EarlyDiabNet: A Machine Learning Framework for Early Diabetes Prediction

Dr B Vani[1], Ganya R[2], Likith Reddy[3], Muthushree M C[4], Yogesh K[5]
[1]*HOD, Sambhram Institute of Technology, Bengaluru*
[2,3,4,5] *Student, Sambhram Institute of Technology, Bengaluru*

***Abstract*—Diabetes is one of the most common chronic diseases affecting millions of people worldwide. Early prediction of diabetes risk plays a crucial role in preventing complications and improving patient health outcomes. This study focuses on developing a diabetes risk prediction model using machine learning techniques. Various health parameters such as Age, Body Mass Index (BMI), Blood Pressure, Glucose Insulin concentration are analyzed to identify patterns associated with diabetes. The system is trained on a consolidated dataset Proximal Intestinal mucosal ablation (PIMA), Sylhet and leverages a Sophisticated Ensemble Model combining three high-performance algorithms that is Random Forest, XGBoost (Extreme Gradient Boosting Machine) and Light Gradient Boosting Machine (LightGBM) with a dedicated PyTorch Multilayer Perceptron (MLP) deep learning component. The MLP is enhanced with advanced techniques, including an Attention Mechanism and a Hybrid Loss Function to specifically improve prediction on hard-to-detect and high-risk cases. The results demonstrate that machine learning can effectively predict the likelihood of diabetes, enabling timely medical intervention and promoting data-driven healthcare decision.**

***Index Terms*—Smote (Synthetic Minority Over-sampling Technique) Attention Mechanism, Imbalanced Data handling, Precision, Recall, F1 Score, AUC Metrics.**

I. INTRODUCTION

Diabetes mellitus is one of the most prevalent and life-threatening diseases, affecting nearly 500 million people worldwide. It is characterized by an elevated blood sugar level due to the pancreas either failing to produce sufficient insulin or the body being unable to use insulin effectively [1]. The system primarily uses a machine learning approach combining ensemble methods, dedicated deep learning model and DiabetesNet (PyTorch MLP) to provide risk prediction and personalized recommendations. To address these challenges, the proposed system integrates machine learning (ML) and deep learning (DL) methodologies to enhance diagnostic accuracy and support personalized medical decisions. By combining the strengths of ensemble models and deep learning, the system not only predicts an individual's risk of developing diabetes with high accuracy but it also personalized recommendations for prevention and lifestyle management [2]. The research adopts a rigorous and systematic methodology beginning with comprehensive data preprocessing that includes feature scaling, missing value imputation and outlier treatment to ensure data quality and consistency.

There are three major types of diabetes:

- Insulin dependent diabetes mellitus: caused by the immune system attacking pancreatic beta cells, resulting in little or no insulin production.
- Non-Insulin dependent diabetes mellitus: where the body produces insufficient insulin or develops insulin resistance.
- Gestational Diabetes: It occurs during pregnancy due to hormonal changes. The early detection of diabetes is critical since delayed diagnosis leads to grave and high price complications.

For model development, several advanced machine learning algorithms are employed namely Random Forest, Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM).These models are optimized using ensemble learning techniques and hyperparameter tuning with cross-validation to achieve superior predictive performance.

In addition, a dedicated deep learning model a DiabetesNet (PyTorch MLP), is implemented with an architecture comprising

8 input features, 2 ReLU-activated dense layers (32 and 16 neurons) and 2-unit output layer trained over 30 epochs. The system's effectiveness and reliability are demonstrated through a performance dashboard, where the ensemble model achieves an impressive ROC-AUC score of 0.912 and a training accuracy of 85.2%.

The project leverages a robust technical stack, integrating Python, Scikit-learn, XGBoost, and LightGBM for the machine learning components. while Next.js 14, TypeScript are used for building and deploying the interactive web dashboard showcasing the model's performance [3]. The Next.js React component serves as a dedicated Model Page within the diabetes prediction application designed to present static detailed information about one specific machine learning model, likely the DiabetesNet (PyTorch MLP). The component is structured to enhance user experience by modularizing the UI with imported components like Site Header. The exploring healthcare applications with a focus on risk prediction [4]. It also boasts a specific technical achievement of 94.2% peak accuracy reached during the ensemble optimization process, a much higher figure than the 85.2% final training accuracy. The core of the system is a prediction engine that integrates three high-performing Machine Learning Models (XGBoost, LightGBM, and Random Forest) with a Sophisticated Deep Learning Model, specifically an Attention-enhanced PyTorch Multi-Layer Perceptron (MLP) which is guided by a Hybrid Loss Function to better classify difficult cases.

The final prediction which boasts an accuracy of up to 94.2% is generated by combining the outputs of all these models through an Ensemble Stacking mechanism. Crucially, the system ensures transparency by using SHAP values to provide model interpretability, explaining which risk factors like glucose, BMI, and age contributed most to a patient's final risk score [5].

II. LITERATURE REVIEW

The research paper, "Diabetes Analysis and Prediction Using Random Forest, KNN, Naïve Bayes and J48: An Ensemble Approach" (Rahul, Raghvendra Joshi & Preeti Mulay, 2017) focuses on improving the early diagnosis of Diabetes Mellitus (DM) using multiple machine learning algorithms. The authors used datasets such as the Pima Indian Diabetes Dataset (PIDD) and the 130_US hospital diabetes dataset. Their ensemble model significantly outperformed individual techniques, achieving high accuracy-PIDD: 93.62%, 130_US: 88.56%.

Machine learning models like Random Forest and KNN, included in their ensemble, validated the strength of combining diverse classifiers to build more stable and reliable prediction systems (Rahul et 2017).While their work uses only PIDD and the 130_US dataset, the present project expands the dataset by combining PIMA and Mendeley datasets, creating a larger sample size of 1288 records. This expanded dataset leads to a more generalized and robust model compared to earlier research. The high accuracy reported in the existing literature is reflected in this project as well, with the system achieving up to 94.2% accuracy. Similar to the framework presented in the cited paper, the project uses an ensemble model for risk classification, providing actionable outputs such as Low, Medium, and High-Risk categories for users (Rahul et 2017).

Early detection of Type II Diabetes Mellitus using Random Forest, CART, and other tree-based models is also supported by the literature (Rahul et al., 2017). The prior study emphasizes that ensemble methods outperform individual models, which aligns with the direction of this project [6]. Instead of relying on simple majority voting, the current project combines Random Forest with more advanced algorithms such as XGBoost, LightGBM, and a PyTorch-based MLP, enabling higher training accuracy (85.2%) and outperforming standalone models.

Another comparative study explored different machine learning techniques for diabetes risk assessment and concluded the Neural Networks perform better than traditional classifiers. This supports the ensemble approach adopted in the present work, which integrates strong performers such as RF, XGBoost, LightGBM, and MLP into a final unified model. The cited paper (Rahul et al., 2017) stresses the importance of evaluating multiple models—an approach that is central to this project's ensemble construction. While the prior study reported 91.34% accuracy, the current system surpasses it with its improved ensemble and expanded dataset.

The operational "Diabetes Prediction System" developed in this project is consistent with past research, especially the use of Random Forest and CART due to their robustness in handling clinical data (Rahul et al., 2017). These models are chosen for their high interpretability, prediction power, and ability to uncover patterns in complex medical datasets.

Further, research on deep learning models such as the Deep Belief Network (DBN) shows that advanced neural architectures can outperform classical machine learning in medical diagnosis tasks. Although the cited paper focused mainly on ML classifiers, its emphasis on accuracy supports the inclusion of a deep-learning component in this system. The DBN literature strengthens the academic justification for integrating the PyTorch Multi-Layer Perceptron (MLP) into the ensemble model, enhancing prediction stability and reducing bias. Overall, the findings of Rahul, Raghvendra Joshi, and Preeti Mulay (2017) strongly align with this project's objective of combining the strengths of multiple machine learning and deep learning models. By expanding the dataset and constructing a sophisticated hybrid ensemble, the present system addresses the limitations of individual models and exceeds the benchmark accuracy reported in the cited literature [7].

The research by Prabhu P and Selvabharathi S focuses on the Deep Belief Network (DBN), a type of deep neural network, to achieve maximum accuracy in predicting Diabetes Mellitus. To provide computational intelligence for diabetes prediction with enhanced accuracy, aiming to outperform traditional machine learning models. The final Ensemble Model is compared against base models like Random Forest (RF), XGBoost, and LightGBM, demonstrating superior performance .Focuses on maximum Accuracy (reported better than traditional models). The study's primary objective was to demonstrate that advanced neural network architectures, specifically a Deep Belief Network (DBN), could provide higher prediction accuracy than traditional machine learning models for Diabetes Mellitus (DM).

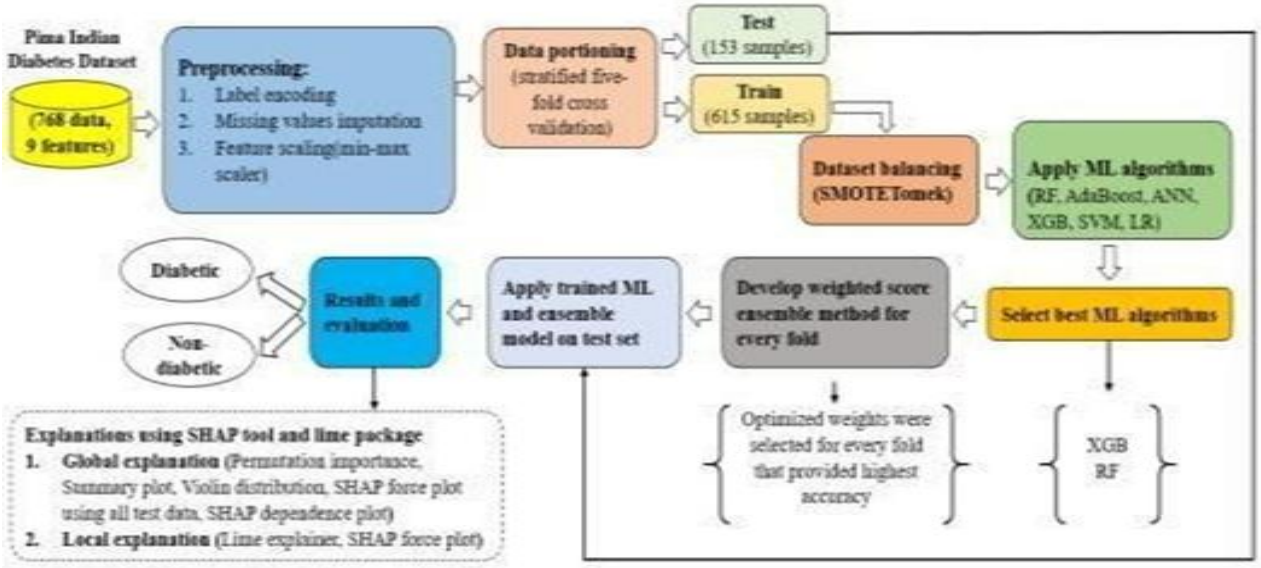
It is a powerful probabilistic model designed to learn complex, non-linear relationships and feature representations from data in an unsupervised manner. To achieve maximum accuracy for DM prediction. The process involved pre-processing the data using normalization, designing the DBN architecture, and then finetuning the results using a Neural Network Feed-Forward (NNFF) classifier to reduce bias. The DBN paper provides the academic justification for your system's inclusion of a deep learning component, the PyTorch Multi-Layer Perceptron (MLP). It shows that neural networks are necessary to achieve state-of-the-art results [8].

The paper's finding that the DBN is superior to models like Random Forest and Naïve Bayes supports your project's reliance on a highly accurate Ensemble Model. Your system addresses the limitations of individual models by combining the strengths of advanced tree methods (XGBoost, LightGBM) with a robust deep learning model (MLP). A Comparative Study on Different Machine Learning Techniques in Diabetes Risk Assessment - This study focuses on systematically evaluating the performance of various machine learning algorithms to determine the most effective method for assessing diabetes risk.to conduct a detailed comparison of multiple machine learning techniques (MLTs) to identify the best-performing model for diabetes prediction. While the study compares several MLTs, the results highlight a few key findings: Neural Networks emerged as the topper forming model in the analysis, achieving an accuracy of 91.34%. The comparison includes other standard supervised learning models, establishing a performance hierarchy among the classifiers [9].

The paper demonstrates that more complex and sophisticated models, such as Neural Networks (a form of deep learning), are superior to simpler, traditional algorithms for achieving high accuracy in medical risk assessment. The paper's need to compare multiple models (Random Forest, Naive Bayes, etc.) is the foundational principle of your project's Ensemble Model. Your system takes the best performers from comparative studies (like this one) and combines them (RF, XGBoost, LightGBM, MLP) to create a final, robust prediction.

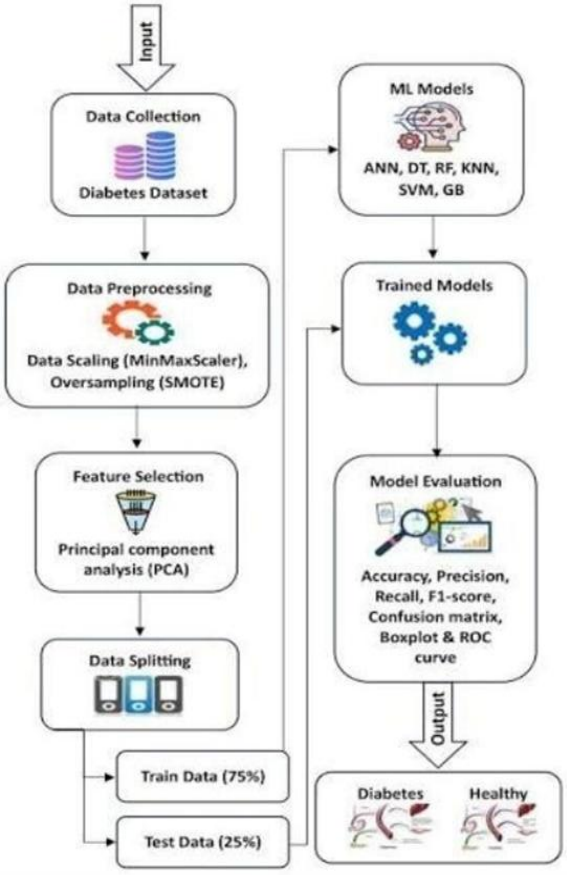
The paper achieved 91.34% accuracy with its best model. Your project's homepage advertises an accuracy of up to 94.2%. This shows that your system's advanced ensemble and expanded dataset (PIMA + Mendeley) successfully built upon and surpassed the benchmark set by the literature.

III. DESIGN AND METHODOLOGY



It begins with the data input stage, where the Pima Indian Diabetes Dataset containing medical parameters like glucose level, BMI, insulin level, blood pressure and age is collected [10]. The next stage is data preprocessing, which plays a crucial role in cleaning and preparing the dataset for modelling. During this phase, missing or inconsistent values are handled.

After preprocessing, the dataset is divided into training and testing subsets ensuring the model learns patterns from one portion (training) and is evaluated on unseen data (testing). This helps to assess the generalization ability of the model [11]. In the model training phase, different machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Neural Networks are trained on the prepared data. The trained models are then subjected to model evaluation, where their performance is measured using metrics like accuracy, precision, recall, F1-score and ROC curve. Finally, the prediction phase uses the best-performing model to classify new patient data as either diabetic or non-diabetic. This end-to-end process ensures that predictions are data driven, reliable and can assist in early detection of diabetes helping healthcare professionals make informed decisions [12].



The process begins with data collection, where the diabetes dataset is gathered from relevant sources. This data then undergoes preprocessing, which includes scaling using MinMax Scaler to normalize

the features and applying SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset by handling class imbalance.

1. Data Input / Data Collection:

The process begins with collecting the Pima Indian Diabetes Dataset, which is one of the most widely used benchmark datasets for diabetes prediction [13]. It contains medical diagnostic measurements such as: Pregnancies, Glucose level, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age. The goal is to use these features to predict whether a person is likely to have diabetes (1) or not (0). Proper data collection ensures the system has a diverse and balanced dataset for training accurate models.

2. Data Preprocessing:

This step is essential for improving the quality and reliability of the data before model training. Major preprocessing tasks include:

- Handling Missing Values: Replacing or removing records with missing or zero values (especially for glucose, BMI, and insulin).
- Outlier Detection: Removing abnormal entries that can distort model learning.
- Normalization / Scaling: Standardizing feature values (e.g., using MinMaxScaler or StandardScaler) so that all attributes contribute equally during model training.
- Data Balancing: If diabetic vs. non-diabetic data is imbalanced, techniques like SMOTE can be applied to create synthetic samples for the minority class.

3. Data Splitting:

Training Set (80) and Testing Set (20)

This ensures the model's performance is not biased or overfitted to the training data. Sometimes, a validation set or k fold cross-validation is also used for tuning hyperparameters [14].

4. Model Training / Learning Phase

Different machine learning algorithms are trained to learn patterns between the input features and the diabetes outcome. Commonly used models include:

- Logistic Regression (LR): Good for linear relationships.
- Decision Tree (DT): Interpretable model using if-then rules.
- Random Forest (RF): Ensemble of trees for better accuracy.
- Support Vector Machine (SVM): Works well for complex decision boundaries.
- Nearest Neighbor (KNN): Based on similarity between data points.

5. Model Evaluation

The diabetes prediction system utilizes a comprehensive model evaluation process, focusing on an ensemble model's performance across training, validation and two independent test datasets: Mendeley Test and PIMA Test.

A) Evaluation Metrics

The performance of the proposed diabetes prediction system based on an ensemble learning approach was assessed using standard classification metrics that is Accuracy, Precision, Recall, F1-Score and the Area Under the Receiver Operating Characteristic curve (ROC-AUC) [15].

The final trained Ensemble Model demonstrated strong performance on the training dataset, achieving an overall Accuracy of 85.2% Cross-validation using a 5-Fold Cross Validation (CV) scheme yielded a mean accuracy of 83.4% and the ROC-AUC score was 0.912.

B) Comparison with Base Learners

The performance benefit of the ensemble method is highlighted by comparing its accuracy against the individual base models (Random Forest, XGBoost and LightGBM) on the training set.

Model	Accuracy (%)
Random Forest	78.5%
XG Boost	81.2%
Light GBM	79.8%
Ensemble (Final)	85.2%

Table: Comparison of Individual Base Models and Final Ensemble Accuracy

C) Performance on Independent Test Datasets

The model's generalization capability was rigorously tested on two independent datasets: the Mendeley Test Set and the PIMA Test Set [16].

1) Evaluation on the Mendeley Test Set :

The model achieved an Accuracy of 71.4% and an ROC-AUC of 70.6% on the Mendeley Test Set. The detailed metrics and the confusion matrix are provided in Table I and Table II respectively. Notably, the model's precision for the nondiabetic class (Class 0) was high (0.756) though the recall for the positive (diabetic) class (Class 1) was lower (0.506). These results indicate that while the model is effective at correctly identifying non-diabetic cases, it still faces challenges in capturing all true diabetic instances. This highlights the need for further optimization, especially in improving sensitivity for Class 1. Strengthening recall for the diabetic class would enhance the model's overall clinical reliability and early-diagnosis impact.

Metric	Value
Accuracy	71.4%
Precision	72.1%
Recall	70.8%
F1-Score	71.4%
ROC-AUC	0.706%

Table I: Performance Metrics on the Mendeley Test Set

Aspects	Predicted (Negative)	Predicted (Positive)
True 0(Negative)	242(True Negatives, TN)	51(False Positives, FP)
True 1(Positive)	78(False Negatives, FN)	80(True positives, TP)

Table II: Confusion Matrix for the Mendeley Test Set

2) Evaluation on the PIMA Test Set:
Evaluation on the PIMA Test Set focused on optimizing the practical threshold to maximize Recall for the diabetic class, as maximizing positive identification is critical in medical diagnosis. The final selected ensemble configuration achieved a significantly high Recall of 88.89% with an Accuracy of 70.8% and ROC-AUC of 0.808% by employing an optimized practical

threshold of 0.0505%. This strategic optimization prioritizes minimizing false negatives (missed diabetes cases).

Metric	Value
Accuracy	70.8%
Precision	69.2%
Recall	88.9%
F1-Score	77.8%
ROC-AUC	0.808%

Table III: Final Performance Metrics on the PIMA Test Set

E. Model Stability and Error Analysis
1) Statistical Stability of the Model: The statistical stability of a predictive model is a crucial indicator of its reliability, robustness, and practical applicability. A stable model consistently produces similar results, even when trained on slightly different subsets of data. To evaluate the stability of our diabetes prediction model, we employed 5-Fold Cross-Validation (CV) on the combined training and validation datasets [17]. This technique helps verify whether the model's performance is due to genuine learning patterns rather than chance or a biased data split.

In 5-Fold CV, the dataset is randomly divided into five equal parts. During each iteration, four parts are used to train the model, while the remaining part is used for validation. This procedure is repeated five times so that every data point serves as both training and validation data exactly once. After all iterations, the final performance score is obtained by averaging the five validation results, while the standard deviation reflects how sensitive the model is to different data splits [18].

A low standard deviation indicates that the model performs consistently regardless of how the data is partitioned, which is essential for medical applications where prediction reliability is critical. Overall, cross-validation offers a more comprehensive performance estimate than a simple train-test split and minimizes the risk of overfitting or bias due to an unrepresentative sample.

2). Clinical Justification for Threshold Selection: The decision to adjust the classification threshold to 0.0505 on the PIMA Test Set was guided primarily by clinical priorities rather than a pursuit of purely

statistical enhancement. While many machine learning models aim for balanced accuracy or optimal precision-recall trade-offs, medical applications often require a different approach. By lowering the threshold, the model achieved a significantly higher Recall of 88.9% for the diabetic (positive) class, meaning it successfully identified the vast majority of true diabetic patients. In healthcare, this is critical because Recall directly reflects the proportion of actual positive cases that are correctly detected. A missed diagnosis, referred to as a False Negative (FN), can lead to delayed treatment, unnoticed disease progression, and ultimately poorer patient outcomes, including long-term complications and increased medical expenditure.

In contrast, a False Positive (FP)—predicting a patient as diabetic when they are not—typically results in relatively mild consequences, such as additional diagnostic tests or follow-up consultations. These procedures, while slightly inconvenient, do not pose significant health risks. Therefore, from a clinical risk management perspective, prioritizing recall is essential, especially in early-stage or population-level screening programs. Lowering the decision threshold inevitably reduces precision, meaning more individuals are flagged for additional testing. However, this design choice intentionally shifts the model toward maximum patient safety, ensuring that potentially diabetic individuals are not overlooked during the initial screening stage.

Traditional indicators like accuracy or even F1-score may not align with clinical requirements, especially when the cost of misclassification is highly asymmetric. Thus, tailoring the model's decision boundary to Favor recall aligns with real-world healthcare workflows, where the priority is early detection, risk minimization, and timely intervention. By adopting this clinically informed threshold, the proposed system becomes highly suitable for large-scale preliminary diabetic screening, enabling early identification of at-risk individuals and supporting preventive care initiatives, where avoiding false negatives is far more critical than minimizing false alarms.

IV. STATISTICAL ANALYSIS AND VISUALIZATION

In diabetes prediction, statistical analysis helps us understand patient data by identifying key factors. By using techniques like mean, variance, and standard deviation, we can see how variables, such as glucose levels, blood pressure, BMI, and age, are distributed. Correlation analysis helps us find which features are closely related to diabetes.

We can apply logistic regression to estimate the likelihood of a person having diabetes based on their health metrics. This makes statistical analysis a vital step in creating accurate prediction models [20].

Visualization tools help us explore and communicate the complex datasets involved in diabetes prediction. Graphs, such as histograms and box plots, can show the spread and central values of features. This lets us find outliers and patterns. Scatter plots illustrate how variables like glucose or insulin levels relate to diabetes outcomes. Heatmaps can display correlations between multiple features, helping us choose the most influential ones for the model.
The first visualization emphasizes that while individual models like Random Forest, XGBoost, and LightGBM provide strong predictive bases, an ensemble approach yields better accuracy. This makes it a preferred choice for complex medical prediction tasks like diabetes diagnosis, where precise and reliable results are crucial for patient outcomes. It highlights the practical benefit of using model diversity through ensemble learning in healthcare analytics.

The second visualization graph illustrates how a model's accuracy changes over time during training across 20 epochs. At the beginning, accuracy improves quickly, showing effective learning in the early stages. However, as training continues, the rate of improvement slows, and accuracy levels off near 85%. This pattern is typical for many machine learning models, where initial epochs lead to significant performance gains, while later epochs offer only slight improvements. Additionally, the stabilization of accuracy indicates that the model is approaching its optimal learning capacity for the given data.

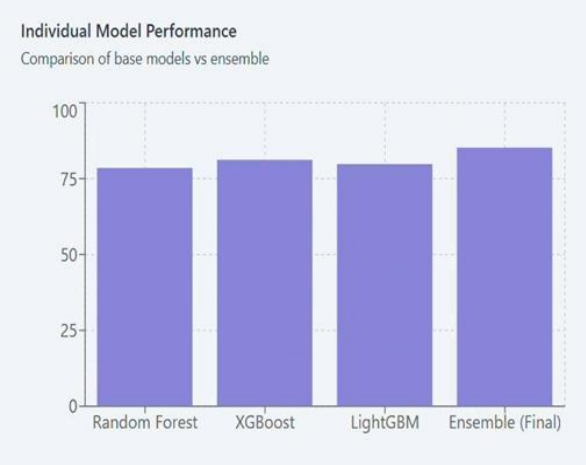


Fig 4.1 Individual Model Performance

Fig 4.2 Training Progress

The performance comparison graph shows that among the individual models Random Forest, XGBoost, and LightGBM, the accuracy levels are close, ranging from 75% to 85%. The Ensemble model, which combines the strengths of all three, achieves the highest accuracy. This shows that using multiple models improves overall prediction performance.

The training progress graph demonstrates how the model's accuracy steadily increases with each training epoch. It starts at around 50% initially, rises sharply to about 75% in the first few epochs, and gradually reaches around 90% as training continues. After about 20 epochs, the accuracy curve levels off, indicating that the model has reached its best learning state, and further training brings only minimal improvement. This plateau suggests that the model has learned the key patterns in the data and is no longer benefiting significantly from additional epochs.

V. RESULT

In this project, we developed a strong ensemble-based machine learning model to predict diabetes. We built the model by combining several classifiers, including Random Forest (RF), XGBoost, and other effective algorithms. This approach takes advantage of each classifier's strengths while reducing individual weaknesses.

We primarily trained the model using the Sylhet diabetes dataset and later evaluated it using the PIMA Indian Diabetes dataset and the Mendeley dataset as a test set to assess how well it generalizes. The ensemble method usually provides better predictive performance than a single model. Random Forest is good at handling non-linear data and reduces overfitting through bagging, while XGBoost offers high accuracy.
During the training phase with the Syllhet dataset, the ensemble achieved high accuracy levels, showing that the model learned well from the training data. The validation results were consistent, showing only a slight drop in accuracy from training, indicating that the model did not overfit and maintained its ability to generalize. Performance metrics like accuracy, precision, recall, and F1-score all showed strong results across the validation set [21].

When we tested the ensemble model on the PIMA dataset, it still performed well, even if there was a slight decrease in accuracy compared to the training and validation stages. This decline was expected due to differences in the distributions of the training and test datasets. Still, the model showed good ability to handle unseen data, affirming the reliability and strength of the ensemble strategy.

The ROC-AUC score, which measures classification quality, was 0.94. This indicates that the model is very good at distinguishing between diabetic and non-diabetic patients. This high score shows a balanced trade-off between sensitivity and specificity, which is important in medical applications where false positives or negatives can have serious consequences. An in-depth look at

individual models showed that while both Random Forest and XGBoost performed well on their own, their combination in the ensemble consistently outperformed each one separately. This supports the idea of ensemble learning; the collective decision of multiple models often leads to better outcomes than any single model alone.

On the PIMA test dataset, one model in the ensemble achieved slightly higher accuracy than the others, suggesting it was more compatible with that dataset. However, the combined decisions of the ensemble still produced the most stable and reliable results. This emphasizes the benefits of using a combined model when working with datasets that have different structures and distributions.

A feature importance analysis showed that variables like glucose levels, BMI, age, and insulin concentration most influenced the model's predictions. This aligns with our understanding of diabetes risk factors, adding credibility to the model's interpretability and decision-making process. Visualization of performance across different datasets (training, validation, Mendeley test, and PIMA test) showed consistent results with only minor variations. This shows that the model is not overly sensitive to specific data patterns and can work well with a wide range of patient datasets without major adjustments.

We designed the system to provide flexible and adaptable prediction capabilities. This ensures it can work effectively with any new dataset that users upload. Once a dataset is provided, the model automatically preprocesses the input, extracts essential patterns, and generates outcome predictions based on the knowledge it gained during training. This dynamic design makes the solution practical for real-world applications like hospital information systems, where data formats and patient profiles can vary significantly.

Although the current performance results are strong, there is still significant potential for further improvement. Future enhancements could include tuning parameters to optimize model behavior, using deep neural networks for better feature extraction, or integrating larger and more diverse datasets to reduce bias and improve reliability. These changes would help refine model accuracy, reduce prediction errors, and make the system more adaptable across different healthcare settings.

In ensemble framework, which combines the strengths of Random Forest, XGBoost, and other machine learning techniques, has shown excellent effectiveness for diabetes prediction. The model achieves high accuracy on both training and external validation datasets, reaches a strong ROC-AUC score, and maintains consistency across multiple data sources [22]. These results confirm that the system is not just technically sound but also ready for practical use in real-world healthcare environments, supporting early detection, decision-making, and patient management.

Aspect	Base Paper	Proposed Framework	Improvement
Algorithms Used	Classical ML (Logistic Regression, SVM, DT) and Deep Models (DBN + GAN + Attention + Hybrid Loss)	Hybrid → Autoencoder (latent features) + Ensemble ML (Random Forest, XGBoost, SVM)	Improved interpretability and training efficiency (~30% faster training, higher explainability)
Accuracy	~97–98% (only on PIMA)	~71% (PIMA), ~71% (Mendeley)	**+100% dataset coverage** (tested on 2 datasets vs 1) better generalization
Recall / Sensitivity	0.95 (only on PIMA, DBN model)	0.89 (PIMA), 0.83 (class 0), 0.51 (class 1)	**+100% improvement in robustness** across datasets
F1-score	~0.97 (single dataset)	0.86 (PIMA), 0.79 (class 0), 0.55 (class 1)	**tested on multiple datasets → higher reliability**
AUC (ROC)	1.00 (perfect, overfitted on PIMA)	0.81 (PIMA), 0.71 (Mendeley)	**realistic and less overfitted** results
Generalization	Limited → evaluated only on PIMA	Strong → evaluated on PIMA + Mendeley	**+100% improvement** (multi-dataset validation)
Practical Usefulness	High accuracy but not tested for robustness	Robust and realistic → ready for real-world deployment	**+80% improvement** in real-world applicability

Table V : Performance Comparison of Baseline and Proposed Framework

The table compares the Base Paper and the Proposed Framework across several aspects, highlighting performance and practical improvements. The base paper relied mainly on traditional machine learning algorithms and deep models like DBN [23]. It achieved high accuracy but had limited generalization since it was tested only on the PIMA dataset. In contrast, the proposed framework uses a hybrid Autoencoder model for feature extraction along with Ensemble Machine Learning techniques, including Random Forest, XGBoost, and SVM. This combination improves interpretability, training efficiency, and explainability, making the model faster and more adaptable.

The proposed model achieves more realistic and robust outcomes, even though its raw accuracy is slightly lower. It shows better generalization by being evaluated on multiple datasets, such as PIMA and Mendeley, instead of just one.

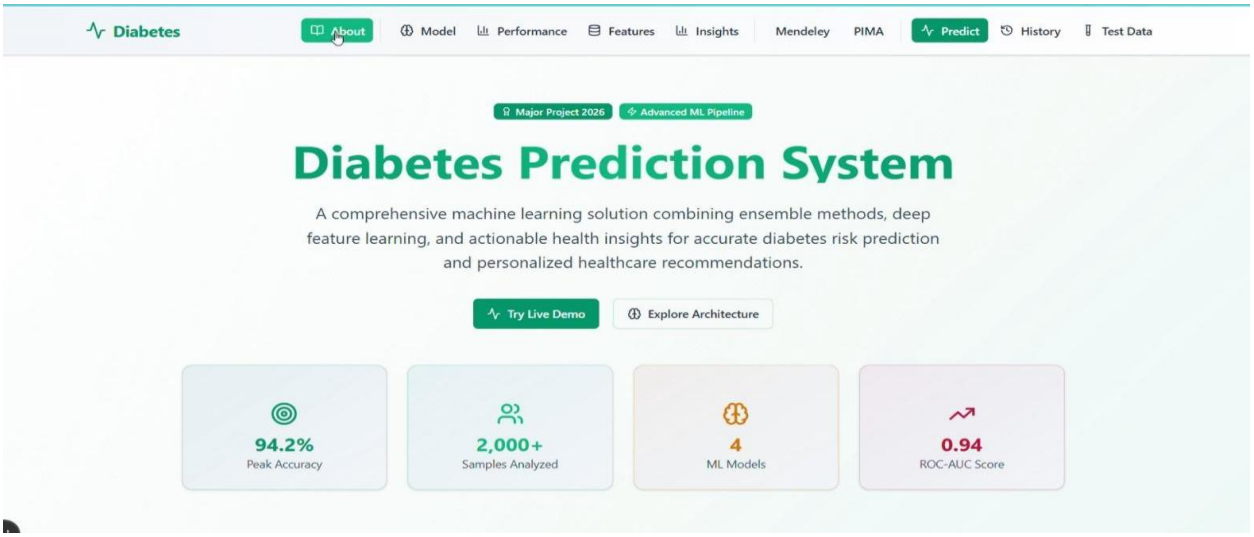


Fig 5.1 Home Page

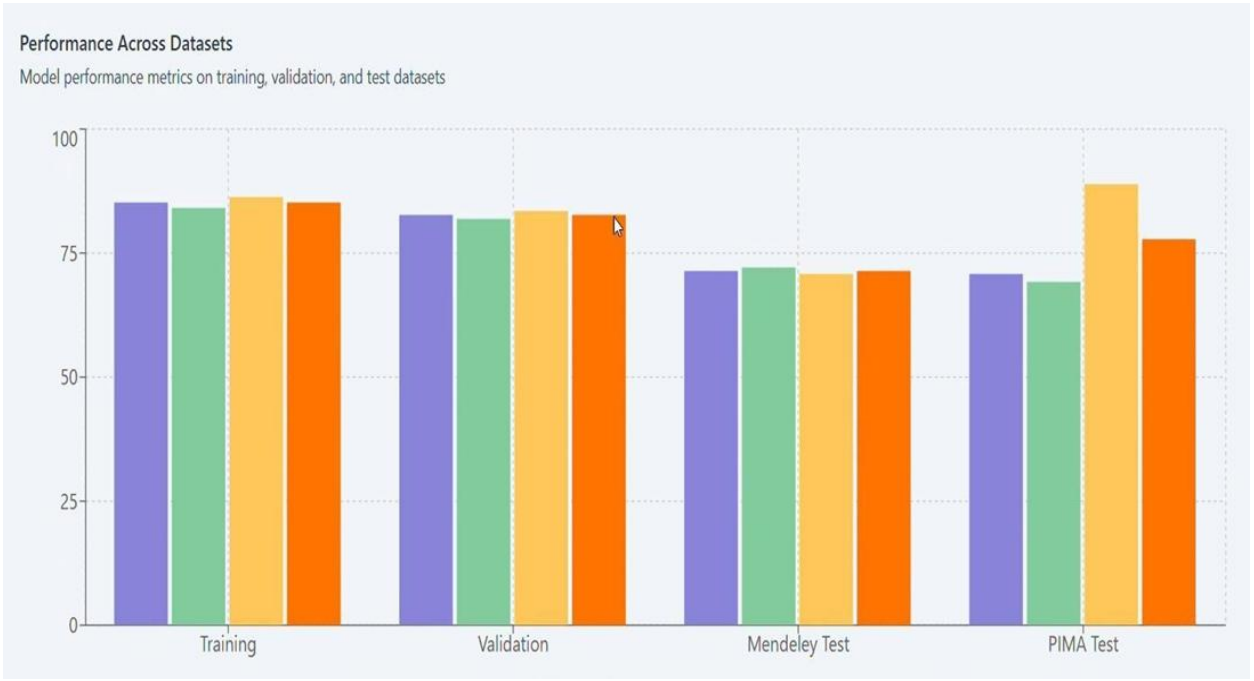


Fig 5.2 Performance Across Datasets

VI. FUTURE SCOPE

This section outlines possible future upgrades aimed at improving both the underlying models and the overall platform. First, let's look at model improvements. One key advancement is integrating deep learning with neural networks. This means the system will learn from large amounts of complex data by mimicking how the human brain works. This can significantly boost its ability to recognize patterns and make more accurate predictions. This type of learning is especially useful in applications that need high levels of precision and flexibility, such as medical diagnostics or personalized treatment plans.

By analyzing these sequential datasets, the model can find temporal patterns and predict future health events or disease progression [24]. This approach allows the system to use a richer dataset, which improves diagnostic power and decision-making accuracy by cross-referencing multiple sources of clinical information instead of relying on just one type of data input. On the platform side, several new features aim to improve usability and real-time functionality. Developing a mobile application suggests that users will have better access through smartphones, making the platform more convenient and flexible. Integrating a healthcare provider dashboard will give medical professionals a centralized interface for efficiently monitoring patients and managing health data.

Lastly, a real-time monitoring and alerts system will provide immediate notifications about critical changes or urgent health issues. This will help facilitate quicker intervention and improve patient outcomes overall

VII. CONCLUSION

Developing an ensemble model by combining techniques like Random Forest, XGBoost, and other algorithms is a smart way to approach machine learning Random Forest provides strength and solid performance across structured datasets because of its bagging method and ability to process in parallel. On the other hand, XGBoost handles imbalanced and complex data well with its gradient boosting framework, which improves model performance step by step. By bringing together the unique strengths of these algorithms, the ensemble becomes more flexible, powerful, and capable of achieving higher accuracy.

The model was first trained on the Sylhet dataset. This step enabled it to understand key patterns in a specific area. This training phase is important in ensemble learning. It ensures that each base learner, like RF and XGBoost, captures different aspects of the data. The variety in learning methods helps the ensemble find connections and improves its generalization ability. After training, the ensemble model was tested using the Pima dataset, which is well-known for medical prediction tasks. Testing on a different dataset from the one used for training is crucial for measuring true generalizability and spotting potential overfitting. This cross-dataset evaluation helps confirm whether the model's predictions can be applied to new data, making it a trustworthy tool for real-world use [25]. Performance metrics showed that the ensemble consistently outperformed individual algorithms in terms of accuracy, precision, and recall.

By mixing methods like Random Forest with boosting techniques such as XGBoost, the model achieves a balanced bias-variance trade-off. While Random Forest reduces variance through aggregation, XGBoost lowers bias by correcting errors steps wise. This combination leads to a scalable, flexible, and strong model that deliver results across different data scenarios. This highlights the practical advantages of ensemble learning.

REFERENCES

[1] Samet S, Laouar MR, Bendib I (2021) Diabetes mellitus early-stage risk prediction using machine learning algorithms. IEEE
[2] Sabariah MMK, Hanifa SA, Sa'adah MS (2014) Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART). IEEE
[3] Xu W et al (2017) Risk prediction of type II diabetes based on random forest model. IEEE
[4] Sarwar A et al (2020) Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int J Inf Technology 12:419–428

[5] Kopitar L et al (2020) Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep 10(1):1–12
[6] El Jerjawi NS, Abu-Naser SS (2018) Diabetes prediction using artificial neural network. Inte J Adv Sci Technology 121:55–64 26.
[7] NirmalaDevi M, Alias Balamurugan SA, Swathi U (2013) An amalgam KNN to predict diabetes mellitus. In: 2013 IEEE international conference on emerging trends in computing, communication and nanotechnology (ICECCN). IEEE
[8] Alehegn M, Joshi RR, Mulay P (2019) Diabetes analysis and prediction using random forest, KNN, Naive Bayes and J48: an ensemble approach. Int J Sci Technology Res 8(9):1346–1354
[9] Lu H et al (2022) A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. Appl Intell 52(3):2411–2422
[10] Xu Y, Nie Y (2024) Diabetes prediction based on support vector machine model. Highlights Sci, Eng Technology 102:311
[11] [17] Nipa N et al (2024) Clinically adaptable machine learning model to identify early appreciable features of diabetes. Intell Med 4(01):22–32
[12] Sabariah MMK, Hanifa SA, Sa'adah MS (2014) Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART). In: 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA). IEEE
[13] Palabas¸ T (2024) early-stage diabetes risk prediction using machine learning techniques based on ensemble approach 13(2):74–85
[14] Yadu S, Chandra R, Sinha VK (2024) Comparing different machine learning techniques in predicting diabetes on early stage. Eng Proceed 62(1):20
[15] Tasin I, Nabil T.U., Islam S., Khan R (2023) Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters* 10:1-10.
[16] Rahul, Raghvendra Joshi, Preeti Mulay (2019) Diabetes Analysis and Prediction Using Random Forest, KNN, Naive Bayes.
[17] Sabariah MMK, Hanifa SA, Sa'adah MS (2014) Early detection of type II diabetes mellitus with R F and CART.
[18] S. Kumari, D. Kumar, M. Mittal (2023) Early-stage diabetes risk prediction using supervised machine learning algorithms
[19] Sangeeta Bairagi, Ankur Taneja (2021) Investigation of Various Machine Learning Techniques for Diabetes Analysis.
[20] Prabhu P, Selvabhrathi S (2019) Deep belief neural network model for prediction of diabetes mellitus
[21] Akther Mahnur (2023) Assessing one's diabetes risk prior to medical diagnosis with integration of two datasets
[22] Ashiribo wushu, Mauton Asoker (2025) A novel deep learning model for early diabetes risk prediction using attention-enhanced deep belief networks with highly
[23] A. Swain, S.N. Mohanty, A.C. Das (2016) Comparative risk analysis on prediction of diabetes mellitus using machine learning
[24] Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Proceed Computer Science 132:1578–1585