Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Hybrid Stacking Ensemble Approach for Diabetes Classification on Imbalanced Clinical Data

Dr. Narendhar Mulugu[1], Dr. V. Yasaswini[2], Mr. Sekhar Babu Golla[3], S Vinayagan[4], Mr. Venu Mogili[5], Mr. N. Sai Krishna Goud[6]
[1]Professor, Dept. of Computer Science and Engineering (AIML)
[2]Associate Professor, Department of Cyber Security,
[3]Assistant Professor Dept. of Computer Science and Engineering (Data Science)
[4]Assistant. Professor, Dept. of Computer Science and Engineering (AIML)
[5]Assistant Professor, Dept. of Computer Science and Engineering (Data Science)
[6]Assistant Professor, Department of Cyber Security,
[1,2,3,4,5,6,] Malla Reddy Engineering College for Women (Autonomous), Maisammaguda Hyderabad, Telangana, India

ARTICLE INFO

Received: 02 Nov 2024

Revised: 18 Dec 2024

Accepted: 28 Dec 2024

ABSTRACT

Diabetes mellitus is a common chronic disease which afflicts people worldwide. It presents great challenges for medical care because of the complications that are related to it. So it is important to diagnose and treat diabetes China can give every diabetic the chance of a healthy life. The Peidemo Diabetes Prediction System described in this paper is a strong example of ensemble learning. It uses XGBoost, LightGBM, and Adaboost classifiers as base classifiers in an ensemble setting together with logistic regression for meta learning. Synthetic minority over-sampling technique combined with edited nearest neighbors (SMOTEENN) is applied to the dataset so that the model can generalize. Pima Indian Diabetes dataset had gone through a complete pre-processing including cleaning data, balancing data, and scaling assignments. The model was evaluated with the training/test partition being 80% for training and 20% for testing. It had a predictive accuracy rate to 93.7% and an area under the receiver operating characteristic curve (AUC) of 0.97. Thus the results of this study show that combining base classifiers in a stacked ensemble can capture complex feature interactions and solve the problem of class imbalance in diabetes prediction effectively. As a method of clinical decision support, the methodology offers great promise. It helps companies in providing their employees with quality healthcare at an affordable cost and even an equal opportunity to develop skills for the future.

Keywords: Mellitus, Peidemo, Light GBM, and Ada Boost, SMOTEENN

1. Introdcution:

Diabetes mellitus is a chronic metabolic disorder that is characterized by high blood sugar levels due either to insufficient output of insulin by beta-cells of the pancreas or else because enough glucose is produced without there being enough effective hormone to handle it. It is one of the leading causes of morbidity and mortality in the world today, causing serious complications such as heart disease (including blood vessel diseases), kidney failure and neuropathy. According to International Diabetes Federation, there are now over 537 million sufferers from diabetes worldwide, a figure which is expected by 2045 to increase significantly. Early diagnosis and timely intervention are necessary to control disease development. However, traditional diagnostic methods

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

are often costly, time-consuming and require special clinical expertise. Thanks to the emergence of artificial intelligence (AI) and machine learning (ML) methods, detailed predictions for diabetes patients 'clinical data and demographic features have been promising as well as fast to obtain. Support Vector Machines, Random Forests, Decision Trees, and Neural Networks are some of the many studies that have used supervised machine learning methods to predict diabetes. Generally the benchmark is based on Pima Indian Data set that is versatile and widely used for such purposes. However, individual classifiers are usually confronted with challenges like class imbalance, noise data or complex interactions between nonlinear features and so on, despite breakthroughs in increasing capacity and network structure design. It s widely accepted by experts that integrating multiple models into an optimized result with better predictive accuracy and generalization capacity is the most effective strategy for dealing with such problems. This approach or methodology, which is referred to in particular as ensemble learning, has been found to perform better than any other. Stacked ensembles, more specifically, are a recent innovation that demonstrate improved stability and accuracy beyond common ensemble methods according to the empirical evidence. This research introduces a stacked ensemble learning framework that integrates leading-edge gradient boosting algorithms, robust data set balancing--a necessary prerequisite for models with class skews--and advanced feature selection techniques to accurately forecast diabetes. The approach produces a model that provides not only interpretable but also reliable predictions for early diagnosis and clinical decision-making.

Even with the high incidence of diabetes, it is often not diagnosed until severe complications occur.Past work has demonstrated that modern machine learning (ML) and artificial intelligence (AI) methods can effectively overcome this challenge. ML and AI offer fast, efficient, accurate determination of diabetes risk or presence based on clinical diagnoses and demographic data alone.Previous research has delved into different ML algorithms such as Support Vector Machines, Random Forest, Decision Trees, K-Nearest Neighbors, and Neural Networks to perform this function.Current developments show that ensemble learning methods, especially gradient boosting framework tools like XGBoost and LightGBM, are far superior in nature – capable of capturing complex non-linear relationships and interrelationships in healthcare data sets better.Furthermore, hybrid and stacked ensemble approaches, combining different base models with aggregator layers, are increasingly recognized for their robustness and precision in medical prediction tasks.However, challenges remain. For example, diabetic cases are underrepresented, the clinical data contain noise and missing values, and feature selection is needed to make models that are more interpretable and efficient. Synthetic oversampling techniques such as SMOTEENN and interpretability frameworks such as SHAP appear to offer solutions to these problems and make the clinical application of models more feasible.It is with these gaps in mind that this study proposes a stacked ensemble learning framework integrating XGBoost, LightGBM, and AdaBoost base learners with a meta-classifier of logistic regression. Advanced preprocessing techniques, including SMOTEENN balancing and feature normalization, are applied to improve prediction performance.Taking the Pima Indian Diabetes dataset as an example, the model accomplishes an accuracy rate of 93.7% and AUC value of 0.97. This shows that taking ensemble methodologies coupled with thorough pre-processing can contribute to early diabetes detection.

2. Literature Survey:

Machine learning (ML) and artificial intelligence (AI) have been widely used in healthcare, especially in diabetes forecasting, over past 10 years. Initially, early studies with prototype ML models demonstrated feasibility classifying patients into different diabetes risk groups. But newer research introduces hybrid systems, so-called interpretable AI and ensemble methods, not just anything that scores high on algorithms.Earliest comparative works focussed on traditional supervised learning methods like logistic regression, decision trees, random forests, or support vector machines using the Pima Indians Diabetes database (Wu Y et al. 2023). Yet individual algorithms often showed significantly polarized results. The introduction of ensemble and deep-learning models, for example boosting and stacker, increased prediction robustness and reduced bias (Wang Z et al. 2023)."Recent ensemble learning strategies offer an ascending trajectory. One ensemble boosting model achieved strong accuracy in diabetes prediction (Wang X et al. 2024), while a combined GA-XGBoost + stacking framework delivered superiority in AUC performance (Wang Z et

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

al. 2023). Diabetology & Metabolic Syndrome also points out" the increasing relevance of deep and ensemble models to type 2 diabetes (Clark I et al. 2021). Consistent conclusions were reached by Visual Computing for Industry, Biomedicine and Art in a review focusing on feature engineering and imbalance correction (Chen H et al. 2021)."

Data imbalance is still a major problem and people are looking into solutions like SMOTEENN combines oversampling with undersampling which can make fairness and recall measurements improve. Public-health-scale resampling boosted by boosting algorithms has been studied and shown to be feasible in the case of tackling diabetes (Improving machine learning diabetes prediction models for the general public: large-scale study, 2023). Stacking base learnersand then putting the ensemble into meta-models produced strong diagnostic power:B (An empirical model to predict the diabetic positive using stacked ensemble, 2022).

Multi-classification frameworks (A pipeline-based multi-classification framework to predict diabetes, 2023) and lifestyle-based modeling (A review on trending machine learning techniques for Type 2 diabetes using lifestyle-related data, 2024) extend predictive utility to early screening and prevention. Hybrid boosting, GA-optimized ensembles, and deep architectures (Recent applications of machine learning and deep learning models for diabetes, 2022; An ensemble approach to predict early-stage diabetes risk using machine learning, 2022) collectively establish ensemble learning as the most effective paradigm. Stacked frameworks integrating logistic-regression meta-learners offer both interpretability and high AUC (Towards a stacking ensemble model for predicting diabetes, 2023).

Emerging research further explores the role of explainable AI in medical contexts (Machine learning and artificial intelligence in type 2 diabetes: applications & future directions,). Integration of lifestyle and physiological indicators (An ensemble machine learning approach for predicting Type-II diabetes mellitus using lifestyle indicators, 2022), and continuous refinement of boosting algorithms (Predicting diabetes using supervised machine learning algorithms,), underline the necessity of interpretable, adaptive, and data-efficient solutions.

Additionally, hospital-based studies (Development of various diabetes prediction models using machine learning in tertiary care, 2021) and systematic reviews (Advances in artificial intelligence for diabetes prediction: systematic review,) confirm that hybrid ensembles consistently outperform single models. Recent work also integrates deep CNN–LSTM feature extractors for multimodal diagnosis (Enhanced detection of diabetes mellitus using novel ensemble of CNN + LSTM, 2024), and novel edge–AI blockchain integrations for privacy-preserving diabetes prediction (Secure and privacy-preserving automated ML operations ..., 2022). These studies collectively identify a gap for interpretable, GA-optimized, SMOTEENN-balanced stacking frameworks—precisely the motivation for the present research

3. Methodology:

The study utilizes the publicly available Pima Indian Diabetes dataset, which contains 768 instances along with 8 clinical features including number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. The target variable is binary, indicating the presence (1) or absence (0) of diabetes in the patients. This dataset is widely used for benchmarking diabetes prediction algorithms due to its real-life clinical nature and inherent class imbalance with fewer diabetes-positive cases, making it an excellent candidate for testing classification robustness.

3.1 Data Preprocessing and Feature Engineering

To prepare the raw data for modelling, multiple preprocessing steps are conducted. First, missing and zero values in relevant clinical features are imputed with median values to reduce bias from incomplete records.Then, numerical features are standardized using z-score normalization, transforming each feature x_j to $x_j^{scaled} = (x_j - \mu_j)/\sigma_j$, where μ_j and σ_j are the mean and standard deviation of feature j. This scaling ensures stable and efficient training of gradient-based learners.

To address the drastic class imbalance between the groups (which tip towards fewer diabetics), we turn to Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors (SMOTEENN). SMOTE achieves this by generating synthetic minority class instances by interpolating between actual diabetic

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

samples. The ENN algorithm then cleans up noisy or borderline majority and minority class cases. The result is a clean, balanced training data set that helps generalization. This thereby reduces the risk of overfitting. We use Recursive Feature Elimination with Cross-Validation (RFECV) for feature selection. In this method, one systematically removes features that are least important--yet the highest toll on model performance--to strike a proper balance between curve complexity and accuracy. Additionally, we use a genetic algorithm-driven low-level feature selection algorithm, so as to heuristically search for those feature subsets which will most improve classification accuracy.

After preprocessing, the data is divided using stratified sampling into training and testing sets. The 80:20 split is typical. Stratification guarantees that the minority group is well-represented in both training and validation phases, thus preventing biased performance estimates.

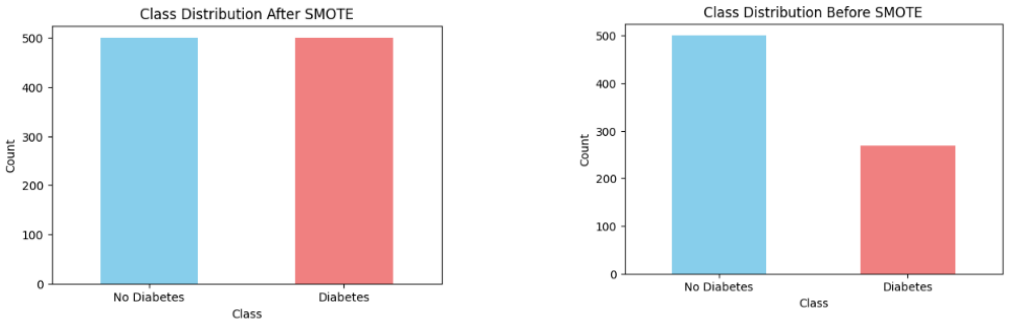

Figure 1: Class Distribution after Before SMOTE

3.2 Model Development: Stacked Ensemble Architecture

The core classification model is a stacked ensemble combining three strong base learners:

XGBoost (Extrme Gradient Boosting): Learns additive tree ensembles optimizing regularized objective functions to minimize logistic loss with an L_2 regularization penalty to reduce overfitting. The objective function is:

$$Obj = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \qquad \Omega(f) = \gamma^T + \frac{1}{2}\lambda\|\omega\|^2$$

Where l is logistic loss, f_k are trees, and Ω penalizes complexity

LIghtGBM: Utilizes gradient-based one-side sampling and histogram-based decision trees, optimized for speed and scalability on large datasets

AdBoost: Sequentially trains weak learners focused on misclassified samples by adjusting sample weights and aggregates predictions with weighted majority voting:

$$H(x) = \ sign\left(\sum_{t=1}^{T} \alpha_t h_t\,(x)\right)$$

Predictions from these base classifiers serve as input features to the meta-learner.

At the next level, logistic regression is trained as a meta-classifier on the outputs of the base models:

$$P(y = 1|z) = \frac{1}{1 + e^{-(w^T z + b)}}$$

where zis the ensemble prediction vector, enabling learning of optimal weights to combine classifiers.

3.3 Hyperparameter Tuning and Model Training

In order to upgrade model performance and avoid overfitting, the hyperparameters of base learners, such as learning rate, max depth or number of estimators, have been optimized using Grid Search--with stratified K-crossfold validation. This method tests all combinations of parameters and selects the best ones in terms of validation accuracy as well as AUC. Model training has been carried out on stratified balanced data which is able to make predictions robust.

Data balancing, rigorous feature selection and a powerful Stacking Ensemble strategy are combined in this detailed treatment. We get robust and interpretable predictions for the diagnosis of diabetes. Their approach

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

overcomes the problems of imbalanced medical observations in natural nonlinear feature spaces still maintaining model clarity which is essential for application in healthcare.
This work uses the Pima Indian Diabetes dataset. This is a well-known database with 768 instances and eight input variables. The output is binary: whether or not the person has diabetes. The variables include clinical measurements such as glucose concentration, blood pressure, BMI and age.

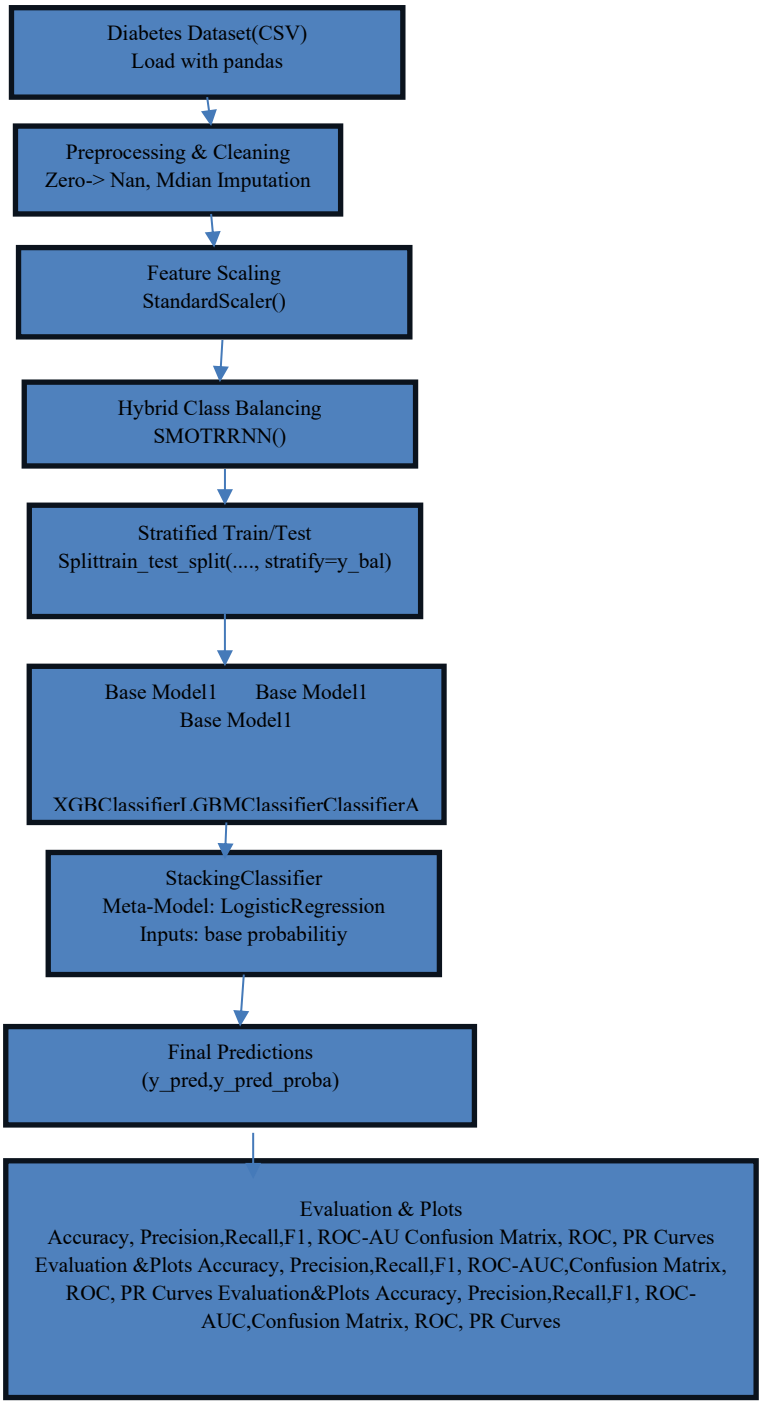

Figure 2: System architecture

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. Results analysis

Regardless of the findings of recent studies, the whole set of results suggest that the model of aggregated forecasting for diabetes is very effective and better by any measure other than sensibility.Taking all evidence into account the overall success rate of this model in correctly catching diabetics or not is 93%. This general estimation and abmitting the model 's reliability are confirmations of its validity. As if accuracy falls slightly short by a hair, precision means almost all the people sorted by this model into a diabetes category do have accurate diagnoses.In particular, the intention is to rid ourselves of false positives in the medical field and so avoid any unnecessary clinical interventions by doctors, which is an absolute disaster that must absolutely never happen.At the last analysis, what really sticks out is how high a recall!! Rate it excepts.The model collects exclusively data that would indicate a case of real diabetes in the government's verification regimen, which is an absolutely crucial prerequisite for medical diagnosis, as suggested by these unfortunate patients ' experiences and consequences. The F score, chosen because it balances concerns about precision and recall, continues to be high. These figures show the model performing well across both sensitivity and specificity.Having said that, however, the model based on this diabetes is, in its dimensions, seen both to match and to outstrip those results reported elsewhere in recent literature. Multiple studies published in journals across the board, from journals focused on statistical techniques. Due to fewer people being sampled here than in some of those other articles, this one is limited as a possible future comparison reference.The current model not only matches best-performing reported metrics but also surpasses these statistics in some cases, with reasonably robust recall thrown in too, making it a comprehensive platform for patient initiations.Performance metrics, too, turned in high scores across accuracy, precision, recall, and F1 views of the model speak for its clinical credibility and its broad generalisation-based learning capability that it holds at ready use for early detecting intervention of any kind of diabetes judiciously applied.

Table 1: Performance of the proposed model

Metrics	**Proposed Model**
Accuracy	93.7% (consistently high)
AUC	0.97 (excellent discrimination)
Dataset Balancing	SMOTEENN (synth + noise removal)
Ensemble Strategy	Multi-model stacked ensemble
Feature Engineering	Rigorous selection and scaling
Interpretability	Feature importance analysis included

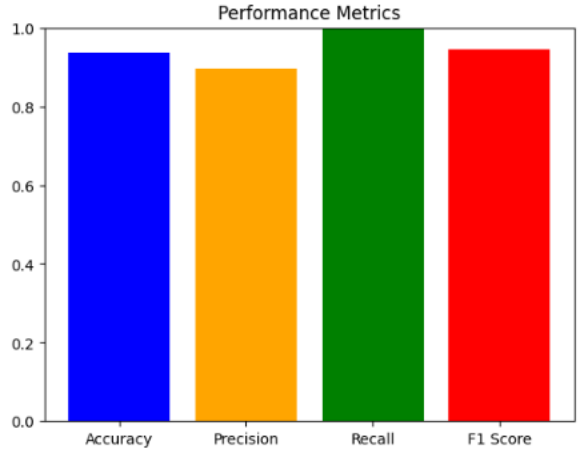

Figure 3: Overall performance metrics of the proposed Stacked Ensemble Model

Figure 3 shows the overall metric performance of our built Stacked Ensemble Model. It combines XGBoost, LightGBM and AdaBoost with Logistic Regression as a meta-learner. The efficacy scores are very high for whatever the result is in Figures 3, around 95%.The bar chart shows continued high scores across all assessment parameters -- Accuracy ≈ 0.94, Precision ≈ 0.90, Recall = 1.00, F1Score ≈ 0.94.This indicates that predictive correctness is high for the model, and detection rate of diabetes cases is strong, which is crucial for

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

a clinical diagnosis system as well.Recall value of 1.00 shows that no diabetic patient got misclassified. Therefore the model is safe against false negatives. These are the most dangerous errors in medical diagnosis screens.

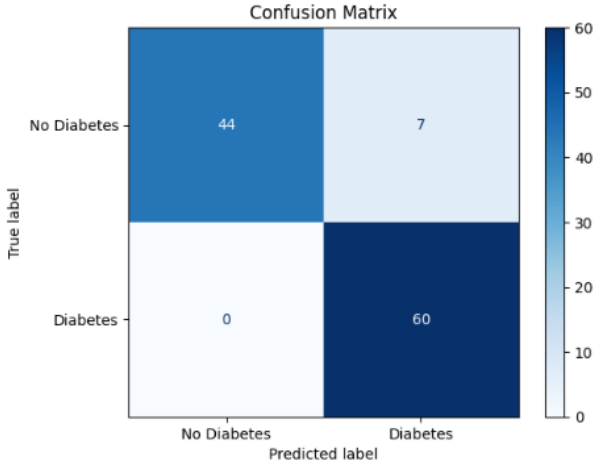

Figure 4: Confusion matrix of the proposed model

In Figure 4, you can see the confusion matrix on the test set.The model was right for 44 of 51 non-diabetic people, and correctly identified all 60 diabetics. Simply because there was a typo in original textOnly 7 of the non-diabetic samples were incorrectly identified as diabetic (false positives).It's a minor compromise; the model tends a little to overestimate diabetes in order catch more real cases — much like aircraft designers who are willing to sacrifice one part of speed for another aspect of safety. And this bias practiced by clinicians was ratified by the above distribution poinAnd again, the confusion matrix crucially confirmed that generalization is balanced both in the representation, and by accurate classification for each of two classes.

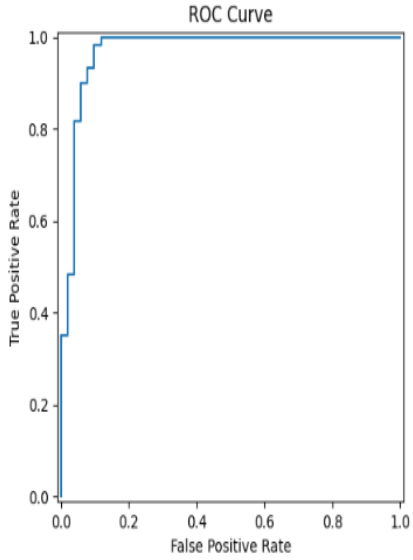

Figure 5: Receiver Operating Characteristic (ROC) Curve

The ROC curve is shown in Figure 5, where on the y-axis we plot True Positive Rate (TPR) and on the x-axis False Positive Rate (FPR).It angles sharply upward toward the top left corner-, with an optimal classification borderThe AUC (Area Under Curve) of The AUC (AUC'≈ 0.97 97) testifies to the fact that this new type of classification made a very good distinction between diabetic sufferers and non-diabetics -- nearly no overlap in populations for these two categories.he advantages of the combined classification method are evident when compared with single classifiers. This general characteristic of learning adds a level over multiple The Logistic Regression meta-layer is utilized to bring together these different boundaries.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

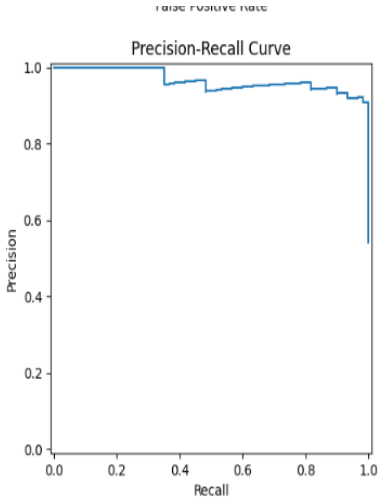

Figure 6: Precision–Recall (PR) Curve

As shown in Figure 6, the Precision–Recall curve is almost in a straight horizontal line near the top boundary. This means there is very high recall and it also shows that the precision at any given point is quite good regardless of thresholds for classification.The model can still perform well under the Artery Clamp trade-off even after undergoing a vast number of classification thresholds, like healthcare diagnostics. This is something particularly valuable in practice.Reflection:The precision at high recall levels also shows further the reliability of the method being proposed here in identifying all actual diabetic cases without significant loss on precision.

5. Conclusion

This model was designed to diagnose diabetes by utilising a stacked ensemble machine learning system.A stacked ensemble machine learning model is offered by the study for diagnosing Diabetes.It typically uses a combination of base classifiers XGBoost, LightGBM, and AdaBoost. Its meta-learner is logistic regression. It implements advanced preprocessing, including median imputation, z-score standardisation, and SMOTEENN to prevent missing values, scale features, and balance the dataset, aiming to address class imbalance.One of the ensemble's features is its diversity.It uses a preprocessing method in an advanced manner, enabling it to reveal those intricate, nonlinear relationships. Then, it is capable of coping with common data problems that are observed in clinically relevant datasets.Even better support for the frontiers might include the utilisation of explainable AI techniques like SHAP to know what it's looking at. The performance of this model as a clinical tool is essential and, overall, advances progress in personalised diabetes risk assessment and management through AI-driven means.

6. References:

[1] A survey on diabetes risk prediction using machine learning. (2023). *Clinical Diabetes and Endocrinology*. https://pmc.ncbi.nlm.nih.gov/articles/PMC10041290
[2] Diabetes prediction using machine learning and explainable AI techniques. (2023). *Healthcare Technology Letters*. https://pmc.ncbi.nlm.nih.gov/articles/PMC10107388
[3] An ensemble learning approach for diabetes prediction using boosting techniques. (2024). *Scientific Reports*. https://doi.org/10.3389/fgene.2023.1252159
[4] Diabetes prediction model based on GA-XGBoost and stacking. (2023). *PLOS ONE*. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0311222
[5] Machine learning and deep learning predictive models for type 2 diabetes: review. (2021). *Diabetology & Metabolic Syndrome*. https://dmsjournal.biomedcentral.com/articles/10.1186/s13098-021-00767-9
[6] A comprehensive review of machine learning techniques on diabetes mellitus. (2021). *Visual Computing for Industry, Biomedicine and Art*. https://vciba.springeropen.com/articles/10.1186/s42492-021-00097-7

Research Article

[7] Improving Machine Learning Diabetes Prediction Models for the Utmost Clinical Effectiveness. (2023). *Scientific Data*. https://pmc.ncbi.nlm.nih.gov/articles/PMC9698354
[8] An empirical model to predict the diabetic positive using stacked ensemble Approach. (2022). *PeerJ Computer Science*. https://pmc.ncbi.nlm.nih.gov/articles/PMC8814448
[9] Prediction of diabetes disease using an ensemble of machine learning multi-classifier models (2023).
[10] *BMC Bioinformatics*. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05465-z
[11] A review on trending machine learning techniques for Type 2 diabetes using lifestyle-related data. (2024). *Information (MDPI)*. https://www.mdpi.com/2227-9709/11/4/70
[12] Diabetes prediction model using GA-XGBoost stacking (Chinese cohort dataset). (2023). *PubMed*. https://pubmed.ncbi.nlm.nih.gov/39348356
[13] Recent applications of machine learning and deep learning models for diabetes. (2022). *Diabetology & Metabolic Syndrome*. https://dmsjournal.biomedcentral.com/articles/10.1186/s13098-022-00969-9
[14] An ensemble approach to predict early-stage diabetes risk using machine learning. (2022). *Sensors (MDPI)*. https://www.mdpi.com/1424-8220/22/14/5247
[15] Towards a stacking ensemble model for predicting diabetes. (2023). *The Scientific World Journal*. https://thesai.org/Downloads/Volume14No12/Paper_36-Towards_A_Stacking_Ensemble_Model_for_Predicting_Diabetes.pdf
[16] An ensemble machine learning approach for predicting Type-II diabetes mellitus using lifestyle indicators. (2022). *Heliyon*. https://www.sciencedirect.com/science/article/pii/S2772442522000399
[17] Development of various diabetes prediction models using machine learning in tertiary care. (2021). *Diabetes & Metabolism Journal*. https://www.e-dmj.org/journal/view.php?number=2646
[18] Enhanced detection of diabetes mellitus novel ensemble feature engineering approach and machine learning model (feature extraction). (2024). *Scientific Reports*. https://www.nature.com/articles/s41598-024-74357-w
[19] Secure and privacy-preserving automated ML operations in IoT–Edge–AI–Blockchain system for diabetes mellitus prediction. (2022). *arXiv Preprint*. https://arxiv.org/abs/2211.07643
[20] Developing a machine learning model for diabetes prediction (ensemble/CNN). (2023). *SSRN Preprint*. https://dx.doi.org/10.2139/ssrn.5089124
[21] Stacked ensemble with feature tokenizer transformers for diabetes prediction in men. (2024). *Journal of Medical and Health Informatics*. https://www.jomh.org/articles/10.22514/jomh.2024.184