MULTIDISCIPLINARY SCIENCE JOURNAL

RESEARCH ARTICLE
Published Online: September 10, 2023
https://doi.org/10.31893/multiscience.2024031

A novel ensemble machine learning framework for early stage diabetes mellitus prediction

Anisha Nagpal[a] | **Munish Sabharwal**[b] | **Rohit Tripathi**[c]

[a]Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India.
[b]School of Computing Science and Engineering, Galgotias University, Greater Noida, India.
[c]Department of Electronics Engineering, JC Bose University of Science and Technology, India.

Abstract Diabetes Mellitus is a harmful condition characterized by elevated blood sugar levels resulting from insufficient insulin production in the body. This disorder gives rise to various health complications affecting the kidneys, heart, nerves, and eyes. If left unidentified and untreated, it can even prove fatal. Therefore, early detection and prediction of chronic diseases are imperative, and advancements in technology have led to a shift towards personalized healthcare. Machine learning offers a viable solution for predicting the likelihood of developing the disease with satisfactory accuracy.The objective of this research is to create a machine learning-based model for accurately predicting early-stage diabetes mellitus. The study employs PCA for dimensionality reduction and the ensemble bagging decision tree classification, thereby, achieving a remarkable accuracy level concerning the disease. The proposed model is evaluated using a publicly available dataset of 520 instances. The dataset utilized in this research includes information such as Polyuria, Polydipsia, sudden_weight_loss and weakness etc. The model achieves an impressive accuracy of 93.26%, precision of 96.9%, recall of 94.1%, and F-Score of 95.4%. Comparative analysis against other existing techniques demonstrates its superior performance in predicting such chronic conditions.

Keywords: Diabetes Mellitus, machine learning, ensemble technique, bagging, decision tree, chronic disease

1. Introduction

Diabetes Mellitus is a critical medical condition caused by an insufficient amount of the hormone insulin in the body, which leads to disturbances in metabolism and affects blood sugar levels (Tomic 2022). It encompasses two main categories: type 1 and type 2 diabetes. Type 1 diabetes occurs when the body fails to produce enough insulin, while type 2 diabetes arises when the body's ability to absorb insulin is impaired.The consequences of diabetes can be severe and life-threatening, including cardiovascular disease, diabetic ketoacidosis, stroke, and nonketotic hyperosmolar conditions (Gabbay 1975; Dobbs 1975). As a result, preventing and safeguarding against this chronic disease has become a paramount concern for public health.

According to the World Health Organization (WHO), approximately 422 million people worldwide suffer from diabetes, making it a significant cause of global mortality (World Health Organization, Global Action Plan on Physical Activity 2018-2030). In the United States alone, the record from 2017 indicates that 30.3 million people have diabetes. Among them, 23.1 million have been diagnosed, while 7.2 million remain undiagnosed due to the asymptomatic phase of the disease, as reported by the CDC (Centers for Disease Control and Prevention).

India is also grappling with a substantial number of diabetes cases, with around 30 million people affected currently. Projections suggest that this number is expected to increase to a staggering 80 million by the year 2030 (Wang 2019). Governments have taken various initiatives to raise awareness about the importance of lifestyle changes and have provided vaccinations for people's welfare. Despite these efforts, diabetes remains a challenging issue that requires concerted attention and action.

The healthcare sector possesses an extensive collection of databases containing diverse types of data, including structured, unstructured, and semi-structured formats. Due to the sheer volume of data, there is a need to uncover hidden patterns and apply predictive analytics to process this information effectively. This has piqued the interest of researchers, leading to the exploration of predictive analysis within the healthcare domain.

Early detection of diseases is crucial in the field of medicine as it increases the likelihood of successful health outcomes. Modern technological advancements offer viable solutions for addressing health issues through data analysis and disease prediction. Machine learning techniques have become prevalent in the healthcare domain, playing a pivotal role in developing early predictive frameworks or models to determine the presence of specific diseases in individuals (Shailaja 2018).

 Received: August 1, 2023 | Accepted: August 14, 2023

The process of machine learning involves learning from historical data to generate valuable insights for current scenarios. In this research article, a new and innovative framework is introduced, utilizing machine learning and ensemble techniques to predict diabetes during its early stages. The framework employs an ensemble bagging decision tree classifier specifically for predicting diabetes mellitus. The performance of this proposed model is then compared with other state-of-the-art techniques used in the field.

1.1 Novelty

The primary goal of this research work is to develop an accurate and reliable predictive model that can identify individuals at risk of developing diabetes mellitus. The purpose is to enhance the effectiveness of traditional machine learning models by integrating ensemble techniques. By achieving this objective, the research seeks to optimize early detection and preventive measures, which can significantly improve patient outcomes. Moreover, the research objective includes evaluating the predictive model's performance, such as sensitivity, specificity, accuracy, and Precision, to ensure its effectiveness and applicability in real-world clinical applications.

2. Literature Review

In recent studies, researchers have made significant efforts to develop accurate predictive models for diabetes using advanced technology. This section provides a concise overview of various techniques applied in the anticipation and prediction of diabetes mellitus.

These approaches encompass a wide range of methods, including machine learning algorithms, statistical analyses, and data mining techniques. Scientists have explored the use of feature selection and extraction methods to identify crucial factors associated with diabetes risk. Additionally, ensemble techniques have been employed to combine multiple models, aiming to achieve improved predictive performance.

Moreover, electronic health records and large-scale healthcare datasets have been utilized in some studies to create robust predictive models. Deep learning and neural networks have also been investigated to capture complex patterns in diabetes data, ultimately leading to more accurate predictions.

The primary objective of these research endeavors is to establish reliable and efficient predictive models that can aid in early diagnosis, prevention, and management of diabetes, ultimately enhancing overall healthcare outcomes for patients."

The research paper El Jerjawi (2018) utilizes an artificial neural network to predict diabetes accurately, aiming to minimize the error function and identify the presence of diabetes in individuals. The model achieves an accuracy of 87.30 percent based on the evaluation of a dataset.

In Sisodia (2018), the focus is on improving the accuracy of diabetes prognosis. The proposed model employs three machine learning algorithms, namely naive bayes, support vector machine, and decision tree, for early detection of diabetes. Among them, naive bayes achieves the highest accuracy of 76.30 percent.

The paper (Alehegn 2018) introduces an ensemble model that combines predictive techniques like naive nets, SVM, and Decision stump to improve predictions. This hybrid ensemble model achieves an accuracy of 90.36 percent when tested on 788 observations.

Using data mining techniques and machine learning, Alam (2019) presents a model incorporating random forest, artificial neural network, and k-means clustering for diabetes prediction. The paper highlights a significant relationship between glucose levels and BMI and achieves an accuracy of 75.7 percent using the artificial neural network.

In Yahyaoui (2019), a decision support system is developed using deep learning and machine learning techniques, including SVM, Random forest, and Convolutional neural networks, to predict diabetes. The proposed system achieves a higher accuracy of 83.67 percent with random forest.

The study of Tripathi (2020) focuses on developing a machine learning-based model employing K-nearest neighbor, Linear discriminant analysis, support vector machine, and random forest for early detection and prediction of diabetes. Random forest outperforms other algorithms, reaching an accuracy of 87.66 percent.

The work of Kumari (2021) suggests an ensemble method for predicting diabetes mellitus using soft voting classifiers and a combination of naive bayes, random forest, and logistic regression. The model achieves an accuracy of 79.04 percent on the diabetes dataset.

In Abdulhadi (2021), a random forest classifier is used to classify the presence of diabetes for early detection and assistance in healthcare treatment. The experiment shows an accuracy of 82 percent.

The study (U. Butt 2021) proposes classification and prediction techniques based on machine learning, using multi-layer perceptron, logistic regression, and random forest classifiers. The model incorporates long short-term memory and linear regression for early prediction and achieves an accuracy of 87.26 percent.

The paper (Naseem 2022) proposes a fusion of the Internet of Things and machine learning to predict diabetes at an early stage. Six machine learning techniques, including artificial neural network, recurrent neural network, convolutional

neural network, support vector machine, Long short-term memory, and logistic regression, are employed, achieving an accuracy of 81 percent with recurrent neural networks.

The study Krishnamoorthi (2022) explores the scope of machine learning and big data analytics in detecting and predicting diabetes. The framework utilizes random forest, support vector machine, and decision tree as learning models, achieving an accuracy of 83 percent.

The paper Mushtaq (2022) presents a predictive model using the voting classification technique. Dataset balancing is handled with SMOTE and Tomek learning. The model achieves an accuracy level of 82 percent with the default dataset and 81.7 percent with the balanced dataset.

In Samet (2023), a comparative analysis is conducted among different classifiers on a dataset for detecting and predicting diabetes mellitus. Random forest achieves the highest accuracy of 93 percent among the supervised learning techniques studied.

3. Proposed Methodology

In the healthcare sector, early disease prediction is crucial for prevention. Today's sedentary lifestyle and poor dietary habits increase the risk of chronic diseases, leading to severe health issues like stroke, diabetes, cancer, and arthritis. Among these, diabetes stands out as a prevalent and critical chronic disease, emphasizing the importance of early diagnosis to avert life-threatening consequences.

Machine learning plays a pivotal role in the early detection and prediction of diseases, and this paper introduces a novel machine learning-based framework for early diabetes prediction, as depicted in Figure 1. Since many individuals with diabetes experience a long asymptomatic phase, understanding the disease's symptoms becomes vital. The proposed model analyzes a dataset comprising 520 observations with 17 attributes to ascertain the accuracy of diabetes diagnoses.

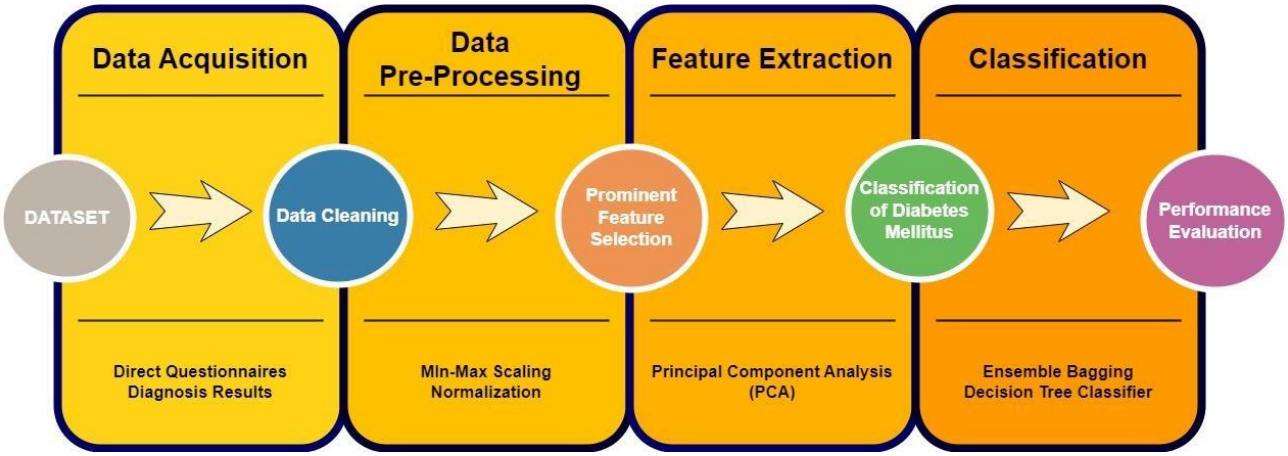

Figure 1 Proposed Framework.

3.1. Data Acquisition and Data Preprocessing

In this research paper, a dataset consisting of 520 observations is collected through direct questionnaires and diagnosis results from patients (M. Islam 2020). The study focuses on 14 attributes for analysis. When preparing Diabetes mellitus data, normalization is be used to scale any numerical characteristics that need it. To prepare the data for analysis, preprocessing techniques like min-max scaling (normalization) are applied to mitigate the negative impact of irregularities in the data. A "sigmoid" activation function is used during the analysis of data. The information is normalized for execution so that they fell within the bounds [0,1]. The level of normality is defined by equation 1, as specified below:

$$N' = \frac{n - \min(n)}{\max(n) - \min(n)'} \qquad (1)$$

Where, n = original value.
min_value= feature's minimum value.
max_value= feature's maximum value.

3.2. Feature Selection

Principal Component Analysis (PCA) is a commonly used technique in machine learning for dimensionality reduction of raw data. Its primary objective is to transform the data into a new set of variables, known as principal components, while

preserving the maximum variance of the original data. PCA achieves this by making the weakest principal component zero, effectively eliminating less significant features.

One of the key challenges in machine learning is dealing with irrelevant and redundant features, which can slow down or even hinder the learning process. By using PCA as a feature selection technique, this article aims to address this issue. PCA retains the essential trends and patterns present in the high-dimensional dataset while simplifying its complexity, making it more amenable for machine learning algorithms to process.

The PCA operation involves transforming the original data into a new coordinate system where the principal components are orthogonal and uncorrelated. Each principal component captures a different aspect of the data's variance, and by selecting the top principal components, we can effectively reduce the dimensionality of the dataset while retaining the most critical information.

$$F = FV^T * OD^T \qquad (2)$$

Where F is Final Dataset, $FV^T$is transpose of the Feature Vector and OD^T is the transpose of the Original Dataset.

3.3. Ensemble Bagging Decision Tree Classifier

Ensemble learning is a powerful technique used to improve the predictive performance of machine learning models. Two popular methods of ensemble learning are Bagging and Boosting, both of which aim to enhance consistency and accuracy. Bagging focuses on reducing variance while avoiding overfitting and loss of decision tree bias. For the current study, the paper employs EBDTC (Ensemble Bagging Decision Tree Classifier) with inputs having minimal bias. The entire dataset is divided into a training set (80%) and a testing set (20%). The training set is then utilized to predict the desired outcome using the classifier, thereby evaluating the model's performance and accuracy.

Algorithm 1: Ensemble Bagging Decision Tree Classifier
Input: A matrix for every training i.e variables and category score
Output: The E-vector to store the estimates of the variables' quality

1. Set E[Ar] =0.0 ∀ values.
2. i=1 to k do start
3. Select an instance randomly Rm_i;
4. Build node N;
5. Tuples in T must be linked to the same class, Cs, otherwise.
   N will be returned as a leaf node having a category Cs label;
6. At any time the attribute _list is empty,
   Return N as T's majority class leaf node; / majority voting:
7. Attribute selection method(T, attribute_list) finds the "best" criterion for splitting;
8. Node N as per Splitting criterion ;
   The split feature is precise and There is no restriction on multi-way splits or binary trees.
9. Splitting attributes in the attribute list; deleting dividing attribute
10. Split the tuple and build sub-trees for every division for each result m of the splitting criteria
11. Allow T_j to be the collection of data tuples in T that meet result m;/ a partition.
12. If T_j is null then
    Join a leaf to node N carrying the majority class's label;
13. If not, connect the node that Generates the decision tree(T_j attribute_list) returned to node N;

$$E[Ar]; E[Ar] \sum_{m=1}^{l} \text{diff} \frac{Ar, Rm_i, H_j}{k.l} \sum_{c \neq class(Rm_i..i} [\frac{P(Cs)}{1 - P(class\ Rm_i)}]$$

$$\sum_{j=}^{l} \text{diff}(Ar, Rm_i, K_j(C))) \ (k.l);$$

14. End

3.4. Performance Evaluation

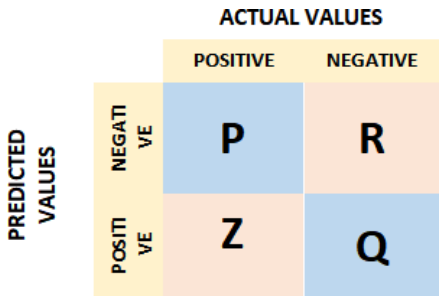

Figure 2 Confusion Matrix.

To access the model efficacy, an N x N confusion matrix is used. Here, N denotes the total number of target groups. This matrix compares the predictive values and actual values together and shows the accuracy of the classification model. The confusion matrix is shown in Figure 2.

Here, P is true positive, R is false positive, Z is false negative and Q is true negative.

The confusion matrix facilitates the calculations of accuracy, precision, recall, and f-score of the model.

4. Implementation

The study utilizes Google Colab and the Python programming language for the implementation of the proposed model. Both the available dataset and the dataset for validation are employed to forecast the presence of diabetes. Subsequently, a comparative analysis is conducted to evaluate the performance of the proposed model against existing machine learning standard algorithms. This approach ensures a comprehensive examination of the model's predictive capabilities and its effectiveness in comparison to established algorithms used in the field.

The parameters taken from the available data set are as follows:

i. Age
ii. Gender
iii. Polyuria
iv. Polydipsia
v. sudden_weight_loss
vi. Weakness
vii. Polyphagia
viii. genital_thrush
ix. visual_blurring
x. Itching
xi. Irritability
xii. delayed_healing
xiii. partial_paresis
xiv. muscle_stiffness
xv. Alopecia
xvi. Obesity

5. Experimental Results

In this section, we assess the performance of the Ensemble Bagging Decision Tree Classifier on the dataset. The model's accuracy is then compared with that of other existing algorithms. Notably, the proposed model attains an impressive accuracy of 93.62% in early predicting diabetes Mellitus. To provide a comprehensive overview, we present the numerical outcomes of the proposed model along with those of the other existing approaches in Table 1. Additionally, the evaluation parameters are graphically depicted in Figure 3, offering a visual representation of the model's performance in comparison to the other algorithms.

Table 1: Numerical Outcomes.

Methods	Accuracy	Precision	Recall	F-Score
KNN	0.764	0.848	0.80	0.823
Logistic Regression	0.823	0.825	0.941	0.88
XG-Boost	0.882	0.891	0.940	0.916
Random Forest	0.921	0.918	0.971	0.944
Proposed Model	0.936	0.969	0.941	0.954

Figure 3 Performance Evaluation.

5.1. Accuracy

The proportion of correctly identified instances with respect to the total available instances is measured as the accuracy of the model. The expression for accuracy is as follows:

$$\text{Accuracy} = \frac{P+Q}{P+Q+R+Z} \qquad (3)$$

P=true positive
Q=true negative
R=false positive
Z=false negative

The accuracy of the proposed system is compared with the existing algorithms and it clearly shows that the proposed model has a higher accuracy of 93.62%.

5.2. Precision

Precision measures the truly relevant data points that are being classified as positive that are actually positive. In Figure 3, it can be seen that the proposed model has high precision of 96.9% as compared to the existing ones. The mathematical expression is as follows:

$$\text{Precision(Pr)} = \frac{P}{P+R} \qquad (4)$$

5.3. Recall

The ability of the model to detect the positive instances is measured as recall. The rise in recall denotes the more positive instances detected within the dataset. The recall of the proposed model is 94.2 percent. However, random forest outperforms with 97.1 percent. The recall can be formulated as follows:

$$\text{Recall(Rc)} = \frac{P}{P+Z} \qquad (5)$$

5.4. F-Score

The Score is the combination of the model's recall and precision results which computes the number of times the model has given the correct prediction. The calculative expression of the F-Score is given as:

$$\text{F-Score} = \frac{Pr*Rc}{Pr+Rc} * 2 \qquad (6)$$

Figure 3 shows that the proposed model has a higher F-Score of 95.4% comparatively. Also, Figure 4 is used to visualize the robustness of relationships among the numerical variables and to understand how the attributes are connected.

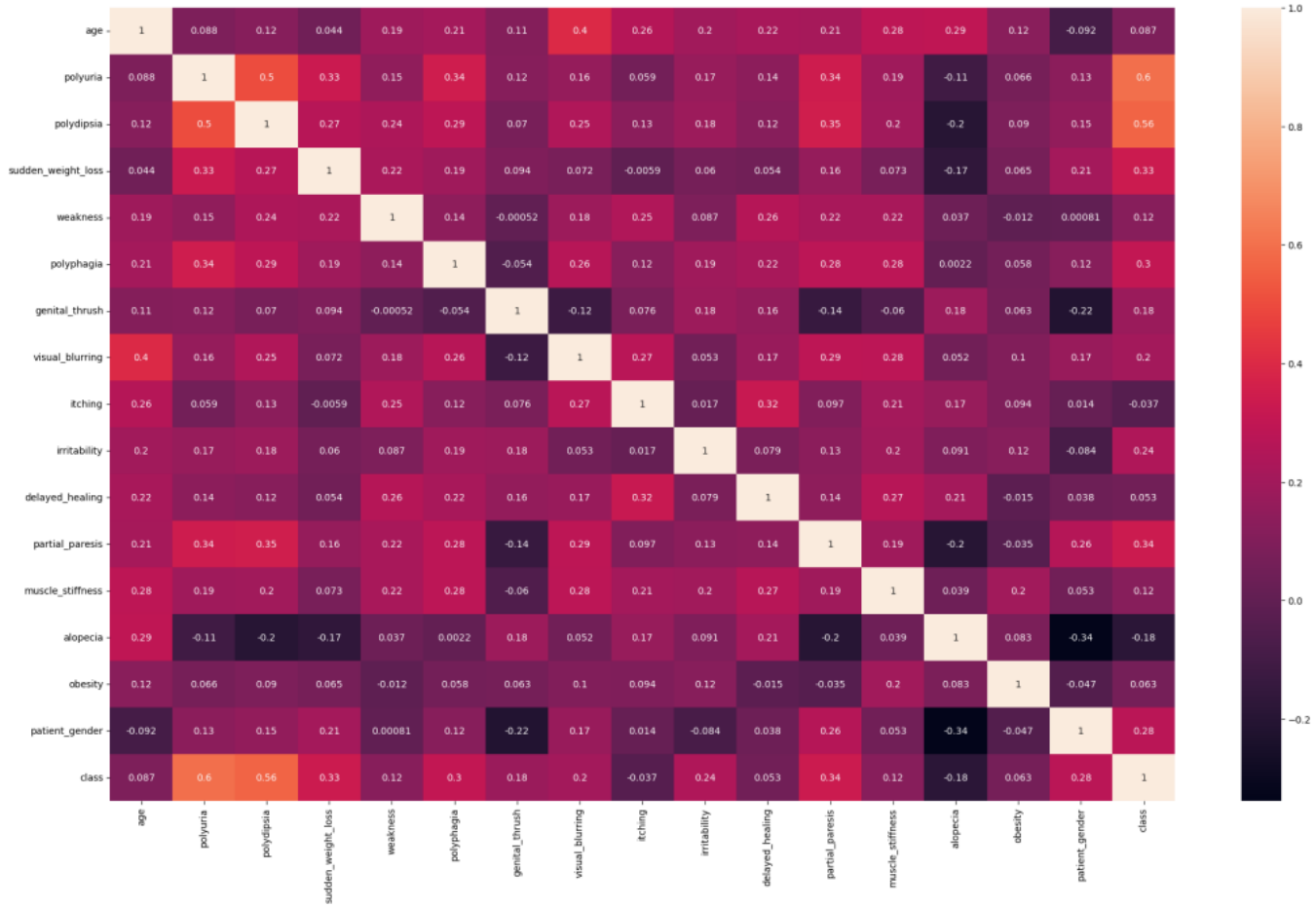

Figure 4 Correlation Analysis.

6. Conclusions

Machine learning has proven to be a suitable approach for detecting and predicting chronic diseases. Among these, diabetes mellitus is one of the most prevalent, and it often remains asymptomatic for an extended period, leading to potentially life-threatening consequences. Diabetes can result in various health issues affecting the heart, kidneys, eyes, and pancreas. Therefore, early prediction is vital to enable timely preventive measures. The paper introduces a novel machine learning-based model designed to facilitate the early prediction of diabetes mellitus. The model exhibits an impressive accuracy of 93.62% on the dataset used, and with training on a larger dataset, it is likely to perform well in a testing phase, further enhancing its predictive capabilities.

6.1. Limitation

While the proposed system shows the capability to have optimal accuracy, it is not without challenges and limitations. One prominent limitation is its heavy dependency on the availability and quality of data. It possess challenge in obtaining high quality features. Also the model does not consider the external factors that may influence the person's health.

7. Future Work

The future work entails testing the proposed model on a larger dataset to achieve early prediction of chronic diseases while enhancing the recall value. Additionally, integrating the Internet of Things (IoT) to collect real-time patient data will facilitate a home-centric environment for disease detection and prediction. The fusion of IoT with machine learning holds promising potential in elevating the accuracy level of the framework. By leveraging the capabilities of IoT to gather continuous and diverse patient data, the model can be further refined and optimized, leading to more accurate predictions and better healthcare outcomes.

Ethical considerations

Not applicable.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

This research did not receive any grant from external funding agencies.

References

Abdulhadi N, Al-Mousa A (2021) Diabetes detection using machine learning classification methods. In 2021 International Conference on Information Technology (ICIT), pp. 350-354. IEEE.

Alam TM, Iqbal MA, Ali Y, Wahab A, Ijaz S, Baig TI, Abbas Z (2019) A model for early prediction of diabetes. Informatics in Medicine Unlocked 16:100204.

Alehegn M, Joshi R, Mulay P (2018) Analysis and prediction of diabetes mellitus using machine learning algorithm. International Journal of Pure and Applied Mathematics 118:871-878.

Butt UM, Letchmunan S, Ali M, Hassan FH, Baqir A, Sherazi HHR (2021) Machine learning based diabetes classification and prediction for healthcare applications. Journal of healthcare engineering 2021.

Dobbs RHHGIDLR, Sakurai H, Sasaki H, Faloona G, Valverde I, Baetens D, Unger R (1975) Glucagon: role in the hyperglycemia of diabetes mellitus. Science 187:544-547.

El_Jerjawi NS, Abu-Naser SS (2018) Diabetes prediction using artificial neural network.

Gabbay KH (1975) Hyperglycemia, polyol metabolism, and complications of diabetes mellitus. Annual review of medicine 26:521-536.

Islam MM, Ferdousi R, Rahman S, Bushra H (2020) Likelihood prediction of diabetes at early stage using data mining techniques. In Computer vision and machine intelligence in medical image analysis, pp. 113-125. Springer, Singapore.

Krishnamoorthi R, Joshi S, Almarzouki HZ, Shukla PK, Rizwan A, Kalpana C, Tiwari B (2022) A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering.

Kumari S, Kumar D, Mittal M (2021) An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering 2:40-46.

Mushtaq Z, Ramzan MF, Ali S, Baseer S, Samad A, Husnain M (2022) Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques. Mobile Information Systems 1-16.

Naseem A, Habib R, Naz T, Atif M, Arif M, Allaoua Chelloug S (2022) Novel Internet of Things Based Approach Towards Diabetes Prediction Using Deep Learning Models. Frontiers in Public Health, 2848.

Wang Q, Cao W, Guo J, Ren J, Cheng Y, Davis DN (2019) DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values. In IEEE Access, vol. 7, pp. 102232-102238.

Samet S, Laouar MR, Bendib I (2023) Comparative Analysis of Diabetes Mellitus Predictive Machine Learning Classifiers. In 12th International Conference on Information Systems and Advanced Technologies "ICISAT 2022" Intelligent Information, Data Science and Decision Support System, pp. 302-317. Cham: Springer International Publishing.

Shailaja K, Seetharamulu B, Jabbar MA (2018) Machine learning in healthcare: A review. In 2018 Second international conference on electronics, communication and aerospace technology (ICECA), pp. 910-914. IEEE.

Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Procedia computer science 132:1578-1585.

Tomic D, Shaw JE, Magliano DJ (2022) The burden and risks of emerging complications of diabetes mellitus. Nature Reviews Endocrinology, 18.9:525-539.

Tripathi G, Kumar R (2020) Early prediction of diabetes mellitus using machine learning. In 2020 8th international conference on reliability, Infocom technologies and optimization (trends and future directions) (ICRITO), pp. 1009-1014. IEEE.

World Health Organization (2019) Global action plan on physical activity 2018-2030: more active people for a healthier world. World Health Organization.

Yahyaoui A, Jamil A, Rasheed J, Yesiltepe M (2019) A decision support system for diabetes prediction using machine learning and deep learning techniques. In 2019 1st International informatics and software engineering conference (UBMYK), pp. 1-4. IEEE.