of rules. Knowledge was represented in contexts, parameters, and rules in this system. For cancer protocol management, contexts refer to medical entities that need to be familiarized with static domain knowledge, whilst parameters refer to the clinical characteristics that must be considered. The logical grouping of similar parameters is accomplished through the usage of data blocks. Rules in ONCOCIN are activated while performing forward or backward reasoning, as with any rule-based medical expert system. As a last phase, control blocks describe the high-level execution steps of an expert system task. Consideration of these data structures allows for a modular approach to knowledge representation, making ONCOCIN more efficient during consultations.

An increasing amount of data in a quickly evolving medical field makes it difficult to analyze it owing to errors or discrepancies in the data [28]. Many medical applications rely on the use of rough sets to deal with these kinds of clinical uncertainty [29]. The repercussions of Egyptian newborn jaundice, such as neurological dysfunction and kernicterus, have been studied by V. Christou et al. [30], who have presented a preliminary set-based paradigm for early management and anticipation. They created a weighted information table, identified relevant qualities from it, decreased the number of weighted attributes, and then came up with a set of diagnostic guidelines for neonatal jaundice that may be used in clinical practise. H. Zhao et al. [31] have developed an ECG categorization decision support system based on rule-based rough-set decision support. Various forms of inconsistencies, such as transcription errors in ECG signals, subjective calculations of attribute values, and lack of information, have been dealt with here using crude set theory. This decision support system is aimed to deliver low-cost, more robust, and patient outcomes. Concept lattice [32], which is based on a lattice structure, can also be used to represent knowledge. Rules, therapeutic decisions, and patient scenarios can all benefit from this method of visual representation. Concept lattice and rough set theory have been successfully combined in certain significant studies [33]. Rough set theory and lattice theory have been merged in several theories [34], however, the combined theory has very few applications in the medical field.

A comparative analysis with considering fuzzy environments is conducted in [35] to develop a model to deal with healthcare supplier problems such as quality, responsiveness, and reliability. It aids in handling varying decision makers' judgments and uncertainty-associated problems. However, it is tested on a low-dimensional healthcare supplier selection problem by considering six evaluation criteria only. To improve the model's performance, similarity measurement is calculated in embedded space using Optical Character Recognition and explicit incorporation of semantic information. The experiment is carried out on different datasets [36]. A comprehensive analysis of Convolutional neural networks and unsupervised pre-trained learning is conducted to propose a hybrid method for the recognition, classification, and detection of images, speeches, texts, and videos. The weight optimization, computational analysis and selection of appropriate hyper-parameters are the main focuses of the proposed approach. The Proposed model outperforms existing state_of_the_arts and is tested on various publicly available datasets [37–40].

To discover the uncertain relationship between data items and clusters and reflect data items into core and fringe regions, three way clustering approach based on neighborhood is applied [41]. It outperforms existing models on real world datasets but the computational load can be reduced by selecting numerous samples at each iteration. A method entitled Mine-First association rule mining using Association rules is presented in [42] to identify Covid patients. A rule base is prepared to explore the association and find frequent patterns between different classes while considering a distributed environment. The experiment is carried out on the COVID-19 symptom checker dataset to generalize the rules for the identification of Covid 19 variants. Table 2 represents the advantages and drawbacks of existing methods and the comparison of the performance of our proposed model with best existing classifiers is given in Table 11.

3. Proposed methodology

This section introduces the proposed framework and methodologies. The proposed framework classifies chronic and acute diseases based on patient attributes using a hybridization of RST and AR. Our proposed methodology comprises four steps. In the first step, the collection of data is designed to collect data from different web resources, including Google Forms and social media platforms. The collected data is stored in a clinical database or data set in the form of CSV and some data like discharge summaries are in the form of text data. In the second step, preprocessing techniques such as binary, on-hot vector, and min–max scale are applied to remove the noise. Then a feature selection technique is used in the third step to remove highly redundant features and reduce the dimension of the datasets. Selected features are stored in the EHR of a concerned patient to analyze the disease. Lastly, in the fourth step, machine learning AR is used to extract the mutually associated symptoms which need to be further tested in the existing database for accurate disease identification. If an incomplete symptom set is provided then finding mutually associated symptoms using AR from any preliminary symptom associated with disease. Only one preliminary symptom is needed to predict the disease. The initial symptom takes out the next probable symptoms from the EHR using AR and the highest possible symptom will be selected to confirm from the patient. After each confirmation, the symptom set will be matched with the clinical knowledge base (CKB). If any existing disease is matched with the same set of symptoms, then the process is stopped, else it tries to get further targeted symptoms which are mutually dependent on the previously selected symptoms. The process continues until we get the confirmed disease from CKB, else the process shows disease not being detected. An Intelligent recommender system is represented in Fig. 1.

We have designed an Associated Symptom Selection (ASS) algorithm which needs only a preliminary symptom and after 'n' number of scans of the EHR, it finds the occurrences of the 'n' co-related symptoms of the preliminary symptom. This assists the patient to recall what other symptoms he has experienced. This also helps the doctors to identify the disease. Before explaining the ASS algorithm used in this work, the feature selection technique needs to be highlighted because the feature selection process plays a vital role in reducing the dimension of the large data and selecting the highly relevant features. We have applied a rough set approach to reduce the dimension and deal with incompleteness, inconsistency, and vagueness.

A rough set-based concept is introduced in [49] which suggests taking a decision more perfectly and finding hidden patterns in historical data. A rough set comprises information system, indiscernibility, set approximation, reducts, core, and rough membership which can be defined as follows:

Information System represents the pair of a non-empty finite set of objects and attributes (U, A), represented as p:UxA→V, where V represents a value of attributes. Table 3, represents information of patient.

Here, V is a set of all attribute values (0/1/2) and Information Function: p:UxA→V.

Decision System: It includes decisional attributes along with conditional attributes, represented as DS: $T = (U, A \cup \{d\})$, where $d \notin A$. Based on this information, a doctor may predict a particular disease. Decisional attributes are included in Table 4.

Table 2
The advantages and disadvantages of current knowledge representation methods.

Author	Existing schemes for knowledge representation	The advantages of the approach in medical expert systems	Drawbacks in the medical expert systems of the scheme
[20]	Rule based knowledge representation scheme	For well-understood medical domains, this method delivers representational simplicity and expressional effectiveness. Individual rules can be readily modified with the addition of new parameters or the removal or modification of current parameters because of their modular structure. To handle uncertainty in knowledge, the standard production rules can be simply adapted to include characteristics such as the 'certainty factor' described in MYCIN.	When it comes to intricate clinical linkages and the underlying causes of disease, the model's representational inadequacy leaves scope for development. A knowledge base with redundant, non-optimal, and inconsistent data can make inference difficult and time consuming.
[25]	Rough set-based representation	A large database may easily and quickly be mined for valuable information using this method, which also yields optimal and consistent decision rules. Against each rule, some numerical measurements like support, strength, certainty factor, and coverage factor are calculated. Ultimately, these measures aid in the inference process for the selection of acceptable rules.	This approach may not always provide a complete knowledge base, in terms of being able to store detailed and up-to-date information, because the knowledge is mostly derived from patient data. In addition, depending solely on patient data may lead to inaccurate conclusions. Rough set theory works well on discrete datasets, but its use in medical domains may be limited because most clinical datasets are real-valued.
[27]	Semantic network-based representation	More unstructured domain knowledge (e.g., dependencies among clinical parameters) can be captured in multiple medical domains with the scheme's improved visualization and representational adequacy. The framework can be simply expanded. It is possible to assure the knowledge base's scalability, robustness, and completeness if the design is considered period out.	A complex medical domain may need a network topology that is difficult to grasp, resulting in increased time and space complexity. There may be a major delay in the clinical decision-making process as a result. Time-consuming construction and modification of a knowledge base in the medical field limits the scheme's usefulness.
[28]	Conceptual lattice-based representation	As a result, the system makes it easier to detect common attributes in medical domains with a limited number of objects and attributes. The reasoning process is time-efficient since the relevant clinical data can be found at any moment, avoiding an exhaustive and sequential search through the knowledge base.	There are many objects and characteristics in a medical domain with a significant number of idea lattices, making it more difficult to store the structure and more time consuming to retrieve information. With regards to how much a given property influences a specific item, there is no way to include this information in the structure of the system.
[43]	Fuzzy based approach	This novel approach is introduced to deal with uncertainty in multi-criteria decision making using fuzzy entropy.	But, it does not perform well on too few or too many criteria. Additionally, it can lead to inaccurate decisions based on inconsistent or incomplete data.
[44]	Multi-attribute decision-making approach	Invariant feature selection is proposed for dimensionality reduction in a more rigorous way. It performs well for ranking, selection, and prioritization problems.	It blocks variables as fixed factors rather than random factors.
[45]	Quartic fuzzy sets approach	It helps to deal with uncertainty in real life multi criteria decision making situations. It also works well to handle ambiguity and intermediate information.	It is computationally expensive to handle high dimensional data.
[46]	Q-RUNG fuzzy set approach	A similarity measure approach based on Hausdorff distance. It is efficient to deal with an uncertain environment and provide decision makers to express information in an effective manner. It works well in pattern recognition, classification, and multi criteria decision making.	Different degrees of uncertainty are not considered during evaluation and computational complexity is very expensive, specifically for high dimensional data.
[47]	Decision Tree approach	A machine learning-based automated diagnosis system based on DT is proposed to diagnose diabetic patients. The model is tested on a clinical dataset and validated using various validity matrices. It performs better than existing systems in terms of accuracy and computational time.	The Proposed model did not consider the overfitting problem.
[48]	Recommender Systems	A survey of existing recommender systems using AI and various ML approaches, including fuzzy techniques, transfer learning, genetic algorithms, evolutionary algorithms, neural networks, deep learning, and active learning is conducted. It supports the researchers to understand and find new trends of RS.	These recommender systems did not deal with incomplete, inconsistent, and vague information on a large scale.

Indiscernibility I(B) is a binary relation on U for x, y∈U iff p(x,a) = p(y,a), for all a ∈ B. A dispensable element is a redundant element in the table and can be identified as if I(B) = I(B- {a}) else 'a' is called indispensable in relation B.

We found the I(B) of all subsets of attributes and observed symptoms that are most often of interest have the same value as the decision attribute. Table 5, represents the indiscernibility relationship of a subset of attributes.

Here, A = (S1, S2,S3, S4, S5), I(A) = {{P1, P2},{P3, P7, P10},{P4}, {P5},{ P6},{P8},{P9}}

From the above indiscernibility relation, we observe that I(B) = I(B-{S1}), I(B) = I(B-{S2}), I(B) = I(B-{S3}), I(B) = I(B-{S4}), I(B) = I(B-{S1,S2}), I(B) = I(B-{S1,S4}), I(B) = I(B-{S2,S4}) and I(B) = I(B-{S3,S4}).

Thus, S1, S2, S3, S4, (S1, S2), (S1, S4), (S2, S4) and (S3, S4) are dispensable in relation and all other subsets are indispensable in relation. This means classification defined by set of 5 equivalence relation S1, S2, S3, S4, S5 is same as the classification defined by relation (S1, S2, S5), (S1, S3, S5), (S2, S3, S5) or (S3, S4, S5).

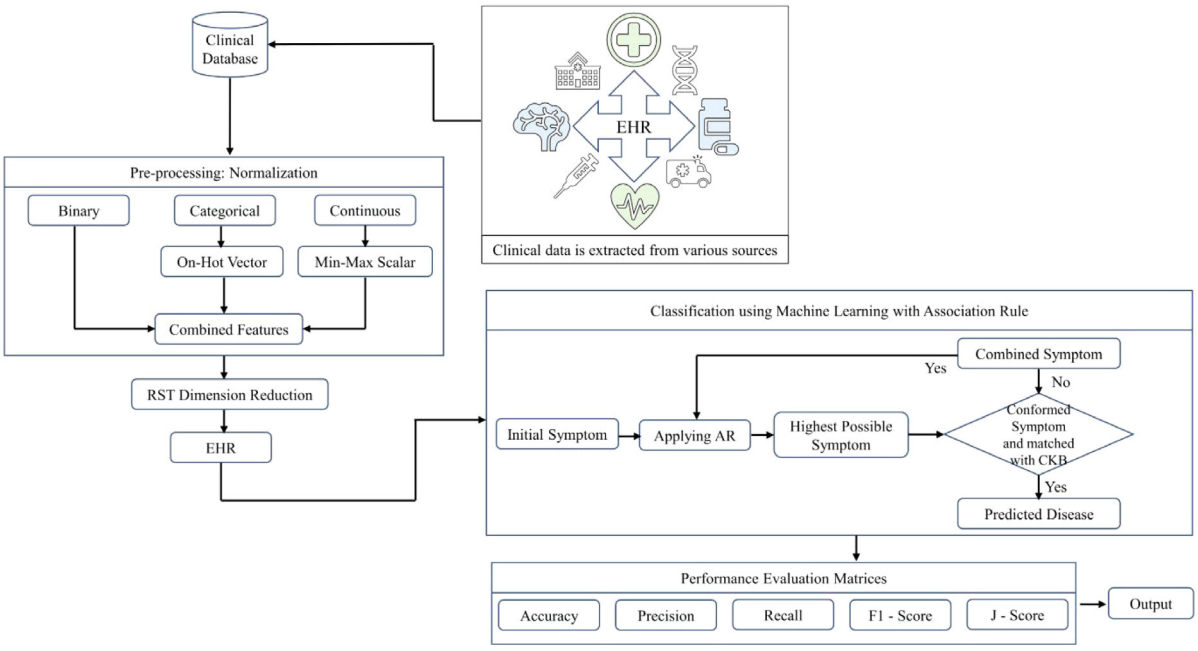

Fig. 1. Intelligent recommender system framework.

Table 3
Information table.

Patient	Attributes				
	S1	S2	S3	S4	S5
P1	1	2	0	1	1
P2	1	2	0	1	1
P3	2	0	0	1	0
P4	0	0	1	2	1
P5	2	1	0	2	1
P6	0	0	1	2	2
P7	2	0	0	1	0
P8	0	1	2	2	1
P9	2	1	0	2	2
P10	2	0	0	1	0

Table 4
Decision table.

	Attributes					Decision
	S1	S2	S3	S4	S5	Disease
P1	1	2	0	1	1	D1
P2	1	2	0	1	1	D1
P3	2	0	0	1	0	D2
P4	0	0	1	2	1	D3
P5	2	1	0	2	1	D4
P6	0	0	1	2	2	D5
P7	2	0	0	1	0	D6
P8	0	1	2	2	1	D7
P9	2	1	0	2	2	D8
P10	2	0	0	1	0	D2

Reduct: Any subset B' of B is called reduct if B' is independent and I(B') = I(B). Here, Reduct1 = {S1, S2, S5}, Reduct2 = {S1, S3, S5}, Reduct3 = {S2, S3, S5}, and Reduct4 = {S3, S4, S5}.

Core: Set of indispensable elements of B is Core of B and calculated as Core(B) = ∩Red(B), where Red(B) is the set of reducts of B. {S1, S2, S5} ∩ {S1, S3, S5} ∩ {S2, S3, S5} ∩ {S3, S4, S5}= S5

Hence, S5 is an important attribute in a decision system which cannot be eliminated. Thus, reduct and core are used to select the features or induce the rules to take further decision.

Set Approximations: In set approximations, we roughly approximate X, from the information contained in B as B-lower, B-upper, boundary region, and outside boundary region represented as follows: B X = {x|[x]B ⊆X}, $\bar{B}$X = {x|[x]B∩X ≠ Φ}, BN(X) = $\bar{B}$X − $\underline{B}$ X and OB(X) = U − $\bar{B}$X, where X ⊆U, B-lower is set of elements of U which surely classify the element of X, B-upper is set of elements which possibly classify, in boundary region, elements may or may not belong to set X and in out-side region data certainly does not belong to set X. The set approximations and regions of set X using RST are presented in Fig. 2.

Here, patients that are definitely identified with their disease are represented as $\underline{B}$ X = { P1, P2, P4, P5, P6, P8, P9 }, patients that are possibly identified with their disease are presented as $\bar{B}$X = { P1, P2, P3, P4, P5, P6, P7, P8, P9, P10}, patients that are identified in the boundary region are represented as BN(X) = $\bar{B}$X − $\underline{B}$ X = { P3, P7, P10}. In the boundary region, the disease of the patient is not identified since they possess the same symptoms but differ in decision. Patients are identified in the outside region, represented as OB(X) = U − $\bar{B}$X = NULL.

Rough Membership is used to validate the accuracy of decisions, defined as $\mu_x^R : U \rightarrow \langle 0,1 \rangle$, where, $\mu_X^R(x) = \frac{|X \cap R(x)|}{|R(x)|}$ and |X| denotes the cardinality of X.

The membership value of lower approximation region's element will be always 1, the membership value of upper approximation set element will be greater than 0, the membership value of boundary region element will lie in between 0 to 1 and the membership value of out side boundary region element will be 0.

Standard clinical and Autism datasets are used in this experiment and the performance of the proposed technique is evaluated with respect to accuracy, precision, recall, F1-score, and J-score. Finally, the outcome in terms of the recommendation will be sent to the patient.

4. Result

In this section, the experimental setup followed by algorithms and results is discussed in detail. Publicly available recent standard clinical high dimensional and binary to multi class datasets are downloaded from Kaggle. (data source: https://www.kaggle.com/code/prashfio/clinical-decision-support-system/data, https://www.kaggle.com/datas

Table 5
Indiscernibility relation of symptoms.

Sr. No	Attributes	Indiscernibility
1	S1	{{P1, P2}, {P3, P5, P7, P9, P10},{P4, P6, P8}}
2	S2	{{P1, P2}, {P3, P4, P6, P7, P10},{P5, P8, P9}}
3	S3	{{P1, P2, P3, P5, P7, P9, P10},{P4, P6},{P8}}
4	S4	{{P1, P2, P3, P7, P10}, {P4, P5, P6, P8, P9 }}
5	S5	{{P1, P2, P4, P5, P8}, {P3, P7, P10 }, {P6, P9}}
6	(S1, S2)	{{P1, P2},{P3, P7, P10},{P4, P6, },{P5, P9},{P8}}
7	(S1, S3)	{{P1, P2}, {P3, P5, P7, P9, P10},{P4, P6, P8}}
8	(S1, S4)	{{P1, P2}, {P3, P5, P7, P9, P10},{P4, P6, P8}}
9	(S1, S5)	{{P1, P2}, {P3, P5, P7, P9, P10},{P4, P6, P8}}
10	(S2, S3)	{{P1, P2}, {P3, P7, P10},{P4, P6},{P5, P9}, {P8}}
11	(S2, S4)	{{P1, P2}, {P3, P7, P10},{P4, P6},{P5, P8, P9}}
12	(S2, S5)	{{P1, P2}, {P3, P7, P10},{P4},{P5, P8}, {P6},{P9}}
13	(S3, S4)	{{P1, P2, P3, P7, P10},{P4, P6},{P5, P9}, {P8}}
14	(S3, S5)	{{P1, P2, P5}, {P3, P7, P10},{P4}, {P6},{P8}, {P9}}
15	(S4, S5)	{{P1, P2}, {P3, P7, P10},{P4, P5, P8}, {P6, P9}}
16	(S1, S2, S3)	{{P1, P2},{P3, P7, P10},{P4, P6 }, {P5, P9},{P8}}
17	(S1, S2, S4)	{{P1, P2},{P3, P7, P10},{P4, P6 }, {P5, P9},{P8}}
18	(S1, S2, S5)	{{P1, P2},{P3, P7, P10},{P4},{P5},{P6 },{P8},{P9}}
19	(S1, S3, S4)	{{P1, P2},{P3, P7, P10},{P4, P6},{P5, P9},{P8 }}
20	(S1, S3, S5)	{{P1, P2},{P3, P7, P10},{P4},{P5},{P6},{P8},{P9 }}
21	(S1, S4, S5)	{{P1, P2},{P3, P7, P10},{P4,P8},{P5},{P6},{P9 }}
22	(S2, S3, S4)	{{P1, P2},{P3, P7, P10},{P4, P6},{P5, P9},{P8}}
23	(S2, S3, S5)	{{P1, P2},{P3, P7, P10},{P4},{P5},{P6},{P8},{P9}}
24	(S2, S4, S5)	{{P1, P2},{P3, P7, P10},{P4},{P5, P8},{P6},{P9}}
25	(S3, S4, S5)	{{P1, P2},{P3, P7, P10},{P4},{P5},{P6},{P8},{P9}}
26	(S1, S2, S3, S4)	{{P1, P2},{P3, P7, P10},{P4, P6 },{P5, P9},{P8}}
27	(S1, S2, S3, S5)	{{P1, P2},{P3, P7, P10},{P4},{P5 },{P6 }, {P8},{P9}}
28	(S1, S2, S4, S5)	{{P1, P2},{P3, P7, P10},{P4},{P5 },{P6 }, {P8},{P9}}
29	(S1, S3, S4, S5)	{{P1, P2},{P3, P7, P10},{P4},{P5 },{P6 }, {P8},{P9}}
30	(S2, S3, S4, S5)	{{P1, P2},{P3, P7, P10},{P4},{P5 },{P6 }, {P8},{P9}}
31	(S1, S2, S3, S4, S5)	{{P1, P2},{P3, P7, P10},{P4},{P5},{P6},{P8},{P9}}

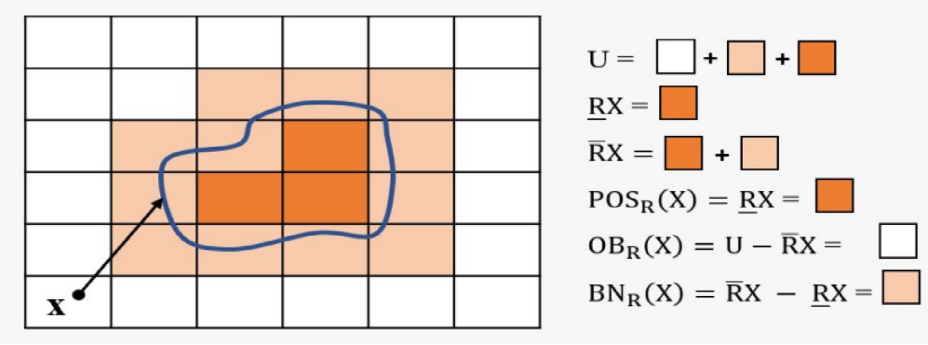

Fig. 2. The approximation regions of set X using RST.

ets/andrewmvd/autism-screening-on-adults) [50,51]. Table 5 represents the description of datasets, and columns 1–5 represent name of the data set, objects (U), conditional features (F), decision classes (C), and missing value (V) respectively. There were 4920 rows and 133 columns in the clinical dataset, row represents patient and column represents patient serial number, symptoms, and disease. Column 1 contains number of object (U), columns 2–132 (F) of the table represents conditional attributes (131 symptoms) and last column (C) contains the decisional attribute (41 diseases). Attributes values are either 0 or 1, 0 represents no and 1 represents yes. There are total 41 types of acute and chronic diseases included, which does not require any test report and 120 patients experienced the same disease. In Autism dataset, there were 704 rows and 21 columns, columns 1–20 (F) represents the conditional attributes (binary, categorical and continuous value) and last column contains the decision class (Yes/No). We have used the standard imputation method (mean/median) approach to replace the missing value of corresponding column. K-NN imputation technique is used to find the k-closest neighbor of missing value. Since most of the data is categorical, we remove the rows with an excessive number of missing values. We have used python 3.10, Pycharm 2019.3.2x 64 and window 11 for implementation. Represented attributes of datasets through Table 6.

4.1. Proposed methods

Data cleaning techniques such as binary, on-hot vector and min–max scale are applied for pre-processing to remove unnecessary comments, tags, noise elements, duplicate entries, etc from the datasets and a rough set approach is applied for feature selection to reduce the dimension of the data and deal with data inconsistency, uncertainty, and vagueness. Table 7 represents the features extracted by the rough set theory on clinical datasets.

For the classification of the disease, machine learning association rules, DT, RF, KNN, LSVM, and NB are used. We compared the performance of classifiers in terms of various validity matrices such as accuracy, precision, recall, F1-score, and J-score and observed that the association rule is the best classifier for these datasets.

Table 6
Attributes of datasets.

Dataset	Object (U)	Conditional Feature (F)	Decisional Feature/Class (C)	Missing Value	Number of missing value
Clinical Data for Disease Prediction	4920	131	41	No	NA
Autism	704	20	2	Yes	192

Table 7
Details of datasets.

Data set	Object (U)	Total Feature in the data set	Features selected by RST	Decisional class	Number of missing value
Clinical Data for Disease Prediction	4920	131	121	41	NA
Autism	704	21	10	2	192

4.2. Association rules for symptoms extraction

The AR is applied to extract the associated symptoms from the EHR which are associated with initial symptoms and help a clinician to predict diseases. The AR can be defined as-

$X_i \rightarrow Y$, where X_i represents the initial symptom of a patient that is identified from an unknown set S and Y represents the target symptom to be extracted from an existing symptom set.

Y resides in the clinical knowledge base, which is previously experienced by various patients. For finding Y, the rule is applied to measure the interestingness of a particular period. Two key terms of interestingness are Support and Confidence [52].

The probability of Support for X_i is

Support (X_i) = Number of times X_i matched during scanning the database

Confidence (Conf) of rule Xi→Y

Probability of Conf(Xi→Y) = Number of times X_i matched with Y/ Number of times X_i matched throughout the database.

The two symptoms are mutually associated on other symptom which may form a symptom set to identify disease such as S_1 and S_5 form symptom set and S_3 and S_5 is a symptom set of a disease, here two symptoms S_1 and S_3 are mutually dependent on S_5. The symptom set will be {S_1, S_3, S_5 }.

4.3. Associated symptom selection (ASS) algorithm

Input: EHR of patient (P_i), disease (D_i), symptom (S_i) (i=1 to n), X: Preliminary

symptom, Y: Targeted Symptom.

Output: Extraction of symptom set and identify disease (D_i)

1. Find Confidence of (X→Si) from P_1 to Pn// Apply AR

2. Repeat step 3 to 5 Until X matches with Di (i = 1 to n) or entire database is scanned

3. Find Y= Si, highest confidence of (X→Si)//confirm highest confidence from

patient, if not then 2nd highest confidence and so on

4. X = {X, Y}

5. If X matched with D_i, then

6. Confirm disease D_i

7. Break

// End of Loop

8. Disease not found

9. End If

The association rule is used to extract the relevant symptom from the symptom set and match it in the clinical database.

For example, the patient symptoms sets for the diseases are represented as follow-

Let d_i, the various types of disease are d_1, d_2, d_3 - - d_n and symptoms s_i are s_1, s_2, s_3 - - s_n and n types of symptom sets are:

$d_1 = \{s_3, s_7, s_1, s_5\}$
$d_2 = \{s_1, s_{15}, s_{13}, s_{11}, s_7\}$
$d_3 = \{s_3, s_9, s_{10}, s_{1,}s_8\}$
$d_4 = \{s_2, s_7, s_{10}, s_5\}$
$d_5 = \{s_4, s_{20}, s_{10}, s_1, s_9\}$
……..
$d_n = \{s_i, s_i, s_i, s_{i,}s_i\}$

Let the EHR contain 10 patient's data p_1 to p_{10} (using for symptom extraction)

$p_1 = \{s_2, s_9, s_8, s_7\}$ d_{10}
$p_2 = \{s_5, s_{10}, s_7, s_2\}$ d_8
$p_3 = \{s_1, s_9, s_3, s_{10}, s_8\}$ d_9
$p_4 = \{s_5, s_{11}, s_3, s_1\}$ d_6
$p_5 = \{s_7, s_5, s_{10}, s_2\}$ d_3
$p_6 = \{s_1, s_3, s_8, s_{10}, s_9\}$ d_4
$p_7 = \{s_3, s_4, s_9, s_5\}$ d_2
$p_8 = \{s_3, s_1, s_{11}, s_5\}$ d_6
$p_9 = \{s_1, s_3, s_6, s_5\}$ d_5
$p_{10} = \{s_5, s_7, s_8, s_6\}$ d_1

Suppose the patient is expressing only one symptom 's_3' about the disease.

According to the above given data set, the support of 's_3' is 6.

By applying AR we observe,

s_3 experienced with s_1 5 times
Confidence of $s_3 \rightarrow s_1 = 5/6 = .8$ similarly
s_3 experienced with s_5 4 times
Confidence of $s_3 \rightarrow s_5 = 4/6 = .6$
s_3 experienced with s_9 3 times
Confidence of $s_3 \rightarrow s_9 = 3/6 = .5$
s_3 experienced with s_{10} 2 times
Confidence of $s_3 \rightarrow s_{10} = 2/6 = .3$
s_3 experienced with s_8 2 times
Confidence of $s_3 \rightarrow s_8 = 2/6 = .3$
s_3 experienced with s_{11} 2 times
Confidence of $s_3 \rightarrow s_{11} = 2/6 = .3$
s_3 experienced with s_6 1 times
Confidence of $s_3 \rightarrow s_6 = 1/6 = .1$
s_3 experienced with s_4 1 times
Confidence of $s_3 \rightarrow s_4 = 1/6 = .1$

The maximum confidence value confirms using top down approach, here

Maximum confidence value is $s_3 \rightarrow s_1 = 0.8$

Now about symptom 's_1' will be confirmed by the patient, if patient confirms 's_1' then 's_1' and 's_3' together are treated as initial symptoms. If 's_1' will not be confirmed by the patient, then either the same priority symptom (top down approach) or the next highest occurrence symptom i.e. 's_{10}' symptom will be asked to the patient for confirmation. If there is no confirmation received for the asked symptom from the patient,

Table 8
Extraction of symptoms based on maximum occurrence of symptom.

Initial symptom	Possibility of target symptoms	Confidence	Selected symptom
S_3	S_1	0.8	S_3S_1
	S_9	0.5	
	S_{10}	0.3	
	S_8	0.3	
	S_5	0.6	
	S_{11}	0.3	
	S_4	0.1	
	S_6	0.1	
S_3S_1	S_9	0.4	$S_3S_1S_5$
	S_{10}	0.4	
	S_8	0.4	
	S_5	0.6	
	S_{11}	0.4	
	S_6	0.2	
$S_3S_1S_5$	S_{11}	0.6	$S_3S_1S_5S_{11}$
	S_6	0.3	

Table 9
Extraction of symptoms from any random symptom of the symptom set.

Initial symptom	Possibility of target symptoms	Confidence	Selected symptom
S_1	S_9	0.4	S_1S_3
	S_3	1.0	
	S_{10}	0.4	
	S_8	0.4	
	S_{11}	0.4	
	S_5	0.6	
	S_6	0.2	
S_1S_3	S_9	0.4	$S_1S_3S_5$
	S_{10}	0.4	
	S_8	0.4	
	S_5	0.6	
	S_{11}	0.4	
	S_6	0.2	
$S_1S_3S_5$	S_{11}	0.6	$S_1S_3S_5S_{11}$
	S_6	0.3	
Identified disease in database is d_6			

then next maximum occurrence symptom i.e. s_8 symptom will be asked to the patient for confirmation.

Now the selected symptom set is s_3, s_1 now I can calculate the mutually dependent symptom of 's_3' and 's_1'.

Support of s_1s_3 is 5 in the given database

Confidence of $(s_1s_3) \rightarrow s_9 = 2/5$=.4

Confidence of $(s_1s_3) \rightarrow s_{10} = 2/5$=.4

Confidence of $(s_1s_3) \rightarrow s_8 = 2/5$=.4

Confidence of $(s_1s_3) \rightarrow s_5 = 3/5$=.6

Confidence of $(s_1s_3) \rightarrow s_{11} = 2/5$=.4

Confidence of $(s_1s_3) \rightarrow s_6 = 1/5$=.2

Here, maximum confidence on $s_1s_3 \rightarrow s_5$(patient confirmed). Now symptom set s_1, s_3, s_5 scan the knowledge base for the identification of disease.

If $\{s_1, s_3, s_5\}$ symptom set does not match with D_i, then again by following the previous same procedure, I can extract the next symptom. Now the total set will be treated as X and the next targeted symptom Y can be extracted from the database.

Support of $\{s_1, s_3, s_5\}$ is 3

Confidence of $s_1s_3s_5 \rightarrow s_{11} = 2/3 = .6$

Confidence of $s_1s_3s_5 \rightarrow s_6 = 1/3 = .3$

Maximum confidence occurs for $s_1s_3s_5 \rightarrow s_{11}$

The highest relevance symptoms are $\{s_1, s_3, s_5, s_{11}\}$

The data set $\{s_1, s_3, s_5, s_{11}\}$ predicted as d_6 type disease.

Randomly any symptom either s_1, s_3, s_5 or s_{11} can be considered as a preliminary symptom and reach the same targeted disease.

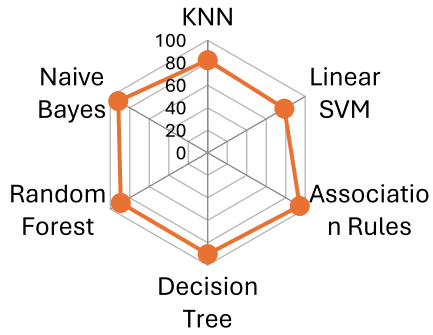

Fig. 3. Accuracy of clinical data.

4.4. Analysis of computational time

The time complexity of this approach depends on the amount of time required to calculate the support, confidence and symptom sets. Finding confidence of an associated symptom requires finding support of the initial symptom first. This step is simply O(n*m) where n is the number of patient's records and m is the maximum number of symptoms for a disease as the algorithm traverses the EHR once in the database. Next, the confidence of the next possible associated symptom takes O(d*m*n) where d is the maximum number of symptoms present in the symptom set. Finally, to find the symptom set, it requires (m*n+m*n*d). Therefore, the time complexity of this approach is O(d*m*n) which is polynomial in nature.

4.5. Identification of disease from any random symptom

Any random symptom (found as a preliminary symptom) from the symptom set (symptom set of any type of disease) can reach to the same targeted disease. Suppose a patient reported symptom s_1, instead of considering s_3 if I consider s_1 as the preliminary symptom, I will find the same disease d_6 by applying my proposed ASS algorithm. After considering one symptom, the support of that symptom throughout the EHR will be the maximum or same support than all other co-related symptoms with it. The symptom extraction procedure for the initial symptom s_1 is presented in Table 8 and the extraction of symptoms from any random symptom is given in Table 9.

This proposed diagnosis system can be applied to detect almost all types of disease but this is best suitable for those diseases that do not require any test report. ASS is especially helpful to detect neurodevelopmental type diseases because the symptoms are usually not detectable by standard tests and testing is usually done by observations of behavioral expressions. The diagnosis procedure of chronic disease is based on patient's confirmation about symptoms, i.e. an immediate response is required to take a decision. As this system depends upon the EHR and CKB, and these databases contain the symptom sets of chronic and acute diseases.

4.6. Performance analysis

A comparison of the performance of NB, DT, AR, KNN, LSVM and RF and corresponding validity measures with respect to accuracy, precision, recall, F1-score and J-Score for clinical data and Autism is presented in Table 10.

We have observed that accuracy, precision, recall, F1-score and J-Score of AR are higher than NB, DT, KNN, LSVM and RF classifiers for clinical data and Autism both. Hence, AR is the best classifier for these datasets.

Radar chart representations of accuracy, precision, recall and F1-score values of NB, DT, AR, KNN, LSVM and RF classifiers respectively for clinical data and Autism are presented in the Figs. 3–10.

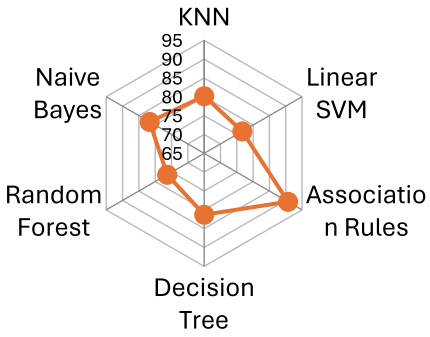

Fig. 4. Precision of clinical data.

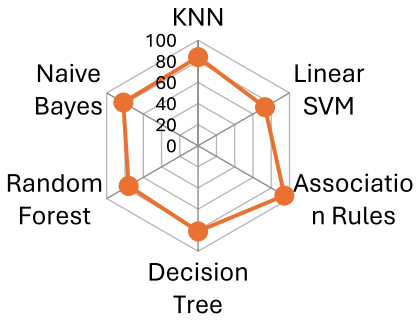

Fig. 5. Recall of clinical data.

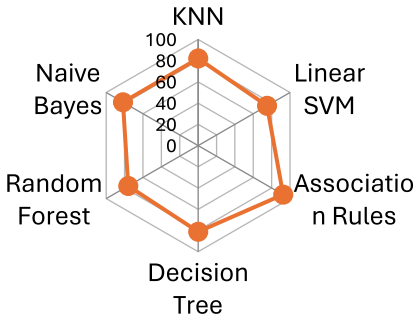

Fig. 6. F1-score of clinical data.

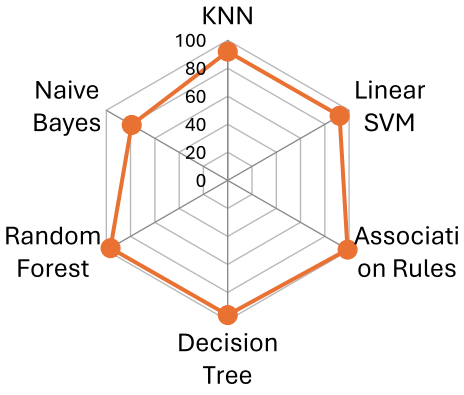

Fig. 7. Accuracy of Autism.

Figs. 3–10. Radar chart representations of the average classification accuracy, precision, recall and F1-Score values of AR, DT, RF, KNN, LSVM and NB classifiers for clinical data and autism respectively.

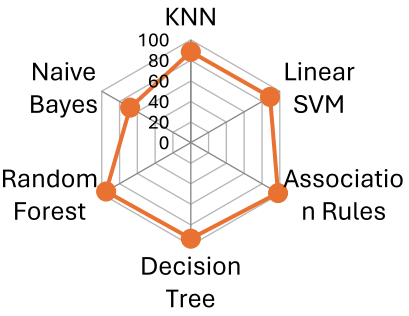

Fig. 8. Precision of Autism.

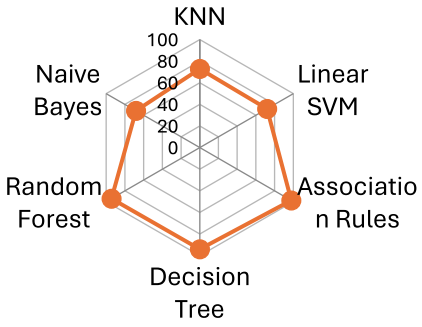

Fig. 9. Recall of Autism.

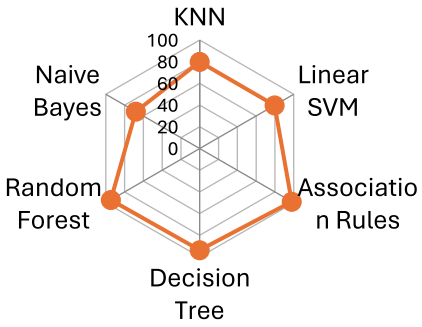

Fig. 10. F1-score of Autism.

The performance analysis of the best classifier for clinical data and autism with respect to accuracy, precision, recall, F1-score and J-score is presented in Table 11.

A comparison of the performance of our proposed model with some of the best previous classifiers is presented in Table 12. Some authors have not computed the F1-Score and J-score values, therefore, these values of corresponding authors are written NA. They computed the specificity, G-mean score, ROC, AUC, and MCC measure. Based on the performance report, we observed that our proposed classifier performs better than existing classifiers.

Values are written in bold indicating the better results of validity matrices.

Figs. 11–12 show the graphical representation of the experimental results of comparative techniques with respect to accuracy, precision, recall, F1-score and J-score values.

Among six different techniques, the machine learning AR algorithm performs better for clinical data and autism both as compared to other ML techniques with 94.40% accuracy, 90.73% precision, 94.45% recall, 92.55% F1-score and 0.951 J-score value and 98.78% accuracy, 98.02% precision, 97.80% recall, 97.91% F1-score and 97.12% J-score value respectively.

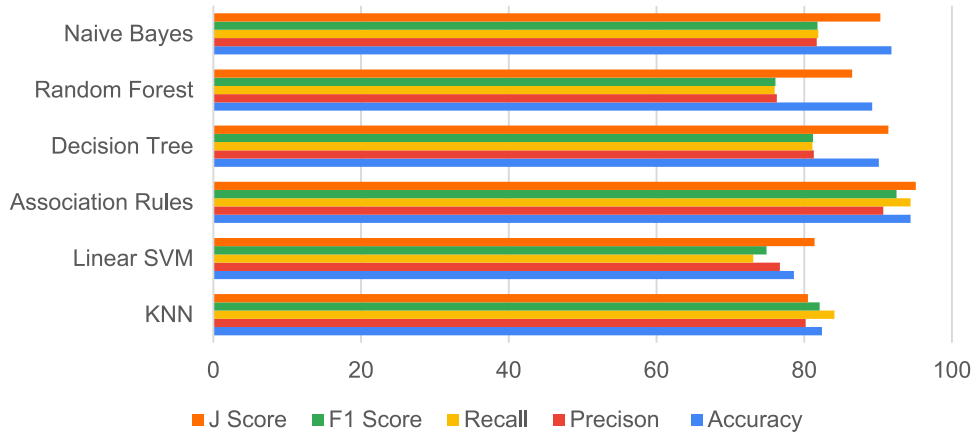

Fig. 11. Performance of classifiers for Clinical Data.

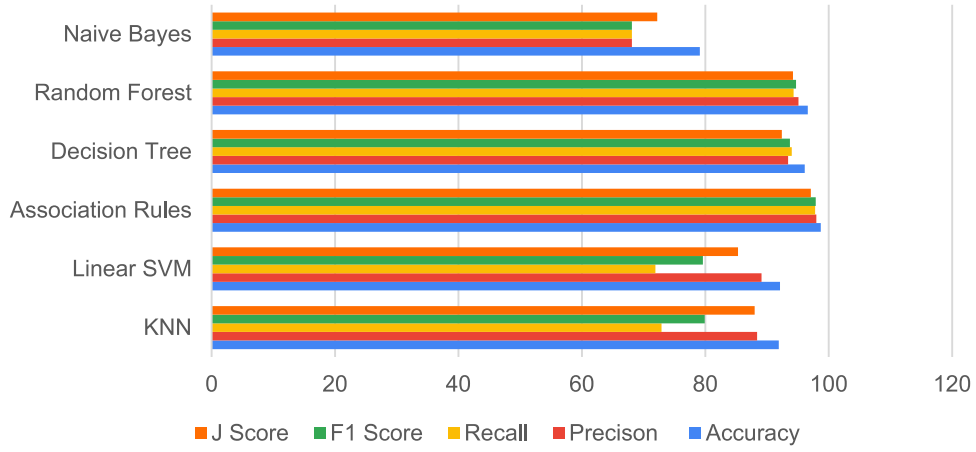

Fig. 12. Performance of classifiers for Autism.

Table 10
Experimental results of AR, DT, RF, KNN, LSVM and NB with respect to accuracy, precision, recall, F1-score and J-score of clinical data and Autism.

Data set	Method	Classifier	Accuracy	Precision	Recall	F1Score	J Score
Clinical data	RST	NB	0.918	0.819	0.817	0.818	0.903
		DT	0.901	0.813	0.811	0.812	0.914
		AR	**0.944**	**0.907**	**0.944**	**0.925**	**0.951**
		RF	0.892	0.763	0.760	0.761	0.925
		KNN	0.824	0.802	0.841	0.821	0.805
		LSVM	0.786	0.767	0.731	0.749	0.814
Autism	RST	NB	0.791	0.681	0.681	0.681	0.722
		DT	0.961	0.934	0.940	0.937	0.924
		AR	**0.987**	**0.980**	**0.978**	**0.979**	**0.971**
		RF	0.966	0.951	0.943	0.947	0.942
		KNN	0.919	0.884	0.729	0.799	0.880
		LSVM	0.921	0.891	0.719	0.796	0.853

Table 11
Performance analysis of proposed classifier for clinical data.

Disease	Best classifier	Performance report				
		Accuracy	Precision	Recall	F1-score	J-score
Clinical Data	AR	0.944	0.907	0.944	0.925	0.951
Autism	AR	0.987	0.980	0.978	0.979	0.971

5. Conclusion

In this research, we proposed a framework based on the hybridization of association rules and rough set theory to classify acute and life threatening diseases. The proposed framework uses an incomplete symptom set which is either only one or more sets of patient data to generate case-specific advice. This approach of diagnosis is a helping hand for a health assistants or less experienced doctors. Our suggested technique also assists rural patients in remembering their symptoms, which will help a doctor in making an appropriate diagnosis. Furthermore, it addresses the issue of a shortage of well qualified professionals and provides health care services to a large population.

Additionally, we focused on preprocessing and feature selection techniques because both are trustworthy processes to increase performance. Data is preprocessed using binary, on-hot vector, and min–max scales to remove noise, null, and duplicate entries. A rough set

Table 12
Comparison of the performance of the proposed classifier with existing classifiers.

Disease	Authors	Feature selection method	Classifier	Performance report				
				Accuracy	Precision	Recall	F1-Score	J-Score
Autism	[14]	KNN	KNN	93.8	94.40	97.43	95.89	NA
	[29]	NB	NB	97.6	98.70	98.07	98.39	NA
	[16]	SVM	SVM	95.2	93.97	100	96.89	NA
	[4]	RST	RF	96.69	95.12	94.34	94.73	94.23
	[53]	XGB	XGB	78.6	82.9	82.6	NA	NA
	[54]	RF	RF	93.33	90	97.29	93.5	NA
	[55]	Chi- Square	LR	97.14	96.67	93.55	95.08	NA
	Proposed	**RST**	**AR**	**98.78**	**98.02**	**97.8**	**97.91**	**97.12**

feature selection technique is applied to deal with inconsistency, incompleteness, vagueness and reduce the dimension of the datasets. ASS algorithm predicts diseases from only one preliminary symptom and extracts associated symptoms which are usually not detectable by standard tests and testing depends on observations of behavioral expressions. Six popular machine learning classifiers such as modified AR, DT, RF, KNN, LSVM and NB were applied to publicly available clinical data and autism downloaded from Kaggle. Performance was compared with proposed classifiers in terms of accuracy, precision, recall, F1-score and J-score. Our proposed classifier machine learning AR outperforms with 94.40% accuracy, 90.73% precision, 94.45% recall, 92.55% F1-score and 95.14% J-score on Clinical data and 98.78% accuracy, 98.02% precision, 97.80% recall, 97.91% F1-score and 97.12% J-score value on Autism respectively as compare to existing techniques. Finally, output in the form of recommendation is provided to the patient. Hence, overall performance strongly suggests that the proposed framework may help medical experts to identify various diseases more accurately and effectively.

The proposed diagnosis model in Section 4.2 used the ASS algorithm for the diagnosis of chronic and acute diseases. This algorithm scans the entire database, i.e., the entire electronic health record (EHR) for the extraction of one co-related symptom or the next targeted symptom for the diagnosis process. Hence, the algorithm scans the total EHR 'n' number of times to extract the targeted n numbers of co-related symptoms for the diagnosis of the disease, where 'n' is the number of co-related symptoms of the given preliminary symptom for the disease prediction. So, it takes much time. To reduce the time and evaluate the significance of the considered symptoms, we will implement other hybrid techniques or find some other ML/DL techniques in the future to predict the disease with diagnosis urgency. Subsequently, we will use the churn technique to know the feedback of patients to improve the patient care and delivery system.

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 is already available on Kaggle. I have also given a link of dataset in reference as well in result part.

References

[1] R. Cerchione, P. Centobelli, E. Riccio, S. Abbate, E. Oropallo, Blockchain's coming to hospital to digitalize healthcare services: Designing a distributed electronic health record ecosystem, Technovation (2022) 102480.

[2] Y. Jiang, X. Ding, D. Liu, X. Gui, W. Zhang, W. Zhang, Designing intelligent self-checkup based technologies for everyday healthy living, Int. J. Hum.-Comput. Stud. (2022) 102866.

[3] K.N. Singh, J.K. Mantri, V. Kakulpati, Churn prediction of clinical decision support recommender system, in: Ambient Intelligence in Health Care, Springer, Singapore, 2023, pp. 371–379.

[4] K.N. Singh, J.K. Mantri, Clinical decision support system based on RST with machine learning for disease prediction, Intell. Med. (2023) (Accepted).

[5] K.N. Singh, J.K. Mantri, Clinical decision support system based on RST with machine learning for medical data classification, Multimedia Tools Appl. (2023) 1–24.

[6] J. Han, M. Kamber, J. Pei, Data Mining Concepts and Techniques, Third ed., Morgan Kaufmann Publishers, New York, 2012.

[7] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognit. Lett. 24 (6) (2003) 833–849.

[8] Z. Pawlak, Rough Sets, Theoretical Aspects of Reasoning About Data, Kluwer Acadmic Publishers Dordrecht, 1991.

[9] Pengfei Zhang, Tianrui Li, Guoqiang Wang, Chuan Luo, Hongmei Chen, Junbo Zhang, Dexian Wang, Zeng Yu, Multi-source information fusion based on rough set theory: A review, Inf. Fusion 68 (2021) 85–117.

[10] E.K. Mohamed, J.U. Shankar, Intelligent diagnostic prediction and classification system for chronic kidney disease, Sci. Rep. 9 (9583) (2019) 1–14.

[11] G. Jothi, H.H. Inbarani, A.T. Azar, K.R. Devi, Rough set theory with Jaya optimization for acute lymphoblastic Leukemia classification, Neural Comput. Appl. (2018) 1–20.

[12] S.R. Dutta, S. Giri, S. Datta, M. Roy, A machine learning-based method for autism diagnosis assistance in children, in: 2017 International Conference on Information Technology, ICIT, IEEE, 2017, pp. 36–41.

[13] P.H. Lu, J.L. Keng, F.M. Tsai, P.H. Lu, C.Y. Kuo, An apriori algorithm-based association rule analysis to identify acupoint combinations for treating diabetic gastroparesis, Evidence-Based Complement. Alternat. Med. (2021) (2021).

[14] D.D. Khudhur, S.D. Khudhur, The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups, Meas.: Sens. 27 (2023) 100774.

[15] T. Bikku, S.R. Nandam, A.R. Akepogu, A contemporary feature selection and classification framework for imbalanced biomedical datasets, Egypt Inform. J. 19 (2018) 191–198.

[16] D.M. Abdullah, A.M. Abdulazeez, Machine learning applications based on SVM classification a review, Qubahan Acad. J. 1 (2) (2021) 81–90.

[17] C. Mohan, S. Nagarajan, Performance analysis of various machine learning techniques to predict cardiovascular disease: An empirical study, Appl. Math. Inform. Sci. 12 (1) (2018) 217–226.

[18] K.N. Singh, J.K. Mantri, V. Kakulpati, C. Misra, Prediction of mental distress about COVID-19 among higher education students in Odisha, India, in: V. Kakulpati (Ed.), Data Science Applications of Post-COVID-19 Psychological Disorders, 2022, pp. 63–83.

[19] N. Suguna, K. Thanushkodi, A novel rough set reduct algorithm for medical domain based on bee colony optimization, J. Comput. 2 (6) (2010) 49–54.

[20] K.B. Nahato, K.N. Harichandran, K. Arputharaj, Knowledge mining from clinical datasets using rough sets and backpropagation neural network, Comput. Math. Methods Med. 2015 (6) (2015) 1–13.

[21] L.S. Riza, A. Janusz, C. Bergmeir, C. Cornelis, F. Herrera, D. Śle, J.M. Benítez, Implementing algorithms of rough set theory and fuzzy rough set theory in the R package RoughSets, Inform. Sci. 287 (2014) 68–89.

[22] Y. Forghani, H.S. Yazdi, Fuzzy Min–Max neural network for learning a classifier with symmetric margin, Neural Process. Lett. 42 (2) (2015) 317–353.

[23] C. Velayutham, K. Thangavel, Unsupervised quick reduct algorithm using rough set theory, J. Electron. Sci. Technol. 9 (3) (2011) 193–201.

[24] B. Sun, W. Ma, Fuzzy rough set model on two different universes and its application, Appl. Math. Model. 35 (2011) 1798–1809.

[25] C. Zhao, Y. Ren, D. Gao, L. Xu, Prediction of service life of large centrifugal compressor remanufactured impeller based on clustering rough set and fuzzy Bandelet neural network, Appl. Soft Comput. J. 78 (2019) 132–140.

[26] Y. Wang, B. Sun, X. Zhang, Q. Wang, BWM and MULTIMOORA-based multi-granulation sequential three-way decision model for multi-attribute group decision-making problem, Internat. J. Approx. Reason. 125 (2020) 169–186.

[27] B. Sun, X. Zhou, N. Lin, Diversified binary relation-based fuzzy multigranulation rough set over two universes and application to multiple attribute group decision making, Inf. Fusion. 55 (2020) 91–104.

[28] K.N. Singh, J.K. Mantri, V. Kakulapati, S. Sharma, S.S. Patra, C. Misra, N. Kumar, Analysis and validation of risk prediction by stochastic gradient boosting along with recursive feature elimination for COVID-19, in: Applications of Artificial Intelligence in COVID-19, Springer, Singapore, 2021, pp. 307–323.

[29] N.A. Mashudi, N. Ahmad, N.M. Noor, Classification of adult autistic spectrum disorder using machine learning approach, IAES Int. J. Artif. Intell. 10 (3) (2021) 743.
[30] V. Christou, M.G. Tsipouras, N. Giannakeas, A.T. Tzallas, Hybrid extreme learning machine approach for homogeneous neural networks, Neurocomputing 311 (2018) 397–412.
[31] H. Zhao, X. Xu, S. Ding, R. Nie, Y. Zhang, Extreme learning machine: Algorithm, theory and applications, Artif. Intell. Rev. 44 (2013) 103–115.
[32] S. Patra, B. Barman, A novel dependency definition exploiting boundary samples in rough set theory for hyperspectral band selection, Appl. Soft Comput. 99 (2020) 106944.
[33] C.X. You, J.Q. Huang, F. Lu, Recursive reduced kernel based extreme learning machine for aero-engine fault pattern recognition, Neurocomputing 214 (2016) 1038–1045.
[34] P. Yang, D. Wang, W.B. Zhao, L.H. Fu, J.L. Du, H. Su, Ensemble of kernel extreme learning machine based random forest classifiers for automatic heartbeat classification, Biomed. Signal Process. Control. 63 (2021) 102138.
[35] S. Chakraborty, R.D. Raut, T.M. Rofin, S. Chatterjee, S. Chakraborty, A comparative analysis of multi-attributive border approximation area comparison (MABAC) model for healthcare supplier selection in fuzzy environments, Decis. Anal. J. 8 (2023) 100290.
[36] F. Masood, W. Boulila, A. Alsaeedi, J.S. Khan, J. Ahmad, M.A. Khan, S.U. Rehman, A novel image encryption scheme based on Arnold cat map, Newton–Leipnik system and logistic gaussian map, Multimed. Tools Appl. 81 (21) (2022) 30931–30959.
[37] S. ur Rehman, S. Tu, M. Waqas, Y. Huang, O. ur Rehman, B. Ahmad, S. Ahmad, Unsupervised pre-trained filter learning approach for efficient convolution neural network, Neurocomputing 365 (2019) 171–190.
[38] S. Tu, Y. Huang, G. Liu, CSFL: A novel unsupervised convolution neural network approach for visual pattern classification, Ai Commun. 30 (5) (2017) 311–324.
[39] S.U. Rehman, S. Tu, O.U. Rehman, Y. Huang, C.M.S. Magurawalage, C.C. Chang, Optimization of CNN through novel training strategy for visual classification problems, Entropy 20 (4) (2018) 290.
[40] S. Tu, S.U. Rehman, M. Waqas, O.U. Rehman, Z. Shah, Z. Yang, A. Koubaa, ModPSO-CNN: An evolutionary convolution neural network with application to visual recognition, Soft Comput. 25 (2021) 2165–2176.
[41] G.A. Khan, J. Hu, T. Li, B. Diallo, Y. Zhao, Multi-view low rank sparse representation method for three-way clustering, Int. J. Mach. Learn. Cybern. 13 (2022) 233–253.
[42] B. Mudumba, M.F. Kabir, Mine-first association rule mining: An integration of independent frequent patterns in distributed environments, Decis. Anal. J. (2024) 100434.
[43] R. Kumar, D.C. Bisht, Picture fuzzy entropy: A novel measure for managing uncertainty in multi-criteria decision-making, Decis. Anal. J. 9 (2023) 100351.
[44] D.P.M. Abellana, D.M. Lao, A new univariate feature selection algorithm based on the best–worst multi-attribute decision-making method, Decis. Anal. J. 7 (2023) 100240.
[45] H.D. Arora, A. Naithani, A new definition for quartic fuzzy sets with hesitation grade applied to multi-criteria decision-making problems under uncertainty, Decis. Anal. J. 7 (2023) 100239.
[46] V. Rani, S. Kumar, An innovative distance measure for quantifying the dissimilarity between Q-Rung orthopair fuzzy sets, Decis. Anal. J. (2024) 100440.
[47] A.U. Haq, J.P. Li, J. Khan, M.H. Memon, S. Nazir, S. Ahmad, et al., Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data, Sensors 20 (9) (2020) 2649.
[48] Q. Zhang, J. Lu, Y. Jin, Artificial intelligence in recommender systems, Complex Intell. Syst. 7 (1) (2021) 439–457.
[49] K.N. Singh, J.K. Mantri, Classifications of COVID-19 variants using rough set theory, in: Ambient Intelligence in Health Care, Springer, Singapore, 2023, pp. 381–389.
[50] Kaggle dataset repository. Kaggle.com, 2023, https://www.kaggle.com/code/prashfio/clinical-decision-support-system/data. (Accessed: 19 March 2023).
[51] Kaggle dataset repository. Kaggle.com, 2024, https://www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults (Accessed: 13 March 2024).
[52] J. Manimaran, T. Velmurugan, Analysing the quality of association rules by computing an interestingness measures, Indian J. Sci. Technol. 8 (15) (2015) 1–12.
[53] K. Xu, Z. Sun, Z. Qiao, A. Chen, Diagnosing autism severity associated with physical fitness and gray matter volume in children with autism spectrum disorder: Explainable machine learning method, Complement. Therapies Clin. Pract. (2023) 101825.
[54] J. Talukdar, D.K. Gogoi, T.P. Singh, A comparative assessment of most widely used machine learning classifiers for analysing and classifying autism spectrum disorder in toddlers and adolescents, Healthcare Anal. 3 (2023) 100178.
[55] R.A. Rasul, P. Saha, D. Bala, S.R.U. Karim, M.I. Abdullah, B. Saha, An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder, Healthcare Anal. 5 (2024) 100293.