Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Exploration of Machine Learning Model for Diabetes Prediction

E. Loganathan[1], X. Arul Bruno[2], Ashiq Hussain Bhat[3],S. Dharun[4]

[1]Assistant Professor, Department of Computer Science and Engineering, Perundurai, Erode, TamilNadu, India.

[2,3,4,5]Student, Department of Computer Science and Engineering, Perundurai, Erode, TamilNadu, India.

Email: [2]arulbruno2003@gmail.com

Abstract

Diabetes is a metabolic disorder characterized by a malfunction in insulin release, resulting in a rise in blood sugar levels in the body. Diabetes diagnosis must be made on time and precisely in order to be effective and enhance patient outcomes. The diabetes management presents a formidable challenge in modern healthcare, demanding a combination of timely interventions, precise data analysis, and personalized medical services. Furthermore, there exists a growing demand for advanced predictive models that not only provide accurate forecasts but also offer transparency and interpretability. The study objectives are to develop an innovative machine learning model for data-driven diabetes prediction and medical services. To identify the machine learning model that best suits the proposed system, a literature review related to different methods used in diagnosing diabetes is conducted. The merits and demerits of each existing system were identified to devise a proposed model that seamlessly integrates data from various sources, including blood glucose levels and patient health information. Based on the review, the study suggests an innovative machine learning model that utilizes the Explainable Decision Tree Model, leveraging cloud computing to analyse diabetes patient data and make predictions. The integration of cloud computing allows for seamless data integration from various sources. This research project represents a significant step forward in personalized diabetes care, enabling patients to proactively manage their condition while providing healthcare professionals with a powerful tool for delivering tailored medical services.

DOI: https://doi.org/10.36548/jaicn.2024.1.005
Received: 21.12.2023, received in revised form: 19.01.2024, accepted: 07.02.2024, published: 21.02.2024

Keywords: Diabetes, Healthcare, Medical services, Decision Tree Model.

1. Introduction

Diabetes is characterized by high blood glucose levels due to inadequate insulin synthesis or poor insulin utilization. The pancreas produces insulin, which regulates glucose metabolism. Without adequate insulin action, glucose accumulates in the blood, causing health problems like as damage to the eyes, kidneys, nerves, and heart, as well as an increased risk of some malignancies. These health hazards can be mitigated by taking preventative actions and managing diabetes properly [16]. Consistently elevated blood glucose levels and the insulin levels can lead to various health complications, including heart disease, nerve damage, and vision impairment [17-19].

1.1 Types of Diabetes

There are several types of diabetes. The most common forms include:

1.1.1 Type 1 Diabetes: It is classified as an autoimmune condition where the body's immune system targets and eliminates the insulin-producing cells in the pancreas, although the exact reasons for this immune response remain unclear. Roughly 10% of individuals diagnosed with diabetes have Type 1. Typically, it can be identified from the children’s and adults, although it can manifest at any age.

Gestational diabetes emerges in some individuals during pregnancy. Generally, this form of diabetes resolves after childbirth. However, those who experience gestational diabetes face an increased risk of developing Type 2 diabetes later in life.

1.1.2 Type 2 Diabetes: It can occur when the body either does not produce sufficient insulin or the body's cells do not appropriately respond to insulin (a condition is called as insulin resistance stage). It stands as the most prevalent form of diabetes and predominantly affects adults and it can also manifest in children.

Prediabetes represents the stage preceding Type 2 diabetes. During this phase, blood glucose levels are elevated beyond normal but haven't reached the threshold for a Type 2 diabetes diagnosis.

Healthcare professionals diagnose diabetes by assessing blood glucose levels through various tests such as Blood glucose test, Random blood glucose test, A1c test (HbA1C or glycated hemoglobin test),

1.2 Problem Definition

As traditional diagnostic methods for diabetes are time-consuming and unreliable, the emergence of machine learning has significantly improved the diagnostic process by reducing time and human reliance. Machine learning is extensively utilized in diabetes diagnosis, excelling in analyzing patient data—demographics, lifestyle factors, and medical history—with precision to predict diabetes risk. Additionally, it offers treatment plans through continuous glucose monitoring and insulin dosage suggestions. Despite its promising capabilities, real-time application faces challenges such as collecting real-world datasets and adapting to new data [11-15].

The proposed study aims to address these challenges by introducing a novel method that integrates machine learning and cloud technology.

1.3 Objective

- To study the literature related to machine learning-based diabetes diagnosis.
- To understand the merits and demerits of the existing system and the challenges faced in real-time integration.
- To suggest a proposed model that addresses the challenges of the existing system.

2. Related Study

The research focuses on the challenges associated with diabetes management and aims to leverage RL (reinforcement learning) to improve aspects such as blood glucose levels, treatment recommendations, and diabetes diagnosis.

The primary objective is to showcase how RL, with its decision-making and adaptive capabilities, can be effectively be applied to enhance diabetes management. This involves improving key parameters related to diabetes care, and the research discusses the advantages

of RL such as capability to handle unpredictable personal details of the individuals suffering from diabetes, suitable for dynamic operating environment, and efficient for real time operations. But they also come with disadvantages that requires more variants as well as enhancements [1].

The DMNet approach efficiently assesses type 2 diabetes risk in the elderly by combining long-term temporal data, synthetic minority over-sampling, and entity embedding.

The study suggests a tailored risk assessment approach called DMNet for senior persons with type 2 diabetes. The approach overcomes the constraints of existing risk prediction methods by collecting long-term temporal data and correlations between diabetes risk factor categories. The research discusses the use of entity embedding to improve feature representations and tandem long short-term memory (T-LSTM) to record correlations between risk factor categories. The suggested DMNet is assessed using real-world data and outperforms baseline approaches in six assessment parameters. The research emphasizes the need of taking into account personal information as well as healthcare system ratings when diagnosing diabetes. Additionally, the study provides insights into the relationships between several risk factor categories, as well as the role of detailed personal and medical information in tailored risk assessment. The suggested DMNet architecture has potential uses in geriatric hospitals and nursing homes, leading to the early identification and accurate diagnosis of type 2 diabetes in the elderly [2] .

The paper authored by Reyazur Rashid Irshad et al [3] introduces a novel hybrid framework called ASM-RF for early disease diagnosis in elderly individuals. The problem addressed is the limitations of traditional predictive diagnosis methods, which often fail to capture inherent data patterns. The objective is to improve accuracy and automate intelligent decisions in healthcare. The ASM-RF framework combines the ASM algorithm with Random Forest, employing Identity-based Encryption for data security. The study uses a dataset from IoT sensor devices and participants for simulation. Merits include high accuracy, precision, AUC, recall, and reduced execution time. Demerits include potential scalability challenges, limited insights into security considerations, and limited generalization beyond the dataset.

The research authored by Usama Ahmed et al [4] focuses on early prediction of diabetes using a fused machine learning approach. The problem addressed is the increasing risk of

diabetes globally, emphasizing the need for accurate disease prediction. The objective is to develop a model combining Support Vector Machine (SVM) and Artificial Neural Network (ANN) models, integrated through fuzzy logic, to determine diabetes diagnosis based on a dataset of symptoms. The proposed model achieves a high prediction accuracy of 94.87%, surpassing existing systems and demonstrating potential for early diagnosis and preventive measures. The paper suggests the need for continued improvement to achieve even higher prediction accuracy in diabetes prediction.

The research authored by Radwa Marzouk et al [5] addresses the significant impact of diabetes worldwide, particularly type-2 diabetes, and proposes a comprehensive solution. The work involves the development of an analytical predictive model based on various machine learning techniques and the creation of a web-based personalized diabetes monitoring system. The primary objectives are early prediction of type-2 diabetes and the establishment of a system for enhanced patient care and treatment. The methodology includes the integration of patient data, QR cards, and the Internet of Things for real-time data sharing. Multiple machine learning algorithms are employed, including Decision Tree, Support Vector Classifier, Random Forest, Gradient Boosting, Multi-layer Perceptron, Artificial Neural Network, k-Nearest Neighbors, Logistic Regression, and Naive Bayes. The model is evaluated using synthetic and PIMA Diabetes datasets. The merits include a predictive model that outperforms others, especially the Artificial Neural Network, and a system offering detailed visualizations and graphs for patient classifications. Future work aims to expand the system to cover other chronic diseases, automate diabetes analysis with various machine learning algorithms, handle diverse data types, and introduce additional features for broader health management.

3. Existing System

There are several machine learning-based Diabetes Prediction Systems available that commonly use the Pima Indian Diabetes Database: The Pima Indian Database is a well-known dataset commonly used for diabetes prediction. Many researchers and developers have built prediction systems using this dataset, employing algorithms such as logistic regression, random forests, support vector machines (SVM), and artificial neural networks [6].

3.1 AdaBoost Algorithm: AdaBoost (Adaptive Boosting) is a popular boosting algorithm that can be used in diabetes prediction systems. AdaBoost combines multiple weak classifiers, such as decision trees, into a strong classifier, iteratively giving more weight to the misclassified instances.

3.2 Random Forest Algorithm: Random Forest is an ensemble learning system that generates numerous decision trees and combines their predictions via voting. It has been employed in diabetes prediction systems to improving the accuracy and handle complex interactions between features.

3.3 Support Vector Machines (SVM): It is a powerful ML algorithm used for classification tasks, including diabetes prediction. SVM aims to finding the optimal hyperplane that separates the data into different classes, maximize the margin between the classes.

3.4 Deep Learning Models: as convolutional neural networks and recurrent neural networks are some of the deep learning models used in the diabetes prediction systems. These models can automatically learn complex patterns and dependencies from raw input data, such as medical images or time-series data.

The commonly used algorithms like AdaBoost Algorithm, reinforcement learning, Support Vector Machines (SVM), and Deep Learning Models in diabetes prediction has both benefits and drawbacks. AdaBoost, while successful at combining weak classifiers, may be susceptible to noisy data and outliers, limiting its robustness [. Reinforcement learning, which focuses on learning optimal choice strategies through trial and error, may necessitate significant computational resources and time-consuming training procedures. Support Vector Machines, which are known for handling high-dimensional data effectively, may struggle with huge datasets, thereby limiting their scalability. Deep Learning Models, with their ability to capture complicated patterns, frequently require large volumes of labeled data and tremendous computer power, making them resource expensive. Furthermore, deep learning models' interpretability is a barrier, limiting their use in sectors where understanding the reasons behind predictions is critical, such as healthcare. Overall, while these techniques have shown promise in diabetes prediction, resolving their limitations is critical for their

successful and responsible deployment in real-world contexts [6-10]. Some of the disadvantages that were gained out the study is listed below

Disadvantages

- Limited generalization to different populations or healthcare settings.
- Data quality and availability issues can affect system accuracy.
- Challenges in feature selection and interpretability of predictions.
- Lack of transparency and explainability in the machine learning models.
- Inadequate incorporation of temporal dynamics in diabetes progression.
- Lack of user-friendly interfaces and interactive features.
- Ethical concerns regarding privacy, security, and biases in predictions.

To overcome the challenges of the existing and address few disadvantages of the existing systems of diabetes prediction the study suggests a proposed method that integrates the machine learning with the cloud computing.

4. Proposed System

The proposed system for diabetes prediction harnesses ensemble learning techniques, merging decision trees with bagging and boosting algorithms to heighten accuracy and stability. It begins with data collection, preprocessing, and splitting for the training and testing. Employing decision trees as a foundational model, bagging creates an ensemble via parallel training, while boosting iteratively refines the model's focus on misclassified instances. Model evaluation utilizes diverse metrics and preforms hyperparameter tuning to optimize the performance. Additionally, the system develops an intuitive interface for users to input data and receive clear, actionable predictions.

4.1 Architecture Diagram

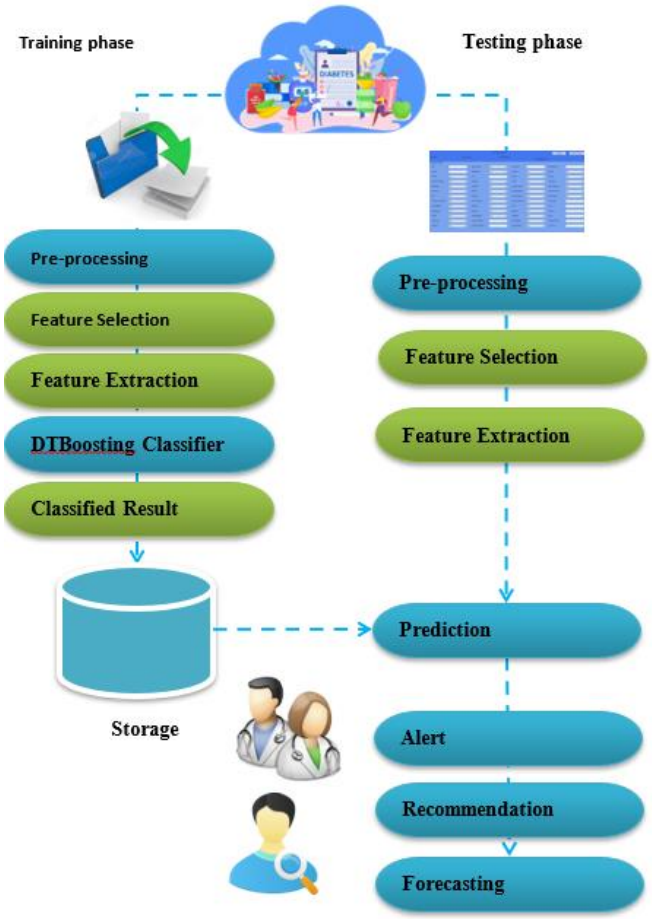

Figure 1. Architecture Diagram

4.2 Methodology

During the data collection and pre-processing phase, a dataset will be compiled that includes relevant parameters such as age, BMI, blood pressure, glucose levels, and insulin levels, as well as the associated target variable indicating the presence or absence of diabetes. Pre-processing procedures will be taken to resolve missing values, handle categorical variables (if any), perform feature scaling, and divide the dataset into testing sets. The decision tree technique will be used as a fundamental classifier to capture correlations between characteristics and target variables. This system, which can handle both numerical and categorical data, is likely to be effective for diabetes prediction. The bagging approach will then be used to generate an ensemble of decision trees, thereby improving prediction accuracy through training on diverse subsets of the data. Similarly, the boosting approach will be used

to generate an ensemble sequentially, with a focus on misclassified cases from earlier trees to iteratively enhance model performance.

Following model development, hyperparameter tuning will be carried out through training and evaluation on the test set. Different hyperparameters for bagging and boosting algorithms, such as the number of decision trees, maximum depth, and learning rate, will be tested to improve model performance using strategies such as grid search or random search. Finally, during the integration and deployment step, the trained model will be incorporated into a usable system. This system will accept user input (such as age, BMI, and blood pressure) and process it, using the trained model to predict the presence or absence of diabetes. The user will be supplied with prediction results that are straightforward and easy to interpret.

5. Modules Description

5.1 Diabetes Monitoring System

The process of designing and developing a website for a diabetes monitoring system in this module encompasses the creation of a comprehensive platform that empowers patients to monitor and manage their diabetes effectively. This is achieved through the utilization of the Python programming language and the Flask framework. The module encompasses the development of the web platform, implementation of programming languages, and management of the database system. It is designed to seamlessly collect essential data such as blood sugar readings, insulin dosages, and other pertinent health metrics from users. The key focus areas within this module include the creation of a user-friendly interface, efficient data collection and analysis mechanisms, as well as robust reporting features.

5.2 Diabetes Model Train and Build

The module involves the training and building of a diabetes prediction model using a dataset with medical predictor variables. These variables include the number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, BMI, diabetes pedigree function, age, and the outcome (0 or 1). The dataset comprises 768 observations.

Data annotation emphasizes the acquisition of large-scale, reliable, and timely medical data using distributed platform-based technologies. A Diabetes Lab Test Sample Data in CSV format is loaded into a web screening tool, featuring nine variables for 400 people.

Data preprocessing is crucial, involving cleaning, handling missing values, and standardizing data using normalization, specifically StandardScaler normalization, to ensure all attribute values are within the range of [-1, 1].

Feature selection is implemented using Principal Component Analysis (PCA) and minimum redundancy maximum relevance (mRMR) methods to reduce dimensionality. Logistic regression identifies significant factors for diabetes prediction, including age, education, BMI, systolic and diastolic blood pressure, direct and total cholesterol.

Data visualization incorporates a heatmap to visualize correlations between different features, aiding in understanding relationships between variables.

5.3 Diabetes Disease Classification

Ensemble learning is a successful machine learning paradigm that combines multiple learners to predict target attributes, improving classification accuracy. This study explores two widely used ensemble learning methods: bagging and boosting.

- **Boosting**

Boosting aims to create a strong classifier from weak learners by reweighting samples during training. AdaBoost, a notable algorithm, increases the weights of misclassified samples in each iteration, improving classification accuracy. MultiBoost combines wagging and AdaBoost to reduce variance and address high bias and variance. Real AdaBoost is an extension that incorporates class probability estimates to generate real-valued contributions.

- **Bagging and Boosting Approaches Using DTB Algorithms**

This study implements bagging and boosting approaches with DTB algorithms to predict early-stage diabetes risk. Bagging involves random forest and classifiers such as C4.5, random tree, REPTree, decision stump, Hoeffding tree, and NBTree. Boosting utilizes AdaBoost, MultiBoost, and real AdaBoost with the same base learners, excluding random

forest. The general structure involves dividing the diabetes training dataset into subsets using bootstrap for bagging, training multiple classifiers, and aggregating them for the ensemble classifier.

In summary, the study explores ensemble learning methods, specifically bagging and boosting, using various DTB algorithms to predict diabetes risk. Boosting algorithms like AdaBoost, MultiBoost, and Real AdaBoost are compared, while bagging incorporates random forest and other classifiers. The general approach involves dividing datasets, training classifiers, and constructing ensemble classifiers for improved accuracy. The following Figure. Depict the modules of the user interface.

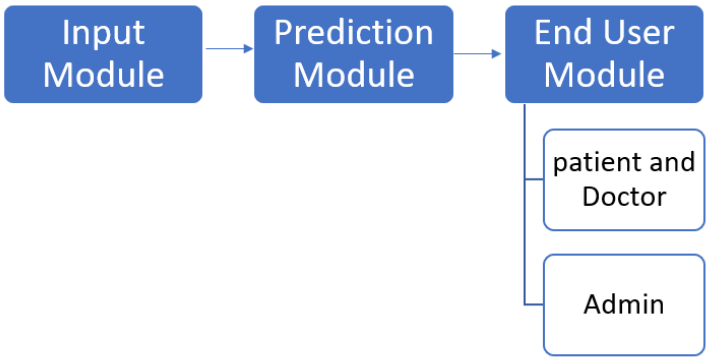

Figure 2. User Interface Modules

5. 4 Future Work

The future work of the study involves the implementation and evaluation of the proposed method for the collected dataset, followed by the design, development, and deployment of the user interface for real time use.

6. Conclusion

The objective of the study is to revolutionize diabetes care by developing an innovative machine learning model that takes advantage of cloud computing. The system employs an Explainable Decision Tree Model to provide precise predictions of blood glucose levels and insulin doses. Emphasizing the need of timely and precise diabetes diagnosis, the proposed effort addresses this gap by delivering actionable suggestions to patients and empowering

healthcare professionals with specialized medical services. The article reviews existing systems and introduces a novel system with benefits such as increased accuracy, robustness, and interpretability. The system's modules include data gathering, model training, disease classification, and prediction. The End User Module is designed for patients, clinicians, and administrators, providing a cost-effective, user-friendly solution for tailored diabetes care. In future the study would leap into the implementation and evaluation of the proposed method for the collected dataset, followed by the design, development, and deployment of the user interface for real time use.

References

[1] Kok-Lim Alvin Yau and Yung-Wey Chong (2023). Reinforcement Learning Models and Algorithms for Diabetes Management.

[2] Yu, Ziyue, Wuman Luo, Rita Tse, and Giovanni Pau. "DMNet: A Personalized Risk Assessment Framework for Elderly People With Type 2 Diabetes." IEEE Journal of Biomedical and Health Informatics 27, no. 3 (2023): 1558-1568.

[3] Irshad, Reyazur Rashid, Shahid Hussain, Ihtisham Hussain, Ahmed Abdu Alattab, Adil Yousif, Omar Ali Saleh Alsaiari, and Elshareef Ibrahim Idrees Ibrahim. "A novel artificial spider monkey based random forest hybrid framework for monitoring and predictive diagnoses of patients healthcare." IEEE Access (2023).

[4] Ahmed, Usama, Ghassan F. Issa, Muhammad Adnan Khan, Shabib Aftab, Muhammad Farhan Khan, Raed AT Said, Taher M. Ghazal, and Munir Ahmad. "Prediction of diabetes empowered with fused machine learning." IEEE Access 10 (2022): 8529-8538.

[5] Marzouk, Radwa, Ala Saleh Alluhaidan, and Sahar A. El_Rahman. "An analytical predictive models and secure web-based personalized diabetes monitoring system." IEEE Access 10 (2022): 105657-105673.

[6] Anjum, R., & Qamar, U. (2016). Diabetes disease prediction using data mining techniques: A systematic literature review. Journal of Medical Systems, 40(8), 207.

[7] Pesaranghader, A., & Aghabozorgi, S. (2017). Ensemble learning for diabetes diagnosis using bagging, boosting, random forests, and rotation forest. Journal of Medical Systems, 41(7), 116.

[8] Al-Mallah, M. H., & Sengupta, P. P. (2017). Challenges in prediction modeling of prevalent cardiovascular disease in diabetes. Cardiovascular Diabetology, 16(1), 1-4.

[9] Olsson, S. J., & Ferreira, D. (2018). The role of decision trees in clinical prediction modeling. European Journal of Epidemiology, 33(7), 645-649.

[10] Sathya, A. P., & Malarvizhi, M. (2019). Comparative analysis of decision tree algorithms in predicting diabetes. International Journal of Electrical and Computer Engineering, 9(1), 548-556.

[11] Nair, S. S., & Sandhya, K. (2020). Diabetes prediction using ensemble of optimized decision tree algorithms. Journal of Medical Imaging and Health Informatics, 10(5), 1104-1109.

[12] Alshammari, R., Alshammari, G., & Alotaibi, B. (2020). An ensemble-based machine learning model for diabetes prediction. International Journal of Advanced Computer Science and Applications, 11(2), 11-18.

[13] Akram, M., Ali, T., Kanwal, N., & Anwar, S. M. (2020). Predicting diabetes mellitus using ensemble learning techniques: A systematic review. Computers in Biology and Medicine, 125, 103982.

[14] Shabut, A. M., Mahmood, W., & Alzahrani, A. I. (2020). Deep ensemble learning architectures for type 2 diabetes prediction. Computers in Biology and Medicine, 120, 103739.

[15] Ayyagari, S., & Sahu, P. K. (2020). An ensemble learning framework for diabetes prediction. In Proceedings of the 11th International Conference on Ambient Systems, Networks and Technologies (ANT) (pp. 268-273).

[16] Farouk, A. M., & Eltoukhy, M. M. (2021). A comparative study of machine learning models for diabetes prediction. International Journal of Artificial Intelligence, 19(2), 1-27.

[17] Islam, M. R., Naim, R., & Hussain, A. (2021). Performance analysis of machine learning algorithms for diabetes prediction: A review. Journal of Healthcare Engineering, 2021, 1-15.

[18] Choi, J. Y., & Choi, M. K. (2021). Type 2 diabetes prediction using machine learning algorithms: A systematic review and meta-analysis. Healthcare Informatics Research, 27(3), 233-243.

[19] Maqbool, M., Raza, B., & Tariq, A. (2021). Predictive analysis for diabetes using machine learning: A systematic review. Applied Sciences, 11(6), 2733.