PENINSULA PUBLISHING LLC

EDRAAK

Vol. **(2024)**, 2024, **pp**. 39–42
ISSN: 3078-8412

Research Article

A Comprehensive Review of Data Mining Techniques for Diabetes Diagnosis Using the Pima Indian Diabetes Dataset

Hadeel M Saleh[1,*]

[1] *Center For Continuing Education*

ARTICLE INFO

Article History

Received 3 Jan 2024

Revised: 28 Feb 2024
Accepted 28 Mar 2024
Published 15 Apr 2024

Keywords

Diabetes diagnosis,

data mining,

Pima Indian Diabetes dataset,

classification algorithms,

machine learning.

ABSTRACT

Diabetes is a major global health concern, and early diagnosis is crucial for effective management and prevention of complications. This paper presents a comprehensive review of various data mining techniques applied to the diagnosis of diabetes, specifically using the Pima Indian Diabetes dataset. The Pima Indian dataset, a widely used benchmark in diabetes research, contains information on various health-related features such as age, body mass index, insulin levels, and glucose concentration, among others, which are crucial for predicting the onset of diabetes. The review explores a range of classification algorithms, including decision trees, support vector machines (SVM), logistic regression, k-nearest neighbors (KNN), and artificial neural networks (ANNs), discussing their performance, strengths, and limitations in predicting diabetes.

In order to increase the models' accuracy and efficiency, the paper also emphasizes the importance of preprocessing procedures like feature selection, data cleaning, and normalization. It also contrasts the evaluation metrics accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) that are used to evaluate the performance of these models. We hope that this review will shed light on the best data mining methods for diagnosing diabetes, with a focus on scalability, interpretability, and model optimization.

The findings of this review suggest that while traditional techniques like decision trees and logistic regression are effective, more complex models such as support vector machines and neural networks tend to yield higher prediction accuracy. However, the trade-off between model complexity and interpretability remains a key challenge in the deployment of these techniques for clinical decision-making. The paper concludes by suggesting future directions for improving diabetes diagnosis through the integration of advanced machine learning methods and big data analytics..

1. INTRODUCTION

Diabetes A group of disorders known as diabetes are characterized by elevated blood glucose levels that arise from problems with the production, function, or both of insulin [1].

Diabetes is usually considered as a modern society illness. Some of the contributing criteria to that disease are widely spread rareness of continuous exercising, increasing rates of obesity, and more munched foods. A great amount of resources is poured in the direction to control this disease. However the average of its happening continues to increase greatly every day. It cannot exist of being single; it even causes some other serious complications like failure of kidneys, detachment of retina, neuropathy, and Cardio-vascular problems, non-traumatic lower-limb amputations and blindness [2].

People of ages more than 65 years are the main victims of this disease. Correct and properly timed measurements must be done in order to control it. All those reasons make this disease one of the basic most important topics in the medical science research field. Association, clustering and classification approaches of data mining are useful tools in this procedure. Complete knowledge of different causes that result in the diabetes is crucial prior to building predictive structures [3, 4].

In addition to being useful for diagnosis and treatment, medical data may also be useful for health care researchers' own education. An important and fascinating aspect of the classification is the interpretation and analysis of medical data [5].

There is a lot of medical data that could be helpful in the healthcare industry; data mining can be used to analyze medical facilities for more adequate sources, to quickly detect and prevent diseases, and to avoid high costs due to unnecessary and expensive medical testing [6]. Many scientists use a variety of data mining techniques to diagnose and treat a wide range of illnesses, including diabetes [7], stroke [8], heart disease [9], and cancer [10].

*Corresponding author email: Haddeel.mohammed@uoanbar.edu.iq
DOI: https://doi.org/10.70470/EDRAAK/2024/006

The process of extracting information from massive databases is known as data mining. There is a lot of data in the medical system. Although data mining on medical records has received a lot of attention, its application is still in its infancy [11]. Classifying and predicting medical data is one of the most crucial data mining techniques, and it greatly contributes to the knowledge extraction process from the existing database. This kind of data analysis could be applied to forecast future data trends or to extract models that characterize important data classes [12].
ANN, genetic algorithms, Bayesian, decision tree, and other classification techniques have been proposed and are being evaluated over the last decade [13].

2. MEDICAL CLASSIFICATION PROBLEM

Data mining has been commonly implemented in several organizations and fields. Whenever mentioning medical diagnosis, data mining is starting to become quite well known. Given its significance in this field and the difficulties it presents, medical data mining has emerged as one of the most important topics in data mining in recent years [14].
Designing any system in the field of medical diagnosis would be very helpful because it is regarded as a very important task that needs to be done correctly and sufficiently.
Not all physicians possess expertise in every subspecialty, and there are also insufficient resources regarding individuals in particular areas. Therefore, by combining them all, an automated medical diagnosis system would likely be very beneficial. Proper computer-based information and/or decision support systems are capable of aiding the achievement of clinical tests with less costs. Sufficient and precise implementing of automated system requires a comparative study of different approaches [15].
The most widely used and significant implementations in data mining possibly include predictive modeling. Classification indicates predicting a target variable which is categorical in its nature, like classifying healthy situations of people [16].
Typically, there are three various expected errors happen, those are:

1. The False -Positive case (FP) that indicates the individual that isn‘t a patient falsely classified as patient. False-positive can happen due to unnecessary processes, worries and financial expenses.
2. The False-Negative case (FN), this one indicates the case where a patient who is falsely diagnosed as healthy. False-negative could happen due to waste of time which might cause life loss or at least cases that are falsely treated.
3. The unclassifiable error, where the system isn‘t capable of classifying a case, probably this situation happens due to the lack of historic data, overlapped information, or probably due to inadequate algorithm.

3. LITERATURE REVIEW

There are several methods of developing in different areas for various reasons. In the following, a set of proposals have been derived from the scientific literature:
In 2004 A. Khan and K.Revett [17], described the way rough set theory is possible to be used as a means to analyze rather complicated decision tables such as the Pima Indian Diabetes Database (PIDD). Reporting using a genetic algorithm-based rough set method for classifying diabetes and it reached a success rate on the test-set equal to 83%.
Not all physicians possess expertise in every subspecialty, and there are also insufficient resources regarding individuals in particular areas. Therefore, by combining them all, an automated medical diagnosis system would likely be very beneficial.
A method based on the neuropathy type of diabetes was proposed by S. Sapna and D.A. Tamalrasi in 2009 [19]. Diabetic mellitus causes nerve disorders. Diabetic neuropathies can easily develop in long-term diabetic patients. There is a 50% chance of developing similar illnesses that affect multiple body nerve systems. For example, somatic nerves, which are found in the body wall and limbs, may also be affected. However, internal organs like the stomach, heart, and so forth are also referred to as automatic nerves.
In 2009 P.Jeatrakul and K.W.Wong [20], a study has been presented where a comparison of NN approaches for binary classification diagnosing of diabetes has been held. The classification performance reached by 5 various kinds of NNs, in other words Back Propagation NN (BPNN), Radial Basis Function NN (RBFNN), General Regression NN (GRNN), Probabilistic NN (PNN), and Complementary NN (CMTNN).
The effectiveness of the data mining approach in the field of health care applications was emphasized by Selvakuberan K and Harini B [21] in 2011. One of the primary causes of early illness and mortality worldwide is diabetes. The use of data mining techniques for diabetes detection in the PIMA Indian Diabetes Dataset (PIDD) was covered in this study. It has suggested a feature selection technique that combines the Ranker Search methodology. The classification's accuracy rate reached 81%, and this approach outperformed earlier approaches.
In 2011, Selvakuberan K and Harini B [21] highlighted the efficacy of the data mining approach in the field of health care applications.
Saba Bashir, 2 Usman Qamar, and 3A study by Farhan Hassan Khan [23] was proposed; it employed multiple categories of classification techniques for diabetes data sets. As base classifiers, three different types of decision trees ID3, C4.5, and CART are implemented. Stacking, Adaboost, Bayesian Boosting, Bagging, and Majority Voting are the ensemble techniques that are used. Two benchmark datasets from the BioStat and UCI repositories have been used.

When compared to single or other ensemble approaches, the evaluation and experiment results showed that the Bagging ensemble approach demonstrated more effective implementation.
The Pima India diabetes dataset from the UCI machine learning repository was used in a 2014 study by Aishwarya S1 and Anto S2 1 PG Scholar [24] that proposed a model for diagnosing diabetes.
The methods that have been used in this work are Extreme Learning Machine for classification and Genetic Algorithm (GA) for feature selection.
The suggested system's implementation has been analyzed based on a number of criteria, including classification accuracy, sensitivity, and specificity using matrix of confusion and 10-fold cross-validation. For the genetic algorithm, the proposed system's precision has increased to 89.54%.
In 2014, Ravi s., Smt.T. [25] conducted a study that tried to diagnose diabetes by analyzing the data bank and applying data mining techniques.
In this study, fuzzy C-Means and support vector machines were also implemented and evaluated on a set of medical data related to the diagnosis of diabetes. Fuzzy C-Means produced the best results, with a precision of 94.3% and a positive predictive value of 88.57%. The precision of the Support Vector Machine is 59.5%, which is not very high.
NNs are one of the soft computing techniques that can be used to make predictions about medical data, as noted by M. Duraraj [26] in 2015. They are referred to as universal predictors.
The disease known as diabetes mellitus, or diabetes, is brought on by elevated blood glucose levels. Diabetes can be diagnosed using a variety of conventional methods that rely on chemical and physical testing. Systems based on artificial neural networks (ANNs) can be used to predict the risk of high blood pressure. The data set is divided into one of the two subsets by this developed system. Doctors have been able to lower the risk of developing a serious illness by using soft computing techniques for early detection.
The Pima Indian Diabetic Set from the (UCI) Repository of Machine Learning data bases is the basis for the data set chosen for classification and simulation testing. A thorough survey on the use of different soft computing techniques for diabetes prediction was conducted in this study. The goal of this study was to identify and recommend an effective method for early disease prediction.

4. CONCLUSION

This review provides an overview of key classification techniques applied to diabetes diagnosis, showcasing their contributions to improving prediction accuracy and clinical decision-making. The use of decision trees, Naïve Bayes, and hybrid methods has demonstrated effectiveness in handling complex datasets like the Pima Indian Diabetes Dataset (PIDD). The findings highlight the necessity of robust preprocessing steps, such as data discretization and feature selection, to enhance model performance.
While each classification technique has its advantages, combining methods through ensemble approaches often yields superior results, as evidenced in several studies. Future research should focus on integrating diverse datasets, developing hybrid models, and leveraging deep learning techniques to further advance diabetes diagnosis. This review underscores the transformative potential of data mining in healthcare, offering insights for researchers and practitioners aiming to optimize diagnostic systems for chronic diseases like diabetes.

Funding:

This research was not funded by any institution, foundation, or commercial entity. All expenses related to the study were managed by the authors.

Conflicts of Interest:

The authors declare that there are no conflicts of interest to disclose.

Acknowledgment:

The authors wish to acknowledge their institutions for their instrumental support and encouragement throughout the duration of this project.

References

[1] National Diabetes Statistics, "National diabetes statistics," 2014. [Online]. Available: www.cdc.gov/diabetes.
[2] Center for Disease Control and Prevention, "National diabetes fact sheet," 2011. [Online]. Available: www.cdc.gov/diabetes.
[3] American Diabetes Association, "Diabetes information," [Online]. Available: www.diabetes.org/diabetes.
[4] I. Yoo, P. Alafaircet, M. Marinov, K. Pen-Hernandez, R. Gopidi, J. F. Chang, and I. Huda, "Data mining in healthcare and biomedicine: A survey of the literature," *Journal of Medical Systems*, 2011.
[5] P. Smitha, L. Shaji, and M. Mini, "A review of medical image classification techniques," in *Proc. International Conference on VLSI, Communication and Instrumentation (ICVCI)*, Karunagapally, 2011.
[6] D. C. J. Ruben, "Data mining in healthcare: Current applications and issues," 2009.

[7] T. Porter and B. Green, "Identifying diabetic patients: A data mining approach," in *Proc. Americas Conference on Information Systems*, 2009.
[8] S. Panzarasa, S. Quaglini, et al., "Data mining techniques for analyzing stroke care processes," in *Proc. 13th World Congress on Medical Informatics*, 2010.
[9] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Systems with Applications*, Elsevier, 2009.
[10] J. Naik and S. Patel, "Tumor detection and classification using decision tree in brain MRI," *International Journal of Engineering Development and Research*, 2013.
[11] Y. Mao, Y. Chen, et al., "Medical data mining for early deterioration warning in general hospital wards," 2014.
[12] M. Esmaeli, "A Scalable Parallel Algorithm for Decision Support from Multidimensional Sequence Data," Ph.D. dissertation, Dept. Comput. Sci., University of Debrecen, Debrecen, Hungary, 2011.