


RESEARCH

Open Access



Building an enhanced case-based reasoning and rule-based systems for medical diagnosis

Eslam M. Mustafa¹, Mahmoud M. Saad¹ and Lydia Wahid Rizkallah^{1*} 

*Correspondence:
lydiawahid@cu.edu.eg

¹ Computer Engineering
Department, Faculty
of Engineering, Cairo University,
Giza, Egypt

Abstract

Expert systems are computer programs that use knowledge and reasoning to solve problems typically solved by human experts. Expert systems have been used in medicine to diagnose diseases, recommend treatments, and plan surgeries. Interpretability of the results in medical applications is crucial since the decision that will be taken based on the system's output has a direct effect on people's health and lives which makes expert systems ideal choices when dealing with these applications in contrast to other machine learning approaches. An expert system has the ability to explain its own line of reasoning providing a robust way of diagnosis. This paper presents two types of expert systems for medical diagnosis. The first system is a case-based reasoning system using a database of previously diagnosed cases to diagnose a new case. The second system is a rule-based expert system that uses a set of if-then rules extracted from a decision tree classifier to make diagnoses. In this paper, machine learning-based similarity functions are proposed and compared with other traditional similarity functions. The results of this study suggest that expert systems can be a valuable tool for medical diagnosis. The two systems presented in this paper achieved competitive results, and they provide diagnoses similar to those made by human experts.

Keywords: Expert systems, Case-based reasoning, Rule-based expert systems, Similarity functions, Medical diagnosis

Introduction

Expert systems are computer programs that mimic problem-solving capabilities of a human expert. Such systems have proved their effectiveness in solving many problems such as COVID-19 diagnosis [1], academic advising [2], network intrusion detection [3], and many more. Expert systems are often used in medical diagnosis to help doctors make more accurate and timely diagnoses. There are many benefits to using expert systems in medical diagnosis. First, they can help doctors to avoid making mistakes. Doctors are humans, and they can make mistakes, especially when they are tired or under pressure. Using these systems can also help reduce the risk of these mistakes by providing doctors with accurate and up-to-date information. Second, they can help doctors to save time. Doctors often have to see many patients in a short period. Expert systems can help them quickly and efficiently gather the information they need to make a diagnosis. Third, they can help doctors provide better patient care. By providing doctors

with accurate and up-to-date information, expert systems can help them to make more informed decisions about the best course of treatment for their patients.

The development of medical diagnosis expert systems has the potential to improve the quality of healthcare. By providing doctors with accurate and reliable information, expert systems can help doctors make better patient care decisions. Some of the pioneering systems that used expert systems in medical diagnosis include the MYCIN system [4], Internist-1 system [5], and DXplain system [6].

There are many different types of expert systems such as rule-based, case-based, semantic networks and fuzzy expert systems. Case-based and rule-based expert systems are the main focus of this paper. Case-based expert systems use a database of past cases to make decisions. When a new case is presented, the system searches its database for similar cases and then uses their information to decide the solution for the new case. Rule-based expert systems use a set of if-then rules to make decisions.

The aim of this paper is to develop a medical diagnosis system that can accurately and reliably diagnose diseases based on their symptoms. Two types of systems are proposed in which case-based reasoning (CBR) and rule-based reasoning (RBR) are used. In this research, we will compare the performance of CBR and RBR for medical diagnosis. We will also evaluate the performance of different similarity functions for CBR systems. Custom machine learning-based similarity functions are also proposed in this paper and compared with the other similarity functions. This paper aims to show the efficiency of using expert systems for medical diagnosis and provide a clear comparison with other machine learning approaches.

The rest of this paper is organized as follows: related work is presented in Sect. 2. The methods and proposed approaches are explained in Sect. 3. Section 4 contains results and discussion, and finally, the conclusions are in Sect. 5.

Related work

In this section, we will give an overview of the approaches found in literature that are used in medical applications. Machine learning approaches as well as expert systems have been used to solve medical problems. The following subsections include examples of using these two types of techniques in medical problems.

Machine learning approaches

Authors in [7] propose a three-step hybrid approach for classifying the breast cancer dataset. The first step involves normalizing the dataset using the MAD normalization technique. The second step involves using K-means clustering-based feature weighting to transform the linearly non-separable dataset into a linearly separable dataset. Finally, the AdaBoostM1 classifier is used to classify the weighted dataset and it shows that AdaBoostM1 is effective in classifying breast cancer.

Authors in [8] discuss using machine learning techniques to predict heart disease. Hybrid random forest with a linear model (HRFLM) is used with different combinations of features to accurately predict the heart disease of patients. The results show that the HRFLM classification method achieves the highest accuracy compared to other methods. The proposed method effectively reduced the critical attribute set and improved accuracy.

The study in [9] used five machine learning models to predict heart disease in patients and analyzed their performance. The models used are decision tree, Naïve Bayes, random forest, support vector machine, and logistic regression. The dataset used was collected from Kaggle and contained 303 patient records. The performance of each model was measured and the results showed that the decision tree classifier has the highest accuracy, while Naïve Bayes has the lowest.

Another method is introduced in [10]. This method aimed to develop a machine learning technique for the early diagnosis of diabetes by selecting the most appropriate algorithms for diabetes classification and reducing the required input attributes to improve classification performance. The study used data from 520 patients from Sylhet Diabetes Hospital. The study used wrapper-based particle swarm optimization, tree seed algorithm, crow search algorithm, slime mould algorithm, and artificial bee colony algorithms for feature extraction and used conventional machine learning algorithms such as decision tree, random forest, support vector machine, K nearest neighbor, and feed forward neural networks for classification.

Expert systems' approaches

A case-based reasoning expert system is proposed in [11]. The system is used in diagnosing heart diseases. The system uses two retrieval approaches namely, induction and nearest-neighbor approaches to find the closest case in the database to the newly presented case in order to find its diagnosis. The nearest-neighbor approach has been proven to produce better results.

Authors in [12] present a fuzzy expert system to diagnose heart diseases. The system takes crisp inputs and performs a fuzzification step. After that, defuzzification is performed to convert the system's fuzzy output to a crisp value which is more readable by humans. The system measures the risk level of heart disease by giving it one of the following outputs: low, high, and risky. The results of the system show that the system is effective in finding the appropriate risk levels.

In [13], authors use a hybrid approach where probabilistic rules are generated from a decision tree and combined with rules provided by experts. The aim of such a combination is to complete the data generated from the decision tree with the experience of experts. The system is used in diabetes diagnosis. The results show that such a combination yields better results than dealing with each set of rules alone.

Methods

This study aimed to build expert systems for medical purposes using two main approaches: case-based reasoning (CBR) and rule-based reasoning (RBR).

Case-based reasoning (CBR) system

We developed a CBR system that uses past cases to solve new problems. Specifically, we built a database of cases consisting of patients' medical records and their diagnoses. When a new patient is presented with symptoms, the system searches its case base for the most similar case. The system then uses the diagnosis of the similar case to generate a diagnosis for the new patient. To compare cases in the CBR system, a similarity function is used. The choice of the similarity function is an important factor that affects the

performance of a CBR system. Therefore, in this study, we use various similarity functions and compare the performance of the system for each function. We also presented new similarity functions which are based on machine learning techniques as will be illustrated in the following subsections.

Similarity functions

To compare cases in the CBR system, we considered various similarity measures [14], including *cosine similarity*, *Euclidean distance*, *Manhattan distance*, *Pearson correlation coefficient*, *Jaccard similarity*, and *Dice coefficient*. We briefly introduce each similarity function below, along with its equation.

Cosine similarity—measures the cosine of the angle between two vectors. The cosine similarity between two vectors x and y is given by:

$$\text{cosine_similarity}(x, y) = (x \cdot y) / (||x|| ||y||).$$

where the dot (.) denotes the dot product, and $||x||$ and $||y||$ denote the lengths of vectors x and y , respectively.

Euclidean distance—measures the straight-line distance between two points in a multi-dimensional space. The Euclidean distance between two points x and y with n dimensions is given by:

$$\text{euclidean_distance}(x, y) = \sqrt{\sum_{i=1}^n |x_i - y_i|^2}$$

Manhattan distance—measures the distance between two points in a grid-like path by summing the absolute differences between their coordinates. The Manhattan distance between two points x and y with n dimensions is given by:

$$\text{manhattan_distance}(x, y) = \sum_{i=1}^n |x_i - y_i|$$

Pearson correlation coefficient—measures the linear correlation between two vectors and is commonly used in statistics. The Pearson correlation coefficient between two vectors x and y is given by:

$$\text{pearson_correlation}(x, y) = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Jaccard similarity coefficient—measures the similarity between two sets by dividing the size of the intersection of the sets by the size of the union of the sets. The Jaccard similarity between two sets A and B is given by:

$$\text{jaccard_similarity}(A, B) = |A \cap B| / |A \cup B|$$

Dice coefficient—measures the similarity between two sets by dividing twice the size of the intersection of the sets by the sum of the sizes of the sets. The dice coefficient between two sets A and B is given by:

$$\text{dice_coefficient}(A, B) = 2|A \cap B| / (|A| + |B|)$$

New machine learning-based similarity functions

This study presents the use of new similarity functions. Such functions are based on machine learning techniques. The main idea of such an approach is to use the decision function of a classifier as an input to the cosine similarity measure. The decision function of a classifier provides a measure of confidence or certainty for each prediction made by the classifier. When using the decision function as an input to the cosine similarity

measure, we can assess the similarity between instances based on their predicted probabilities or distances from the decision boundaries.

In the case of binary classification, where there are only two possible classes, the decision function or predicted class probabilities provide information about the confidence or likelihood of a sample belonging to each class. By comparing the decision function scores or the predicted probabilities between two samples, we can estimate their similarity. For example, if two samples have similar probabilities for the positive class and dissimilar probabilities for the negative class, it indicates a higher similarity between them in terms of class membership.

For multiclass classification, where there are more than two classes, the predicted class probabilities represent the likelihoods of a sample belonging to each class. Similar to binary classification, we can use these probabilities to estimate the similarity between two samples.

Using the classifier's decision function or predicted class probabilities can provide a more nuanced and informative measure of similarity compared to directly using the raw feature vectors, especially when dealing with complex classification problems. The classifier has learned from the training data and can capture meaningful relationships among the classes, leading to better similarity estimation. In this study, we used the decision functions of three classifiers, namely, *logistic regression*, *random forest classifier*, and *Gaussian Naive Bayes classifier*.

Logistic regression estimates the probability of an instance belonging to a specific class. During training, logistic regression seeks to find the optimal coefficients that minimize the disparity between the predicted probabilities and the actual class labels within the training data. Typically, this involves leveraging methods such as maximum likelihood estimation. The learned coefficients offer insights into the impact of each feature on the predicted probability.

Random forest classifier is an ensemble learning algorithm widely employed for classification tasks due to its robustness and accuracy. It harnesses the power of decision trees by creating a collection of trees and combining their predictions to make reliable classifications.

Gaussian Naive Bayes classifier is a probabilistic algorithm which applies Bayes' theorem with the feature independence assumption, making it suitable for continuous input features following a Gaussian distribution. During training, the algorithm estimates the probability distribution for each class by calculating the mean and standard deviation of each feature within the class.

Rule-based reasoning (RBR) system

In addition to the CBR system, we developed a rule-based expert system that relies on a set of rules to make diagnoses. Decision tree (DT) algorithms are used to extract rules from a dataset for constructing a rule-based expert system. The basic idea is to learn a set of decision rules from the data that can be used to make decisions on new instances.

A DT algorithm works by recursively partitioning the input space based on the values of different attributes. This process creates a tree-like structure, where each node corresponds to a test on an attribute, and each leaf node corresponds to a class label or a decision. Once the decision tree has been constructed from the dataset, the rules

can be extracted by traversing the tree from the root node to the leaf nodes. Each path from the root node to a leaf node corresponds to conditions that must be satisfied to reach that node. These conditions can be expressed as rules as “if–then” statements.

One advantage of using a DT algorithm to extract rules for a rule-based ES is that it can handle noisy and incomplete data. The DT algorithm can identify the most important attributes for making decisions and can also handle missing or inconsistent values by inputting missing values or treating them as a separate category. Another advantage is that the resulting rules are transparent and easy to understand, which can be helpful for domain experts who need to validate or modify the rules. Furthermore, the DT algorithm can also select the most relevant features for the classification task. This ability can reduce the input space’s dimensionality and improve the resulting rules’ accuracy and interpretability.

Results and discussion

In this section, the experimental results and their analysis are presented. To evaluate the performance of the expert systems, we used a range of metrics, including accuracy, precision, recall, and F1-score to compare the performance of the systems with each other and also with methods presented in literature.

We conducted experiments on various medical datasets, including the Heart Disease dataset [15], The Breast Cancer Coimbra dataset [16], The Early-Stage Diabetes Risk Prediction dataset [17], and The WISDM Smartphone and Smartwatch Activity and Biometrics dataset [18]. The datasets are obtained from publicly available sources and are preprocessed to remove noise and missing values. The description of each dataset is as follows:

- *Heart Disease Dataset:* This dataset classifies the presence of heart disease in a certain patient. The dataset can be used to predict which patients are more likely to suffer from heart disease given the indicated features. Some of the features that the dataset uses are resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate achieved, and others.
- *Breast Cancer Coimbra Dataset:* The goal of this dataset is to indicate the presence or absence of breast cancer in the patient. The predictors are anthropometric data and parameters. Classification models built using this dataset can be used as a biomarker of breast cancer.
- *The Early-Stage Diabetes Risk Prediction dataset:* This dataset is used to classify whether a person has diabetes. This dataset can help in predicting diabetes disease in patients at an early stage and hence can help in early treatment of the disease and avoiding its complications. The features include polyuria, sudden weight loss, visual blurring, partial paresis, and others.
- *WISDM Smartphone and Smartwatch Activity and Biometrics dataset:* This dataset includes accelerometer and gyroscope time-series sensor data which are gathered from smart devices such as phones and watches. This dataset is used for activity recognition and in building behavioral biometric models.

Table 1 Results of CBR and DT RBR on heart disease dataset

Approach	Accuracy	Precision	Recall	F1-Score
Cosine similarity	100%	100%	100%	100%
Euclidean distance	98.54%	97%	100%	99%
Manhattan distance	98.54%	97%	100%	99%
Pearson correlation coefficient	100%	100%	100%	100%
Jaccard similarity	84.39%	86%	82%	84%
Dice coefficient	84.39%	86%	82%	84%
Logistic regression similarity	100%	100%	100%	100%
Random forest similarity	100%	100%	100%	100%
GaussianNB similarity	98%	97%	100%	98%
DT RBR	100%	100%	100%	100%
HRFLM [8]	88.4%	90.1%	92.8%	90%
SVM approach [9]	92.3%	92.55%	92.43%	N/A

Table 2 Results of CBR and DT RBR on the breast cancer coimbra dataset

Approach	Accuracy	Precision	Recall	F1-Score
Cosine similarity	62.5%	62%	67%	64%
Euclidean distance	54.167%	57%	33%	42%
Manhattan distance	58.33%	62%	42%	50%
Hamming distance	50%	50%	100%	67%
Pearson correlation coefficient	62.5%	62%	67%	64%
Jaccard similarity	50%	50%	100%	67%
Dice coefficient	50%	50%	100%	67%
Logistic regression similarity	71%	71%	71%	71%
Random forest similarity	92%	92%	92%	92%
GaussianNB similarity	75%	78%	75%	74%
DT RBR	83%	88%	81%	82%
KMC AdaBoost [7]	91.37%	91.9%	91.4%	91.4%

In the results tables CBRS variations are referred to with the name of the used similarity function and the RBRS based on decision tree extracted rules is referred to as DT RBRS. The comparisons of approaches' results using the Heart Disease Dataset, Breast Cancer Coimbra Dataset, Early-Stage Diabetes Risk Prediction Dataset, and WISDM Smartphone and Smartwatch Activity and Biometrics Dataset are presented in Tables 1, 2, 3, and 4, respectively. The tables also include the results of other approaches found in literature for comparison purposes. The best value of each metric is written in bold.

The results of the approaches on the Heart Disease Dataset indicate that the two proposed expert systems achieved the best results. For the case-based reasoning system, it is found that cosine similarity, Pearson correlation coefficient, logistic regression similarity, and random forest similarity are the similarity measures that are best suited for such problems.

The results of the approaches on breast cancer coimbra dataset indicate that the best accuracy, precision, and F1-score were achieved using case-based reasoning systems using the random forest similarity measure. It is also noted that KMC AdaBoost

Table 3 Results of CBR and DT RBR on the early-stage diabetes risk prediction dataset

Approach	Accuracy	Precision	Recall	F1-Score
Cosine similarity	95.19%	91%	94%	92.5%
Euclidean distance	91.35%	79%	100%	88%
Manhattan distance	92.31%	80%	100%	89%
Hamming distance	50.96%	47%	51%	49.25%
Pearson correlation coefficient	92.31%	84%	94%	89%
Jaccard similarity	94.23%	89%	94%	91.5%
Dice coefficient	94.23%	89%	94%	91.5%
logistic regression similarity	91.35%	82%	94%	87%
random forest similarity	99.04%	97%	100%	98.5%
GaussianNB similarity	94.23%	86%	97%	91%
DT RBR	96%	95%	96%	95%
FFNN-TSA [10]	96.15%	95.86%	96.03%	N/A
FFNN-CSA [10]	99.04%	99.03%	99%	N/A

Table 4 Results of CBR and DT RBR on the WISDM smartphone and smartwatch activity and biometrics dataset

Approach	Accuracy	Precision	Recall	F1-Score
Cosine similarity	43.98%	32%	32%	32%
Euclidean distance	48.43%	34%	34%	34%
Manhattan distance	50.86%	38%	37%	37%
Hamming distance	25.28%	15%	12%	13%
Pearson correlation coefficient	43.88%	32%	32%	32%
Jaccard similarity	54.19%	50%	46%	45%
Dice coefficient	54.19%	50%	46%	45%
Logistic regression similarity	73.21%	55%	49%	47%
Random forest similarity	89.48%	88%	85%	86%
GaussianNB similarity	75.83%	70%	69%	68%
DT RBR	86%	84%	83%	84%
CNN [19]	N/A	86.8%	N/A	70.25%
ConvLSTM [19]	N/A	83.8%	N/A	69.6%

[7] accuracy, precision, and F1-score are in the second position. The best recall is achieved by the case-based reasoning system using Hamming distance, Jaccard similarity, and dice coefficient similarity measures.

The results of the approaches on the Early-Stage Diabetes Risk Prediction Dataset indicate that the case-based reasoning system using the random forest similarity measure and FFNN-CSA [10] achieved the best accuracy values. FFNN-CSA [10] achieved the best recall. It is also noted that the case-based reasoning system recall is in the second position. The best recall is achieved by the case-based reasoning system using Euclidean distance, Manhattan distance, and random forest similarity measures. The case-based reasoning system using random forest similarity achieved the best F1 score.

The results of the approaches on the WISDM dataset indicate that the case-based reasoning system using random forest similarity measure achieved the best values for the metrics.

In medical diagnosis, datasets often need to be balanced. Imbalance means that the number of cases with a particular diagnosis is much smaller than the number of cases with other diagnoses. This imbalance can lead to problems with the performance, which may become biased towards the majority classes. To address this problem, we used oversampling techniques to balance the datasets. Specifically, we used the Synthetic Minority Oversampling Technique (SMOTE) algorithm to generate synthetic examples of the minority class, which are then added to the training set. We evaluated the impact of oversampling on the performance of CBR and RBR systems and compared it with the system's performance without oversampling. As the WISDM dataset is an imbalanced dataset, the SMOTE is applied to it in order to balance the dataset. Table 5 presents the results of WISDM before and after applying SMOTE. As shown in Table 5, it can be concluded that most of the approaches show an enhancement in the results after applying SMOTE.

In order to assess the effectiveness of the three new similarity measures which are based on three machine learning classifiers, namely, logistic regression, random forest and GaussianNB, we compare the results of those three machine learning classifiers when their output is used directly to classify the examples and when their output is used as an input to the similarity function in order to get the closest similar example. The results of the Breast Cancer Coimbra and the Early-Stage Diabetes Risk Prediction datasets are found in Tables 6 and 7, respectively. All the results for the Heart Disease Dataset are around 100% for both classifier and classifier-based similarity measures.

The results show that using the classifier-based similarity functions is more effective than using the classifiers directly. It is found that random forest similarity achieved the best results on the Breast Cancer Coimbra Dataset while it achieved the best accuracy and recall on the Early-Stage Diabetes Risk Prediction Dataset. In addition, when comparing the machine learning classifier with its respective classifier-based similarity in Tables 6 and 7, it is found that in most cases the classifier-based similarity functions achieve better recall. Recall measures how many actual positives are predicted as positive while precision measures how many of the predicted positives are true positives. Higher recall means that the system outputs less false negatives. False negatives

Table 5 Comparing the results of CBR and DT RBR on WISDM smartphone and smartwatch activity and biometrics dataset before and after SMOTE

Approach	Accuracy	Precision	Recall	F1-Score
Dice coefficient (before SMOTE)	54.19%	50%	46%	45%
Jaccard similarity (before SMOTE)	54.19%	50%	46%	45%
Logistic regression similarity (before SMOTE)	73.21%	55%	49%	47%
Random forest similarity (before SMOTE)	89.48%	88%	85%	86%
GaussianNB similarity (before SMOTE)	75.83%	70%	69%	68%
DT RBRS (before SMOTE)	86%	84%	83%	84%
Dice coefficient (after SMOTE)	65.54%	57%	50%	53%
Jaccard similarity (after SMOTE)	61.23%	65.17%	49.41%	51%
Logistic regression similarity (after SMOTE)	64.00%	61%	54%	57%
Random forest similarity (after SMOTE)	94.19%	94%	93%	93%
GaussianNB similarity (after SMOTE)	70.98%	69%	66%	67%
DT RBRS (after SMOTE)	89%	87%	88%	86%

Table 6 Results of classifier and classifier-based similarity measures on the breast cancer coimbra dataset

Approach	Accuracy	Precision	Recall	F1 score
Logistic regression classifier	71%	72%	72%	71%
Random forest classifier	90%	89%	91%	90%
GaussianNB classifier	75%	78%	75%	74%
Logistic regression similarity	71%	71%	71%	71%
Random forest similarity	92%	92%	92%	92%
GaussianNB similarity	75%	78%	75%	74%

Table 7 Results of classifier and classifier-based similarity measures on the early-stage diabetes risk prediction dataset

	Accuracy	Precision	Recall	F1 score
Logistic regression classifier	91%	92%	90%	91%
Random forest classifier	99%	99%	99%	99%
GaussianNB classifier	91%	90%	90%	90%
Logistic regression similarity	91.35%	82%	94%	87%
Random forest similarity	99.04%	97%	100%	98.5%
GaussianNB similarity	94.23%	86%	97%	91%

in medical diagnosis can lead to delayed treatments and hence can cause further medical complications. Therefore, higher recall is preferable than higher precision in medical diagnosis where early treatment is crucial.

Overall, the results suggest that the random forest similarity function is the most accurate function for the given datasets. SMOTE can be used to improve the results. Finally, the RBR system is also a good choice for tasks where the knowledge base is represented as rules since it achieves comparable results.

Conclusions

This paper presented two types of expert systems for medical diagnosis: Case-based and Rule-based expert systems. The two systems were evaluated on four different datasets, and they achieved competitive results when compared to other approaches in literature. The results of this study suggest that expert systems can be a valuable tool for medical diagnosis. The performance of similarity functions in a case-based reasoning system for medical diagnosis was investigated, and their impacts on accuracy, precision, recall, and F1 score were evaluated. The study found that the choice of a similarity function significantly affected the system's results. The random forest similarity function was found to be the most accurate function for the given datasets. In addition, RBR system achieved comparable results making it a good choice for tasks where the knowledge base is represented as rules. The study highlights the importance of carefully selecting a similarity function when building a case-based reasoning system for medical diagnosis. Overall, the findings of this study can contribute to the development of more accurate and efficient expert systems for medical diagnosis.

Abbreviations

CBR	Case-based reasoning
CBRS	Case-based reasoning system
CNN	Convolutional neural network
convLSTM	Convolutional long short-term memory
DT	Decision tree
GaussianNB	Gaussian Naive Bayes
HRFLM	Hybrid random forest with a linear model
MAD	Median absolute deviation
RBR	Rule-based reasoning
RBRS	Rule-based reasoning system
SMOTE	Synthetic minority oversampling technique
SVM	Support vector machine
WISDM	Wireless sensor data mining

Acknowledgements

Not applicable.

Authors' contributions

EM contributions: methodology, software, validation, resources, data curation, and writing—original draft preparation. MS contributions: methodology, software, validation, resources, data curation, and writing—original draft preparation. LR contributions: conceptualization, methodology, formal analysis, investigation, supervision, and writing—review and editing. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets used during the current study are available in the UCI machine learning repository, <https://archive.ics.uci.edu/>.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 23 August 2023 Accepted: 7 November 2023

Published online: 14 November 2023

References

- Fawzi R, Ghazy M, Rizkallah LW (2022) Designing knowledge-based systems for COVID-19 diagnosis. In: et al. Hybrid Intelligent Systems. HIS 2021. Lecture Notes in Networks and Systems, vol 420. Springer, Cham. https://doi.org/10.1007/978-3-030-96305-7_7
- El-Sayed R, Seddik S, Rizkallah LW (2022) Expert systems in academic advising. In: Hassanien, A.E., Snášel, V., Chang, K.C., Darwish, A., Gaber, T. (eds) Proceedings of the international conference on advanced intelligent systems and informatics 2021. AISI 2021. Lecture Notes on Data Engineering and Communications Technologies, vol 100. Springer, Cham. https://doi.org/10.1007/978-3-030-89701-7_18
- Galal O, Nasr A, Rizkallah LW (2023) A rule learning approach for building an expert system to detect network intrusions. *Int J Intell Comput Inform Sci* 23(1):106–114
- Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN (1975) Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Comput Biomed Res* 8(4):303–320
- Miller RA, McNeil MA, Challinor SM, Masarie FE Jr, Myers JD (1986) The INTERNIST-1/quick medical REFERENCE project—Status report. *West J Med* 145(6):816
- Barnett GO, Cimino JJ, Hupp JA, Hoffer EP (1987) DXplain: an evolving diagnostic decision-support system. *JAMA* 258(1):67–74
- Polat K, Sentürk U (2018) A Novel ML Approach to prediction of breast cancer: combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier. In 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE, Ankara, p 1–4
- Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 7:81542–81554
- Alotaibi FS (2019) Implementation of machine learning model to predict heart failure disease. *Int J Adv Comput Sci Appl* 10:6
- Yasar A (2021) Data classification of early-stage diabetes risk prediction datasets and analysis of algorithm performance using feature extraction methods and machine learning techniques. *Int J Intell Syst Appl Eng* 9(4):273–281
- Salem ABM, Roushdy M, HodHod RA (2005) A case based expert system for supporting diagnosis of heart diseases. *AIML Journal* 5(1):33–39
- Mazhar T, Nasir Q, Haq I, Kamal MM, Ullah I, Kim T, Alwadai N (2022) A novel expert system for the diagnosis and treatment of heart disease. *Electronics* 11(23):3989

13. Aguilera-Venegas G, Roanes-Lozano E, Rojo-Martínez G, Galán-García JL (2023) A proposal of a mixed diagnostic system based on decision trees and probabilistic experts rules. *J Comput Appl Math* 427:115130
14. Prasath VB, Alfeilat HAA, Hassanat A, Lasassmeh O, Tarawneh AS, Alhasanat MB, Salman HSE (2017) Distance and similarity measures effect on the performance of K-nearest neighbor classifier--a review. arXiv preprint arXiv:1708.04321
15. Cleveland Heart Disease Data Set. (1988). UCI machine learning repository. Retrieved from <https://archive.ics.uci.edu/ml/datasets/heart+disease>
16. Cortez P, Silva C (2008) Breast cancer Coimbra data set. UCI Machine Learning Repository. Retrieved from <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra>
17. Islam MF, Alam MS, Hasan MM (2020) Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis*. Springer, Singapore, pp 113–125
18. Weiss GM (2019) Wisdm smartphone and smartwatch activity and biometrics dataset. UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set 7:133190–133202
19. Oluwalade B, Neela S, Wawira J, Adejumo T, Purkayastha S (2021) Human activity recognition using deep learning models on smartphones and smartwatches sensor data. arXiv preprint arXiv:2103.03836

Publisher's Note

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

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
