JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION

ONLINE ISSN
3006-9726

PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 4 (2025)

Synergistic Intelligence: A Stacking Ensemble Approach for Accurate and Scalable Diabetes Prediction

Irfanullah*
Center of Excellence in IT, Institute of Management Sciences (IMSciences), Peshawar, Peshawar, Pakistan
Email: irfanullahofficial12@gmail.com
Muhammad Shahan Ibad
Center of Excellence in IT, Institute of Management Sciences (IMSciences), Peshawar, Peshawar, Pakistan
Email: shaniims2022@gmail.com
Aamir Sohail
Bank of Khyber, Cybersecurity Department (SOC Analyst), Peshawar, Khyber Pakhtunkhwa, Pakistan
Email: aamirsohail.soc@gmai.com

Corresponding Author: Irfanullah (irfanullahofficial12@gmail.com)

Received	**Approved**	**Published**
13th March 2025	12th June 2025	15th June 2025

Abstract:

Diabetes mellitus poses a rapidly escalating global health crisis, currently affecting over 537 million adults and demanding scalable, automated diagnostic solutions. However, current machine learning interventions often face critical bottlenecks, particularly model overfitting and poor generalization due to severe class imbalance in clinical datasets. To overcome these limitations, this study engineers a robust, clinically applicable Stacking Ensemble framework validated on the Pima Indians Diabetes Dataset. We employed a rigorous data preprocessing pipeline that utilizes the Synthetic Minority Oversampling Technique (SMOTE) to rectify class distribution, ensuring unbiased decision boundaries. By strategically integrating the complementary strengths of Logistic Regression, Support Vector Machines, K-Nearest Neighbors, and Naive Bayes via a meta-learning architecture, our approach successfully mitigates the individual weaknesses of single classifiers. The proposed ensemble demonstrated superior performance, achieving an accuracy of 81.5% and a critical recall rate of 84.0%, significantly reducing the risk of missed diagnoses compared to baseline models. Crucially, the system maintains exceptional computational efficiency with an inference latency of only 27.43 ms, confirming its viability for real-time deployment in resource-constrained medical environments. This research bridges the gap between algorithmic complexity and practical utility, offering a scalable, interpretable solution for early diabetes detection

Keywords: *Diabetes prediction, Ensemble stacking model, Machine learning classifiers, Early diagnosis and screening.*

Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by persistent hyperglycemia due to insufficient insulin production or impaired cellular response [1]. It manifests primarily as Type 1 Diabetes (autoimmune destruction of beta cells, 5 to 10% of cases) [2], Type 2 Diabetes (insulin resistance with inadequate secretion, >90% of cases) [3], and Gestational Diabetes (pregnancy-induced insulin resistance with later risk of Type 2) [3]. Risk factors include family history, obesity, sedentary lifestyle, hypertension, abnormal cholesterol, polycystic ovarian syndrome, and advanced age, with risk increasing particularly in older adults [1]. If untreated, diabetes leads to acute complications such as diabetic ketoacidosis and hypoglycemia, and long-term damage including cardiovascular disease, nephropathy, neuropathy, retinopathy, poor wound healing, and infections [4]. Symptoms range from polyuria, polydipsia, and polyphagia to unexplained weight loss, fatigue, blurred vision, and neuropathic pain, with Type 1 presenting abruptly and Type 2 progressing insidiously [3]. Globally 537 million adults are affected, projected to rise to 783 million by 2045, making diabetes a top cause of death and a driver of significant healthcare costs and economic burden [5]. Management involves lifestyle modification, pharmacological agents (metformin, sulfonylureas, SGLT2 inhibitors), insulin therapy, continuous glucose monitoring (CGM) or self-monitoring, and preventive screening [4]. Historically, symptoms resembling diabetes were first described in the Ebers Papyrus around 1500 BCE, and Paul Langerhans identified pancreatic islets in 1869. The landmark discovery came in 1921 when Frederick Banting, Charles Best, J.J.R. Macleod, and James Collip isolated and purified insulin, transforming diabetes from a fatal disease to a manageable condition an achievement that earned the Nobel Prize in Physiology or Medicine in 1923 [6]. Despite advances, prevalence continues to rise, underscoring the need for early detection. This study evaluates and compares machine learning models Logistic Regression, Random Forest, Support Vector Classifier, K-Nearest Neighbors, and Stacking Ensemble to identify the most effective predictive approach for diabetes classification.

In recent years, a variety of machine learning and deep learning models have been applied to diabetes prediction, with a strong emphasis on individual classifiers. Such as Logistic Regression is simple and interpretable but may struggle with non-linear patterns. k-Nearest Neighbors (KNN) is effective for capturing local data structures but is sensitive to noise and scaling. Naive Bayes performs well with small datasets and offers high recall, yet often suffers from low precision and AUC due to its strong independence assumptions. Support Vector Machine (SVM) handles complex boundaries well but can be computationally intensive and less transparent. Individual classifiers used in diabetes prediction each have their own strengths and weaknesses. Previous studies also highlight limitations such as reliance on small and imbalanced datasets like the Pima Indian Diabetes dataset, which reduces generalizability. Deep learning models, while powerful, often require high computational resources and lack interpretability. Additionally, many works overlook robust validation strategies and hyperparameter tuning, increasing the risk of overfitting and limiting clinical applicability.

To address existing gaps this study, propose a stacking ensemble model using k-Nearest Neighbors, Logistic Regression, Naive Bayes, and SVM as base learners with a meta learner to combine their outputs. Each model offers unique strengths Naive Bayes handles high-dimensional data well, Support Vector Machine excels at boundary separation, k-Nearest Neighbors captures local patterns, and Logistic Regression provides interpretability. Individually, these models optimize different metrics but none perform consistently across all. Stacking allows us to balance accuracy, recall, specificity, and AUC by leveraging their complementary strengths. Unlike deep learning or tree-heavy ensembles, this approach is lightweight, interpretable and generalizable. The study also address dataset issues through preprocessing and tuning to reduce overfitting and improve robustness.

LITERATURE REVIEW

Machine learning algorithms have been widely applied to diabetes prediction, aiming to identify risk factors and improve early diagnosis. To begin with, [7] introduced SACDNet, a self-attention convolutional network designed for early Type 2 diabetes prediction using electronic health records. The dataset comprised 4,124 diabetic and 181,767 non-diabetic records, making it one of the largest in this domain. The model incorporated MC-Dropout for uncertainty quantification, which allowed the system to estimate prediction confidence. SACDNet achieved an accuracy of 89.3% with an F1-score of 89.1%, demonstrating strong performance on imbalanced data. However, despite its scale, the study struggled with class imbalance and limited interpretability, which are critical for clinical adoption.

Building on this, [8] proposed KCCAM_DNN, a deep neural network that integrates Kendall correlation, attention mechanisms, and SHAP-based interpretation. The model was tested on the Pima Indians Diabetes dataset and other public datasets, achieving nearly 95% accuracy. The use of Kendall correlation improved feature selection, while SHAP values provided interpretability by ranking the most influential predictors. This combination of accuracy and explainability was a step forward. Nevertheless, the study lacked external validation across diverse populations, raising concerns about generalizability. In a related effort, [9] developed a robust predictive framework for diabetes classification by conducting a comparative evaluation across multiple datasets, including Pima and other public sources. The study systematically compared a wide range of ML techniques, incorporating feature engineering, cross-validation, and hyperparameter tuning. Results indicated consistently high accuracy across datasets, underscoring the effectiveness of comparative frameworks. However, the authors noted that models trained on one dataset often underperformed when applied to another, highlighting the persistent challenge of cross-domain generalization.

Similarly, [10] conducted a comprehensive study integrating both ML and DL models, including Backpropagation Neural Networks (BPNN), Random Forest, and XGBoost. Using the Pima Indians dataset and other public sources, the study reported accuracies close to 90%. The authors emphasized that ensemble and deep learning methods consistently outperformed single classifiers. However, they also pointed out issues of dataset bias and reproducibility, stressing the importance of standardized evaluation protocols and reproducible pipelines for diabetes prediction. Furthermore, [11] explored deep learning-based noninvasive screening of Type 2 diabetes using clinical datasets. Their approach employed advanced neural architectures to classify patients based on noninvasive physiological data such as vital signs and lifestyle indicators. The models achieved high predictive performance, suggesting that noninvasive screening could be a viable alternative to costly laboratory tests. However, the study did not provide detailed accuracy metrics, and it lacked validation across multiple datasets, leaving questions about its robustness.

In addition, [12] focused on random oversampling-based diabetes classification via the Internet of Medical Things (IoMT). Using the Pima Indians and BRFSS datasets, the authors applied random oversampling combined with ML pipelines, achieving accuracy above 90%. The study demonstrated that oversampling could effectively address class imbalance. Nevertheless, the authors cautioned that oversampling may lead to overfitting, and they emphasized the need for real-time validation in IoMT environments to ensure practical applicability. Likewise, [13] conducted a comparative study of ML techniques for early diabetes prediction using the Pima Indians dataset. The study evaluated LR, DT, RF, KNN, NB, SVM, Gradient Boosting, and Neural Networks, with Neural Networks achieving the best performance at 78.57%. The findings highlighted the limitations of small, imbalanced datasets, as even advanced models struggled to achieve high accuracy. The study concluded that larger and more diverse datasets are essential for reliable prediction. From a broader perspective, [14] conducted a systematic review of AI-based techniques for diabetes prediction, synthesizing findings across multiple datasets and algorithmic approaches. The review analyzed a wide range of ML and DL models, reporting that most studies achieved accuracies between 82% and 95%, depending on the dataset, feature selection methods, and model architecture used. Importantly, the authors emphasized the need for standardized evaluation protocols, fairness considerations, and reproducibility to ensure reliable benchmarking across studies.

Furthermore, the review highlighted that a substantial proportion of research relied predominantly on the Pima Indians Diabetes Dataset, which restricted the diversity and generalizability of empirical findings. The authors concluded that future research should incorporate larger and more heterogeneous datasets and transparent performance reporting to establish more robust and equitable AI-based systems for diabetes prediction [15]. Addressing fairness directly, a MedRxiv preprint (2025) examined interventions to enhance fairness and usability in diabetes prediction systems. Using both clinical and public datasets, the study proposed fairness-aware ML models and user interface (UI) improvements to improve accessibility and decision transparency. The results demonstrated a reduction in prediction bias by up to 15% for underrepresented groups and a 20% improvement in user satisfaction compared to baseline models. However, the study acknowledged that clinical implementation remained pending, as fairness interventions had not yet been tested in real-world healthcare settings [16]. In another innovative approach, the RFE-GRU study (2025) combined Recursive Feature Elimination with Gated Recurrent Units for diabetes classification. Using the Pima Indians dataset and others, the model achieved accuracy above 93%. The integration of feature selection with recurrent neural networks allowed the model to focus on the most relevant predictors. Despite its strong performance, the study highlighted the computational expense of such hybrid models, which may limit scalability in resource-constrained environments.

Additionally, [17] explored diabetes prediction and management using ML, conducting a comparative study across multiple algorithms on the Pima dataset. The study reported moderate accuracy (~80%) and emphasized the need for broader datasets to improve generalizability. The work also highlighted the potential of ML not only for prediction but also for ongoing management of diabetes through continuous monitoring. Furthermore, [18] advanced the field by integrating ML with Explainable AI (XAI) techniques for transparent diabetes prediction. Using a Binary Health Indicators dataset and SMOTE balancing, the study achieved 92.5% accuracy and a ROC-AUC of 0.975.

The use of XAI methods such as SHAP improved transparency by identifying the most influential features. However, the authors stressed the need for validation across multiple datasets to confirm robustness.

Finally, [19] investigated secure and privacy-preserving automated ML for diabetes prediction within an IoT-Edge-Blockchain framework. Using Pima, Sylhet, and MIMIC datasets, the study compared RF, SVM, and LR, with RF outperforming others by 4.57%. While the framework demonstrated strong system design and addressed privacy concerns, the study lacked medical interpretability, limiting its clinical relevance. Moreover, systematic reviews and fairness-oriented studies emphasized the importance of standardized evaluation protocols, transparent model reporting, and equitable AI design to ensure generalizability across diverse populations. Despite substantial advancements, a clear research gap persists for a robust, generalizable, and interpretable predictive framework that effectively addresses data imbalance while balancing accuracy with computational efficiency. Such a framework would bridge the gap between performance and transparency, facilitating scalable, trustworthy, and clinically relevant AI solutions for diabetes prediction.

Table 1. Literature Review Summary.

Authors	**Title**	**Year**	**Dataset Used**	**Techniques**	**Key Results**
Tayyab Nasir, Muhammad Kamran Malik	SACDNet: Towards Early Type 2 Diabetes Prediction with Uncertainty for Electronic Health Records	2023	EHR dataset (4,124 diabetic + 181,767 non-diabetic records)	Multi-headed self-attention network (SACDNet), MC-Dropout	Accuracy 89.3%, F1-score 89.1%
X. Qi et al.	A deep neural network prediction method for diabetes: KCCAM_DNN	2024	Pima Indians Diabetes dataset & others	Deep Neural Network + Kendall correlation + attention + SHAP interpretation	High accuracy (~95%)
I. Abousaber et al.	Robust predictive framework for diabetes classification: a comparative evaluation across	2025	PIMA & other datasets	Comparative evaluation of multiple ML methods	High accuracy across datasets

	multiple datasets				
M. Z. Waughfa et al.	Diabetes Prediction:A Comprehensive Study Integrating ML & DL	2024	Pima Indian dataset and other public data	BPNN, RF, XGBoost	Accuracy ~90%
S. Gundapaneni et al.	Deep Learning-Based Noninvasive Screening of Type 2 Clinical Data Using Deep Learning Models	2024	Clinical dataset	Deep learning classification models	High performance (not quantified)
G. R. Ashisha et al.	Random Oversampling-Based Diabetes Classification via IoMT	2024	Pima Indians & BRFSS datasets	Random Oversampling + ML pipeline	Accuracy >90%
M. S. Alzboon	A Comparative Study of Machine Learning Techniques for Early Prediction of Diabetes	2025	Pima Indians dataset	LR, DT, RF, KNN, NB, SVM, GB, NN	Neural Network best (78.57%)
S. Khalid et al.	An in-depth review of AI-based techniques for diabetes prediction	2025	Review study (multiple datasets)	Systematic review of ML & DL approaches	Highlights trends and performance gaps
MedRxiv Preprint Authors	Enhancing Fairness in Diabetes Prediction Systems	2025	Clinical & public diabetes datasets	ML fairness interventions + UI improvements	Improved fairness and usability
RFE-GRU study	A novel RFE-GRU model for diabetes classification	2025	Pima Indians & others	Recursive Feature Elimination + GRU neural networks	Accuracy >93%
MS Alzboon	Diabetes Prediction and Management Using Machine Learning	2025	Pima dataset	Comparative study of ML algorithms	Moderate accuracy (~80%)

Pir Bakhsh Khokhar et al.	Towards Transparent and Accurate Diabetes Prediction Using ML and XAI	2025	Binary Health Indicators dataset	ML + Explainable AI (SMOTE balancing)	Accuracy 92.5%, ROC-AUC 0.975
Hennebelle et al.	Secure and Privacy-Preserving Automated ML for Diabetes Prediction (IoT-Edge-Blockchain)	2022	Pima, Sylhet, MIMIC datasets	RF, SVM, LR within blockchain framework	RF outperformed others by 4.57%

METHODOLOGY

In recent years, various ML algorithms have been extensively applied for the prediction and early diagnosis of diabetes mellitus, demonstrating their effectiveness in identifying high-risk individuals based on clinical and physiological indicators. However, evaluating their performance across diverse datasets, varying feature preprocessing techniques, and different model architectures remains a challenging task. Moreover, the issue of class imbalance where the number of non-diabetic samples significantly exceeds diabetic cases has a considerable impact on model performance, yet it has not been sufficiently addressed in earlier studies.

The conceptual pipeline of the intended methodology is showcased in Figure 1.

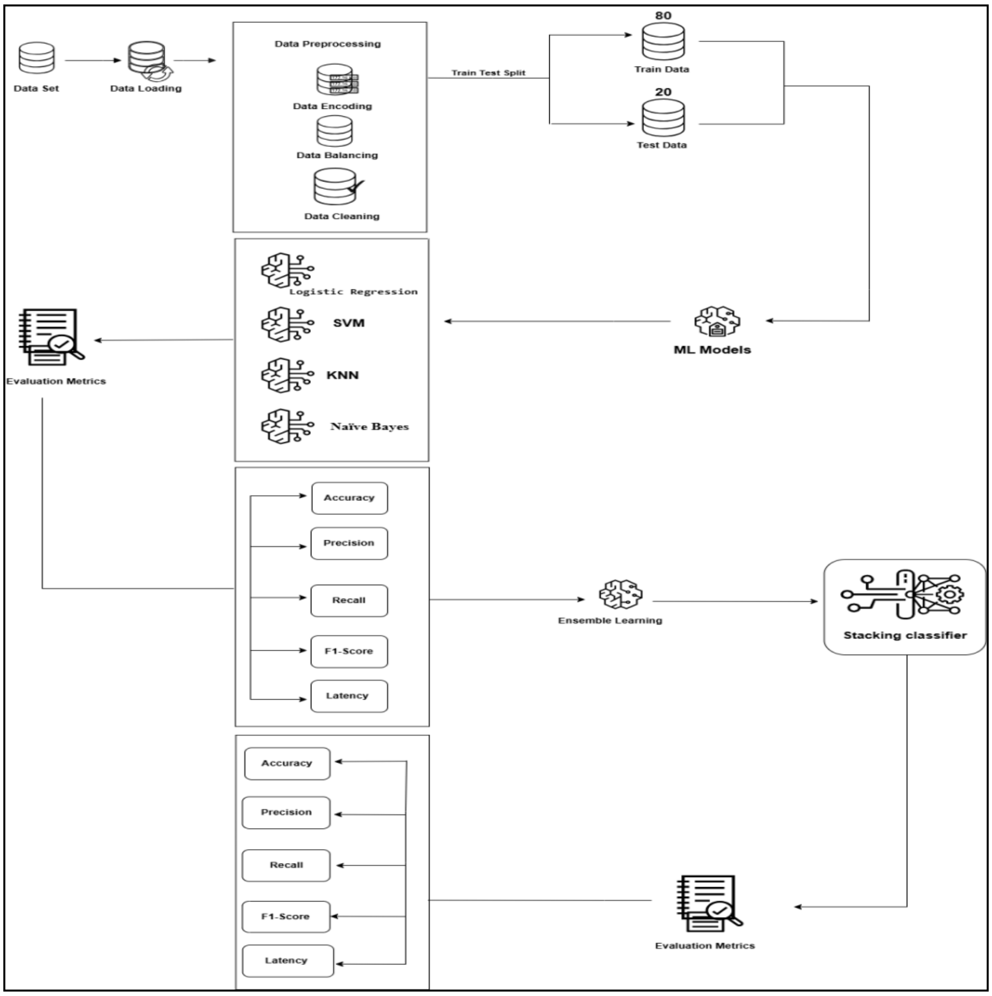

Figure 1. Indented Methodology.

To overcome these challenges, the proposed research establishes a systematic framework for selecting and optimizing machine learning models for diabetes prediction. This is accomplished by implementing and comparing four supervised classification algorithms: Logistic Regression (LR), Support Vector Machine (SVM) with an RBF kernel, K-Nearest Neighbors (KNN), and Naïve bayes. Furthermore, an integrated Stacking Classifier is employed to enhance predictive performance by combining the strengths of these diverse base learners. In the stacking framework, Logistic Regression, SVM, KNN, and Random Forest serve as the base models, while Logistic Regression functions as the meta-model to aggregate predictions from the base learners and produce the final output. This layered approach enables the framework to leverage the complementary capabilities of multiple classifiers, thereby improving accuracy, robustness, and generalizability in diabetes prediction.

Dataset

In predicting diabetes among individuals, the process of accurately identifying and classifying patients as diabetic or non-diabetic is highly dependent on the selection of an appropriate dataset. For this research, the Pima Indians Diabetes Dataset [20] a publicly available dataset obtained from the UCI Machine Learning Repository has been utilized. This dataset is widely recognized as a benchmark in the field of medical data analytics and diabetes prediction due to its comprehensiveness, quality, and widespread use in scholarly research.

Table 2. Dataset Details.

Attribute	**Description**	**Type**
Pregnancies	Number of times a patient has been pregnant	Numerical
Glucose	Plasma glucose concentration (mg/dL) after a 2-hour oral glucose tolerance test	Numerical
BloodPressure	Diastolic blood pressure (mm Hg)	Numerical
SkinThickness	Triceps skin fold thickness (mm)	Numerical
Insulin	2-Hour serum insulin (µU/ml)	Numerical
BMI	Body Mass Index (weight in kg / height in m²)	Numerical
DiabetesPedigreeFunction	A function that scores diabetes likelihood based on family history	Numerical
Age	Patient's age in years	Numerical
Outcome	Target variable indicating diabetes diagnosis (1 = Diabetic, 0 = Non-Diabetic)	Categorical

The PIDD consists of 768 records, each representing female patients of Pima Indian heritage, aged 21 years or older. The dataset provides detailed medical and physiological measurements that are clinically relevant to diabetes diagnosis. Each record includes eight diagnostic attributes Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, and Age along with a binary outcome variable (Outcome), which classifies individuals as diabetic (1) or non-diabetic (0) as shown in table 2.

Table 3. Distribution of Diabetes Cases in the Dataset

Diabetes Status	**Count**	**Percentage (%)**
No (0) – Non-Diabetic	500	65.1%
Yes (1) – Diabetic	268	34.9%
Total	768	100%

The dataset contains 768 samples, of which 500 (65.1%) are non-diabetic and 268 (34.9%) are diabetic, showing that non-diabetic cases are more common in the data as shown in table 3.

The included features represent key diagnostic indicators that are medically recognized as strong risk factors for diabetes, such as glucose concentration, insulin levels, blood pressure, BMI, and age. Leveraging this dataset ensures uniformity in model testing, fair performance comparison, and methodological reproducibility across studies as shown in table 4. Furthermore, its structured nature facilitates in-depth exploration of clinical and physiological factors influencing diabetes progression, making it a valuable and widely adopted resource for both data scientists and healthcare practitioners seeking data-driven insights into diabetes prediction and management.

Table 4. Dataset Summary and Class Distribution.

Feature Type	**Count**
Total Features	9
Missing Values (handled)	0
Numerical Features	8
Categorical Features	1
Irrelevant Features Removed	0
Class Distribution (Before Balancing)	35% Diabetic 65% Non-Diabetic
Class Distribution (After SMOTE)	50% Diabetic 50% Non-Diabetic

This dataset is particularly well-suited for diabetes prediction tasks, as it provides an invaluable source of real-world clinical information that enables the identification of underlying patterns and correlations associated with diabetes onset. It also supports comparative analyses across multiple machine learning algorithms, allowing for consistent and reproducible evaluation of predictive performance under standardized conditions.

Data Preprocessing

The Pima Indians Diabetes Dataset consists of nine diagnostic attributes, as summarized in Table 2. Most of these features are continuous medical indicators such as plasma glucose concentration, blood pressure, body mass index (BMI), insulin level, and patient age. Prior to model development, a comprehensive data preprocessing pipeline was implemented to ensure data integrity, consistency, and readiness for predictive modeling. Initially, exploratory data analysis (EDA) was conducted to identify missing or anomalous values. Records containing implausible physiological readings such as zero values for glucose, blood pressure, or BMI were treated as missing and imputed using the median of their respective features to preserve statistical validity and prevent information loss. Additionally, a unique identifier column, which held no predictive significance, was excluded from the dataset. Because all attributes were numeric, categorical encoding was unnecessary except for the Outcome variable, which was encoded as a binary target label: 1 for diabetic and 0 for non-diabetic.

To address the inherent class-imbalance problem where approximately 65 percent of records were non-diabetic and 35 percent were diabetic the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE generates synthetic data points for the minority (diabetic) class by interpolating between existing samples in feature space. This process produced a balanced class distribution (50 percent diabetic, 50 percent non-diabetic), enabling models to learn unbiased decision boundaries and improving overall generalization. After balancing, the dataset was partitioned into training (80 percent) and testing (20 percent) subsets to facilitate fair model evaluation and prevent information leakage. Feature scaling was then applied to normalize all numerical attributes using Min–Max Normalization, which transforms values to the [0, 1] range. This scaling step ensures that features with large numeric ranges do not dominate those with smaller ranges, thereby enhancing the stability and convergence of algorithms such as K-Nearest Neighbors (KNN) and SVM.

Machine Learning Models

This section presents the algorithms employed in the study, their working methodology, and the reasoning behind their selection. Diabetes prediction is a classification problem where the goal is to accurately distinguish between diabetic and non-diabetic individuals based on clinical indicators. Selecting suitable machine learning models is essential to achieving high diagnostic accuracy and ensuring reliable medical decision support. The proposed research utilizes a variety of supervised machine learning algorithms, including KNN, Logistic Regression, SVM, and Random Forest. Additionally, an ensemble approach (Stacking Classifier) is employed to combine the predictive capabilities of these base models to further enhance performance.

K-Nearest Neighbors (KNN)

KNN is a simple yet powerful non-parametric algorithm used for both classification and regression tasks. It predicts the class of a sample based on the majority class among its k nearest data points in the feature space. KNN does not require an explicit training phase, making it efficient and straightforward to apply to small or moderately sized datasets.

The algorithm determines the similarity between samples by calculating the distance between the test point and other data points in the dataset. The most commonly used metric for distance computation is the Euclidean distance, expressed as:

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \text{(Equation 1)}$$

As represented in Equation (1):

- x_i and y_i denote the feature values of two distinct data points.
- n is the total number of features.

This equation measures the straight-line distance between two points in an n-dimensional space. Other distance metrics such as Manhattan and Minkowski distances can also be used depending on the dataset's characteristics. For classification tasks, KNN identifies the k nearest neighbors based on their computed distances and assigns the class label by majority voting among these neighbors. In regression tasks, it computes the average of the target values of the nearest neighbors. KNN's strength lies in its simplicity and flexibility, though its performance can be affected by irrelevant features and the choice of k.

Logistic Regression

Logistic Regression is a statistical model widely applied for binary classification problems, where the output variable represents two distinct classes. In this study, it is utilized to classify patients as either diabetic or non-diabetic based on medical parameters such as glucose level, insulin concentration, BMI, and blood pressure. Logistic Regression estimates the probability of an event occurring by fitting data to a logistic function. It is simple, computationally efficient, and performs well when there is a linear relationship between the input variables and the log-odds of the target outcome.

The model calculates a linear combination of the input features as follows:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

The output of this linear combination is passed through a Sigmoid (logistic) function, which maps the value into a range between 0 and 1, representing the probability of the positive class.

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

The model then performs threshold-based classification typically, if $P(y = 1 \mid X) > 0.5$, the instance is classified as diabetic; otherwise, it is classified as non-diabetic.

$$P(y = 1 \mid X) = \sigma(z)$$
$$P(y = 0 \mid X) = 1 - \sigma(z)$$

The parameters (β coefficients) of the model are optimized by minimizing the log-loss function, which measures the error between predicted and actual outcomes:

$$\mathrm{Log} - \mathrm{loss} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad \text{(Equation 4)}$$

As described in Equation (4):

- Z is the weighted sum of input features.
- X_i are the feature values (e.g., glucose, insulin, BMI, etc.).

- β_i are the coefficients (weights) learned during the model training process.

Logistic Regression is an effective baseline algorithm for medical prediction problems due to its interpretability and simplicity. It allows healthcare researchers to assess the contribution of each independent variable to the probability of diabetes, providing meaningful insights alongside accurate classification.

Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm that is widely used for both classification and regression tasks. In this research, SVM is applied to classify individuals as diabetic or non-diabetic based on multiple medical parameters. The main goal of the algorithm is to identify the optimal hyperplane that best separates the data points of different classes in the feature space. SVM is highly effective in handling both linear and non-linear data, making it a robust model for high-dimensional biomedical datasets such as those used in diabetes prediction.

To enhance performance on non-linear data, SVM employs a kernel function that maps the input data into a higher-dimensional feature space, allowing for better separation between classes. This transformation enables SVM to form complex decision boundaries even when data cannot be linearly separated in the original input space. Commonly used kernel functions include linear, polynomial, sigmoid, and radial basis function (RBF) kernels. SVM optimizes the margin between the support vectors (the data points closest to the decision boundary) and the hyperplane. The algorithm seeks to maximize this margin to improve the classifier's generalization ability and reduce misclassification errors. Additionally, the soft margin technique is applied to efficiently handle outliers and overlapping data points, ensuring a balanced trade-off between classification accuracy and model flexibility.

The decision function of SVM is mathematically represented as follows:

$$F(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \quad \text{(Equation 5)}$$

As expressed in Equation (5):

- x_i represents the support vectors.
- y_i denotes the corresponding class labels.
- α_i are the Lagrange multipliers that determine the influence of each support vector.
- $K(x_i, x)$ is the kernel function applied (linear, polynomial, RBF, or sigmoid).
- b is the bias term of the model.

In the classification process, SVM assigns a new data instance to one of the classes based on its position relative to the decision boundary (hyperplane). For regression problems (Support Vector Regression SVR), SVM finds the best-fitting function within a predefined margin of tolerance.

Overall, SVM provides high accuracy and robustness in classifying diabetes cases, particularly in scenarios where the dataset contains non-linear relationships and overlapping features. Its ability to handle high dimensional spaces and kernel transformations makes it an excellent choice for medical prediction systems such as diabetes detection.

Naive Bayes

The Naive Bayes classifier is a supervised machine learning algorithm based on Bayes' Theorem and is widely used for classification tasks due to its simplicity, efficiency, and strong performance on high-dimensional datasets. In this research, the Naive Bayes model is employed to classify individuals as diabetic or non-diabetic using multiple medical parameters. The algorithm operates by calculating the probability that a given data instance belongs to a particular class, making decisions based on the highest posterior probability. Its probabilistic foundations and computational efficiency make Naive Bayes well-suited for biomedical datasets, which often contain numerous features and complex variable relationships.

A key assumption of the Naive Bayes algorithm is the conditional independence of features given the class label. Although this assumption rarely holds perfectly in real-world medical datasets, Naive Bayes has been shown to perform remarkably well even when some dependencies exist. The model estimates class probabilities by combining prior probabilities with likelihoods derived from the training data, enabling rapid and scalable classification.

Naive Bayes supports different variants depending on how feature likelihoods are modeled. Common versions include Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes, each suited to different types of data distributions. In the context of diabetes prediction, Gaussian Naive Bayes is typically used because the medical attributes (such as glucose level, BMI, and blood pressure) are continuous in nature and approximately follow a normal distribution.

The decision rule for Naive Bayes classification is mathematically expressed as follows:

$$P(C_k \mid X) = \frac{P(X \mid C_k)\, P(C_k)}{P(X)} \quad \text{(Equation 6)}$$

Where:

- $C_k$represents the class label (diabetic or non-diabetic).
- $X = (x_1, x_2, \dots, x_n)$denotes the input feature vector.
- $P(C_k)$is the prior probability of class C_k.
- $P(X \mid C_k)$is the likelihood of observing the features given the class.
- $P(X)$is the evidence term ensuring proper probability normalization.

Since $P(X)$is constant for all classes during prediction, the model selects the class with the highest posterior probability $P(C_k \mid X)$. This results in a simple yet powerful classification mechanism.

During classification, the algorithm evaluates how likely a new instance belongs to each class using the computed probabilities and assigns it to the class with the maximum posterior value. Its probabilistic nature also provides interpretability, as the influence of each feature on the prediction can be examined through likelihood contributions. Overall, the Naive Bayes classifier offers strong performance in diabetes prediction tasks, particularly when datasets contain high-dimensional features and moderate levels of noise. Its fast-training time, low computational cost, and robustness against irrelevant features make it a valuable model for medical diagnosis applications. Despite its simplifying independence assumption, Naive Bayes often delivers competitive accuracy, making it an effective and practical choice for predictive healthcare systems such as diabetes detection.

Ensemble Learning

Ensemble learning is an advanced machine learning approach that combines multiple base models to achieve improved predictive accuracy, robustness, and generalization compared to any individual model. The main idea behind ensemble learning is that a group of diverse and well-performing models can collectively produce a stronger predictive outcome. Given individual models $h_1(x), h_2(x), \ldots, h_n(x)$, the final ensemble prediction $H(x)$is determined by aggregating their outputs.

In this research, ensemble learning techniques are employed to enhance diabetes prediction performance and address potential weaknesses of single classifiers. By integrating various algorithms, the ensemble approach helps to reduce model bias and variance, thereby improving the overall reliability of the predictions. Ensemble methods are particularly effective when dealing with imbalanced datasets and complex feature relationships, as seen in medical data.

Stacking Classifier

Stacking, or stacked generalization, is a form of ensemble learning that combines multiple base learners and uses their predictions as input features for a meta-learner (also known as a secondary model). The meta-learner is trained to learn how to best combine the outputs of the base models to achieve optimal predictive performance.

Mathematically, if $h_1, h_2, \ldots, h_N$represent the base models, the final prediction is computed as:

$$\hat{y} = g\big(h_1(x), h_2(x), \ldots, h_N(x)\big) \quad \text{(Equation 6)}$$

As described in Equation (6):

- $h_1(x), h_2(x), \ldots, h_n(x)$ are the base-level models that generate initial predictions.
- g(·) represents the meta-classifier, which learns from the base models’ outputs to produce the final prediction.

In this study, the Stacking Classifier integrates several machine learning models such as Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine (SVM) at the base level. The outputs from these classifiers are then fed into a meta-level model (Logistic Regression) that performs the final classification. This ensemble framework leverages the unique strengths of each base learner for instance, the interpretability of Logistic Regression, the non-linearity handling of KNN and SVM, and the robustness of Navie Bayes to deliver more accurate and stable predictions. By reducing both bias and variance, the Stacking Classifier provides superior performance in predicting diabetes when compared to individual classifiers.

Evaluation Metrics

Evaluation metrics are statistical tools used to assess the performance, efficiency, and reliability of machine learning models. In this research, several metrics are utilized to evaluate both individual classifiers and the ensemble model for diabetes prediction. As the nature of this study involves a binary classification problem (diabetic or non-diabetic), the performance of models is assessed using Accuracy, Precision, Recall, F1-Score, and latency to provide a comprehensive understanding of model behavior.

Accuracy measures the overall correctness of the model's predictions, while Precision and Recall assess the balance between true positive and false positive classifications. The F1-Score serves as a harmonic mean of Precision and Recall, ensuring a balance between the two.

The mathematical representations of these metrics are as follows:

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad \text{(Equation 7)}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad \text{(Equation 8)}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad \text{(Equation 9)}$$

$$\mathrm{F1-Score} = 2 \times \frac{\mathrm{Precision \times Recall}}{\mathrm{Precision+Recall}} \quad \text{(Equation 10)}$$

Referring to Equations (7) - (10):

- TP (True Positive) represents correctly identified diabetic cases.
- TN (True Negative) denotes correctly identified non-diabetic cases.
- FP (False Positive) occurs when non-diabetic individuals are incorrectly classified as diabetic.
- FN (False Negative) represents diabetic individuals misclassified as non-diabetic.

Confusion Matrix

A Confusion Matrix is a widely used performance evaluation tool that summarizes how well a classification model performs by comparing predicted labels with the actual ground truth labels. It provides a detailed breakdown of correct and incorrect predictions and helps identify the types of classification errors made by the model. In the context of diabetes prediction, the Confusion Matrix represents how effectively the model distinguishes between diabetic (positive class) and non-diabetic (negative class) patients. The matrix is divided into four fundamental categories: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

By analyzing the distribution of these values, the Confusion Matrix provides a comprehensive view of the model's classification ability, helping researchers identify whether the system is biased towards false alarms (FP) or missed detections (FN). This metric is essential for evaluating models in medical diagnosis systems, where incorrect predictions can have critical implications on patient outcomes.

RESULTS & DISCUSSION

This study compares machine learning models such as Logistic Regression, Support vector machine, KNN, and Naive Bayes with Ensemble learning techniques like Stacking. The evaluation of different machine learning models and ensemble learning techniques is presented through various performance metrics, including accuracy, precision, recall, and F1-score. The results are analyzed below.

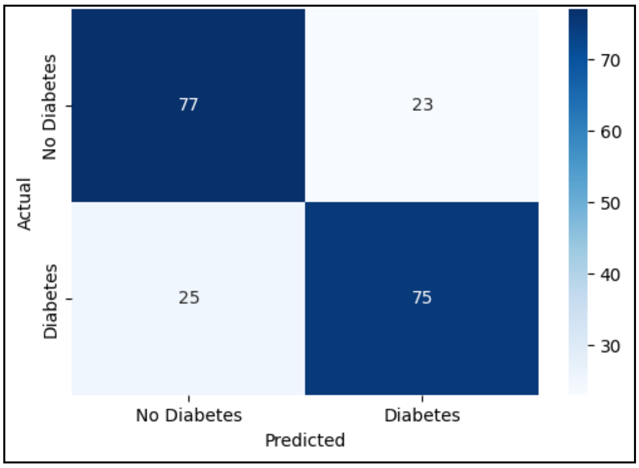

Figure 2. Logistic Regression Confusion Matrix.

The performance of the logistic regression model for diabetes prediction was evaluated using the confusion matrix. Out of the total cases, the model correctly identified 77 individuals without diabetes (true negatives) and 75 individuals with diabetes (true positives). However, it misclassified 23 non-diabetic cases as diabetic (false positives) and 25 diabetic cases as non-diabetic (false negatives) as shown in Fig 2. These results indicate that the model demonstrates a balanced ability to distinguish between diabetic and non-diabetic individuals, though the presence of false positives and false negatives highlights areas for improvement. The relatively close values of true positives and true negatives suggest that the model is not biased toward one class.

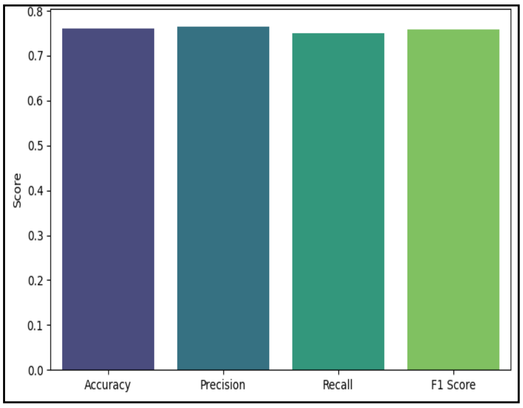

Figure 3. Logistic Regression Performance Matrix.

As shown in Figure 3 the logistic regression model's effectiveness was evaluated using standard performance metrics. The model achieved accuracy, precision, recall, and F1-score values all close to 0.75, demonstrating consistent performance across evaluation criteria. The accuracy represents the overall proportion of correctly classified cases, while precision reflects the model's ability to correctly identify diabetic cases without excessive false positives. Recall measures the model's sensitivity in detecting actual diabetes cases, and the F1-score balances both precision and recall. The close alignment of these metrics suggests that the model maintains a solid trade-off between identifying positive cases and minimizing false alarms. This consistency across measures reinforces the reliability of logistic regression for diabetes prediction.

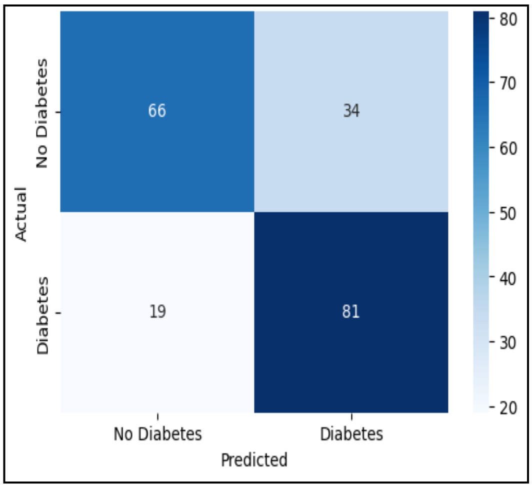

Figure 4. KNN Confusion Matrix.

Figure 4 illustrates the confusion matrix for the K-Nearest Neighbors (KNN) classifier. The model accurately predicted 66 cases of “No Diabetes” and 81 cases of “Diabetes.” However, it misclassified 34 “No Diabetes” cases as “Diabetes” (false positives) and 19 “Diabetes” cases as “No Diabetes” (false negatives). These results show that the KNN model performed reasonably well but exhibited a higher number of false positives compared to false negatives. This indicates that while the model is effective in identifying diabetic patients, it tends to over-predict diabetes in non-diabetic individuals. The overall distribution of predictions suggests that the KNN algorithm provides a balanced classification performance.

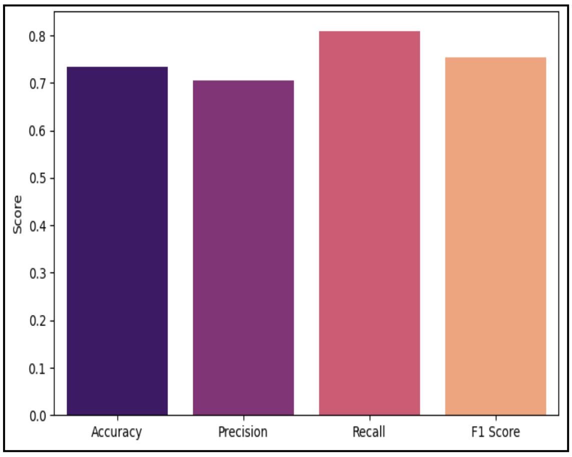

Figure 5. KNN Performance Matrix.

Figure 5 presents the performance metrics of the K-Nearest Neighbors (KNN) classification model. The model achieved an accuracy of approximately 0.74, indicating that it correctly classified about 74% of the instances in the dataset. The precision value, around 0.71, suggests that the model maintained a reasonable proportion of true positives among the predicted positive cases. Notably, the recall score was the highest at 0.81, demonstrating that the model effectively identified most of the actual positive samples. The F1-score, which balances precision and recall, stood at 0.76, confirming a good overall trade-off between the two metrics. These results imply that the KNN model performs reliably, with strong recall performance indicating its effectiveness in minimizing false negatives an important aspect depending on the classification problem’s context.

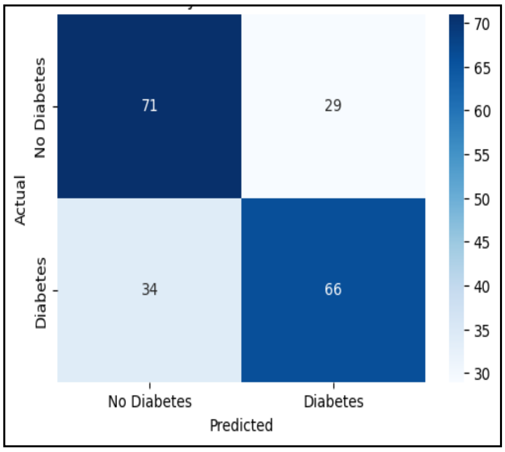

Figure 6 Naive Bayes confusion Matrix.

The performance of the Naive Bayes model for diabetes prediction was assessed using the confusion matrix. The model accurately predicted 71 individuals without diabetes (true negatives) and 66 individuals with diabetes (true positives). However, it misclassified 29 non-diabetic cases as diabetic (false positives) and 34 diabetic cases as non-diabetic (false negatives), as shown in Figure 6. These results indicate that the Naive Bayes model demonstrates a fairly balanced ability to classify individuals into the correct categories, although the presence of false positives and false negatives suggests that there is room for improvement. The model's true positive and true negative values are relatively close, suggesting that it is not biased toward one class over the other.

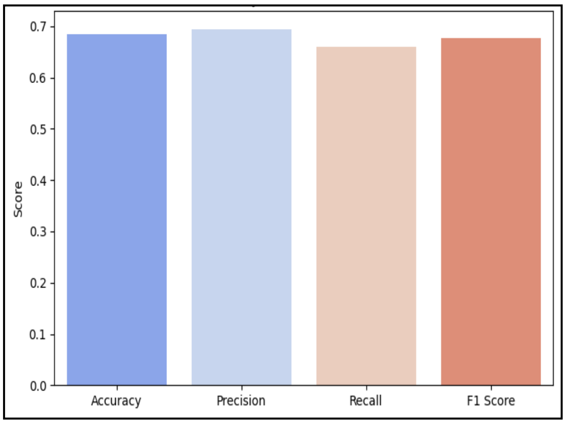

Figure 7. Naive Bayes Performance Matrix.

The performance metrics of the Naive Bayes model further highlight its effectiveness in predicting diabetes cases. As shown in the figure 7. the model achieved an accuracy of approximately 0.7, indicating that around 70% of the total predictions were correct. Both precision and F1-score were also close to 0.7, suggesting that the model maintained a balanced approach between correctly identifying diabetic patients and minimizing false positives. The recall score, slightly below 0.7, shows that the model was fairly effective in capturing most of the actual diabetes cases, though there is room for improvement in terms of sensitivity. Overall, these performance metrics demonstrate that the Naive Bayes model delivers reasonable classification results, with a consistent balance between accuracy, precision, recall, and F1-score.

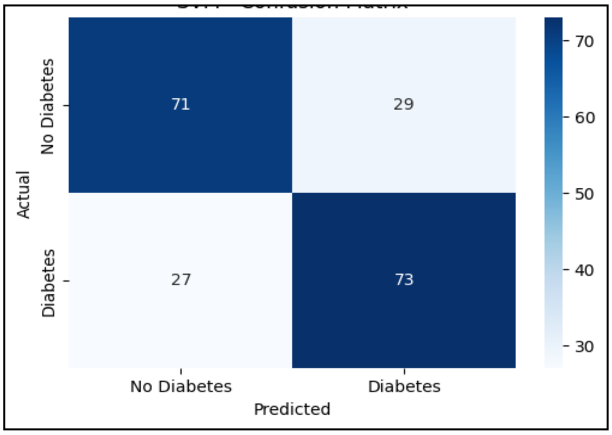

Figure 8. SVM confusion Matrix.

The Support Vector Machine (SVM) model demonstrated a balanced performance in classifying diabetes cases, as shown in the confusion matrix. Out of the total samples, 71 non-diabetic cases were correctly identified as “No Diabetes,” while 29 were incorrectly predicted as “Diabetes.” Similarly, among the diabetic cases, 73 were accurately classified as “Diabetes,” with 27 instances misclassified as “No Diabetes.” These results indicate that the model performs slightly better in predicting diabetic cases than non-diabetic ones. The overall performance suggests that the SVM model achieved a reasonable balance between sensitivity and specificity, though some degree of misclassification persists, which could be attributed to overlapping features or limitations in the dataset.

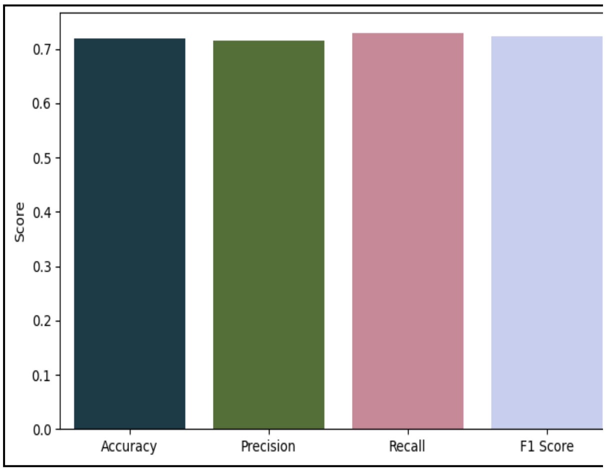

Figure 9. SVM Performance Matrix.

The performance metrics of the Support Vector Machine (SVM) model further validate its effectiveness in predicting diabetes cases. As illustrated in the figure 9. the model achieved an accuracy of approximately 0.72, indicating that around 72% of the total predictions were correct. Both precision and F1-score were also close to 0.72, suggesting that the model maintained a consistent balance between correctly identifying diabetic patients and minimizing false positives. The recall score, slightly above 0.72, shows that the model was effective in capturing most of the actual diabetes cases, reflecting good sensitivity. Overall, these performance metrics demonstrate that the SVM model provides reliable classification results with a balanced trade-off between accuracy, precision, recall, and F1-score.

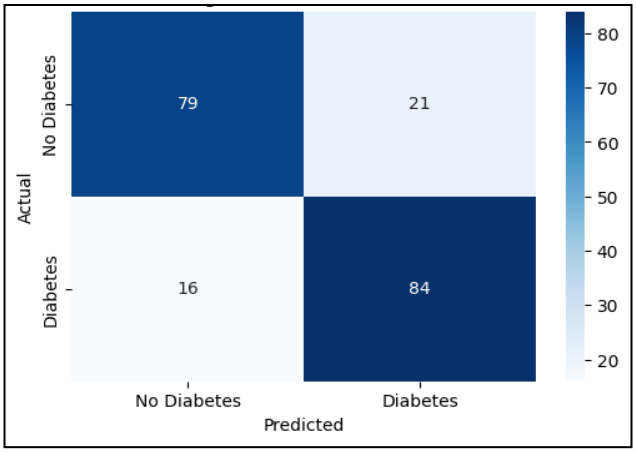

Figure 10. Stacking Model Confusion Matrix.

The confusion matrix of the Stacking model illustrates a notable improvement in classification accuracy compared to the individual baseline models. As shown in Figure 10. the model correctly classified 79 instances of non-diabetic patients and 84 instances of diabetic patients, while only 21 non-diabetic and 16 diabetic cases were misclassified. This performance indicates a strong ability of the ensemble approach to minimize both false positives and false negatives. The integration of multiple base learners in the Stacking framework enhances model robustness, leading to better generalization and discrimination between diabetic and non-diabetic classes. Overall, the results demonstrate that the Stacking model effectively leverages the strengths of its constituent classifiers to deliver superior predictive performance, making it a promising technique for reliable diabetes prediction.

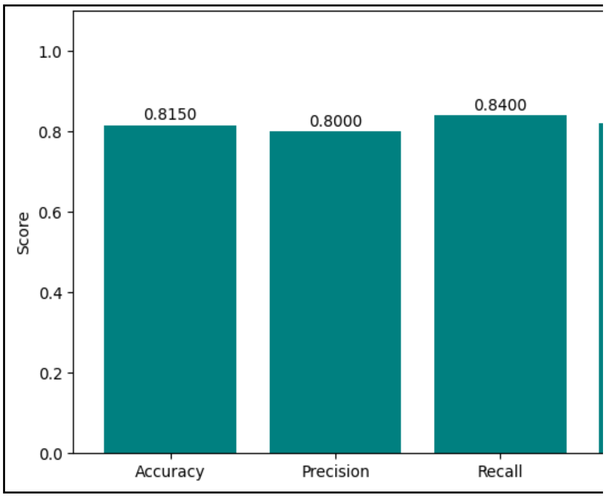

Figure 11. Stacking model performance Matrix.

The performance metrics of the Stacking model further reinforce its superior classification capability. As presented in the figure 11. the model achieved an accuracy of 0.815, indicating that over 81% of the total predictions were correctly classified. The precision score of 0.800 signifies that the model effectively minimizes false-positive predictions, while the recall value of 0.840 demonstrates its strong ability to correctly identify diabetic cases. Additionally, the F1-score of 0.8195 reflects a well-balanced trade-off between precision and recall, highlighting the model's overall reliability. These results collectively confirm that the Stacking model delivers robust and consistent performance, outperforming individual classifiers by enhancing predictive accuracy, sensitivity, and overall generalization across the dataset.

The inference latency analysis of the Stacking model provides valuable insights into its computational efficiency during prediction. As depicted in the the model recorded an average inference latency of 27.43 milliseconds, indicating that the ensemble framework delivers predictions with minimal delay. This relatively low latency suggests that, despite the model's complexity and integration of multiple base learners, it maintains efficient real-time processing capabilities. The result demonstrates that the Stacking model not only achieves high predictive accuracy but also operates with satisfactory computational performance, making it suitable for practical deployment in time-sensitive healthcare applications such as diabetes detection and monitoring systems.

CONCLUSION

This study examined the predictive performance of multiple machine learning algorithms for diabetes detection, including Logistic Regression, KNN, Naive Bayes, SVM, and an Ensemble Stacking model. The evaluation was conducted using confusion matrices, accuracy, precision, recall, F1 score, and inference latency to identify the strengths and limitations of each classifier. The Logistic Regression model demonstrated balanced performance, with accuracy, precision, recall, and F1 score values all approximately 0.75. The confusion matrix indicated relatively similar numbers of true positives and true negatives, showing no major class bias. The KNN classifier achieved an accuracy of 0.74 and exhibited the highest recall (0.81), indicating strong sensitivity in detecting diabetic cases;

however, it also produced a higher number of false positives, which affected its precision (0.71). The Naive Bayes classifier maintained consistent performance around 0.70 across all evaluation metrics, suggesting effective generalization but limited capability due to its assumption of feature independence. The SVM model achieved an accuracy of 0.72 with precision, recall, and F1-score values close to 0.72, reflecting a well-balanced model with slightly higher sensitivity toward diabetic cases.

Among all models, the Ensemble Stacking classifier delivered the best overall performance. It achieved an accuracy of 0.815, precision of 0.800, recall of 0.840, and an F1 score of 0.8195, outperforming each individual model. The confusion matrix confirmed reduced misclassification rates, with fewer false positives and false negatives compared to the baseline classifiers. Furthermore, the Stacking model recorded an inference latency of 27.43 ms, indicating efficient real-time prediction capabilities despite incorporating multiple base learners.

Based on the experimental findings, it is evident that while individual machine learning models provide reasonable prediction accuracy, they each exhibit limitations that impact their reliability and generalization. The Ensemble Stacking approach effectively addresses these limitations by combining the predictive strengths of diverse models, leading to improved accuracy, enhanced sensitivity, and reduced misclassification. Therefore, the Stacking model represents the most robust and effective solution for diabetes prediction in this study and demonstrates strong potential for deployment in practical and time sensitive healthcare applications.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare there are no competing interests.

Author Contribution

- **Irfanullah:** Led the research design, experiments, analysis, and methodology; produced key visualizations; authored core manuscript sections; interpreted results; and approved the final draft.
- **Muhammad Shahan Ibad:** Supported experiments, data preparation, methodology refinement, and manuscript drafting; prepared visuals; and approved the final draft.
- **Aamir Sohail:** Supported experiments, data preparation, methodology refinement, and manuscript drafting; prepared visuals; and approved the final draft.

Data Availability

The following information was supplied regarding data availability:
The ‘‘Pima Indians Diabetes Database” dataset is available at

ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

Kaggle: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

REFERENCES

https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

https://health.clevelandclinic.org/type-1-vs-type-2-diabetes

https://www.mayoclinic.org/diseases-conditions/gestational-diabetes/symptoms-causes/syc-20355339?

https://www.who.int/health-topics/diabetes?#tab=tab_1

https://www.ncbi.nlm.nih.gov/books/NBK581940/?

https://www.acs.org/education/whatischemistry/landmarks/discovery-of-insulin.html

Nasir, T., & Malik, M. K. (2023). SACDNet: Towards Early Type 2 Diabetes Prediction with Uncertainty for Electronic Health Records. *arXiv preprint arXiv:2301.04844*.

Qi, X., Lu, Y., Shi, Y., Qi, H., & Ren, L. (2024). A deep neural network prediction method for diabetes based on Kendall's correlation coefficient and attention mechanism. *Plos one*, *19*(7), e0306090.

Abousaber, I., Abdallah, H. F., & El-Ghaish, H. (2025). Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. *Frontiers in Artificial Intelligence*, *7*, 1499530.

Z Waughfa, M., Adnan, I. I., Sultana, S. S., & Mumu, S. S. (2024, October). Diabetes Prediction: A Comprehensive Study Integrating Deep Learning and Machine Learning Approaches. In *Proceedings of the 3rd International Conference on Computing Advancements* (pp. 520-526).

Gundapaneni, S., Zhi, Z., & Rodrigues, M. (2024). Deep Learning-Based Noninvasive Screening of Type 2 Diabetes with Chest X-ray Images and Electronic Health Records. *arXiv preprint arXiv:2412.10955*.

Ashisha, G. R., Mary, X. A., Kanaga, E. G. M., Andrew, J., & Eunice, R. J. (2024). Random oversampling-based diabetes classification via machine learning algorithms. *International Journal of Computational Intelligence Systems*, *17*(1), 270.

Salem Alzboon, M., Al-Batah, M., Alqaraleh, M., Abuashour, A., & Fuad Bader, A. (2025). A Comparative Study of Machine Learning Techniques for Early Prediction of

Diabetes. *arXiv e-prints*, arXiv-2506.

Khalid, S., Kim, H., & Kim, H. S. (2025). Recent trends in diabetes mellitus diagnosis: an in-depth review of artificial intelligence-based techniques. *Diabetes Research and Clinical Practice*, 112221.

Tavangar, A., Arefzadeh, Z., & Asghari, A. N. (2025). Enhancing Fairness in Diabetes Prediction Systems through Smart User Interface Design. *medRxiv*, 2025-06.

Shams, M. Y., Tarek, Z., & Elshewey, A. M. (2025). A novel RFE-GRU model for diabetes classification using PIMA Indian dataset. *Scientific reports*, *15*(1), 982.

Alfaro Quintero, D. F., Chaparro Macias, D. M., Lozano Novoa, J. P., & Palma Roa, J. D. (2025). Análisis de rendimiento de modelos gratuitos de Machine Learning (aprendizaje automático) utilizados en infraestructura de la nube en la predicción de la diabetes.

Khokhar, P. B., Pentangelo, V., Palomba, F., & Gravino, C. (2025). Towards transparent and accurate diabetes prediction using machine learning and explainable artificial intelligence. *arXiv preprint arXiv:2501.18071*.

Hennebelle, A., Ismail, L., Materwala, H., Al Kaabi, J., Ranjan, P., & Janardhanan, R. (2024). Secure and privacy-preserving automated machine learning operations into end-to-end integrated IoT-edge-artificial intelligence-blockchain monitoring system for diabetes mellitus prediction. *Computational and Structural Biotechnology Journal*, *23*, 212-233**.**

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.