International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 967 - 975
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0135

An Analytical Predictive Model and Secure Wed Based Personalized Diabetes Monitoring System using Stacking Ensemble Classification

Harshini[1], *Srithar*[2]
[1,2]*Department of Computer Science and Applications, Periyar Maniammai Institute of Science & Technology (Deemed to be University), Vallam, Thanjavur, Tamil Nadu, India.*
Email ID: *harshiniharshu33@gmail.com*[1]*,sritharbalaji22@gmail.com*[2]

Abstract

As the number of people diagnosed with diabetes continues to rise, this study takes a groundbreaking approach by developing a secure web-based personalized diabetes monitoring system that incorporates analytical prediction models. This groundbreaking technology was developed in response to the critical need for sophisticated monitoring solutions that address the unique demands of each patient. The suggested method aims to transform diabetes treatment by using predictive modeling to anticipate diabetic trends and possible consequences. The study demonstrates a strong commitment to security by creating a web-based platform that handles patient data with the highest level of care. By resolving important privacy problems and creating a trustworthy environment for users, this secure framework guarantees the protection of sensitive health information. This approach takes use of several models' characteristics to make diabetes trend estimates more accurate and reliable. In addition to improving the system's prediction skills, stacking ensemble classification helps it adapt to different patient profiles. Because of the importance of accessibility and usability in encouraging patient participation, the suggested solution revolves around the creation of a user-friendly online interface. The interface is a living, breathing platform that allows for frictionless communication between healthcare practitioners and their patients. With the help of tailored insights and trend predictions, patients are better able to manage their diabetes. At the same time, doctors and hospitals have access to all the patient information they need, which allows them to take better, more preventative measures.

Keywords: *Diabetes, Monitoring System, Personalized, Predictive Modeling, Web-Based Platform.*

1. Introduction

Diabetes, or diabetes mellitus, has recently emerged as a major health concern on a global scale. Type 1 diabetes, in which the body does not make enough insulin, and Type 2 diabetes, in which the body is unable to use the insulin it does create, are the two main causes of the long-term metabolic condition known as diabetes, which causes fluctuations in blood glucose (BG) levels [1-3]. Both type 1 and type 2 diabetes have been on the rise, but type 2 has been more prevalent, making it a rising pandemic that is putting a heavy strain on healthcare systems worldwide, particularly in poorer nations [4, 5]. In 2014, over 13.7% of Koreans (4.8 million individuals) had diabetes, whereas in 2015, 9.4% of Americans and 30.3% of persons of all ages in the US had diabetes. From 171 million in 2000 to 366 million in 2030, the global amount of individuals with diabetes is expected to climb [7]. Problems including hypertension and stroke, which are cardiovascular illnesses, are triggered by poorly managed diabetes [8]. On the other hand, a major factor in lowering or preventing diabetic complications is routine blood glucose monitoring [9–11]. Innovative biosensors and new advances in information and communication technology have given rise to a new perspective on diabetes treatment by allowing for the continuous monitoring of a patient's condition in real time. Persons with diabetes are able to respond quickly and appropriately by tracking glucose fluctuations with the use of portable self-monitoring of blood glucose (SMBG) devices [12–14] and continuous glucose monitoring (CGM) sensors [3]. The findings demonstrate that keeping track of patients' glucose levels can help them gain control of their condition [12-14] and enhance the effectiveness of diabetes therapy [15,16]. The perfect solution would have sensors, a gateway (smart phone), and a cloud-based glucose monitoring platform to improve diabetes care [17]. It collects sensor data from a body-affixed node by using a smartphone as a gateway [18]. Bluetooth Low Energy (BLE) is the best solution for wireless connection between the smart phone and the sensor node [19, 20]. As a result, the sensor node can function with minimal battery usage. The main contribution of the paper is:

- Dataset preprocessing using Mean Imputation
- Feature Selection using Recursive Feature Addition
- Classification using Stacking ensemble

What follows is the outline for the rest of the article. In Section 2, a number of writers discuss different approaches to diabetes diagnosis. In Section 3, the suggested model is shown. The findings of the inquiry are reviewed in Section 4. Discussion of the outcome and plans for further research constitute Section 5's last section.

1.1. Problem Definition

A growing number of people are living with diabetes, which has prompted researchers to consider developing more sophisticated monitoring systems to meet the unique requirements of each patient. The study's overarching goal is to transform diabetes care by bringing analytical prediction models that are securely embedded within a web-based personalized diabetes monitoring system. With an emphasis on privacy and security, this study uses several models to improve the accuracy and reliability of diabetes trend forecasts while also protecting the confidentiality of patient data. Healthcare practitioners are able to take a more proactive role in patient care because to the intuitive interface, which encourages patient participation and provides personalized insights for informed decision-making.

2. Background Study

- Chou, C.Y., et al. [2] these authors research delves into a broad range of novel hybrid classifiers. A number of preprocessing procedures were used, including the prediction of missing values and the reduction of the dimension of the incoming data. These authors' experimental results show that the suggested model improves data quality and outperforms affined techniques. The current technique consists of two components. The first step was to compile the Diabetes Type dataset, and the second was to choose the Pima Indian Diabetes Dataset for practical use.

- Geetha, G., and K. Mohana Prasad [4] Type 1 diabetes was a metabolic disorder characterized by hyperglycemia, a condition that worsens over time due to insulin action resistance and decreased insulin production. Predictions of diabetes mellitus using the proposed CGRNN Meta model approach were more accurate than those using current approaches. First, the author uses the Gaussian distribution approach to identify and delete data points that were thought to be outliers. The preprocessing method mainly centers on assessing the standard deviation, mean, and probability values for every parameter in order to complete the missing value.

- Jaiswal, S., et al. [6] the investigation's focal point was on new hybrid classifiers. The dimensionality of the raw data was reduced and missing value forecasting was done utilizing a number of preprocessing approaches. The developed model outperformed affined techniques in terms of data quality and performance, as measured by these authors' practical measures. The present approach was really the result of two separate developments. Gathering the Diabetes Type information was the first step. Second, the author focused on a real-world example by selecting the Pima Indian Diabetes Dataset.

- Kasula, Balaram Yadav. [8] A potential area in diabetic healthcare was the improvement of predictive modeling, early diagnosis, and individualized patient treatment via the integration of machine learning (ML) methods. An in-depth examination of ML models' capabilities for predicting diabetic complications and managing patient outcomes was the driving force for this study.

- Nagpal D, et al. [10] It was encouraged that automated screening and diagnosis of DR be developed, since retinopathy was very prevalent among the large populations of people who were diabetic or hypertensive. Early detection and screening can lessen the likelihood of blindness. The article focuses on the HR and DR grading systems that have been effective. Ophthalmologists can now track the disease's course and lessen the chances of blindness using a mass screening approach made possible by a number of procedures. This study uses the MESSIDOR and ODIR datasets to test and validate the proposed methodologies and pseudo code.

- Rghioui, A., et al. [14] Predictive analytics can help academics and healthcare practitioners better sort through patient information, spot trends, and make educated decisions. This study set out to demonstrate a diabetes monitoring system that makes use of 5G networks and machine learning algorithms. Using artificial intelligence and big data, the author built a system that can keep an eye on diabetic patients' records and notify the proper authorities if anything goes wrong.

- Vehí, et al. [18] People with type 1 diabetes now have a better way to predict when they can have a hypoglycemic episode. By using machine learning techniques on different datasets, the author was able to perform patient health assessments, forecast continuous glucose levels, and anticipate postprandial and nocturnal hypoglycemia episodes. Different studies have ignored the systems' efficient operation in favor of CSII treatment data. However, most of the methods can be adjusted to work with MDI.

2.1. Problem Definition

In light of the alarming rise in diabetes rates, this study highlights the importance of developing more sophisticated monitoring systems. Its primary goal is to provide a safe, individualized system for tracking diabetic trends and complications via the use of predictive modeling. Secure handling of sensitive patient information is a primary goal of the system, which places a premium on privacy and security. Improved accuracy and dependability, more patient involvement, and better informed healthcare practitioner decision-making are all outcomes of this system's utilization of several models and an intuitive user interface.

3. Materials and Methods

In this section, methodology including data collecting, preprocessing, model training, system architecture design, and security procedures is necessary for the creation and implementation of the analytical prediction model and the safe web-based personalized diabetes monitoring system. This part gives a comprehensive rundown of the resources and techniques used at each stage of the project, explaining how everything came together to produce the expected results, as shown in Figure 1.

3.1. Dataset Collection

The dataset was collected from Kaggle website https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

3.2. Dataset Preprocessing Using Mean Imputation

One common method for filling up datasets with missing data is mean imputation. This technique takes the mean of all accessible values in a feature and uses it to fill in any missing values in a column or feature.

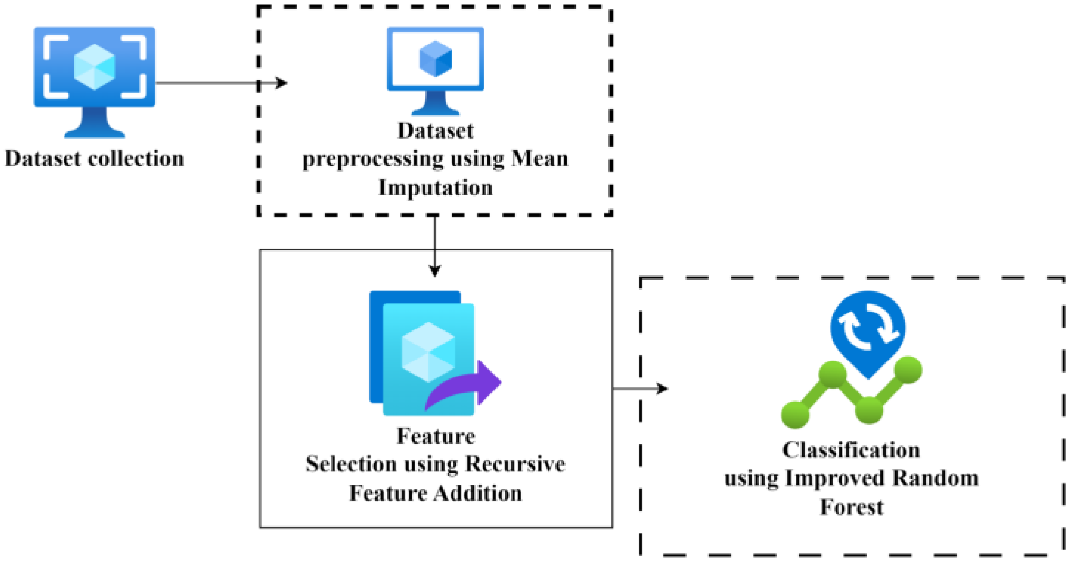
Dataset collection
Dataset preprocessing using Mean Imputation
Feature Selection using Recursive Feature Addition
Classification using Improved Random Forest

Figure 1 Overall Architecture

To fill in the missing entries while keeping the general distribution of the data intact, this strategy is especially helpful when the missing values are dispersed randomly throughout the dataset referred by Geetha, G. et al. (2023). The dataset is kept intact by replacing missing values with the mean, which allows for analysis and modeling to be done later on without major distortion. Nevertheless, it is crucial to think about how mean imputation might affect the data's variability and associations, because it could generate biases in certain cases. Validation of the imputation result is a crucial step in proving the correctness and reliability of the imputation process. To begin, there are two main types of validation methods: external and internal. Internal validation, which relies only on data set information, is used to validate the imputation process. In order to do external validation, one must examine the impact of imputation on the analysis of future biological data. Instead than relying on the data sets' internal information, external validation makes use of outside expertise. Normalized root mean square error (NMRSE) is the most used metric for evaluating the efficacy of missing value imputation algorithms. Generally speaking, a lower NMRSE indicates a more accurate missing value imputation technique. Here is the equation for NMRSE:

$$NMRSE = \sqrt{\frac{\sum_{i=1}^{m}\sum_{k=1}^{n}(g_{ik}-g_{ik})^2}{\sum_{i=1}^{m}\sum_{k=1}^{n}(g_{ik})^2}} \text{ --------- } (1)$$

This is where g and $\hat{g}$ stand for the actual and imputed values, respectively, and bik is the kth experiment for dataset value bi. As an example of an external validation strategy for missing value imputation, one could look at the effects of functional annotation of route information to downstream biological analysis. Searching for statistically significant enrichment of keywords inside gene clusters is one way to assess the effectiveness of a missing value imputation technique. This is done during a microarray experiment, the objective of which is to discover groupings of genes that have shared functional roles.

$$p = \sum_{i=k}^{min(b,T)} \frac{\binom{t}{i}\binom{B-T}{b-i}}{\binom{B}{b}} \text{ ------------- } (2)$$

Where b=number of gene in cluster.

3.3. Feature Selection using Recursive Feature Addition

An approach for selecting features that repeatedly improves a model's performance is known as Recursive Feature Addition (RFA), and it uses wrappers to do so. Using cross-validation or similar validation approach, the model's performance is tested once a feature is added. Feature additions are kept in the feature set if they increase performance above a certain threshold or criteria. In such case, it is eliminated, and the algorithm moves on to the next potential feature referred by Rghioui, A. et al. (2020).

$$j = \left(\frac{1}{2}\right) \alpha^T H \alpha - \alpha 1 \text{ ---------------- } (3)$$

Where H is the matrix that can be calculated as:

$$H = y_h y_k K(X_h X_k) \text{ ------------ } (4)$$

With x_h and x_k serving as training examples, and 1 being an n-dimensional vector with a single element. Equation 5 uses a kernel function K to measure the similarity between two samples xh andx_k, where h ranges from 1 to N and k ranges from 1 to. N features, where N is the feature count and y is the class label vector. This approach makes use of the RBF kernel function, which can be computed as:

$$K(X_h, X_k) = EXP\left(-r\left|\left|x_h - x_k\right|\right|^2\right) \text{ ------------ } (5)$$

The change in the cost function due to adding feature i requires recalculating the H matrix; hence, it is denoted$H(+i)$, with the notation $(+i)$ corresponding to feature i. The calculation $K(x_h(+i), x_k(+i))$ must be performed. To get the ultimate ranking coefficient, DJ, we need:

$$Dj = \left(\frac{1}{2}\right) \alpha^T H \alpha - \left(\frac{1}{2}\right) \alpha^T H(+i) \alpha \text{ -------------- } (6)$$

Since no features have been chosen at the beginning of the algorithm, the ranking coefficient DJ is only reduced to the second term. The prioritized feature list is supplemented with the feature that corresponds to the largest difference $DJ(i)$. a process called Recursive Feature Addition (RFA) is carried out by repeatedly running the algorithm. An ordered set of characteristics, from most essential to least, will be the output of the algorithm.

3.4. Classification using Stacking Ensemble

By incorporating state-of-the-art techniques like balanced sampling strategies to tackle class imbalance, grid search for hyper parameter fine-tuning, and Bayesian optimization for feature sampling, Stacking ensemble enhances the standard Random Forest algorithm. More precise evaluations of the model's performance can also be achieved by using out-of-bag error estimates. Stacking ensemble is a powerful tool for classification tasks in machine learning and data analysis because it uses these changes to obtain improved predicted accuracy, more resilience against over fitting, and better handling of unbalanced datasets.

As an ensemble learning method, Random Forest trains a large number of decision trees to achieve a certain objective (classification, regression, etc.), and then combines their predictions into one output. Decision trees, which can be either binary or non-binary, are diagrams that show possible actions in the shape of a tree. The decision tree starts at the root node and works its way to the leaf node by assessing the target categories related feature characteristics and selecting the relevant branches for output. In the end, the decision is dependent on the leaf node's reported categorization. A stacking model consists of unconnected decision trees. The algorithm's d parameter determined the depth of each branch, and the Gini index was used to segregate the decision tree's attributes.

Calculating the Gini Index at an internal tree node is as follows: For a potential (nominal) split attribute $XX\#$, define possible levels as $LL\&$;…;'. The Gini Index for this characteristic is computed as:

$$G(X_i) = \sum_{j=1}^{j} pr\left(X_i = L_j\right)\left(1 - pr\left(X_i = L_j\right)\right) \text{ ------ } (7)$$

$$= 1 - \sum_{j-1}^{j} pr\left(X_i = L_j\right)^2 \text{ -------- } (8)$$

Our decision to choose Random Forest was based on its superior performance compared to other machine learning methods: Even when there is a lot of missing data, it's still a wonderful way to make predictions.

```
Algorithm 1: Stacking ensemble
Input:
  • Dataset with features and target labels
Steps:
1. Initialization: Set parameters including the number of decision trees, feature sampling method, sampling strategy, and hyper parameters.
2. Training:
    • For each decision tree:
        o Sample features using the chosen feature sampling method.
        o Sample instances using the chosen sampling strategy.
```

- Train a decision tree with the sampled data.

$$G(X_i) = \sum_{j=1}^{j} pr\left(X_i = L_j\right)\left(1 - pr\left(X_i = L_j\right)\right)$$

- Aggregate the predictions of all decision trees into a single output.

3. Evaluation (Optional):

- If out-of-bag error estimation is used:
  - For each instance not used in the training of a particular decision tree, aggregate the predictions from the decision trees in which it was out-of-bag.

    $= 1 - \sum_{j-1}^{j} pr\left(X_i = L_j\right)^2$
  - Calculate the error using these aggregated predictions and the true labels.

Output:

- Trained Stacking ensemble model.

4. Results and Discussion

This ground-breaking study's findings show that by combining predictive modeling with a safe web-based monitoring system, diabetes treatment has advanced significantly. Not only does this revolutionary method priorities the privacy and security of patient data, but it also predicts trends in diabetes and its consequences.

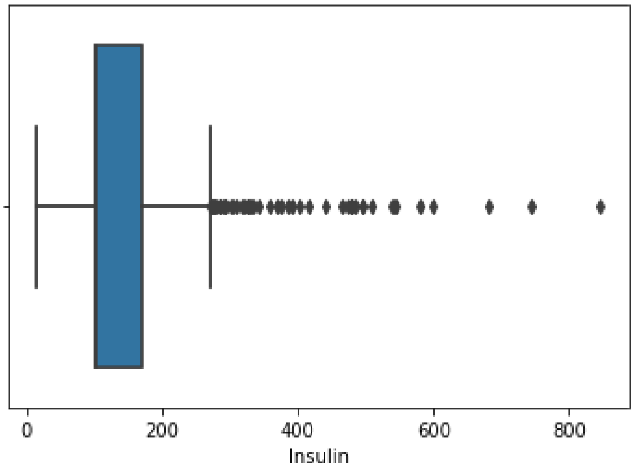

Figure 2 Insulin Levels

Insulin levels throughout a certain time period are shown graphically in Figure 2. The vertical axis shows insulin levels in units per milliliter (u/ml) or another appropriate unit, while the horizontal axis shows time or another pertinent variable. Depending on the data, the graph might depict how insulin levels fluctuate in response to things like eating, working out, taking medicine, or other physiological events. And it might show how insulin levels change throughout the day, drawing attention to patterns like diurnal changes or irregular spikes and dips that could be caused by insulin resistance or insufficient insulin synthesis.

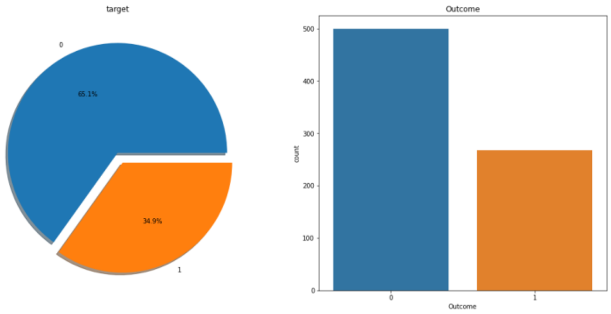

Figure 3 Analysis of Target and Outcome Variables

The distribution of a given feature among a population and the accompanying results of an event or scenario for distinct groups within that population can be understood via the study of target and outcome variables, as shown in Figure 3.

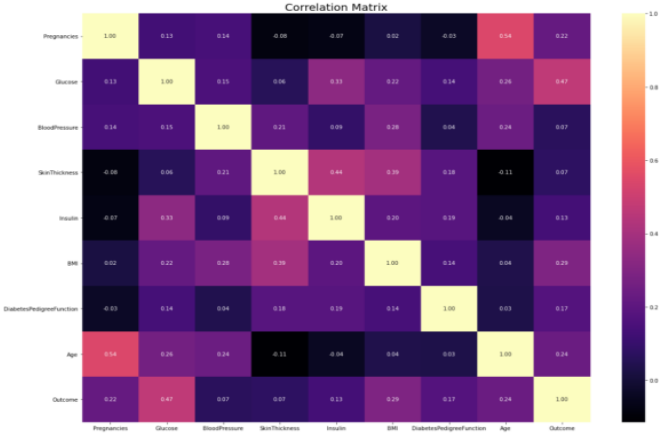

Figure 4 Correlation Matrix

Figure 4 shows a correlation matrix, which is a graphical depiction of the dataset's variables' correlation coefficients. The matrix cells provide the correlation coefficients between the two variables, which can take values between -1 and 1, and Confusion Metrics in Figure 5.

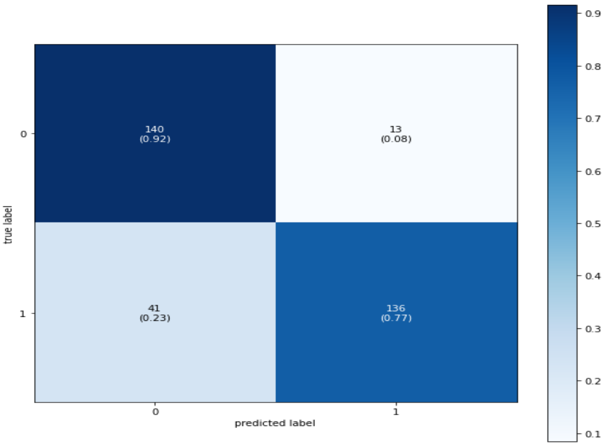

Figure 5 Confusion Metrics

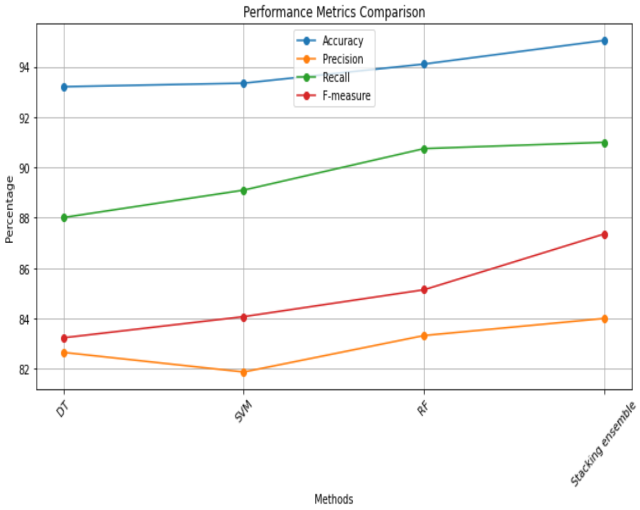

Figure 6 Performance Metrics Comparison Chart

Table 1 Performance Metrics Comparison

	Algorithm	Accuracy	Precision	Recall	F-measure
Existing methods	DT	93.21	82.65	88.01	83.23
	SVM	93.35	81.87	89.10	84.07
	RF	94.11	83.32	90.75	85.14
Proposed methods	Stacking ensemble	95.05	84.00	91.00	87.36

Significant variations in performance metrics are shown in Table 1 and Figure 6, which compare the new technique, Stacking ensemble, to three current methods: Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF). In comparison to DT's 93.21% accuracy, SVM's 83.10% recall and 81.87% precision and 84.07% F-measure score are somewhat lower. With a precision of 83.32%, an F-measure of 85.14%, and an accuracy of 94.11%, RF shows increased performance across the board. With an F-measure score of 87.36%, a precision of 84.00%, and a recall of 91.00%, the suggested stacking ensemble approach outperforms the alternatives. Its maximum accuracy is 95.05%. In terms of accuracy and F-measure in particular, these findings indicate that STACKING ENSEMBLE provides a substantial improvement over current approaches, demonstrating its promise for more efficient categorization jobs.

Conclusion

Finally, by incorporating predictive modeling into a safe, user-friendly web-based platform, this groundbreaking study presents a thorough and revolutionary method to diabetes treatment. In addition to forecasting diabetes trends and possible problems, the suggested system places a premium on the confidentiality and privacy of patient data, thereby meeting the urgent demand for customized monitoring solutions. On the other hand, the suggested stacking ensemble technique shows even more improvement in performance, with F-measure scores of 87.36%, recall of 91.00%, and accuracy of 95.05%. The system improves accuracy, dependability, and flexibility by using ensemble classification and a user-centric interface. It caters to the different demands of healthcare practitioners and patients alike. This novel approach promises to greatly enhance patient outcomes and quality of treatment with its focus on accessibility, patient participation, and informed decision-making. It signals a new era in diabetes management.

Reference

[1]. Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A. and Sherazi, H.H.R., 2021. Machine learning based diabetes classification and prediction for healthcare applications. Journal of healthcare engineering, 2021.

[2]. Chou, C.Y., Hsu, D.Y. and Chou, C.H., 2023. Predicting the onset of diabetes with machine learning methods. Journal of Personalized Medicine, 13(3), p.406.

[3]. Daza, Alfredo, Carlos Fidel Ponce Sánchez, Gonzalo Apaza-Perez, Juan Pinto, and Karoline Zavaleta Ramos. "Stacking ensemble approach to diagnosing the disease of diabetes." Informatics in Medicine Unlocked 44 (2024): 101427.

[4]. Geetha, G., and K. Mohana Prasad. "Stacking Ensemble Learning-Based Convolutional Gated Recurrent Neural Network for Diabetes Miletus." Intelligent Automation & Soft Computing 36.1 (2023).

[5]. Hacisoftaoglu, Recep E., Mahmut Karakaya, and Ahmed B. Sallam. "Deep learning frameworks for diabetic retinopathy detection with smartphone-based retinal imaging systems." Pattern recognition letters 135 (2020): 409-417.

[6]. Jaiswal, S., Gupta, P., Prasad, L.N. and Kulkarni, R., 2023. An Empirical Model for The Classification of Diabetes and Diabetes_Types Using Ensemble Approaches. Journal of Artificial Intelligence and Technology.

[7]. Kangra, Kirti, and Jaswinder Singh. "Comparative analysis of predictive machine learning algorithms for diabetes mellitus." Bulletin of Electrical Engineering and Informatics 12, no. 3 (2023): 1728-1737.

[8]. Kasula, Balaram Yadav. "Machine Learning Applications in Diabetic Healthcare: A Comprehensive Analysis and Predictive Modeling." International Numeric Journal of Machine Learning and Robots 7, no. 7 (2023).

[9]. Langarica, Saul, et al. "A meta-learning approach to personalized blood glucose prediction in type 1 diabetes." Control Engineering Practice 135 (2023): 105498.

[10]. Nagpal D, Alsubaie N, Soufiene BO, Alqahtani MS, Abbas M, Almohiy HM. Automatic Detection of Diabetic Hypertensive Retinopathy in Fundus Images Using Transfer Learning. Applied Sciences. 2023 Apr 7;13(8):4695.

[11]. Oikonomou EK, Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovascular Diabetology. 2023 Sep 25;22(1):259.

[12]. Pina, A. F., Meneses, M. J., Sousa-Lima, I., Henriques, R., Raposo, J. F., & Macedo, M. P. (2023). Big data and machine learning to tackle diabetes management. European Journal of Clinical Investigation, 53(1), e13890.

[13]. Rahim, Md Abdur, Md Alfaz Hossain, Md Najmul Hossain, Jungpil Shin, and Keun Soo Yun. "Stacked Ensemble-Based Type-2 Diabetes Prediction Using Machine Learning Techniques." Annals of Emerging Technologies in Computing (AETiC) 7, no. 1 (2023): 30-39.

[14]. Rghioui, A., Lloret, J., Sendra, S. and Oumnad, A., 2020, September. A smart architecture for diabetic patient monitoring using machine learning algorithms. In Healthcare (Vol. 8, No. 3, p. 348). MDPI.

[15]. Saihood, Q., & Sonuç, E. (2023). A practical framework for early detection of diabetes using ensemble machine learning

models. Turkish Journal of Electrical Engineering and Computer Sciences, 31(4), 722-738.

[16]. Sowah, Robert A., Adelaide A. Bampoe-Addo, Stephen K. Armoo, Firibu K. Saalia, Francis Gatsi, and Baffour Sarkodie-Mensah. "Design and development of diabetes management system using machine learning." International journal of telemedicine and applications 2020 (2020).

[17]. Tariq, Maria, Vasile Palade, YingLiang Ma, and Abdulrahman Altahhan. "Diabetic retinopathy detection using transfer and reinforcement learning with effective image preprocessing and data augmentation techniques." In Fusion of Machine Learning Paradigms: Theory and Applications, pp. 33-61. Cham: Springer International Publishing, 2023.

[18]. Vehí, Josep, Iván Contreras, Silvia Oviedo, Lyvia Biagi, and Arthur Bertachi. "Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning." Health informatics journal 26, no. 1 (2020): 703-718.

[19]. Vettoretti, Martina, Giacomo Cappon, Andrea Facchinetti, and Giovanni Sparacino. "Advanced diabetes management using artificial intelligence and continuous glucose monitoring sensors." Sensors 20, no. 14 (2020): 3870.

[20]. Zhu, Taiyu, Kezhi Li, Pau Herrero, and Pantelis Georgiou. "Basal glucose control in type 1 diabetes using deep reinforcement learning: An in silico validation." IEEE Journal of Biomedical and Health Informatics 25, no. 4 (2020): 1223-1232.