scientific reports

Check for updates

OPEN

Identifying diseases symptoms and general rules using supervised and unsupervised machine learning

Fatemeh Sogandi

The symptoms of diseases can vary among individuals and may remain undetected in the early stages. Detecting these symptoms is crucial in the initial stage to effectively manage and treat cases of varying severity. Machine learning has made major advances in recent years, proving its effectiveness in various healthcare applications. This study aims to identify patterns of symptoms and general rules regarding symptoms among patients using supervised and unsupervised machine learning. The integration of a rule-based machine learning technique and classification methods is utilized to extend a prediction model. This study analyzes patient data that was available online through the Kaggle repository. After preprocessing the data and exploring descriptive statistics, the Apriori algorithm was applied to identify frequent symptoms and patterns in the discovered rules. Additionally, the study applied several machine learning models for predicting diseases, including stepwise regression, support vector machine, bootstrap forest, boosted trees, and neural-boosted methods. Several predictive machine learning models were applied to the dataset to predict diseases. It was discovered that the stepwise method for fitting outperformed all competitors in this study, as determined through cross-validation conducted for each model based on established criteria. Moreover, numerous significant decision rules were extracted in the study, which can streamline clinical applications without the need for additional expertise. These rules enable the prediction of relationships between symptoms and diseases, as well as between different diseases. Therefore, the results obtained in this study have the potential to improve the performance of prediction models. We can discover diseases symptoms and general rules using supervised and unsupervised machine learning for the dataset. Overall, the proposed algorithm can support not only healthcare professionals but also patients who face cost and time constraints in diagnosing and treating these diseases.

Keywords Diseases symptoms, Classification methods, Association rules, Apriori algorithm, Machine learning algorithms

Advancements in healthcare analytics can benefit both doctors and patients, as they can help detect and diagnose diseases early on, ultimately improving healthcare quality and patient outcomes. The use of Machine Learning (ML) techniques to predict disease symptoms in patients is both promising and challenging in the field of Artificial Intelligence (AI). AI enables the analysis of vast medical datasets, enhancing diagnostics, predicting disease outcomes, and optimizing treatment plans. AI methods can discover patterns in patient data, aiding in early detection of illnesses and personalizing medical interventions. Additionally, AI contributes to operational efficiencies, streamlining administrative tasks and improving resource allocation. The continuous evolution of AI in healthcare analytics holds great promise for improved decision-making, patient outcomes, and overall healthcare system optimization. The reliability of AI has significantly benefited medical diagnostics in the modern era. AI has extended the capabilities of human vision, and utilized in medical research.

The application of ML in discovering diseases symptoms has the potential to revolutionize diagnostics, treatment, and patient care, but further research and development are needed to overcome the existing challenges. Some ML algorithms have been used in healthcare field to predict different diseases like heart disease[1]. Additionally, Association Rules (AR) have been employed for knowledge extraction. These algorithms analyze data to identify patterns and make predictions, offering the use of ML techniques in predicting disease symptoms among patients has the potential to enhance patient outcomes, and enhance the efficiency of the health centers[2]. Despite the potential benefits, the integration of ML in healthcare is still in its infancy, and there are several challenges

Department of Industrial Engineering, University of Torbat Heydarieh, Torbat Heydarieh, Iran. email: f.sogandi@torbath.ac.ir

to overcome before widespread adoption can occur primarily due to the lack of user-friendly ML systems that cater to non-technical users. Therefore, the development of a model or system that facilitates the diagnosis of diseases using ML techniques is a promising and challenging aspect of AI. With this objective in mind, we have conducted a study to establish an AI-based methodology for the initial diagnosis of these symptoms. The application of ML algorithms has been evident in numerous recent healthcare works[3]. Several literature reviews have been conducted on the use of ML algorithms in diagnosing diseases. These reviews cover a comprehensive range of diseases and the application of various ML techniques for disease diagnosis. The research work[4] conducted a comprehensive review of ML-based disease diagnosis, examining the most recent trends and approaches in ML for disease diagnosis. Some of the main findings from these reviews include the use of ML algorithms like Naïve Bayes, Support Vector Machine (SVM), K Nearest Neighbor (KNN), and Random Forest (RF) for disease diagnosis[5]. Also, the authors of[6] focused on the most common ML methods applied to extend AI applications, including neural networks, SVM, ANN, RF, Decision Trees (DT), Logistic Regression (LR), and Neural-boosted (NB). Woodman and Mangoni[7] also discussed the growing application of ML in diagnosing both common and rare diseases. Additionally, Poudel[8] provided a perfect overview of the most frequently used ML algorithms in disease diagnosis, along with a focus on the clinical challenges involved in relying on these algorithms. Furthermore, the research[9] highlighted the benefits, methodologies, and functionalities of using ML algorithms in disease diagnosis in the healthcare sector. Ferdous et al.[10] provided a literature survey on them in healthcare with the best accuracy in diagnosing diseases. Fatima and Pasha[11] highlighted the advantages and disadvantages of these methods and provided a comparative analysis of different ML techniques for disease diagnosis. Overall, these literature reviews offer valuable insights into the use of ML algorithms for disease diagnosis and provide a comprehensive understanding of the current trends and future research directions.

Supervised ML algorithms demonstrate impressive results when dealing with well-labeled datasets, and they are widely employed in various fields. The application of supervised ML in healthcare analytics empowers clinicians, administrators, and policymakers to make data-driven decisions, enhance patient care, and optimize healthcare delivery[12]. Supervised ML plays a pivotal role in healthcare analytics too, particularly in predictive modeling. In healthcare, Supervised ML methods can be used to identify diseases and diagnose them, predict patient outcomes, and optimize treatment plans. Predictive modeling in supervised ML algorithms is the process of building a model that can predict future outcomes using historical data. After the in-depth search, Kumar et al.[13] found that 85% of the supervised learning methods characterized the study, while the remaining 15% were characterized by unsupervised learning methods. In this regard, the Flores et al.[14] surveyed the application of unsupervised ML methods in discovering latent disease clusters using electronic health records. The authors used Latent Dirichlet Allocation, and suggested a new model named Poisson Dirichlet model. The research effort[15] showed that K-Mean and SVM have also diagnosed and evaluated diabetes as an amalgamation of supervised and unsupervised ML techniques. In addition, Lim et al.[16] provided an unsupervised ML model for discovering latent infectious diseases using social media data. The research[17] focused on an unsupervised ML algorithm for detecting patient clusters using genetic signatures. The authors could assign high-risk and chronic disease patients into a detected cluster using their genomic makeup. The study by Bose and Radhakrishnan[18] employed unsupervised ML techniques to categorize patients with heart failure who utilized telehealth services in the home health setting. The researchers analyzed the differences between these subgroups in terms of patient characteristics, such as symptoms.

Predictive analytics is used to forecast future events by examining the correlation between input and output variables. The increasing availability of electronic clinical data in the U.S. healthcare system has led to the growing popularity of predictive systems in healthcare[19]. Some common predictive algorithms include ML and deep learning, which are subsets of AI. These algorithms use historical data to train algorithms that can predict future outcomes. For example, predictive models help assess the risk of patient readmission. Hospitals can use predictive analytics to estimate the length of a patient's hospital stay. This aids in resource planning, bed management, and improving overall operational efficiency. On the other hand, predictive modeling is applied to identify fraudulent activities in healthcare billing too. Also, physicians can benefit from predictive models that offer insights into potential diagnoses based on patient data. Besides, predictive models are utilized to forecast the likelihood of diseases and adverse events. These models can analyze patient data to predict the likelihood of developing diseases symptoms[20]. Unsupervised learning and predictive modeling are both important techniques in ML, each serving different purposes in data analysis and pattern recognition.

In unsupervised learning, the model works on its own to discover patterns and information in unlabeled data. AR learning is a type of unsupervised learning that investigates for the dependency of one data item on another and is used to extract hidden patterns from data. Additionally, AR mining can empower clinicians to make quick and automatic decisions, extract valuable information. The study's findings are crucial for understanding disease symptoms, which is critical in initial triage to distinguish the severity of cases. Hence, this study aims to use AR mining to identify symptom in the patients and explore these patterns based on explanatory variables. Some notable papers on AR in healthcare. For the first time, Brossette et al.[21] discussed the use of AR for discovering new patterns in hospital infection control and public health surveillance data. The authors proposed a process for analyzing surveillance data by comparing their confidences across different data partitions. The study of[22] used AR mining to extract hidden patterns and relationships between diagnostic test requirements in real-life medical data. After that, Happawana et al.[23] explored the use of AR mining techniques for generalizing diagnoses from a public health dataset based on techniques for reducing the search space. Additionally, Miswan et al.[24] presented a case study on using AR mining to analyze hospital readmission data. The authors discussed various related studies and techniques, such as data mining for hospital readmission. In COVID-19 epidemic, Tandan et al.[25] discovered symptom patterns of COVID-19 patients using AR mining. In a similar way, the symptom patterns of COVID-19 from recovered and deceased patients are extracted by work[26] using Apriori AR mining. In another view point, Khafaga et al.[27] constructed a prediction system for predicting diabetes by AR algorithm.

More recently, Cui et al.[28] proposed the weighted Apriori algorithm for discovering AR from disease diagnostic data. The authors also designed an improved KNN algorithm as a pre-step to obtain more accurate associations on a higher level.

Now, we perform a detailed comparison of this work with the relevant prominent studies in the field of hybrid supervised and unsupervised ML. As a pioneer, Péran et al.[29] focused on the classification of Parkinson's disease and multiple system atrophy using supervised and unsupervised learning techniques applied to MRI data. After that, Ma et al.[30] leveraged the phenotyping structure using the integrated of unsupervised and supervised ML methods for phenotyping complex diseases with a unique application. Also, Cai et al.[31] presented an approach that combines unsupervised and supervised learning techniques to detect self-reported COVID-19 symptoms on Twitter. More recently, Sáiz-Manzanares et al.[32] explored the application of supervised and unsupervised ML techniques in therapeutic interventions for children. In comparison to these existing methodologies, our study aims to identify patterns of symptoms and general rules regarding symptoms among patients using a combination of supervised and unsupervised ML techniques. This study utilizes the Apriori algorithm to identify frequent symptoms and patterns, which is a unique approach compared to the other studies. In other words, the study integrates a rule-based ML technique and classification methods to extend a prediction model. This approach is different from the studies that focus on a single algorithm or a specific type of disease. Additionally, our study applies several ML models, including Stepwise Regression (SR), SVMs, Bootstrap Forest (BF), Boosted Trees (BT), and NB methods, to predict diseases, demonstrating the versatility of our approach.

To the best of our knowledge, there is no work to utilize supervised and unsupervised ML algorithms to extract the common symptoms of the mentioned diseases. As aforementioned it is necessary to extend an integrated diagnosis system of diseases using a suite of AI. In summary, our approach is distinct from other studies in several ways:

- Integration of rule-based ML and classification methods: Our study combines the Apriori algorithm with classification methods to identify frequent symptoms and patterns, which is a novel approach compared to other studies that focus on a single algorithm or a specific type of disease.
- Versatility of ML Models: We applied a range of ML models, including SR, SVMs, BF, BT, and NB methods, to predict diseases. This demonstrates the versatility of our approach and the ability to adapt to different disease scenarios.
- Real-world dataset: We used a real-world dataset from the Kaggle repository, which is not typically used in other studies. This allows our findings to be more generalizable and applicable to real-world healthcare settings.
- Survey of diseases and symptoms: Our study takes a unique perspective by surveying diseases and symptoms from multiple angles, rather than focusing on a single disease or specific type of symptom. This comprehensive approach allows us to identify patterns and general rules that can be applied across various diseases.
- Novel Approach to Disease Prediction: Our study integrates classification algorithms with ARs for extracting relationships between diseases, symptoms, and improving disease prediction. This approach has not been introduced before in the literature, making our study a significant contribution to the field.

Therefore, this paper aims to model and predict disease symptoms using classification and AR methods. In this regard, we distinguish the most significant risk variables and the correlation between them after data preparation. Moreover, we compare the predictive performance of a range of different ML models to determine the best solution for diseases symptoms diagnosis. Also, AR mining has been used to extract symptom patterns of the diseases set to conduct intelligent diagnosis by extract valuable rules in this paper. The remainder of this paper is structured as follows: Section "Methodology" discusses related work. Section "Results" presents the details of the research methodology and dataset. Section "Conclusion and future research" covers the results and discussion. In the final section, we conclude the study with objectives, limitations, and research contributions.

Methodology

The proposed method is given in this section. The main goal is to exploring the relationship that exist between the disease symptoms and implement different types of ML techniques in discovering diseases symptoms to predict the diseases. In this regard, the proposed methods of supervised and unsupervised ML are explained for discovering diseases symptoms. The selection of algorithms was based on a thorough review of the literature and consideration of the specific research question and data characteristics. For supervised learning, we chose to use linear regression, SVMs, BF, BT, and NB methods because these algorithms have been widely used and shown to be effective in predicting diseases in various studies[4]. For unsupervised learning, we chose to use the Association rule algorithm because it is a well-established method for discovering frequent patterns and rules in data[21–28]. The Apriori algorithm is particularly useful for identifying patterns in large datasets and can be used to identify both frequent and rare events. Additionally, the Apriori algorithm can be used to identify patterns that are not necessarily linear or continuous, making it a useful tool for identifying complex relationships in data. We believe that the combination of these algorithms provides a comprehensive approach to identifying patterns of symptoms and general rules regarding symptoms among patients. The use of multiple algorithms allows us to leverage the strengths of each method and to identify patterns that may not be apparent using a single algorithm.

The study leverages a combination of programming languages and libraries to facilitate data analysis and ML tasks. Specifically, JMP scripting language, JMP data tables, and JMP modeling and ML are employed to streamline data preprocessing, model training, and evaluation. The data preprocessing process involves several key steps, including data import, data cleaning, and data transformation. Furthermore, the model training and evaluation phases utilize a range of techniques, including model screening, association analysis, model training,

and model evaluation. Additionally, various tools are utilized, such as JMP modeling and ML and JMP association analysis, to support these tasks.

The approach starts with a data preparation, and the AR method. Then, classification algorithms are used and compared with each other to predict disease models. Figure 1 is a schematic overview of the proposed approach. The following subsections present the dataset description and data preparation, applied unsupervised and supervised ML methods as well.

Data preprocessing

Firstly, this subsection goes into the specifics of the disease dataset that was utilized, and then data preprocessing is performed in this research. Our dataset provides a comprehensive compilation of symptoms and patient profiles for a range of diseases. The mysteries of diseases can be unveiled with this disease symptom and patient profile dataset. The analytic results can show intricate relationship between patients and diseases. In other words, the proposed system can assist in the extracting the AR and development of predictive models for disease diagnosis and monitoring based on symptoms and patient characteristics. On this subject, we utilized an available online data set by the Kaggle Repository. In our study, we used a publicly available Kaggle dataset that does not contain personally identifiable information (PII). To handle the sensitive health data responsibly, we consider ensured data anonymization and maintained compliance with data privacy regulations, such as HIPAA and GDPR. By taking these measures, we aim to protect the privacy and confidentiality of the patient information, comply with relevant data protection regulations, and conduct the research in an ethical and transparent manner, prioritizing the rights and well-being of the study participants. The dataset offers a detailed examination of the intricate relationships between patients and diseases, comprising over 100 distinct medical conditions and featuring 3490 records. The dataset offers a treasure trove of information including fever, cough, fatigue, and breathing difficulty, intertwined with age, gender, blood pressure, and cholesterol levels revealing the fascinating connections between symptoms, demographics, and health indicators. We aim to explore the hidden patterns, and uncover unique symptom profiles. The dataset has 10 attributes, which are given in Table 1.

A classifier can be ineffective in processing raw data in some cases due to features such as incompleteness, noise, and inconsistency[33]. Data preprocessing is necessary for preparing a dataset to improve prediction accuracy in data mining and ML. The data was preprocessed before analysis, which included which included label encoding, data transformation, and handling outliers. During data preprocessing, label encoding is conducted to transform the data into numerical format. Categorical variables, which include symptoms, gender, blood pressure, cholesterol level, and the outcome variable (disease), are often non-numeric and represent various categories or groups. During data preprocessing, label encoding is conducted to transform the data into numerical format. When categorical data is transformed into numerical data, predictive modeling and classification

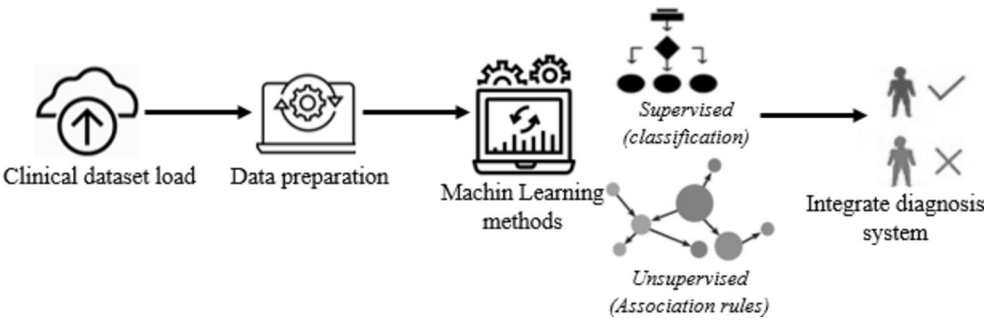

Figure 1. Pipeline of proposed research methodology.

Attribute name	Description	Data type
Disease	The name of the disease or medical condition	Nominal
Fever	Indicates whether the patient has a fever (Yes = 1/No = 0)	Binary
Cough	Indicates whether the patient has a cough (Yes = 1/No = 0)	Binary
Fatigue	Indicates whether the patient experiences fatigue (Yes = 1/No = 0)	Binary
Breathing difficulty	Indicates whether the patient has breathing difficulty (Yes = 1/No = 0)	Binary
Age	The age of the patient in years	Numeric
Gender	The gender of the patient (Male = 1/Female = 0)	Binary
Blood pressure	The blood pressure level of the patient (High/Normal/Low)	Nominal
Cholesterol level	The cholesterol level of the patient (High/Normal/Low)	Nominal
Outcome variable	The outcome variable indicating the result of the diagnosis or assessment for the specific disease (Positive = 1/Negative = 0)	Binary

Table 1. Explanation of dataset.

algorithms can effectively process and learn from the data. This transformation avoids misleading orderings, and enables the algorithms to create more accurate models, better generalize to new data, and ultimately improve model performance[34]. On the other hand, In JMP software, to perform AR mining, the data needs to be in list format, and then it should be transformed to nominal format type. This process allows the software to analyze transactional data and identify items that have an affinity for each other, a technique frequently used in market basket analysis.

One of the main issues in ML is dealing with outliers. An outlier is a data point that deviates from the typical behavior exhibited by other data points. The presence of outliers can impact the performance of AI-based forecasting methods and the discovery of diseases symptoms. Therefore, ensuring that the dataset is free of outliers is a critical task for achieving superior prediction results.

Note that the limitations of the dataset and acknowledge any biases or incompleteness that may affect the interpretation and application of the findings. These limitations include sample size and representation, data collection methodology, missing or incomplete data, geographical and demographic limitations, and temporal limitations. To minimize the influence of these constraints and improve the applicability of the outcomes, it is essential to carefully interpret the findings, and recognize the potential limitations and the importance of thoughtful consideration when extrapolating the results to particular patient groups or clinical situations.

Applied unsupervised ML method

AR learning in unsupervised ML algorithms is a valuable method for uncovering interesting connections among features in a dataset. The Apriori, Eclat, and FP-growth algorithms are widely used for AR learning and are instrumental in identifying patterns and associations in large datasets, offering valuable insights for various applications such as market basket analysis, customer segmentation, and recommendation systems. These algorithms play a crucial role in fields like retail, healthcare, and finance, where they help in understanding customer behavior, optimizing product offerings, and improving business strategies. The Apriori algorithm, for instance, is essential for data scientists and businesses seeking to extract meaningful patterns and associations from their data. We used the Apriori algorithm, which has a computational complexity of O(nlog n), where n is the number of data points. The time complexity is O(nlog n) for training and O(1) for prediction. The space complexity is O(n) for storing the model coefficients. Extracted AR are valuable for predicting class values in early-stage diseases. However, different criteria can be used to measure the strength of these rules. Some of these criteria are described below:

Support: This metric indicates how often a given rule appears in the database being mined.

$$Support(X \rightarrow Y) = \frac{\text{Patients having both } X \text{ and } Y}{\text{Total number of patients}}. \tag{1}$$

Confidence: This metric refers to the number of times a given rule turns out to be true in practice.

$$Confidence(X \rightarrow Y) = \frac{\text{Patients having both } X \text{ and } Y}{\text{Patients having } X}. \tag{2}$$

Lift: This metric is utilized to compare the confidence of a rule with the expected confidence, or how many times an if-then statement is expected to be found true.

$$Lift(X \rightarrow Y) = \frac{\left(\text{Patients having both } X \text{ and } Y\right)\left(\text{Patients having } X\right)}{\text{Proportion of patients having } Y}. \tag{3}$$

These metrics help assess the effectiveness of AR in predicting class values for early-stage diseases.

Applied supervised ML method

By harnessing the power of basic health indicators, we can improve the understanding of diseases and their progression, ultimately leading to better patient care and more effective interventions. Predictive modeling in supervised ML algorithms aims to extend a model that can accurately predict the value of a target variable based on one or more input variable. In this context, we will briefly discuss several popular supervised ML algorithms, including SR, SVM, BF, BT, and NB methods.

SR

Sequential Regression (SR) is a statistical method utilized for feature selection within predictive modeling. It involves a systematic approach to identifying the optimal subset of predictors that exhibit the strongest correlation with the target variable. Through an iterative process, predictors are added or removed from the model based on their statistical significance and impact on the model's overall performance. This iterative refinement continues until a feature set that maximizes model performance is determined. However, the computational demands of SR can be substantial, particularly with large datasets. The selection of the best feature subset entails evaluating numerous feature combinations, leading to potential time constraints. Moreover, as the dataset's feature count grows, the computational complexity escalates, rendering SR less feasible for certain scenarios.

SVM method

SVM are a supervised ML method that can be employed for both classification and regression tasks. SVM operates by identifying a hyperplane that best separates the data points, effectively finding a decision boundary that distinguishes the classes with the largest margin. In this context, SVM algorithms are used to create models that can make predictions based on known relationships between input and target variables, such as in classification

problems, or continuous predictions in regression tasks. In the context of predictive modeling, SVM can be used to find a function that best predicts the value of the target variable using the input features. The computational complexity of SVMs is $O(n^2)$, where n is the number of data points. The time complexity is $O(n^2)$ for training and O(1) for prediction. The space complexity is O(n) for storing the model coefficients.

BF method

BF is a term that combines bootstrap aggregation (bagging) and RF. Bagging is a method that aims to enhance the accuracy and robustness of ML models by training multiple models on various subsets of the training data and then combining their predictions. RF is an ensemble learning method that creates multiple DT during training and outputs the class that is the mode of the classes of the individual trees. This method is used in supervised ML for both classification and regression tasks Therefore, BF refer to an ensemble learning method that combines the principles of bagging and RF. The computational complexity of BF is O(nlog n), where n is the number of data points. The time complexity is O(nlog n) for training and O(1) for prediction. The space complexity is O(n) for storing the model coefficients.

BT method

BT, such as Gradient Boosting are ensemble learning methods that combine the predictions of multiple weak learners to create a strong learner. These algorithms work by iteratively training and combining the predictions of weak learners, such as DT or linear regression models, to improve the overall accuracy and reduce overfitting. In the context of predictive modeling, BT can be applied to predict the value of a target variable using the input features. The computational complexity of BT is O(nlog n), where n is the number of data points. The time complexity is O(nlog n) for training and O(1) for prediction. The space complexity is O(n) for storing the model coefficients.

NB method

NB is a ML approach that combines neural networks and boosting algorithms to enhance prediction accuracy. Boosting is an ensemble learning method that merges weak learners to create a strong learner, reducing training errors. In contrast, neural networks are ML algorithms capable of discerning intricate patterns in data. NB integrates these techniques by training a neural network on a subset of the training data and then using boosting to combine multiple neural networks, creating a more precise model. The process involves iteratively training a neural network on a subset of the training data and then adding the network to the ensemble. The weights of the neural network are adjusted to minimize the error of the ensemble. NB has demonstrated effectiveness in various applications, such as image classification, speech recognition, and natural language processing. However, NB has limitations, including the potential for overfitting and the requirement for large amounts of training data. Despite these limitations, NB is a potent ML technique that can enhance prediction accuracy across a variety of applications. The computational complexity of NB methods is $O(n^2)$, where n is the number of data points. The time complexity is $O(n^2)$ for training and O(1) for prediction. The space complexity is O(n) for storing the model coefficients.

To enhance the description of the proposed method, we have outlined the key steps of the implementation of our methodology using JMP software.

Key steps for the proposed methodology.

1. Load and preprocess the dataset
Import the dataset into JMP
Handle missing values
Encode categorical variables
2. Apply association rule mining using the Apriori algorithm
Use the Apriori node in JMP to extract association rules
Set the minimum support and confidence thresholds
Analyze the generated rules to identify frequent symptoms and patterns
3. Fit various machine learning models
Use the Fit Model node in JMP to apply different models
Stepwise Regression:
Select the stepwise method and appropriate model type (e.g., Generalized Linear Model)
Specify the response variable and predictor variables
Perform stepwise selection based on the criteria
Support vector machines (SVMs):
Select the SVM model type (SVM Classifier)
Set the kernel function and other relevant parameters
Train the SVM model using the training data
Bootstrap forest:
Select the Bootstrap Forest model type
Set the number of trees and other parameters
Train the Bootstrap Forest model using the training data

Boosted trees:
Select the Boosted Tree model type
Set the number of trees, learning rate, and other parameters
Train the Boosted Tree model using the training data
Neural-boosted methods:
Select the Neural Network model type
Set the number of hidden layers, activation functions, and other parameters
Train the Neural Network model using the training data
4. Evaluate model performance
Use cross-validation to assess the performance of each model
Calculate relevant metrics for each model
Compare the performance of different models and select the best-performing one
5. Extract significant decision rules
Use the Decision Tree node in JMP to generate decision rules
Analyze the decision rules to identify relationships between symptoms and diseases
Assess the significance and interpretability of the extracted rules
6. Interpret results and draw conclusions
Summarize the key findings, including the performance of the best-performing model and the significant decision rules
Discuss the implications of the results for improving disease prediction and clinical decision-making

The implementation of the methodology is available in the following GitHub repository: https://github.com/fsogandi/disease-symptoms.git

Results

In this section, we analyze data to investigate disease symptoms using AR and predictive modeling.

Data preparation

As shown in Table 1, blood pressure and cholesterol level characteristics are nominal data type. Hence, we use the transformation method and encoding to have Binary variables that are then treated as numeric. On the other hand, in JMP software, to perform AR mining, the data needs to be in list format, and then it should be transformed to nominal format type. In this respect, we treated each patient as a single transaction. Then, we divided the dataset into three groups based on the patient's age to transform a list format including: young adult, middle-age adults, and older adults. We initially applied AR mining to symptom data and identified symptom rules. Additionally, to identify and mange outliers, we apply the KNN (K = 8), robust principal component analysis (with lambda = 0.107 and outlier threshold = 2), T^2, Mahalanobis, and Jackknife distances methods. Generally, results show that the rows of 1, 81, 122, 213 are outlier and should be excluded. Note that KNN identifies outliers based on distance to each observation nearest neighbors for theses rows as well as 39 rows that we ignore it. Figure 2 shows outliers using T^2, Mahalanobis, and Jackknife distances for instance.

After data preparation phase, we perform a descriptive statistical analysis to help more ML methods. Some of these investigations are provided here. In this regard, the dataset does not exhibit significant skewness, with only a few outliers present, and the gender distribution in the dataset is relatively balanced. Figure 3 shows that individuals have a higher likelihood of testing positive for diseases, in older age. Additionally, Fig. 4 shows fever is a main symptom of these diseases. This figure demonstrates many individuals, regardless of the type of experience (positive or negative), report coughing.

The more analytics using pie chart shows the majority of the individuals in the study have high blood pressure and cholesterol. Additionally, out of 348 patients, 185 tested positive for a disease. Only 23 of the positive cases developed all symptoms. The average age of the patients is 46, with the majority being middle-aged. However, positive cases are proportionally higher in older adults. A violin plot indicates that older adults have high blood pressure, but older adult to middle-aged patients also exhibit high blood pressure. The most common symptoms were fatigue (139 cases), fever (109 cases), breathing difficulty, and cough (both seen in 88 cases). Females are more prone to the diseases than Males.

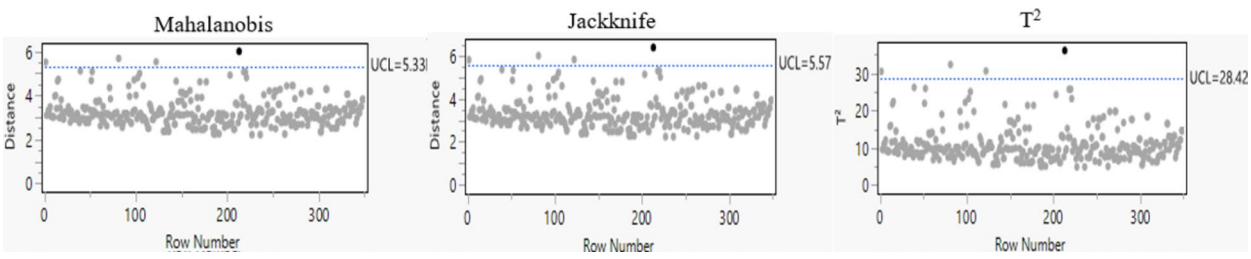

Figure 2. Outlier plot for the T^2, Mahalanobis, and Jackknife distances methods.

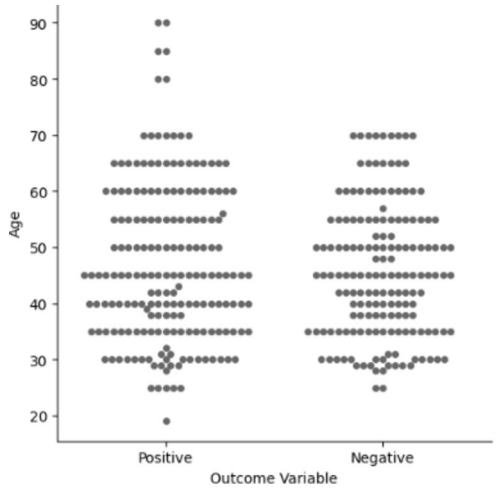

Figure 3. Plot of outcome results in terms of age factor.

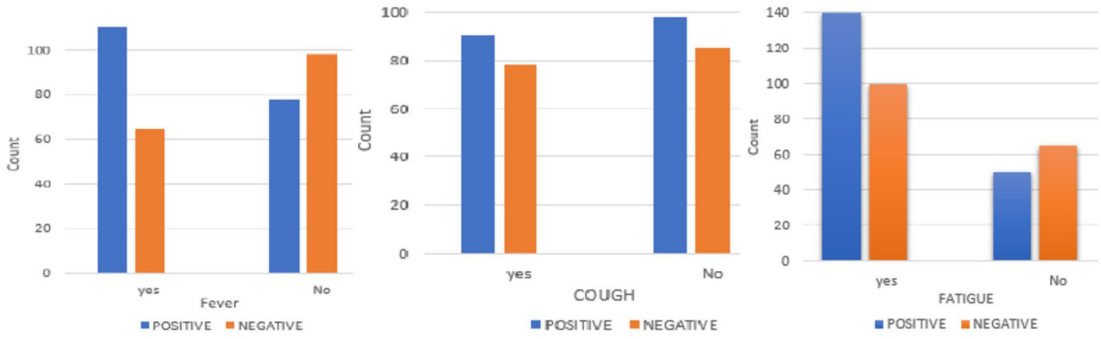

Figure 4. Bar graph of some symptoms of the diseases.

AR in unsupervised ML

We used an a Apriori method to extract lift matrix-based strong rules. Symptom transactions are part of the AR mining which aims to identify frequent item sets that meet a minimum threshold. To achieve this, we set the minimum confidence level to 1, ensuring that all generated rules have a 100% confidence level. Additionally, we establish a minimum support threshold above 0.01 and a lift greater than 4 for positively correlated rules. This means that the rules generated must have a support value greater than 1% and a lift value greater than 4, indicating a strong positive correlation between the antecedent and consequent items. Furthermore, we limit the maximum number of antecedents to 3 and the maximum rule size to 4, ensuring that the generated rules are concise and interpretable. To do so, we discover many significant AR for the data, and the top 20 symptom rules by highest lift values are given in Table 2. Table 2 concentrates on the antecedents (diseases) associated with the consequents (symptoms) to predict asymptotes of diseases.

Table 2 shows diseases strongly linked to symptoms with a confidence of 100% (except for rule 18) and a lift greater than 1. A confidence level of 100% indicates a high degree of certainty. Lift measures the performance of an AR as a response enhancer. Lift values greater than 1 indicate interdependence between conditions and their outcomes, emphasizing positive relationships. Based on rule 2, if a patient had Chronic Obstructive Pulmonary Disease (COPD) (condition), then this patient had a higher confidence for breathing difficulty in older adults' group (consequent). Specifically, Rule 1 suggests a positive association between Typhoid fever, high cholesterol, and fatigue, while rule 10 indicates that Hepatitis B increases the likelihood of coughing, fatigue, and high cholesterol. The results also demonstrate that demographic factors impact the relationships between symptom patterns and disease types. Additionally, the proposed model seeks to predict the potential disease of a patient based on their specific symptoms. In this regard, the 20 top rules are given in Table 3.

The associations in Table 2 exhibit a high confidence and a lift greater than 1, indicating a positive links. For example, Table 3 shows 5 rules related to COPD with a 100% confidence level and a notably high lift. According to these rules, older adults experiencing breathing difficulty, fever, high blood pressure, high cholesterol, and fatigue have a 67% chance of having COPD. Similarly, the presence of high cholesterol, high blood pressure, and breathing difficulty in older adults may indicate a higher likelihood of Rheumatoid Arthritis. Additionally, the model aims to predict potential diseases based on specific symptoms, while also considering the influence of demographic factors on symptom-disease associations. Furthermore, the unsupervised algorithm can identify relationships between symptoms and various attributes, aiding in the discovery of symptom relationships. In this respect, 25 rules are extracted in Table 4.

Row	Rule		Confidence %	Lift
	Condition	Consequent		
1	Typhoid fever	High cholesterol, fatigue	100	93
2	COPD	Breathing difficulty, older adults	100	93
3	COPD	Breathing difficulty, older adults	100	62
4	Typhoid fever	High cholesterol, fatigue	100	62
5	COPD	Breathing difficulty, older adults, fatigue	100	62
6	COPD	Breathing difficulty, older adults, fever	100	62
7	COPD	Breathing difficulty, older adults, high blood pressure	100	62
8	COPD	Breathing difficulty, older adults, high cholesterol	100	62
9	COPD	Breathing difficulty, older adults, cough	100	62
10	Hepatitis B	Cough, fatigue, high cholesterol	100	62
11	Typhoid fever	Cough, high cholesterol, fatigue	100	62
12	Ebola virus	Breathing difficulty, middle age group, high blood pressure	100	62
13	Ebola virus	Breathing difficulty, middle age group, high cholesterol	100	62
14	Hepatitis B	Fatigue, high cholesterol, middle age group	100	62
15	Hepatitis B	Fatigue, high cholesterol, fever	100	62
16	Typhoid fever	High cholesterol, middle age group, fatigue	100	62
17	Parkinson's disease	Middle age group, fatigue, high cholesterol	100	62
18	Rheumatoid arthritis	High cholesterol, high blood pressure, older adults	67	62
19	Hepatitis B	Cough, high cholesterol, fever	100	46.5
20	Lyme disease	Middle age group, cough, fatigue	100	46.5

Table 2. Top extracted rules for predicting asymptotes of diseases.

Row	Rule		Confidence %	Lift
	Condition	Consequent		
1	High cholesterol, male, fatigue	Typhoid fever	100	93
2	High cholesterol, fatigue, fever	Typhoid fever	100	93
3	High cholesterol, high blood pressure, older adults	Rheumatoid arthritis	100	62
4	High cholesterol, fatigue	Typhoid fever	67	62
5	Breathing difficulty, older adults, fatigue	COPD	67	62
6	Breathing difficulty, older adults, fever	COPD	67	62
7	Breathing difficulty, older adults, high blood pressure	COPD	67	62
8	Breathing difficulty, older adults, high cholesterol	COPD	67	62
9	Breathing difficulty, older adults, cough	COPD	67	62
10	Cough, fatigue, high cholesterol	Hepatitis B	67	62
11	Cough, high cholesterol, fatigue	Typhoid fever	67	62
12	Breathing difficulty, middle age group, high blood pressure	Ebola virus	67	62
13	Breathing difficulty, middle age group, high cholesterol	Ebola virus	67	62
14	Fatigue, high cholesterol, middle age group	Hepatitis B	67	62
15	Fatigue, high cholesterol, fever	Hepatitis B	67	62
16	High cholesterol, middle age group, fatigue	Typhoid fever	67	62
17	High cholesterol, breathing difficulty, fatigue	Typhoid fever	67	62
18	High cholesterol, fatigue, high blood pressure	Typhoid fever	67	62
19	Middle age group, fatigue, high cholesterol	Parkinson's disease	67	62
20	High cholesterol, breathing difficulty, older adults	Rheumatoid arthritis	67	41.3

Table 3. Top extracted rules for predicting diseases types conditional on observed asymptotes.

According to Table 4, among all rules, fever was the most common consequent. To describe the extracted rules, we focus on one rule for instance. Based on rule 1, if a patient has breathing and coughing problems and high blood pressure there is a 100% confidence that he or she had a fever. Similarly, Rule 2 highlights that when a patient experience both fatigue and high cholesterol, they will also have a fever. Moreover, the last rule shows that male patient with breathing difficulty who are strongly associated with fever, with a confidence of 90%. In general, our analysis from Tables 2–4 shows that the older adults age group strongly correlated with diseases

Row	Rule		Confidence %	Lift
	Condition	Consequent		
1	Cough, breathing difficulty, high blood pressure	Fever	100	1.706422
2	Fatigue, high cholesterol	Fever	100	1.706422
3	High blood pressure, no cough, no fever	High cholesterol	97	1.563478
4	Cough, breathing difficulty, older adults age	Fever	96	1.643221
5	Breathing difficulty, high blood pressure	Fever	96	1.635321
6	Male, high cholesterol	Fever	96	1.635321
7	Cough, breathing difficulty, high cholesterol	Fever	96	1.63223
8	Breathing difficulty, male, older adults age	Fever	95	1.628857
9	Cough, high blood pressure, high cholesterol	Fever	95	1.625164
10	Breathing difficulty, high blood pressure, older adults age	Fever	95	1.625164
11	Breathing difficulty, high blood pressure, high cholesterol	Cough	95	2.007955
12	Breathing difficulty, high blood pressure, high cholesterol	Fever	95	1.621101
13	High blood pressure, breathing difficulty, no fever	High cholesterol	94	1.522251
14	Female, high blood pressure, no fever	High cholesterol	93	1.509565
15	Older adults	High cholesterol	93	1.505847
16	Breathing difficulty, high cholesterol	Fever	93	1.584535
17	Breathing difficulty, older adults age	Fever	92	1.575159
18	Cough, high cholesterol, older adults age	Fever	92	1.569908
19	Breathing difficulty, high cholesterol, older adults age	Fever	92	1.569908
20	Cough, breathing difficulty	Fever	92	1.568063
21	Breathing difficulty, fever, high blood pressure	Cough	91	1.929842
22	Cough, high cholesterol	Fever	91	1.555855
23	Cough, breathing difficulty, high blood pressure	High cholesterol	90	2.758782
24	Cough, high blood pressure, high cholesterol	Breathing difficulty	90	3.116402
25	Breathing difficulty, male	Fever	90	1.53578

Table 4. The AR for different symptoms.

Transaction(age)	Item sets (diseases)
19	Influenza
25	Asthma, common cold, eczema, influenza
28	Asthma, hyperthyroidism
29	Allergic rhinitis, anxiety disorders, common cold, depression, diabetes, gastroenteritis, liver cancer, pancreatitis, rheumatoid arthritis, stroke, urinary tract infection
30	Asthma, bipolar disorder, bronchitis, cerebral palsy, colorectal cancer, dengue fever, eczema, gastroenteritis, hepatitis, hypertensive heart disease, influenza, kidney cancer, migraine, multiple sclerosis, muscular dystrophy, myocardial infarction, sinusitis, ulcerative colitis, urinary tract infection
31	Asthma, common cold, migraine, osteoporosis
32	Pneumonia
35	Allergic rhinitis, asthma, atherosclerosis, chronic obstructive pulmonary…, cirrhosis, common cold, conjunctivitis (pink eye), depression, eczema, epilepsy, gastroenteritis, hypertension, hyperthyroidism, kidney cancer, liver cancer, liver disease, malaria, migraine, pancreatitis, pneumonia, psoriasis, rheumatoid arthritis, rubella, spina bifida, ulcerative colitis, urinary tract infection, urinary tract infection
38	Allergic rhinitis, anxiety disorders, depression, diabetes, gastroenteritis, influenza, kidney disease, liver cancer, liver disease, migraine, osteoarthritis, osteoporosis, pneumonia, stroke
39	Klinefelter syndrome
40	Acne, asthma, brain tumor, bronchitis, chickenpox, coronary artery disease, cystic fibrosis, diabetes, eating disorders (anorexia,…, fibromyalgia, gastroenteritis, glaucoma, hemophilia, hyperthyroidism, hypoglycemia, lymphoma, osteoarthritis, pneumonia, psoriasis, rabies, tuberculosis
42	Anxiety disorders, common cold, depression, diabetes, hypothyroidism, influenza, kidney cancer, kidney disease, liver cancer, liver disease, lung cancer, migraine, osteoarthritis, stroke, urinary tract infection

Table 5. Categorized diseases based on age using transaction list in AR method.

occurrence. Another analytic can be achieved from AR mining, gives in Table 5 in which the disease may be occurred in specified age are shown. To sum up, we provide some of the ages in this respect.

Moreover, we can identify diseases that have an affinity for each other using Singular Value Decomposition (SVD). Diseases that exhibit overlap, based on the SVD method, can be identified by leveraging the SVD technique. This approach decreases the dimensionality of the data, allowing for the grouping of similar diseases and the extraction of relevant information. Figure 5 and Table 6 show points or diseases that are close to each other.

Predictive modeling in supervised ML

Now, we aim to develop a model that can accurately predict diseases using the disease symptoms and patient profile dataset. As aforementioned, this dataset contains valuable information on symptoms, demographics, and health indicators, which can be used to reveal fascinating connections and patterns. After examining the "Disease" column, we found that many unique diseases have only 1 to 5 samples, which is insufficient for a reliable disease prediction model. Predicting diseases with such limited information could lead to inaccurate results and misdiagnosis, which we want to avoid. Therefore, we will focus only on the diseases that have 10 or more samples to ensure the robustness of our model. This decision will reduce the number of cholesterol asses we are predicting down to 6, making our model more accurate. On the other hand, using checking for and handling missing values and identifying and removing duplicate entries we can ensure that our data is accurate, complete, and ready for further analysis or model building. After cleaning our data, we have focused on diseases with 10 or more samples. Understanding the balance of cholesterol asses is crucial as it can impact the performance of our ML model. To visualize this, we have utilized a pie chart in Figure 6. This step is essential for ensuring that our model is trained on a well-balanced dataset, which can ultimately enhance its predictive accuracy and reliability.

The pie chart shows that the classes are imbalanced, and we need to handle class imbalance. Before that, we need to process our categorical variables to perform a univariate analysis. This analysis will help us understand the distribution of our variables and their individual impact on disease prediction. We will start with the age variable, followed by other variables like symptoms, gender, blood pressure, and cholesterol level.

The univariate analysis of the age variable in Fig. 7 reveals that age is a valuable feature for predicting certain diseases. For instance, if the age is greater than 80, the disease is likely to be a stroke. However, the dataset has limited samples, especially for ages greater than 80, which could make predicting new values in this age range challenging. The analysis also shows that some diseases like Migraine and Hypertension are not present in ages between 20 and 30, suggesting that these conditions are more prevalent in older age groups. Hypertension and Osteoporosis appear more frequently as the age increases, indicating a potential correlation between these diseases and age. Also, cholesterol levels and blood pressure, significantly influence disease prediction. For example, High blood pressure is associated with the absence of stroke, which is crucial for stroke prediction. These observations emphasize the importance of these variables in predicting diseases. The next step is to examine

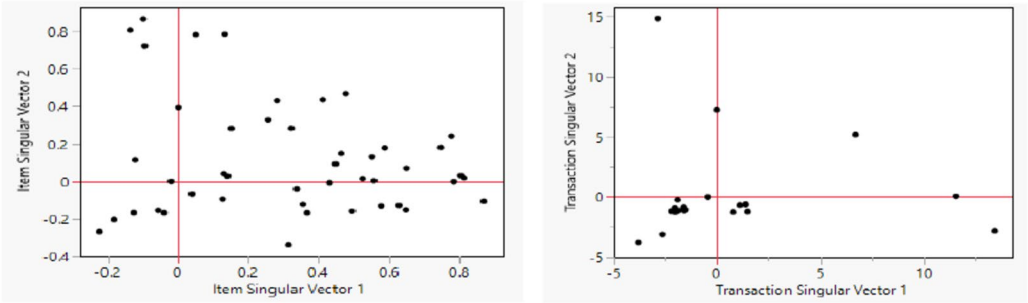

Figure 5. Item SVD plots for the data set.

Lung cancer	Influenza
Common cold	Urinary tract infection
Allergic rhinitis	Urinary tract infection
Sleep apnea	Liver disease
Zika virus	Anxiety disorders
Migraine	Multiple sclerosis
Diabetes	Eczema
Hypertension	Ulcerative colitis
Rheumatoid arthritis	Kidney cancer
Rabies	Tuberculosis

Table 6. Diseases that exhibit overlap, based on the SVD method.

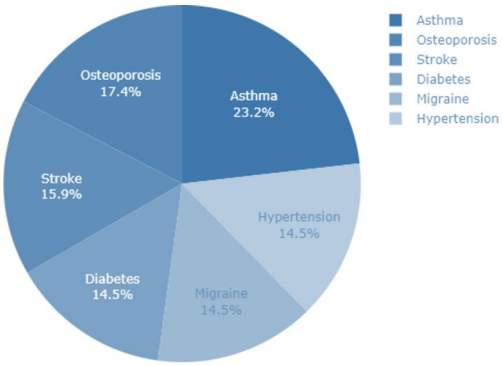

Figure 6. Pie chart for initial diseases classification.

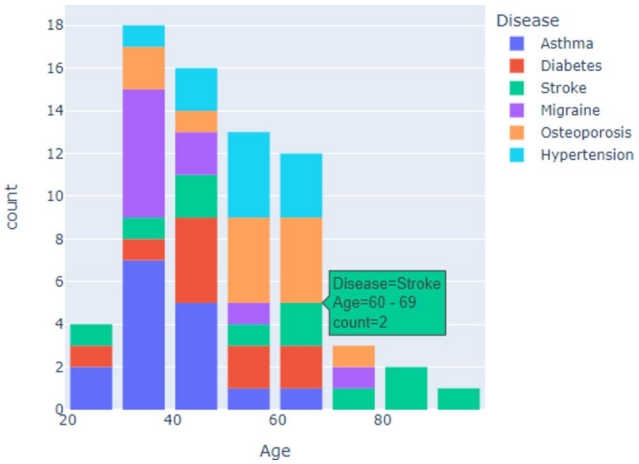

Figure 7. Bar graph of age-diseases.

how these variables correlate with each other, which can help identify patterns and potential multicollinearity, ultimately influencing the model's performance.

Figure 8 shows that none of the variables have a strong correlation with the "Disease" variable. The most correlated variables are "Age" and "Difficulty Breathing", with scores of 1 and – 1, respectively. In situations where there are multiple variables with high correlation scores, ML can be a viable alternative for prediction tasks. However, it's essential to consider that ML algorithms, typically require large amounts of data to perform optimally. In our case, we have only 79 data points, which is relatively small.

For hyperparameter tuning, we used the Grid Search method in JMP. Grid search is a simple and effective method for finding the optimal combination of hyperparameters by systematically varying each hyperparameter over a range of values and evaluating the performance of the model at each combination. We used a grid search with 10 iterations to find the optimal combination of hyperparameters for each model. For example, for the SR, we used a grid search to optimize the following hyperparameters:

- Stepwise selection: We used a grid search to optimize the stepwise selection method. We varied the number of features to include in the model from 1 to 10, and evaluated the performance of the model at each combination.
- Lambda: We used a grid search to optimize the lambda value, which is a hyperparameter that controls the strength of the regularization term in the model. We varied the lambda value from 0.1 to 1.0, and evaluated the performance of the model at each combination.

For the SVMs, we used a grid search to optimize the following hyperparameters:

- Kernel: We used a grid search to optimize the kernel function, which is a hyperparameter that controls the shape of the decision boundary in the model. We varied the kernel function between linear, polynomial, and radial basis functions, and evaluated the performance of the model at each combination.
- Gamma: We used a grid search to optimize the gamma value, which controls the width of the kernel function. We varied the gamma value under (0.1–1) and evaluated the performance of the model at each combination.

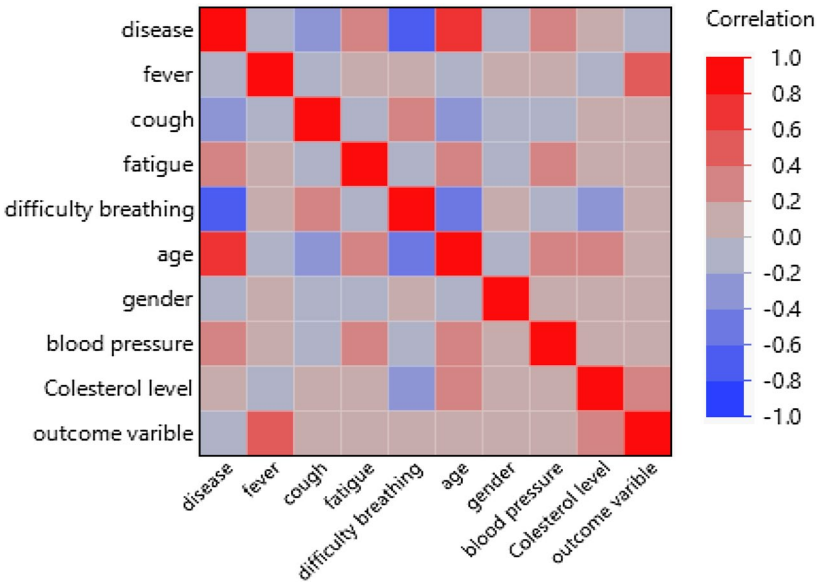

Figure 8. Correlation of each feature in the dataset using the heat map generated by JMP Pro 17 (version 17.2.1, available at https://www.jmp.com/).

For the BF model, we used a grid search to optimize the following hyperparameters:

- Number of trees: We used a grid search to optimize the number of trees in the forest that controls the complexity of the model. We varied the them under (10–100), and assessed the performance of the model at each combination.
- Max depth: We used a grid search to optimize the maximum depth of the trees, which is a hyperparameter that controls the complexity of the model. We varied the maximum depth from 5 to 10, and evaluated the performance of the model at each combination.

For the BT model, we used a grid search to optimize the following hyperparameters:

- Number of iterations: We used a grid search to optimize the number of iterations in the boosting algorithm, which is a hyperparameter that controls the complexity of the model. We varied the number of iterations from 10 to 100, and evaluated the performance of the model at each combination.
- Learning rate: We used a grid search to optimize the learning rate that controls the step size in the boosting algorithm. We varied it under (0.1–1) and evaluated the performance of the model at each combination.

For the NB methods, we used a grid search to optimize the following hyperparameters:

- Number of hidden layers: We used a grid search to optimize the number of hidden layers in the neural network, which is a hyperparameter that controls the complexity of the model. We varied the number of hidden layers from 1 to 3, and evaluated the performance of the model at each combination.
- Number of neurons: We used a grid search to optimize the number of neurons in each hidden layer, which is a hyperparameter that controls the complexity of the model. We varied the number of neurons from 10 to 100, and evaluated the performance of the model at each combination.

To conduct a fair comparison between different classifiers and identify the superior model with the best performance, we have considered and calculated several evaluation metrics that are well-suited for our specific case and dataset. The evaluation metrics we have included are:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \tag{4}$$

$$\text{Precision} = \frac{TP}{(TP + FP)}, \tag{5}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}, \tag{6}$$

$$\text{F1} - \text{score} = \frac{2.TP}{(2.TP + FP + FN)}, \tag{7}$$

$$\text{MCC} = \frac{2^{(TP.TN - FP.FN)}}{\sqrt{(TP + FP).(TP + FP).(TN + FP).(TN + FN)}}. \tag{8}$$

This is a crucial measure for evaluating imbalanced multi-class classification problems. A comparative assessment of most common used ML classifiers is performed in Table 7 for analyzing and classifying diseases.

We used the confusion matrix to calculate different metrics, and the best results are marked in bold. As illustrated by Table 7, SR method is the superior model leading to the best performance with the accuracy of 86.73% (95% CI 82.69–90.71) and the precision of 75.36%. Besides, the corresponding criteria of recall and F1- measure and (Matthews Correlation Coefficient) MCC are 77.87, 81.31, and 54.02%, respectively. Based on these metrics, the "SR" model consistently performs well across evaluation criteria. To avoid additional complexity and keep this table simple to read, we preferred to exclude the standard deviation of each result metrics.

Overall, researchers focused on specific diseases or conditions mentioned in the dataset can utilize it to explore relationships between symptoms, age, gender, and other variables. Also, healthcare technology companies can use the proposed method based on ML methods for developing healthcare diagnostic tools. It is worth mentioning that the model shows strong performance in predicting asthma cases but struggles to predict other conditions, suggesting its potential use in a one-vs-all approach for asthma diagnosis. Notably, the training data is imbalanced, with asthma being the most frequent class. To address this, data augmentation techniques such as rotation, scaling, or adding noise could be implemented to improve the model's accuracy in predicting less frequent diseases.

The study aims to identify common patterns and general rules across various diseases using ML techniques. By analyzing a diverse dataset, the research uncovers connections between symptoms, demographics, and health indicators, providing valuable insights for developing predictive models and early warning systems applicable to multiple diseases. It is worth noting that the decision to generalize the study across various diseases is grounded in several key considerations, including identifying common patterns, improving early detection, enhancing understanding, and practical implications. While the generalized approach offers several advantages, it is important to acknowledge that the study may not capture disease-specific nuances or rare symptoms that are unique to particular diseases. Future research could focus on validating the identified patterns and rules in specific disease contexts or exploring the applicability of the findings to rare or understudied diseases. In conclusion, the decision to generalize the study across various diseases is justified by the potential benefits of identifying common patterns, improving early detection, enhancing understanding, and providing practical implications for healthcare professionals. However, the limitations of this approach should be considered, and further research is needed to validate and refine the findings in specific disease contexts.

To improve the model's ability to adapt to new, emerging diseases or changes in symptom presentation, the following strategies can be implemented in our approach:

We can easily implement a system to continuously collect and integrate new patient data into the training dataset, including information on emerging diseases and changing symptom patterns. The models can then be retrained on a regular cadence (e.g., monthly, quarterly) to ensure they remain up-to-date and can adapt to evolving disease landscapes. Additionally, we can monitor model performance on a holdout test set to identify when retraining is necessary due to degradation in predictive accuracy. This will help ensure the models can adapt to new, emerging diseases and changing symptom presentations. As a future research direction, we recommend exploring the use of ensemble learning techniques. Specifically, we suggest investigating the application of various ensemble methods to further enhance the ability of the proposed models.

Statistical significance

Now, we use a statistical test to compare the proposed ML to ensure the statistical significance of the results and provide a robust comparison. Overall, the non-parametric tests are safer than parametric tests since they do not assume normal distributions or homogeneity of variance. In the case where multiple algorithms are to be compared, Friedman's test is the most interesting non-parametric statistical test. In Friedman test, the blocks of data, are considered independent. The underlying variables in the data are typically numeric in nature. The goal of this test is to determine whether there are significant differences among the algorithms considered over given sets of data. Training/Test set is generated as random sample from the population. The Friedman rank test can determine if there are significant differences in variation, central tendency, or shape among at least one

Method	Accuracy (%)	Precision (%)	Recall (%)	F1- measure (%)	MCC (%)
SR	86.73	75.36	77.87	81.33	54.02
SVM	73.48	74.51	77.86	76.16	52.72
BF	84.36	71.36	80.24	70.23	45.68
BT	81.62	69.23	75.39	86.2	42.67
NB	78.92	59.42	55.69	52.68	36.75

Table 7. Comparing the performance of different classifiers.

pair of the populations being compared. The test determines the ranks of the algorithms for each individual data set, i.e., the best performing algorithm receives the rank of 1, the second-best rank 2, etc.; in the case of ties average ranks are assigned. The Friedman test is performed in respect of average ranks, which use χ_F^2. Consider r_i^j be the rank of the jth of k ML algorithms on ith of n data sets. The Friedman test compares the average ranks of algorithms, R_j. The null hypothesis states that all algorithms perform equivalently. Under this hypothesis the Friedman statistics is as follows:

$$\chi_F^2 = \frac{12n}{k(k+1)}\left(\sum_j R_j^2 + \frac{k(k+1)^2}{4}\right),$$

in which χ_F^2 is distributed with k-1 degrees of freedom, when n and k are large enough. We can understand with comparing the corresponding statistics and $\chi_F^2(4)$ with $\alpha = 0.05$, the null hypothesis is rejected. In this regard, average rankings of the ML algorithms over the data sets by the Friedman test are shown in Table 8.

Feature importance and scoring

In the literature, two primary strategies for feature selection are Forward Selection (FS) and Backward Elimination (BE) for our classifier. FS starts by selecting the best single feature and then iteratively adds the feature that improves performance the most. Conversely, the BE begins with all considered features and repeatedly removes the feature that reduces performance the most. We conducted a series of experiments using fivefold cross-validation. The dataset was divided into 80% training cases and 20% test cases. In each fold, the training data was used to calculate the accuracy of a random forest classifier using different sets of features. The set of features that yielded the best accuracy was retained. The results are presented in Table 9.

The features were ranked incrementally based on their importance, with the most important feature labeled as one, the next most important feature labeled as two, and so on. Features with the "ignored" tag were removed from the dataset.

In the Forward Selection (FS) and Backward Elimination (BE) methods, we observe that the "age" and "breathing difficulty" features are consistently ranked as the most important, indicating its significant contribution to the model. Furthermore, we note that the "fatigue" feature is ranked last in both FS and BE, suggesting its relatively low relevance. Additionally, the "Blood pressure" feature is either ignored or ranked last in both methods, implying its minimal impact on the model. This further validates the effectiveness of our algorithm in ranking features.

Deployment challenges

Successful deployment of the ML models developed in this study necessitates careful consideration of the challenges to ensure effective implementation and adoption in real-world healthcare settings. The integration of these models into existing healthcare systems can pose significant challenges. Healthcare organizations often have complex and diverse IT infrastructures, with various systems and platforms in place. Seamless integration of the ML models into these existing systems is crucial for ensuring efficient data flow, accurate predictions, and effective decision support. Key considerations for integration include data compatibility, security and privacy, and scalability. Additionally, effective clinician training and adoption are crucial. Clinicians may be hesitant to rely on automated decision support systems, especially if they lack understanding of how the models work or have concerns about their accuracy and reliability. To address these challenges, the following strategies can be employed:

- Comprehensive training: Providing comprehensive training to clinicians on the use and interpretation of the ML models, including their strengths, limitations, and appropriate applications.
- Transparency: Ensuring that the ML models are as transparent and explainable as possible, allowing clinicians to understand the reasoning behind the predictions and build trust in the system.

1-st	2-nd	3-rd	4-th	5-th
SR	BF	BT	NB	SVM
1.54	2.22	3.22	4.75	5.36

Table 8. Average rank position of ML algorithms determined during the Friedman test.

Method	Rank							
	Fever	Cough	Fatigue	Breathing difficulty	Age	Gender	Blood pressure	Cholesterol level
FS using univariate feature selection	3	7	Ignore	2	1	4	6	5
Recursive BE with cross-validation	3	6	Ignore	2	1	5	Ignore	4

Table 9. Results of feature selection on the dataset.

• Continuous feedback and improvement: Establishing mechanisms for clinicians to provide feedback on the performance and usability of the ML models, enabling continuous improvement and adaptation to user needs.
• Incentives and support: Providing incentives and support for clinicians to adopt and integrate the ML models into their daily workflows, such as through performance metrics or dedicated support staff.

Successful deployment of ML models in healthcare requires careful consideration of integration challenges with existing systems and effective clinician training and adoption strategies. By addressing these challenges, healthcare organizations can effectively leverage the power of ML to improve patient outcomes and enhance clinical decision-making.

Conclusion and future research

Early disease prediction significantly enhances healthcare quality and can avert serious health complications. This proactive approach is particularly crucial due to the rise of new disease variants and the increasing availability of healthcare data. This study proposed an AI-based disease detection system for predicting diseases. Our results show several important results that enhance our diagnosing. In this regard, firstly we conduct data processing including data transformation and outlier detection, and then many significant AR was extracted based on Apriori algorithm. Generally, our research shows strong correlations between different variables, the occurrence of the diseases and medical conditions. For example, our study found that individuals in the older adults age group, those experiencing symptoms such as high cholesterol coughing and breathing difficulty have a strong relationship with Rheumatoid Arthritis. Additionally, various classification methods were applied to determine the best performing classifier, of the models investigated, SR method significantly outperformed the others. The proposed method can be used for medical practitioners, doctors, clinical analysis, and epidemiological investigations related to different diseases. It also can aid in understanding the prevalence and patterns of symptoms among patients with specific medical conditions.

We acknowledge the limitations of their research, which was based on a provided dataset that may not fully represent the diversity of patient populations. We recognize the need for larger-scale studies to validate the generalizability of their findings to other settings and populations. The authors emphasize the importance of future research to confirm their findings and investigate underlying mechanisms in more detail.

Additionally, we suggest that future work could involve the use of other types of ARM methods, such as the Frequent Pattern Growth to discover patterns. Future studies could consider using multiple datasets to improve the robustness of the findings. Overall, exploring different approaches, including data augmentation, is crucial to enhance the model's accuracy and enable more precise predictions across a wider range of conditions.

Data availability

The dataset used in this study is publicly available in the Kaggle repository https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset.

Received: 2 March 2024; Accepted: 30 July 2024
Published online: 02 August 2024

References

1. Yan, H., Jiang, Y., Zheng, J., Peng, C. & Li, Q. A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Syst. Appl.* **30**, 272–281. https://doi.org/10.1016/j.eswa.2005.07.022 (2006).
2. Manikandan, K. Diagnosis of diabetes diseases using optimized fuzzy rule set by grey wolf optimization. *Pattern Recogn. Lett.* **125**, 432–438. https://doi.org/10.1016/j.patrec.2023.03.011 (2019).
3. Bajwa, J., Munir, U., Nori, A. & Williams, B. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthc. J.* **8**, 188–194. https://doi.org/10.7861/fhj.2021-0095 (2021).
4. Ahsan, M. M., Luna, S. A. & Siddique, Z. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare* **10**, 541. https://doi.org/10.3390/healthcare10030541 (2022).
5. Ali, O. *et al.* A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *J. Innov. Knowl.* **8**, 100333. https://doi.org/10.1016/j.jik.2023.100333 (2023).
6. Mirbabaie, M., Stieglitz, S. & Frick, N. R. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. *Health Technol.* **11**, 693–773. https://doi.org/10.1007/s12553-021-00555-5 (2021).
7. Woodman, R. J. & Mangoni, A. A. A comprehensive review of machine learning algorithms and their application in geriatric medicine: Present and future. *Aging Clin. Exp. Res.* **35**, 2363–2397. https://doi.org/10.1007/s40520-023-02552-2 (2023).
8. Poudel, S. A study of disease diagnosis using machine learning. *Med. Sci. Forum* **10**, 8–20. https://doi.org/10.3390/IECH2022-12311 (2022).
9. Kumar, Y., Koul, A., Singla, R. & Ijaz, M. F. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J. Ambient Intell. Humaniz. Comput.* **1**, 1–28. https://doi.org/10.1007/s12652-021-03612-z (2022).
10. Ferdous M., Debnath J. and Chakraborty N.R., (2020). Machine learning algorithms in healthcare: A literature survey. In *2020 11th International conference on computing, communication and networking technologies* 1–6. https://doi.org/10.1109/ICCCNT49239.2020.9225642
11. Fatima, M. & Pasha, M. Survey of machine learning algorithms for disease diagnostic. *J. Intell. Learn. Syst. Appl.* **9**, 1–16. https://doi.org/10.4236/jilsa.2017.91001 (2017).
12. Burkart, N. & Huber, M. F. A survey on the explain ability of supervised machine learning. *J. Artif. Intell. Res.* **70**, 245–317. https://doi.org/10.1613/jair.1.12228 (2021).
13. Dowdell, J. *et al.* Intervertebral disk degeneration and repair. *Neurosurgery* **80**, S46. https://doi.org/10.1093/neuros/nyw078 (2017).
14. Flores, A. M. *et al.* Unsupervised learning for automated detection of coronary artery disease subgroups. *J. Am. Heart Assoc.* **10**, e021976. https://doi.org/10.1161/JAHA.121.021976 (2021).
15. Chauhan T., Rawat S., Malik S. and Singh P., (2021). March. Supervised and unsupervised machine learning based review on diabetes care. In *2021 7th International Conference on Advanced Computing and Communication Systems,* 1, 581–585. IEEE.https://doi.org/10.1109/ICACCS51430.2021.9442021

16. Lim, S., Tucker, C. S. & Kumara, S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. *J. Biomed. Inform.* **66**, 82–94. https://doi.org/10.1016/j.jbi.2016.12.007 (2017).
17. Shomorony, I. *et al.* An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. *Genome Med.* **12**, 1–14. https://doi.org/10.1186/s13073-019-0705-z (2020).
18. Bose, E. & Radhakrishnan, K. Using unsupervised machine learning to identify subgroups among home health patients with heart failure using telehealth. *CIN Comput. Inform. Nurs.* **36**, 242–248. https://doi.org/10.1097/CIN.0000000000000423 (2018).
19. Callahan, A. & Shah, N. H. Machine learning in healthcare. In *Key Advances in Clinical Informatics* (eds Callahan, A. & Shah, N. H.) 279–291 (Elsevier, 2017).
20. Talukdar, J., Gogoi, D. K. & Singh, T. P. A comparative assessment of most widely used machine learning classifiers for analysing and classifying autism spectrum disorder in toddlers and adolescents. *Healthc. Anal.* **3**, 100178. https://doi.org/10.1016/j.health.2023.100178 (2023).
21. Brossette, S. E. *et al.* Association rules and data mining in hospital infection control and public health surveillance. *J. Am. Med. Inform. Assoc.* **5**, 373–381. https://doi.org/10.1136/jamia.1998.0050373 (1998).
22. Sarıyer, G. & Öcal, T. C. Highlighting the rules between diagnosis types and laboratory diagnostic tests for patients of an emergency department: Use of association rule mining. *Health Inform. J.* **26**, 1177–1193. https://doi.org/10.1177/1460458219871135 (2020).
23. Happawana, K. A. & Diamond, B. J. Association rule learning in neuropsychological data analysis for Alzheimer's disease. *J. Neuropsychol.* **16**, 116–130. https://doi.org/10.1111/jnp.12252 (2022).
24. Miswan, N. H., Sulaiman, I. M., Chan, C. S. & Ng, C. G. Association rules mining for hospital readmission: A case study. *Mathematics* **9**, 2706. https://doi.org/10.3390/math9212706 (2021).
25. Tandan, M., Acharya, Y., Pokharel, S. & Timilsina, M. Discovering symptom patterns of COVID-19 patients using association rule mining. *Comput. Biol. Med.* **131**, 104249. https://doi.org/10.1016/j.compbiomed.2021.104249 (2021).
26. Dehghani, M. & Yazdanparast, Z. Discovering the symptom patterns of COVID-19 from recovered and deceased patients using Apriori association rule mining. *Inform. Med. Unlocked* **42**, 101351. https://doi.org/10.1016/j.imu.2023.101351 (2023).
27. Khafaga, D. S., Alharbi, A. H., Mohamed, I. & Hosny, K. M. An integrated classification and association rule technique for early-stage diabetes risk prediction. *Healthcare* **10**, 2070. https://doi.org/10.3390/healthcare10102070 (2022).
28. Cui, J., Zhao, S. and Sun, X., (2022). An association rule mining algorithm for clinical decision support. *In Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, 1, 137–143. https://doi.org/10.1145/3532213.3532234.
29. Péran, P. *et al.* MRI supervised and unsupervised classification of Parkinson's disease and multiple system atrophy. *Mov. Disord.* **33**(4), 600–608. https://doi.org/10.1002/mds.27307 (2018).
30. Ma, E. Y. *et al.* Combined unsupervised-supervised machine learning for phenotyping complex diseases with its application to obstructive sleep apnea. *Sci. Rep.* **11**(1), 4457. https://doi.org/10.1038/s41598-021-84003-4 (2021).
31. Cai, M., Li, J., Nali, M., & Mackey, T. K. (2021, June). Evaluation of hybrid unsupervised and supervised machine learning approach to detect self-reporting of COVID-19 symptoms on Twitter. In *2021 IEEE International Conference on Communications Workshops (ICC Workshops)* (pp. 1–6). https://doi.org/10.1109/ICCWorkshops50388.2021.9473830.
32. Sáiz-Manzanares, M. C. *et al.* Use of digitalisation and machine learning techniques in therapeutic intervention at early ages: Supervised and unsupervised analysis. *Children* **11**(4), 381. https://doi.org/10.3390/children11040381 (2024).
33. Ahmed, K. *et al.* Early detection of lung cancer risk using data mining. *Asian Pac. J. Cancer Prev.* **1**, 595–598. https://doi.org/10.7314/APJCP.2013.14.1.595 (2013).
34. Hasan, S. M. M., Mamun, M. A., Uddin, M. P. & Hossain, M. A. Comparative analysis of classification approaches for heart disease prediction. *Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng.* https://doi.org/10.1109/IC4ME2.2018.8465594 (2018).

Author contributions
The Authors have the same contributions in this study.

Funding
This work was financially supported by the Research Deputy of Education and Research, University of Torbat Heydarieh. Grant number: 212.

Competing interests
The author declares no competing interests.

Additional information
Correspondence and requests for materials should be addressed to F.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.