EVALUATION OF DIVERGENT DIABETES PREDICTION MODELS THROUGH COMPARATIVE ANALYSIS

K. Kanmani, Research scholar, Dr. Ambedkar Govt Arts College, Vyasarpadi, Chennai- 600039, India, Assistant professor, Department of Computer Applications, College of Science and Humanities, SRM Institute of Science and Technology, kattankulathur- 603203. Chennai, TN, India
Dr. A. Murugan, Associate Professor and Head, PG and Research Department of Computer Science, Dr. Ambedkar Govt. Arts College (Autonomous), Vyasarpadi, Chennai – 600039, TN, India

Abstract

One of the largest health issues affecting millions of people worldwide is diabetes. Diabetes that is not under control increases the risk of cancer, renal damage, heart attack, blindness, and other diseases. Healthcare providers must use diverging algorithms if they are to be more successful at diagnosing diseases. This research compared multiple algorithms that can detect diabetes risk early in order to enhance the medical diagnosis of diabetes. A clinical dataset from an authorised dataset is one of the real datasets that were analysed for this study. Here, numerous diabetes disease prediction systems that have been put into practise were discussed. Six proposed algorithms have been successfully used in the experimental study such as Decision Tree (DT), Optimized Support Vector Machine (OSVM), Point- Based Algorithm (PBA), Euclidean Distance with Manhattan Metric Measures (EDMM), Hybrid Pruning with MD Measures (HPMMM), Enhanced Sunflower Algorithm (ESA). The obtained results showed that the proposed method based on the ESA technique provides great performances with an accuracy of 98.8%. ESA is the most accurate Methodology with the uppermost accuracy rate with the balanced data set finally, this research makes it possible for us to accurately determine the level of prevalence and prognosis of diabetes.

Keywords: Diabetes, Decision Tree, HPMMM, Enhanced Sunflower

1.Introduction

High blood sugar levels are one of the main causes of diabetes. The body needs glucose to function properly. In people with type 2 diabetes, the lack of exercise and unhealthy lifestyles are also contributing factors. A large amount of sugar in the blood can lead to diabetes. The body cannot properly convert food into insulin, which results in the sugar remaining unabsorbed. This condition can affect various organs and blood vessels. Type 1 diabetes is a type of disease that affects children [1]. It can be caused by the destruction of the cells that produce insulin in the body. Also known as juvenile diabetes, this condition usually occurs after the age of 40. With proper diet and regular exercise, people with diabetes can manage their condition. Type 2 diabetes is referred to as insulin-dependent or insulin-free diabetes. This condition occurs when the body does not produce enough insulin to properly convert glucose into energy [2]. On the other hand, in type 1 diabetes, the body makes enough insulin to treat the condition. Gestational diabetes is a type of disease that occurs when the body does not produce enough insulin to properly convert glucose into energy. This condition can be triggered by the changes in hormones during pregnancy [3]. Moreover here, analysed the effectiveness of the DT, OSVM, PBA, EDMM, HPMMM, and ESA. The various steps involved in this process given below.

(i) Six Various Algorithms were used, namely, DT, OSVM, EDMM, PBA, HPMMM, and ESA
(ii) The goal of feature selection is to determine the most important variables.
(iii) For the purpose of finding the highest level of accuracy, data balancing is used.
(iv) Proposed a Methodology for Enhanced Sunflower Algorithm for Risk Prediction

2. Literature Review

The techniques used in this study are geared toward improving the accuracy of the diagnosis and categorizing the data. Machine learning is mainly focused on learning about the patterns in the data collected from the diabetes patient. In order to help individuals with diabetes manage their condition, healthcare systems offer various services. These include treatment plans, education, and support services.

One of the most common issues that the medical profession faces is diabetes. The goal of this study is to use divergent techniques to identify people with diabetes based on their clinical and personal information. The following section summarizes the various studies that were conducted on the use of machine learning in the diagnosis of diabetes. It is useful to note the areas of weakness and potential improvement in the treatment regimens of diabetic patients with machine learning.

In the previous section, Sun and Zhang et.al methodology about the various types of deep learning techniques that can be used to classify and improve the accuracy of the diagnosis of diabetes.[4-6]. Qawqzeh et al. analysed the data collected from 459 individuals with diabetes. They found that the classification accuracy of the model was 92%, but it could not be validated due to its lack of comparison with other models.[7]

Tafa et al. divided the collected data into two groups: a training set and a testing set. They used a combination of machine learning and naive Bayes for their proposed model. The data collected by the researchers included eight attributes. Out of these, 80 patients were type-2 diabetics. The combination of the two techniques performed well in achieving an accuracy of 97.6%. Karan et al. A new method for diagnosing diabetes using a dispersed end-to-end system architecture that utilizes AI neural network computing techniques [8].

3. Methodology

3.1 Decision Tree

The concept of a decision tree is a representation of a possible outcome of a decision, which can be accomplished through the use of R programming. It is commonly used in the healthcare industry to perform quality assessments and solve various tedious tasks. After the process for collecting the data, we used the R programming language to create a decision tree. R- programming which is used for developing a decision tree. Figure 1 illustrates the decision tree's implementation for predicting the clinical presentation of diabetes. The root node, which is the insulin level, determines if a person is affected by the condition or not. The other nodes, such as age, BMI, and skin thickness, are used to test the patient's condition. If the glucose level is good, the tree predicts that there is no diabetes. On the other hand, if it is below the constraints, it will perform the next test.

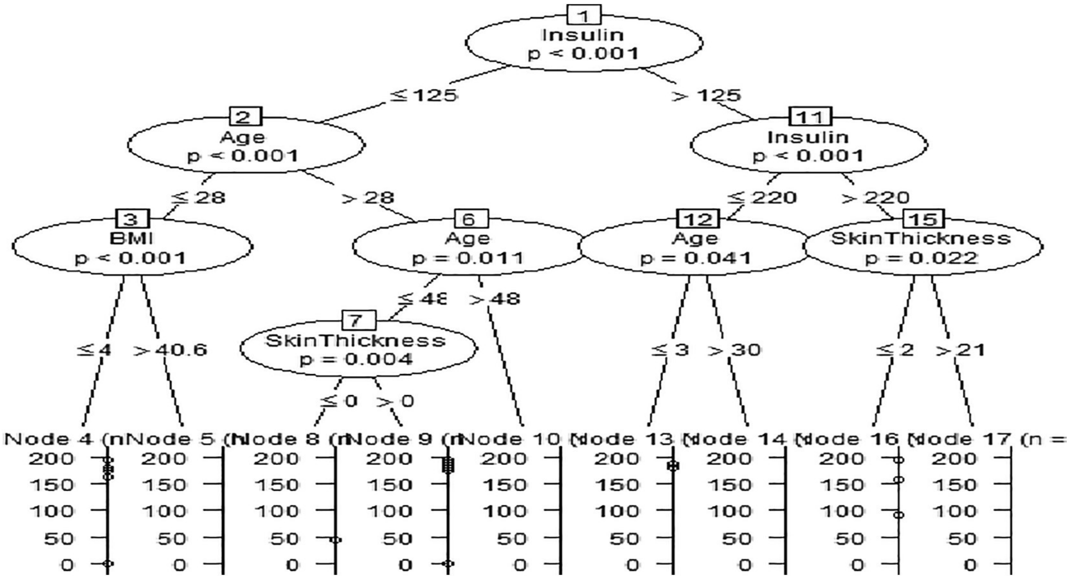

Figure1: Implementation of Decision Tree using R - Programming

Confusion matrix measurement analysis

In the context of this study, cases that are actually non-diabetes (negative) or diabetes (positive), respectively, but were categorised as either of the two groups, X = non-diabetes (negative), or Y = diabetes, are represented by the letters X, Y, in the confusion matrix (positive). ‘‘The number of cases that were incorrectly labelled as positive (diabetes) and equals is known as the false-positive value. Pre-processed data is utilised to improve the calibre of the outcomes. It displays a confusion matrix and an accuracy value in addition to the sample code used in the development of the decision tree. Additionally, practical use yields a rating of 78% accuracy.

3.2 Optimized Support Vector Machine Algorithm

OSVM technique was used with the goal of removing outliers by learning from the pre-processing of the data. For classification systems, learning from unbalanced and aberrant data presents significant challenges. This study's initial phase was devoted to figuring out the best strategy to deal with missing data. The effectiveness of various data imputation techniques was examined in the second phase. This approach combines outlier detection and removal strategies with the Optimal Support Vector Machine (OSVM) to increase the precision of classification systems. It accomplishes this by accurately mapping the features of the data.

Data mining methodology using an Optimized SVM model exhibits an accuracy value of 86%. Outlier detection method is used to remove instances that were incorrectly categorised for more effective data categorization. When compared to the prior SVM method (without outlier detection), which had a 78% success rate, this result is noticeably superior. The scientific community has accepted outlier detection, and it is altering how research is conducted. Many applications in healthcare and other industries can be made use of because to its capacity to gather and analyse enormous volumes of data.

3.3 & 3.4. Point Based Algorithm & Euclidian distance with Manhattan Metric Measures

The experimental results showed that combining Manhattan Metrics with the Point Based Algorithm improves accuracy. R programming is used to implement this PBAMM, which offers accuracy of 92.5 percent. Thus, experimental results demonstrated that the created prediction model performed reasonably well at predicting the occurrence ofT2D in the global population using a limited set of characteristics and the PBAMM. The model can give doctors and patients useful insight on the likelihood ofT2D in the future, enabling patients to take proactive steps to reduce their risk for the disease's onset, progression, and complications.

3.5 Hybrid Pruning with Manhattan Metric Measures

The HPMMM approach is founded on the idea that medical practitioners should take a patient's unique health problems into account when making a diagnosis. For the experiment, two open data sets with a larger number of entries for each set, the Pima Indians Diabetes dataset and the Mendeley Data for Diabetes dataset were examined. Precision, specificity, and accuracy were used to assess the prediction performance produced using the HPMMM approach. Implementation of this HPMMM using R programming which provides 97% accuracy.

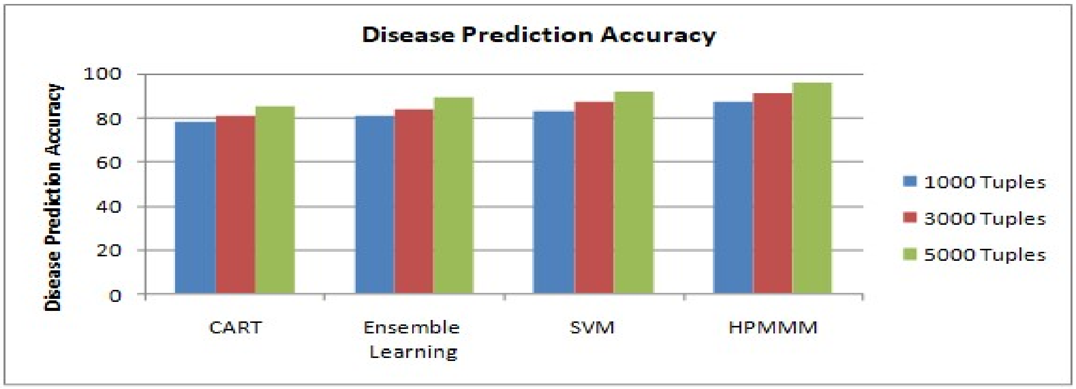

Figure 2: Analysis on Disease Prediction Accuracy

The performance in disease prediction produced by various approaches were measured and presented in Figure 2. The proposed HPMMM algorithm has produces higher accuracy than other methods.

3.6 Enhanced Sunflower Algorithm

The Enhanced Sunflower Optimization Custering algorithm initializes k number of clusters based on the number of disease classes considered. Also, the method initializes different parameters. For each tuple available a sunflower is generated and initialized with the features of the tuple. Third for each tuple available in the data set, the method computes Multi Feature Morality Value (MFMV) and Multi Feature Polination Rate (MFPR) towards different sunflowers present in different clusters. According to the morality value, the method identifies the group for a tuple and indexed. Fourth, the method generates initial population from each cluster of sunflowers, and computes MFMV and Multi Feature Polination Rate (MFPR) using the objective function towards various sunflower groups. According to the value of MFMV, the method assigns the populations to new clusters. This is iterated for all the populations.

Considering that type 1 diabetes and type 2 diabetes have many differences in their treatment methods, this method will help to provide the right treatment for the patient. A comparison is also shown in each case. The highest accuracy obtained was around 97% for Dataset 1, after employing NAN removal and it was around 98.8 % for Dataset 2, after using the Sunflower Algorithm with pollination.

4. Result of all Implemented Algorithms

S. No	Methodology	Accuracy
1.	Attribute Cleansing Decision Tree Construction	78%
2.	Optimized SVM algorithm with Outlier detection and removal	86%
3.	Portending Diabetes using Euclidian distance with Manhattan Metric Measures	92.5%
4.	Diabetes Prediction using Point Based Algorithm	96%
5.	Hybrid Pruning with MMD Measures	97%
6.	Prediction of Diabetes using Enhanced Sunflower Algorithm (Multiple Dataset) Data set 1: (Which has around 6145 data) Data set 2: (Which has around 2000 data)	98% 97% 98.8%

Table 1: Applied Algorithms with Results

The accuracy of the classifiers was evaluated following dataset balancing. The Results of various implemented algorithms are given in Table 1. It has been suggested that using contemporary algorithms for anticipatory prediction can help to reduce the upward diabetes tendency. For this reason, six algorithms including DT, OSVM, PBA, EDMM, HPMMM, and ESA were used.

5. Conclusion

Diabetes is a severe, ongoing disease. If diabetes is identified early enough, a more successful course of treatment may be possible. In order to predict a patient's diabetic condition at the earliest possible stage, this research also compares different classification models based on divergent algorithms. Of all these algorithms, it has been assessed that ESA has the highest accuracy of 98.8%. The best and most effective diabetes prediction algorithm can be found by comparing a wide range of algorithms and algorithm combinations.

6. References

[1]. B. S. Kim, M. Ahn, W.-W. Cho, G Gao, J. Jang, and D.-W. Cho, “Engineering of diseased human skin equivalent using 3D cell printing for representing pathophysiological hallmarks of type 2 diabetes in vitro,” Biomaterials, vol. 272, Article ID 120776, 2021.

[2]. K. E. Minges, P. Zimmet, D. J. Magliano, D. W. Dunstan, A. Brown, and J. E. Shaw, “Diabetes prevalence and determinants in Indigenous Australian populations: a systematic review,” Diabetes Research and Clinical Practice, vol. 93, no. 2, pp. 139–149, 2011.

[3]. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting diabetes mellitus with machine learning techniques,” Frontiers in Genetics, vol. 9, p. 515, 2018.

[4] Y. L. Sun and D. L. Zhang, “Machine Learning Techniques for Screening and Diagnosis of Diabetes: A Survey,” Technical Gazette, vol. 26, pp. 872–880, 2019.

[5] S. Malik, S. Harous, and H. El-Sayed, “Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women,” Modelling and Implementation of Complex Systems, Springer, in Proceedings of the International Symposium on Modelling and Implementation of Complex Systems, pp. 95–106, October 2020.

[6] Y. K. Qawqzeh, A. S. Bajahzar, M. Jemmali, M. M. Otoom, and A. Thaljaoui, “Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modelling,” BioMed Research International, vol. 2020, Article ID 3764653, 2020.

[7] Z. Tafa, N. Pervetica, and B. Karahoda, “An Intelligent System for Diabetes Prediction,” IEEE Explore, in Proceedings of the 4th Mediterranean Conference on Embedded Computing (MECO), pp. 378–382, Budva, Montenegro, June 2015.

[8] O. Karan, C. Bayraktar, H. Karlık, and B. Karlik, “Diagnosing diabetes using neural networks on small mobile devices,” Expert Systems with Applications, vol. 39, no. 1, pp. 54–60, 2012.