Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model

Ms. A. Pavani [1], Ms. SK. Rakheeba [2],

Assistant Professor [1][2], Department of Computer Science and Engineering, Geethanjali Institute of Science and Technology, Nellore, Andhra Pradesh-524137

Abstract: Diabetes ranks as the most prevalent ailment in developing nations, making early detection and expert medical intervention crucial to mitigating its impact. A highly effective approach for identifying diabetes involves evaluating specific indicators related to the condition. However, a common challenge in automated diabetes detection is the presence of gaps in data, which can significantly affect the performance of machine learning models. To address this, the study leverages two techniques: WithKNN-Imputer and WithoutKNN-Imputer, to handle missing values in the diabetes dataset. By employing a Stacking Classifier, which combines the predictions of a BaggingClassifier with Random Forest as the estimator and Decision Tree with LightGBM as the secondary estimator, the proposed model aims to enhance the prediction accuracy. The techniques are compared based on their ability to effectively handle missing data and generate reliable predictions. The results show that the Stacking Classifier outperforms other methods in terms of prediction accuracy and robustness, providing an effective tool for automated diabetes detection. This approach ensures that missing data does not significantly hinder model performance, offering a valuable solution for early diabetes diagnosis.

“Index Terms *- Diabetes detection, ensemble learning, missing values, KNN Imputer, healthcare”.*

1. INTRODUCTION

Healthcare professionals play a pivotal role in diagnosing and treating various medical conditions, including diseases, injuries, and physical or mental impairments. This group encompasses doctors, dentists, nurses, optometrists, physical therapists, and pharmacists, among others. The healthcare system is fundamental to maintaining both physical and mental well-being. Early disease detection is crucial, particularly for conditions like Diabetes Mellitus (DM), commonly known as diabetes. Diabetes is a condition where the body either produces insufficient insulin or fails to effectively manage it. Insulin regulates blood sugar levels, and when diabetes goes uncontrolled, it leads to hyperglycemia (high blood sugar), which can severely damage organs and systems, particularly the nervous and blood vessel systems [1]. In 2014, diabetes affected 8.5% of the global population over the age of 18, and in 2012, it was responsible for 2.2 million deaths worldwide, a toll that rose to 1.6 million in 2016 [1][2][3]. By 2019, the death toll exceeded 1.5 million, and the increasing number of diabetes patients has become a significant economic burden, with global costs approaching 825 billion dollars annually for diabetes care [6]. Projections suggest that by 2045, the number of diabetes sufferers could reach 629 million globally [7].

Diabetes Mellitus is categorized into four types: Type 1 (juvenile or insulin-dependent diabetes), Type 2 (non-insulin-dependent), Gestational Diabetes (GD), and impaired glucose regulation (pre-diabetes, Type 4). Type 1 diabetes is characterized by the body ' s failure to produce insulin, requiring external insulin injections. Type 2 diabetes, however, does not require external insulin supplementation but results from the body ' s inability to properly use insulin. Gestational diabetes occurs during pregnancy, where blood sugar levels rise in women who were previously not diabetic. Type 4, or pre-diabetes, is characterized by blood sugar levels higher than normal but not high enough to be classified as Type 2 diabetes. The risk factors for DM include elevated blood glucose levels, high fasting blood sugar, high triglycerides, lack of physical activity, age, high blood pressure, obesity, family history, and pregnancy [8].

The rising incidence of DM in both developing and developed countries is linked to a sedentary lifestyle, poor dietary habits, and other socio-economic factors like stress and lack of healthcare knowledge. To address this growing issue, technologies such as meal suggestion systems, activity trackers, drug warning systems, and interactive chatbots are being used to manage and treat diabetes more effectively. Data mining and machine learning (ML) are also essential tools in healthcare, enabling faster and more accurate diagnoses. However, data used for diabetes prediction often contains incomplete or missing values, which can affect the performance of ML models, making it crucial to address these gaps for better accuracy and reliability in diagnosis [9][10].

2. RELATED WORK

Diabetes Mellitus (DM) continues to be a major health concern globally, with increasing prevalence rates and serious consequences for individuals and healthcare systems alike. As the number of diabetes cases continues to rise, there is a pressing need for early detection and effective management strategies to mitigate the impacts of this chronic condition. Machine learning (ML) and data mining have emerged as critical tools in healthcare for automated disease diagnosis, particularly for diabetes prediction. A growing body of research has explored various ML algorithms and methodologies to enhance the accuracy and efficiency of diabetes prediction systems.

In recent years, multiple studies have focused on utilizing machine learning algorithms for the prediction of diabetes. For example, Deberneh and Kim [11] proposed a machine learning-based prediction model for Type 2 diabetes, leveraging algorithms like Random Forest and Support Vector Machines (SVM) to improve prediction accuracy. The authors found that these models could identify potential diabetes patients with high accuracy, underlining the importance of leveraging machine learning in healthcare applications. Moreover, Rupapara et al. [12] utilized Chi-square and Principal Component Analysis (PCA)-based feature selection methods to improve diabetes detection. Their study employed an ensemble classifier combining multiple models to enhance performance, showcasing the benefits of feature selection techniques in diabetes classification tasks. By reducing the dimensionality of the dataset through PCA, the models were able to focus on the most relevant features, improving classification accuracy and robustness.

Another important advancement in diabetes prediction has been the incorporation of deep learning techniques. Deng et al. [13] demonstrated the application of deep transfer learning and data augmentation techniques for improving glucose level predictions in Type 2 diabetes patients. The study emphasized that the use of transfer learning, which involves pre-training models on large datasets and fine-tuning them for specific applications, can significantly improve prediction performance. This approach was further enhanced by data augmentation, a technique that generates synthetic data to overcome the challenges posed by limited or imbalanced datasets, making it particularly valuable for medical predictions where data availability may be a limiting factor. Their work illustrates the power of deep learning techniques in diabetes prediction, especially when working with complex, non-linear data patterns.

Similarly, Butt et al. [14] employed various machine learning models to classify and predict diabetes outcomes. Their study focused on utilizing a variety of algorithms, including decision trees, Random Forests, and SVMs, to determine the best approach for healthcare applications. The research found that ML models could effectively be used to analyze patient data and identify individuals at risk of developing diabetes, thus enabling early intervention. Their study underscored the importance of selecting the right algorithms for diabetes prediction and emphasized the growing significance of using ML in healthcare settings to support decision-making processes.

Pethunachiyar [15] presented another approach, utilizing kernel-based support vector machines for classifying diabetes patients. This technique, which uses a non-linear decision boundary to separate different classes, has been widely used in classification tasks due to its robustness and effectiveness in handling high-dimensional data. In this study, the kernel-based SVM outperformed other models in diabetes classification, demonstrating the efficacy of SVMs in managing complex medical datasets with numerous features. The results highlighted the relevance of SVMs for automated diabetes detection systems, where the data often contains numerous features and may require advanced techniques to ensure accurate classification.

Laila et al. [16] explored an ensemble approach to predicting early-stage diabetes risk using machine learning. The ensemble method involved combining multiple models to leverage their individual strengths and improve overall prediction performance. The authors tested a range of models, including decision trees and logistic regression, and combined them into an ensemble learning framework. This approach was successful in identifying high-risk individuals, demonstrating that ensemble methods could significantly improve prediction accuracy, especially when dealing with heterogeneous data sources. Their study emphasized the benefits of ensemble techniques, which combine the predictions of multiple models to reduce the likelihood of errors and improve robustness, making it a powerful tool for diabetes prediction.

Furthermore, Madan et al. [17] explored the use of optimization-based diabetes prediction models, combining Convolutional Neural Networks (CNNs) with Bi-directional Long Short-Term Memory (Bi-LSTM) networks in a real-time environment. The hybrid CNN-Bi-LSTM model was designed to handle both spatial and temporal patterns in diabetes-related data, including patient medical records and sensor data. This combination enabled the model to effectively capture complex dependencies between different variables, which is critical in diabetes prediction. By optimizing the network architecture and training process, the authors were able to achieve impressive results, demonstrating the potential of deep learning models in predicting diabetes in real-time settings. The use of CNNs for feature extraction and Bi-LSTMs for sequence modeling in this context highlights the growing trend of incorporating deep neural networks into healthcare applications to manage complex, multi-dimensional data.

In addition to the above studies, several other works have focused on integrating advanced data preprocessing techniques to enhance the performance of ML models. One of the key challenges in diabetes prediction is the presence of incomplete or missing data in medical records, which can significantly affect model accuracy. Approaches such as KNN imputation have been used to address missing data, allowing models to better handle incomplete datasets. Other studies have explored the use of feature selection and dimensionality reduction techniques to improve model performance by focusing on the most important predictors of diabetes. These techniques aim to reduce noise and irrelevant information, allowing models to learn more effectively from the data and make more accurate predictions.

Moreover, researchers have also focused on making diabetes prediction models more interpretable. As the use of machine learning in healthcare increases, it becomes essential to ensure that the decisions made by these models can be understood and trusted by healthcare professionals. Techniques such as Explainable AI (XAI) are gaining traction in the field of diabetes prediction, helping to provide transparent and interpretable results. By making the reasoning behind predictions clear, these approaches aim to increase the adoption of ML models in clinical settings, where trust and transparency are paramount.

While these advances in machine learning and data mining have led to significant improvements in diabetes detection, challenges remain. One of the key limitations is the availability of high-quality data. Diabetes prediction models require large, well-annotated datasets that include a variety of patient characteristics, medical history, and lab results. In many cases, data can be incomplete or noisy, and this can affect the model's ability to make accurate predictions. To address these challenges, researchers are increasingly turning to synthetic data generation, data augmentation, and transfer learning techniques, which can help mitigate the impact of limited or missing data.

3. MATERIALS AND METHODS

The proposed system aims to develop an advanced diabetes prediction model by leveraging a variety of machine learning algorithms and techniques. We will begin by preprocessing the dataset using KNN-Imputer [18] to handle missing data, comparing the results with a version that does not use KNN-Imputer for imputation. The dataset will be based on real-world medical data, featuring attributes such as patient demographics, medical history, and lab results. We will apply several algorithms, including Logistic Regression, Decision Tree [20], Random Forest [21], Stochastic Gradient Descent (SGD), ExtraTree, XGBoost, Support Vector Machine (SVM), and Naive Bayes, to assess their predictive capabilities. For ensemble methods, we will utilize a Voting Classifier [19] combining ExtraTree, XGBoost, and Random Forest, along with a Stacking Classifier that combines BaggingClassifier with Random Forest as an estimator and Decision Tree with LightGBM as an estimator. The models will be evaluated using K-Fold cross-validation to ensure robust performance and reliable predictions. The goal is to develop a high-accuracy, scalable diabetes prediction system.

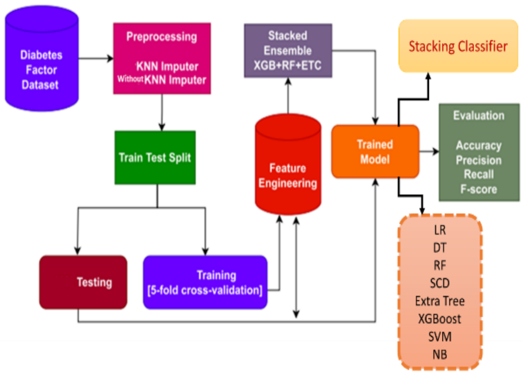
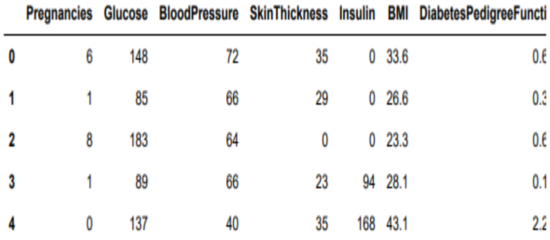


Fig.1 Proposed Architecture

The system architecture involves preprocessing diabetes factor data, splitting it into training and testing sets, and applying feature engineering. It then employs a stacking ensemble technique with various base models (LR, DT [20], RF [21], etc.) to create a stacked model. The trained model is evaluated using metrics like accuracy, precision, recall, and F-score.

i) Dataset Collection:

The diabetes dataset comprises key clinical and physiological attributes to predict diabetes onset. It includes features such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age, along with the target variable Outcome (0 for non-diabetic, 1 for diabetic). The data reflects critical indicators influencing diabetes diagnosis, enabling detailed analysis and machine learning-based predictions. Collected from healthcare records, this dataset provides a reliable foundation for evaluating diverse machine learning algorithms and improving diabetes detection accuracy.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFuncti
0	6	148	72	35	0	33.6	0.6
1	1	85	66	29	0	26.6	0.3
2	8	183	64	0	0	23.3	0.6
3	1	89	66	23	94	28.1	0.1
4	0	137	40	35	168	43.1	2.2

Fig.2 Dataset Collection Table – Diabetes

ii) Pre-Processing:

Pre-processing transforms raw data into a clean and structured format, ensuring quality for analysis. It includes data cleaning, exploratory data analysis (EDA), and feature selection to improve model performance.

a) Data Processing: Data processing begins by removing duplicate records to eliminate redundancy and ensure data integrity. Next, irrelevant or null entries are dropped during the cleaning phase to reduce noise and enhance the dataset’s quality. Categorical features, if present, are converted into numerical representations using label encoding, enabling compatibility with machine learning algorithms. These steps collectively refine the dataset, making it suitable for effective analysis and predictive modeling. Clean, duplicate-free data forms the foundation for accurate machine learning outcomes.

b) EDA: EDA involves understanding data distributions, relationships, and patterns. A correlation matrix is utilized to identify interdependencies among features, helping highlight strongly correlated attributes. Visualizations and sample outcome analysis provide insights into the target variable distribution (diabetic vs. non-diabetic). EDA enables detecting anomalies, missing values, and trends in the dataset, offering a comprehensive understanding of its structure. These insights are crucial for informed decision-making in feature engineering and selection for optimal model performance.

c) Feature Selection: Feature selection identifies the most relevant attributes contributing to the predictive task. Techniques like correlation analysis, mutual information, or statistical tests filter out redundant or less informative features. This reduces dimensionality, enhances model interpretability, and prevents overfitting. By selecting impactful features such as Glucose, BMI, and Insulin, the model focuses on critical determinants of diabetes. Feature selection streamlines the dataset, ensuring efficient computation while improving the accuracy and generalizability of machine learning models.

iii) Training & Testing:

The training phase involves preparing the model by feeding it with the labeled dataset, allowing it to learn patterns and relationships within the data. During testing, the trained model is evaluated on a separate dataset to assess its generalization ability and performance. The testing phase provides insights into how well the model can make predictions on unseen data, helping to ensure its robustness and reliability. This process is crucial for determining the model's effectiveness in real-world scenarios.

iv) Algorithms:

LR: Logistic Regression is used to model the probability of diabetes based on the features in the dataset. It uses a logistic function to estimate the outcome, providing a binary classification (diabetic or non-diabetic). Both WithKNN-Imputer and WithoutKNN-Imputer techniques are applied to handle missing values and improve model accuracy by imputing missing data before training.

DT: The Decision Tree algorithm creates a model by splitting the data at each node based on the most informative feature. It is used for classifying diabetes cases, offering an interpretable decision process. WithKNN-Imputer and WithoutKNN-Imputer techniques are used to fill in missing values, ensuring robust splits during tree construction, which optimizes model performance [20].

RF: Random Forest is an ensemble method that uses multiple decision trees for classification. Each tree contributes to the final classification decision, making it more accurate and robust. It is applied to classify diabetes cases and handles missing data efficiently using WithKNN-Imputer and WithoutKNN-Imputer techniques to improve prediction reliability and prevent overfitting [21].

SGD: Stochastic Gradient Descent is a gradient-based optimization technique used to minimize the loss function and improve model performance for large datasets. It is applied for binary classification in diabetes detection. WithKNN-Imputer and WithoutKNN-Imputer techniques handle missing values to optimize the learning process and ensure that the model learns effectively despite data gaps.

ExtraTree: ExtraTree is an ensemble algorithm that builds multiple decision trees using random feature selection and splits. It is used to improve classification accuracy for predicting diabetes. Missing data is addressed through WithKNN-Imputer and WithoutKNN-Imputer techniques to enhance the decision-making process, ensuring that incomplete data does not degrade model performance.

XGBoost: XGBoost is a gradient boosting algorithm that improves the classification performance by sequentially combining weak learners into a stronger model. It is utilized to detect diabetes based on feature patterns. WithKNN-Imputer and WithoutKNN-Imputer techniques are used to handle missing data, improving the stability and accuracy of the final boosted model.

SVM: Support Vector Machine creates a hyperplane that maximizes the margin between classes. It is applied to classify individuals as diabetic or non-diabetic. WithKNN-Imputer and WithoutKNN-Imputer techniques handle missing data, ensuring that the algorithm can create a clear decision boundary despite gaps in the dataset, thus improving model accuracy.

NB: Naive Bayes uses probabilistic approaches to classify diabetes by calculating the likelihood of each outcome given the features. It is particularly effective for handling large datasets. WithKNN-Imputer and WithoutKNN-Imputer techniques are employed to fill in missing values, allowing the model to make informed predictions even when data is incomplete.

VC: The Voting Classifier combines the predictions of ExtraTree, XGBoost, and Random Forest models, using a majority vote to classify diabetes. This ensemble method improves classification accuracy. Missing data is handled using WithKNN-Imputer and WithoutKNN-Imputer techniques to ensure that each model in the ensemble has complete data, leading to more reliable predictions [19].

SC: The Stacking Classifier uses an ensemble approach by combining BaggingClassifier with Random Forest as an estimator, along with Decision Tree and LightGBM for stacking. It leverages multiple base learners to improve performance in diabetes classification. WithKNN-Imputer and WithoutKNN-Imputer techniques address missing data across all models in the ensemble, ensuring the overall system's robustness and prediction accuracy.

4. RESULTS & DISCUSSION

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} (1)$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

$$Precision = \frac{\text{True Positive}}{\text{True Positive } + \text{False Positive}} (2)$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP}\ +\ \mathrm{FN}} (3)$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$F1\ Score = 2 * \frac{Recall \text{ X } Precision}{Recall + Precision} * 100(1)$$

Tables (1 & 2) evaluate the performance metrics—F1-score, precision, recall, and accuracy—for each algorithm. Across all metrics, the Stacking Classifier consistently outperforms all other algorithms. The tables also offer a comparative analysis of the metrics for the other algorithms.

Table.1 Performance Evaluation Metrics for With KNN Imputator

ML Model	Accuracy	f1_score	Recall	Precision
LogisticRegression	0.773	0.775	0.773	0.780
DecisionTree	0.740	0.739	0.740	0.739
RandomForest	0.760	0.756	0.760	0.758
SGD	0.669	0.663	0.669	0.707
ExtraTree	0.662	0.657	0.662	0.658
XGBoost	0.708	0.704	0.708	0.704
SVM	0.649	0.781	0.649	0.994
Naive Bayes	0.727	0.724	0.727	0.724
Voting Classifier	0.727	0.723	0.727	0.725
Stacking Classifier	**0.974**	**0.974**	**0.974**	**0.975**

Graph.1 Comparison Graphs for With KNN Imputator

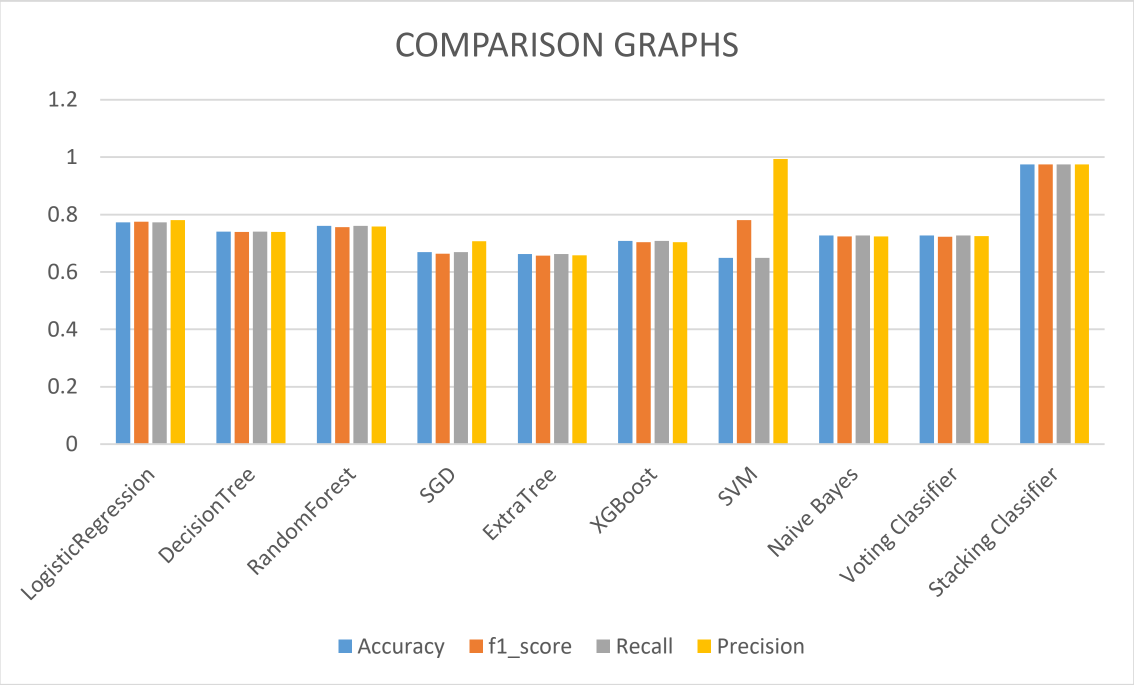
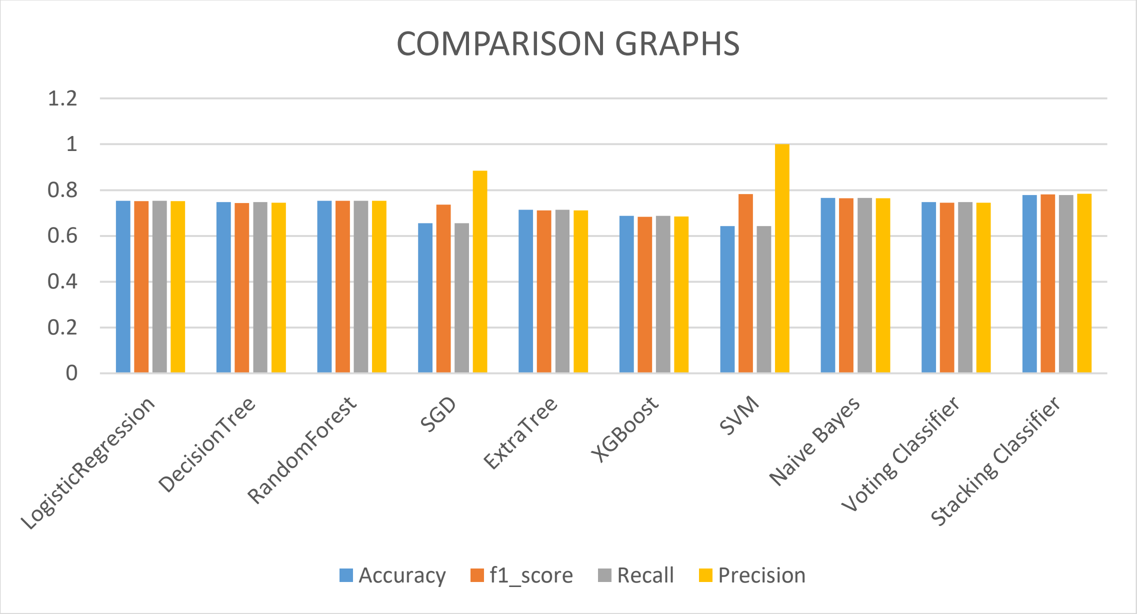


Table.2 Performance Evaluation Metrics for Without KNN Imputator

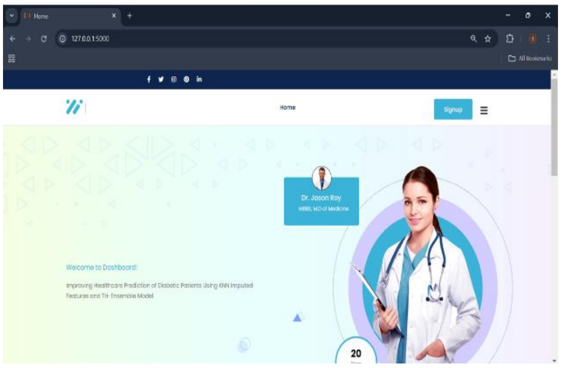
ML Model	Accuracy	f1_score	Recall	Precision
LogisticRegression	0.753	0.752	0.753	0.752
DecisionTree	0.747	0.743	0.747	0.745
RandomForest	0.753	0.753	0.753	0.753
SGD	0.656	0.736	0.656	0.885
ExtraTree	0.714	0.711	0.714	0.711
XGBoost	0.688	0.684	0.688	0.685
SVM	0.643	0.783	0.643	1.000
Naive Bayes	0.766	0.765	0.766	0.764
Voting Classifier	0.747	0.745	0.747	0.745
Stacking Classifier	**0.779**	**0.781**	**0.779**	**0.784**

Graph.2 Comparison Graphs for Without KNN Imputator



Accuracy is represented in blue, F1-score in orange, recall in grey, and precision in light yellow in ***Graphs (1 & 2)***. In comparison to the other models, the Stacking Classifier shows superior performance across both techniques, achieving the highest values. The graphs above visually illustrate these findings.

Dashboard – User Interface

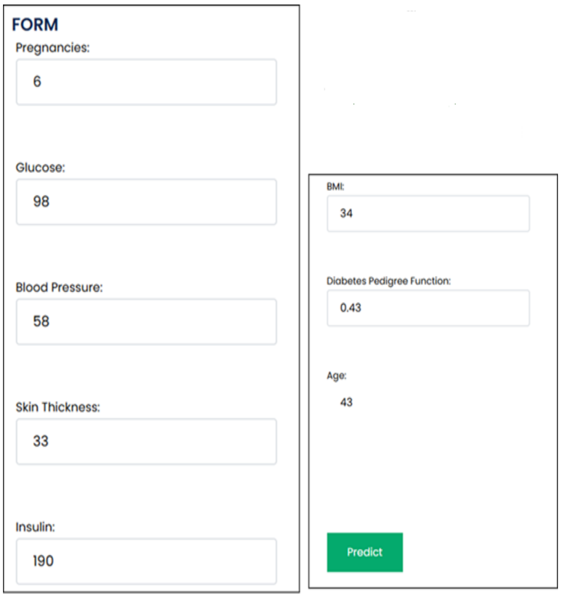


Welcome to Dashboard!

Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model

Fig.3 Home Page

In the above figure 3, this is a user interface dashboard, it is a welcome message for navigating page.



FORM

Pregnancies: 6

Glucose: 98

Blood Pressure: 58

Skin Thickness: 33

Insulin: 190

BMI: 34

Diabetes Pedigree Function: 0.43

Age: 43

Fig.4 User input Page

In the above figure 4, this is a user input page, using this user can upload data for testing.

OUTCOME

PREDICTED WITH NO DIABETIC!

Fig.5 Classification result

In the above figure 5, this is a result screen, in this user will get output for loaded input data.

5. CONCLUSION

In recent times, there has been a noticeable surge in the prevalence of diabetes, impacting millions of individuals worldwide. Timely interventions hold the key to mitigating the intricate complications associated with diabetes. Among the various machine learning algorithms evaluated for predicting diabetes, the Stacking Classifier, which combines BaggingClassifier with Random Forest as the estimator and Decision Tree with LightGBM as the stacking estimator, emerged as the highest-performing model. This algorithm achieved an impressive accuracy of 97.4% when employing the WithKNN-Imputer technique, demonstrating its superior capability to handle missing data and deliver highly accurate predictions. In contrast, when the WithoutKNN-Imputer technique was applied, the model showed a lower accuracy of 77.9%. The significant performance improvement observed with the WithKNN-Imputer technique underscores the importance of imputing missing data for maximizing prediction accuracy. The Stacking Classifier’s exceptional performance highlights its potential for providing reliable and effective diabetes risk predictions, ultimately contributing to better early detection and management of diabetes-related complications.

Future research aims to integrate deep learning models into diabetes prediction to achieve even greater accuracy and robustness. By leveraging advanced neural network architectures, the model is expected to handle larger and more complex datasets effectively, enhancing performance and versatility. This direction promises to unlock significant improvements in prediction capabilities, addressing the challenges of high-dimensional data and contributing to more resilient and precise diabetes detection systems in future applications.

REFERENCES

[1] Diabetes Gojka. (Jul.2019). Diabetes: World Health Organization (WHO). Accessed: May 25, 2023.

[2] S. El-Sappagh, F. Ali, S. El-Masri, K. Kim, A. Ali, and K.-S. Kwak, “Mobile health technologies for diabetes mellitus: Current state and future challenges,’’ IEEE Access, vol. 7, pp. 21917–21947, 2019.

[3] L. Mertz, “Automated insulin delivery: Taking the guesswork out of diabetes management,’’ IEEE Pulse, vol. 9, no. 1, pp. 8–9, Jan. 2018.

[4] H. A. Klein and A. R. Meininger, “Self management of medication and diabetes: Cognitive control,’’ IEEE Trans. Syst., Man, Cybern., A, Syst. Hum., vol. 34, no. 6, pp. 718–725, Nov. 2004.

[5] WHO. (Apr. 2023). Diabetes: World Health Organization (WHO). Accessed: May 25, 2023.

[6] A. A. Al Jarullah, “Decision tree discovery for the diagnosis of type II diabetes,’’ in Proc. Int. Conf. Innov. Inf. Technol., Apr. 2011, pp. 303–307.

[7] G.D.Kalyankar, S.R.Poojara, and N.V.Dharwadkar, “Predictive analysis of diabetic

patient data using machine learning and Hadoop,'' in Proc. Int. Conf. I-SMAC, Feb. 2017, pp. 619–624.

[8] B.S.Ahamed, M.S.Arya, and A.O.V.Nancy, "Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation,'' Adv. Hum.-Comput. Interact., vol. 2022, pp. 1–14, Sep. 2022.

[9] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data miningclassification techniques to predict diabetes,'' Proc. Comput. Sci., vol. 82, pp. 115–121, Jan. 2016.

[10] I. Kavakiotis, O. Tsave, and A. Salifoglou, "Machine learning and data mining methods in diabetes research,'' Comput. Struct. Biotechnol. J., vol. 15, no. 9, pp. 104–116, 2017.

[11] H. M. Deberneh and I. Kim, ''Prediction of type 2 diabetes based on machine learning algorithm,'' Int. J. Environ. Res. Public Health, vol. 18, no. 6, p. 3317, Mar. 2021.

[12] V. Rupapara, F. Rustam, A. Ishaq, E. Lee, and I. Ashraf, ''Chi-square and PCA based feature selection for diabetes detection with ensemble classi f ier,'' Intell. Autom. Soft Comput., vol. 36, no. 2, pp. 1931–1949, 2023.

[13] Y. Deng, L. Lu, L. Aponte, A. M. Angelidi, V. Novak, G. E. Karniadakis, and C. S. Mantzoros, ''Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients,'' NPJ Digit. Med., vol. 4, no. 1, p. 109, Jul. 2021.

[14] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, ''Machine learning based diabetes classification and prediction for healthcare applications,'' J. Healthcare Eng., vol. 2021, pp. 1–17, Sep. 2021.

[15] G. A. Pethunachiyar, ''Classification of diabetes patients using kernel based support vector machines,'' in Proc. Int. Conf. Comput. Commun. Informat. (ICCCI), Jan. 2020, pp. 1–4.

[16] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, ''An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study,'' Sensors, vol. 22, no. 14, p. 5247, Jul. 2022.

[17] P. Madan, V. Singh, V. Chaudhari, Y. Albagory, A. Dumka, R. Singh, A. Gehlot, M. Rashid, S. S. Alshamrani, and A. S. AlGhamdi, ''An optimization-based diabetes prediction model using CNN and bi directional LSTM in real-time environment,'' Appl. Sci., vol. 12, no. 8, p. 3989, Apr. 2022.

[18] A.Juna, M.Umer, S.Sadiq, H.Karamti, A.A.Eshmawi, A.Mohamed, and I. Ashraf, "Water quality prediction using KNN imputer and multilayer perceptron,'' Water, vol. 14, no. 17, p. 2592, Aug. 2022.

[19] Y. Zhang, H. Zhang, J. Cai, and B. Yang, "A weighted voting classifier based on differential evolution,'' Abstract Appl. Anal., vol. 2014, pp. 1–6, Jan. 2014.

[20] M. Brijain, R. Patel, M. R. Kushik, and K. Rana, "A survey on decision tree algorithm for classification,'' Int. J. Eng. Develop. Res., vol. 2, no. 1, pp. 1–5, 2014.

[21] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests,'' Statist. Comput., vol. 27, no. 3, pp. 659–678, May 2017.