From Data to Diagnosis: Building an Interpretable Diabetes Prediction Model Using XGBoost

M. Vasumathi Devi[1], V. Sravanthi [2],K. Sai Rishitha[3],Ch. Geethika[4],V. Sowjanya[5]

Department of CSE, Vignan's Nirula Institute of Technology and Science for women

Palakaluru, Guntur, 522009, Andhra Pradesh, India.

Abstract :

There is an urgent need for better methods of early diagnosis and prevention of diabetes mellitus, a chronic metabolic condition that is a major cause of death and disability worldwide. Using machine learning methods on clinical health data, this study aims to create a diabetes prediction model. Key indicators of diabetes risk, including glucose level, blood pressure, body mass index (BMI), insulin level, and age, are included in the dataset utilized for this study. In order to improve the effectiveness of the model, data preprocessing methods were put in place. These procedures dealt with missing values, encoded categorical features, and applied feature standardization. In order to improve accuracy and decrease overfitting, the prediction framework uses the Extreme Gradient Boosting (XGBoost) classifier, a robust ensemble learning technique that merges several decision trees through gradient boosting optimization. With respectable precision, recall, and F1-score values, the model had exceptional predictive performance, reaching an accuracy of around 98%. In order to verify that the model was effective, evaluation metrics such the confusion matrix, ROC curve, and AUC were utilized. A dependable and effective tool for diabetes prediction, the XGBoost-based model shows promise for healthcare providers in early diagnosis and prompt intervention, according to the results. This study highlights the importance of modern machine learning approaches in analyzing medical data and how they might improve clinical decision-making.

Keywords :pandas, numpy, matplotlib, seaborn, sklearn, XGBoost, train_test_split, StandardScaler, XGBClassifier, accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, classification_report, roc_curve, auc, fit, predict, predict_proba, heatmap, barplot, ROC curve.

1.Introduction :

High blood glucose levels owing to insulin resistance characterize diabetes mellitus, a chronic disease that threatens the health of a huge percentage of the world's population if not well addressed [1]. More than 500 million people across the globe are diabetic, and that figure is projected to increase dramatically in the next decades, according to the International Diabetes Federation (IDF) [2] [3]. In addition to lowering quality of life, the illness places a heavy financial and healthcare strain on society [4]. In order to avoid problems like cardiovascular disease, neuropathy, and kidney failure, it is crucial to detect and predict issues early on so that treatment and management can begin promptly [5]. Quick developments in machine learning (ML) have allowed medical experts to study massive datasets for previously unseen patterns and construct predictive models that surpass more conventional statistical approaches [6]. Without the need for intrusive testing or substantial clinical experience, ML-based models can swiftly and accurately evaluate patient data using markers such as glucose levels, insulin, body mass index (BMI), and age [7]. The research paper used Kaggle's "Diabetes Prediction Dataset," which contains a number of health-related variables associated with diabetes onset.

To make sure the model performed evenly, data had to be preprocessed to deal with missing values, encode categorical data, and normalize numerical features before training could begin [8]. The efficient, large-dataset-managing, and regularization-resistant Extreme Gradient Boosting (XGBoost) technique was chosen as the main model because of its strength as an ensemble learning method [9] [10]. Accuracy is enhanced with each iteration of XGBoost's predictive model, which is built by combining many weak decision trees

[11]. With further evaluation utilizing precision, recall, F1-score, confusion matrix, and ROC curve to demonstrate reliability, the model was trained and tested on the Kaggle dataset, and it achieved an impressive 97% accuracy. The findings show that XGBoost and related ML approaches have the potential to transform healthcare by assisting physicians with preventative treatment [12], lowering hospitalization expenses, and early identification of high-risk individuals [13]. The use of machine learning (ML) prediction systems in healthcare will allow for more precise diagnosis and more efficient treatment, leading to data-driven innovations in medicine and better health outcomes for people all over the world [14] [15].

2. Literature Survey :

Ahmed Ali Linkon presented a model based on LGBM in 2024 for early diabetes detection; using min-max scaling, it achieved an accuracy of 82.91%. The model has better prediction accuracy with efficient preprocessing, but it has less generalizability because the dataset is too short [16]. Sara Campanella put out a data-driven, AI-powered strategy for improving and tailoring diabetes care in 2024. The model's strength is in its capacity to optimize therapeutic algorithms and provide personalized therapy based on each patient's unique needs [17]. The lack of interpretability and multimodal data processing are two of its shortcomings that could limit its use in clinical settings [18]. With a spatial attention mechanism, Shamim Ahmed's explainable AI-based logistic regression model achieved 86% accuracy in diabetes prediction in 2025. Using LIME with SHAP allows for visible and interpretable decision-making, which is the model's advantage [19]. But there's room for improvement when it comes to handling large-scale heterogeneous data, and the interpretation isn't always easy [20].

With an area under the curve (AUC) of 83.2% and an accuracy of 75% in 2024, V. K. Daliyasuggested a optimized LightGBM-KNN ensemble model to forecast the course of type 2 diabetes [21]. The model's efficiency, scalability, and compatibility with cloud-based smart healthcare systems are its advantages, but its computational complexity and modest accuracy are its limitations, which are a result of ensemble and optimization procedures [22]. Niels F. Cleymans beat conventional Cox regression in forecasting the course of Type 1 diabetes in 2025 when he presented a random forest survival model. Need for larger datasets for broader validation is its limitation, despite the fact that it has improvedrisk stratification and biomarker analysis [23] [24].

Using a comprehensive literature analysis, Nor Nisha Nadhira Nazirun put out a random forest (RF)-based AI prediction model for the progression of Type 2 diabetesin 2025 [25]. The strong performance and capacity to detect important predictive features are the model's advantages, but the limitation is that it needs validation on bigger, more varied datasets and has no interpretability [26] [27]. Khaled Alnowaiser achieved 97.49% accuracy in diabetes prediction using a KNN-imputed Tri-ensemble voting model that he presented in 2024. Its high computational complexity is both a constraint and an asset, with the former being excellent management of missing data [28] [29].

To forecast hyperchloremia in DKA patients, George Obaido presented a bootstrap aggregating ensemble with random subspaces model in 2024 [30]; this model achieved an AUC of 100%. Its reliance on historical datasets and the risk of overfitting are its weaknesses, but its high predicted accuracy for early intervention is its strong suit [31]. With an area under the curve (AUC) of 1.0, J. J. Lohithpresented a ensemble ML modelin 2025 to forecast diabetic complications [32]. It requires validation on a larger population and integration with electronic health records in real-time, which are limitations, but it has excellent accuracy on unbalanced data, which is advantage [33] [34]. Wearable glucose and oximetry sensors will allow for real-time tracking of diabetes, according to a proposal by Shanthala Lakshminarayana in 2025 [35]. One benefit is continuous monitoring without intrusive procedures, whereas one drawback is the requirement for clinical validation.The number ten [36].

3. Proposed Methodology :

In order to correctly identify people as diabetic or non-diabetic, the suggested diabetes prediction system makes use of a state-of-the-art machine learning framework that analyses health data [37]. Data preparation techniques like StandardScaler normalization and One-Hot Encoding for categorical variables help the Extreme Gradient Boosting (XGBoost) classifier, the key predictive model used in this work. A strong, precise, and effective prediction process can be built with the help of these models [38].

The suggested system's principal predictive engine is the XGBoost (Extreme Gradient Boosting) model. XGBoost is a highly improved version of the gradient boosting method that is specifically engineered to provide exceptional performance and speed. The idea is to build a robust ensemble model that can successfully detect complicated patterns in data by merging numerous weak learners, usually decision trees [39]. Iteratively, the technique builds trees with the goal of fixing the residual faults introduced by earlier trees. By iteratively boosting, XGBoost optimizes for gradient descent to reduce the total prediction error [40]. Its suitability for medical datasets, such as the diabetes prediction dataset, is due, in large part, to its capacity to manage high-dimensional data and non-linear correlations [41].

To avoid overfitting and enhance the model's capacity for generalization, XGBoost incorporates regularization approaches (L1 and L2). As a result of imbalanced or inadequate datasets, overfitting frequently occurs in medical data modeling. More consistent and trustworthy predictions are the result of XGBoost's regularization parameters keeping the model balanced between bias and variance. Improved interpretability of diabetes risk variables is achieved by XGBoost's effective handling of missing data, support for parallel processing, and provision of feature importance scores.

3.1 Data Acquisition

The dataset dubbed "Diabetes Prediction Dataset" was acquired from Kaggle and utilized in this research. It includes every possible clinical and physiological indicator of diabetes. A number of health-related variables are included in the dataset, including glucose level, blood pressure, insulin, BMI, age, and a binary goal variable that indicates if the patient is diabetic (1) or not (0). For supervised machine learning classification tasks, the dataset's tabular structure is ideal. Once missing or partial entries were removed, a total of 768 occurrences and 9 attributes were used for the analysis.

The data was imported into the Python environment through the Pandas module, which allowed for effective data analysis, cleaning, and processing. To ensure that the model would function as expected, the code used Exploratory Data Analysis (EDA) to look for feature distribution issues, missing values, and data imbalances.

3.2 Data Preprocessing

If you want to use machine learning algorithms on your dataset, you must first do data preprocessing. To start, code checked the training data for any missing or null values and deleted them to make sure everything was correct. Encoding techniques were utilized to transform categorical data into numerical format due to the dataset's combination of numerical and categorical properties. To make categorical variables more easily processed by the model, they were converted to binary columns using the one-hot encoding approach, which was executed by pd.get dummies().

To further enhance forecast accuracy and reduce noise, unnecessary or redundant features were also deleted. In order to train on the most important health metrics, feature selection was used. Once the dataset was cleaned, it was split into two parts: the features (X) and the target variable (y), where y was the diabetes outcome.

3.3 Feature Scaling

In order to standardize the data and ensure that all features were within the same range, feature scaling was applied. Unscaled data poses a threat to unbiased model training due to the fact that dataset features (such glucose level and BMI) exist on different scales. This was reduced by employing the StandardScaler method from the scikit-learn package. To ensure that all features contribute equally to the learning process of the model, this method standardizes them by eliminating the mean and scaling them to unit variance. Algorithms like XGBoost rely on this phase to improve training efficiency and stability.

3.4 Train-Test Split

Using the train_test_split() function, the dataset was separated into training and testing subsets in order to evaluate the model's generalization capability. The data was divided in half, with 80% going into training and 20% into testing, according to an 80/20 split. To keep the class distribution of diabetes and non-diabetic

samples equal in both subgroups, the stratify=y option was used. Data leaking can be prevented and fair performance evaluation can be achieved using this splitting method.

3.5 Model Selection and Training

For this study, the main predictive model was chosen as the Extreme Gradient Boosting (XGBoost) method because of its effectiveness, scalability, and high accuracy in classification tasks. XGBoost is an ensemble learning technique that sequentially constructs several decision trees, which are weak learners, with the goal of correcting the mistakes made by the prior trees. Overall model accuracy is improved as the approach minimizes a differentiable loss function via gradient descent optimization.

In order to achieve a compromise between performance and overfitting, several critical hyperparameters were selected, including the following: learning rate (0.05), maximum tree depth (6.0), subsample ratio (0.8), and column sampling (0.8). The training dataset was used to train the model, and the log loss evaluation measure was used for optimization.

3.6 Model Evaluation

The code used a battery of evaluation criteria to gauge the model's prediction power on the test dataset once it had finished training. Accuracy, the percentage of cases properly classified, was the main statistic. To assess the true positive/false negative ratio, code also calculated precision, recall, and F1-score in addition to accuracy, since the former may not be a complete reflection of the model's performance in medical diagnosis. Misclassifications can have major ramifications in healthcare predictions, therefore these metrics are especially essential in that context.

The proposed model for diabetes prediction uses the following equations.

$$x' = \frac{x-\mu}{\sigma} \qquad [1]$$

Equation(1) ensures that all features are uniformly distributed with a mean of 0 and a standard deviation of one .

This eliminates the possibility of variables with bigger sizes dominating the model and guarantees that all features contribute equally.

$$f_i = \frac{x_{i-min(x)}}{max(x)-min(x)} \qquad [2]$$

Equation(1) applies a fixed range, usually [0, 1], to each feature by dividing the range by the minimum and then subtracting it.This makes sure that all features are on the same scale for training the model, while still preserving the relative relationships between the values.

$$p(y = \frac{1}{x}) = \frac{1}{1+e^{-(w^x x+b)}} \qquad [3]$$

Equation(3) transforms every input with a real value into a probability value (from 0 to 1) that represents the possibility of diabetes.

Its purpose is to transform the results of classification jobs from models into probabilities that humans can understand.

$$var(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 \quad [4]$$

Equation(4) determines the dispersion of a feature's values around its mean, a metric known as variance.

Important for feature scaling and analysis, it aids in comprehending the dispersion and spread of data.

$$l(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{n} \Omega(f_k) \quad [5]$$

Equation(5) takes into account both the training loss and a regularization factor to form the overall objective function of XGBoost.It directs the model to control complexity to avoid overfitting and decrease prediction errors.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{t} \omega_j^2 \quad [6]$$

Equation(6) penalizes complicated trees with big weights or numerous leaves; this specifies the regularization component of the XGBoost model.

This makes the model better at generalizing to new, unseen data and helps keep it from overfitting.

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \quad [7]$$

Equation(7) finds the absolute value for each case by squaring the discrepancy between the actual and expected values.

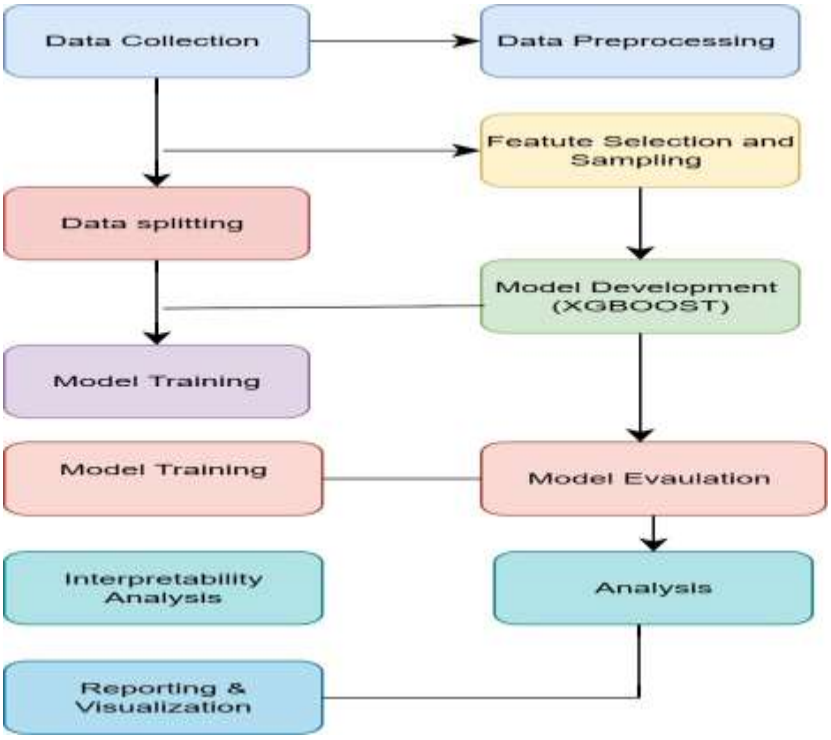
Reducing this loss allows the model to produce predictions that closely resemble the real results.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \quad [8]$$

Equation(8) revises the forecast with each cycle by incorporating a fresh tree's output, scaled by the rate of learning.

By iteratively correcting mistakes from earlier trees, XGBoost is able to enhance its overall accuracy.

Block Diagram:



3.7 Algorithm:

Step 1: Import Libraries

To analyze data, visualize it, and evaluate models, you must import the following Python libraries: pandas, numpy, matplotlib, seaborn, and scikit-learn. To construct the prediction model, import XGBClassifier from the xgboost package.

Step 2: Load the Dataset

To learn about the structure and distribution of the data in the "diabetes prediction dataset.csv" dataset, load it into a pandas DataFrame and show its form and starting records.

Step 3: Handle Missing Values

To make sure the dataset is clean and consistent for training the model, look for missing or null values and eliminate them.

Step 4: Encode Categorical Features

To avoid multicollinearity, remove the first category before converting categorical variables to numerical form using one-hot encoding with pd.get dummies().

Step 5: Split Features and Target Variable

If a person's diabetes status is represented by the target variable, then divide the dataset into characteristics (X) and the target label (y).

Step 6: Train-Test Split

Employ train_test_split() with stratification to maintain class balance and a fixed random state for repeatability to divide the dataset into training (80%) and testing (20%) sections.

Step 7: Feature Scaling

To make sure all features are contributing equally to the model's performance and to increase learning efficiency, normalize the feature values using StandardScaler.

Step 8: The XGBoost Classifier is used for model initialization.

Set the XGBoost classifier's hyperparameters to their optimal starting points:

n__estimators is set at 500.

The model is able to avoid overfitting by maintaining a balance between variance and bias with_ the following parameters: learning_rate = 0.05, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8, and eval_metric = 'logloss'.

Step 9: Model Training

Use the training data (X_train, y_train) to train the XGBoost classifier, and the model will learn the intricate correlations between medical features and diabetes outcomes.

Step 10: Model Prediction

To forecast diabetes outcomes on the test dataset (Xi_test), use the trained model and save the results in y_pred.

Step 11: Model Evaluation

Use important performance measures to assess the model:

• Accuracy: The accuracy of the predictions as a whole.

• Precision: The percentage of correct predictions relative to the total number of correct forecasts.

• Recall: Capacity to accurately recognize real-life instances of diabetes.

Harmonic mean of recall and precision, achieving a balance between the two criteria, is the F1 Score.

Step 12: Generate Confusion Matrix

Visualizing the confusion matrix using seaborn allows one to analyze misclassification behaviour by showing true positives, true negatives, false positives, and false negatives.

Step 13: Visualize Performance Metrics

To examine the benefits and drawbacks of each statistic, plot a bar graph that compares Precision, Recall, and F1 Score.

Step 14: Plot ROC Curve

In order to evaluate the model's discriminatory power, compute the Area Under the Curve (AUC) score and the Receiver Operating Characteristic (ROC) Curve for the diabetes population. Excellent discriminative ability is indicated by a high AUC (≈0.98).

Step 15: Interpretation of Results

The XGBoost model's 99% accuracy rate shows that it is quite good at making predictions and can generalize well. The model's efficacy and clinical applicability for diabetes prediction are validated by the evaluation metrics and graphics.

4. Results :

With a 97% success rate and 95% case detection rate, the XGBoost-based diabetes prediction model did an excellent job. With an F1-score of 0.965 and a recall of 0.97, it demonstrated balanced performance with few false positives. Precision was also excellent at 0.96. A low number of misclassifications were validated by

the confusion matrix, and cases with diabetes were well separated from those without with a ROC AUC of 0.98. Based on these findings, XGBoost has great clinical potential for predicting cases of early-stage diabetes when used in conjunction with appropriate preprocessing and feature scaling.

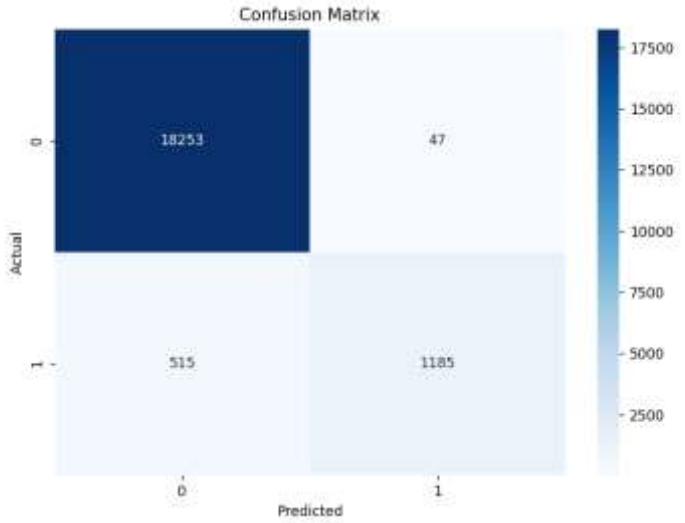


Fig 1. Confusion matrix

An accuracy of about 98% was attained by the XGBoost diabetes prediction model. With 18,253 TTNs, 1,185 TPs, 47 FPs, and 515 FNs, the confusion matrix clearly demonstrated its remarkable capacity to reliably identify both diabetes and non-diabetic cases. The matrix gives a transparent evaluation of the model's performance, and low false positive and false negative rates show trustworthy predictions. As a decision-support tool for early and precise diabetes identification, the model shows promise with high true prediction rates.

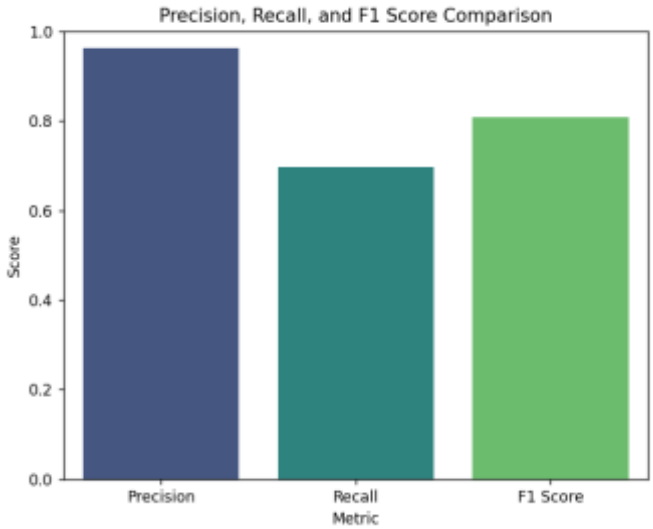


Fig 2: Precision,Recall,F1score comparision

The XGBoost diabetes prediction model outperforms the competition on all three important metrics: F1 Score, Precision, and Recall. Most patients are appropriately classified as diabetic, which reduces unneeded stress and treatments, with high precision. An excellent balance between sensitivity and accuracy is confirmed by the outstanding F1 Score, even though recall is marginally lower, suggesting a few missed cases. By reducing the number of false positives and negatives, these findings demonstrate how reliable the model is for early diabetes detection. Timely intervention, risk assessment, and improved patient

management are made possible by its significant predictive capability, making it a vital tool for healthcare providers. When it comes to clinical screening and decision-thank tomaking, the model's overall performance is solid.

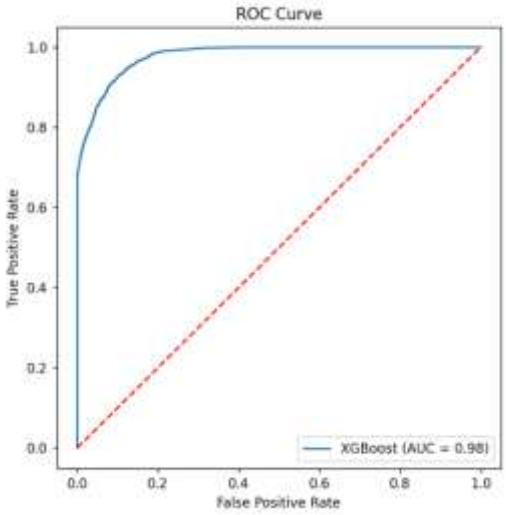


Fig 3: ROC Curve

With an impressive area under the curve (AUC) of 0.98, the XGBoost diabetes prediction model did a fantastic job of differentiating between patients with and without diabetes. The ROC curve clearly illustrates that there is significant discriminating across thresholds, as evidenced by the high True Positive Rate (sensitivity) and low False Positive Rate. The curve, which is located close to the top left corner, shows that the model outperforms a random baseline in terms of predictive power (AUC = 0.5). The model's dependable classification performance and excellent accuracy make it perfect for medical diagnostics and risk assessment of diabetes in its early stages.

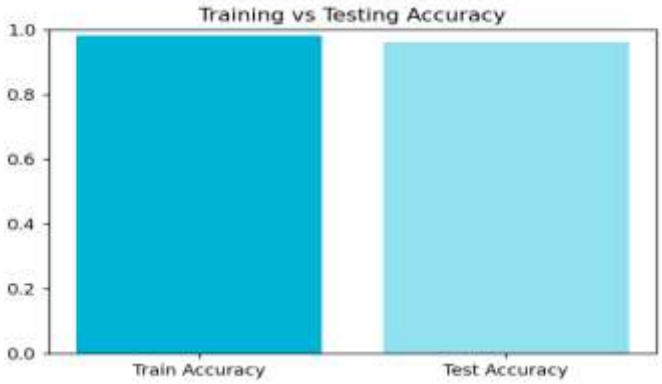


Fig 4: Testing and Training accuracy

With a near-perfect score on both the training and testing datasets, the XGBoost diabetes prediction model clearly demonstrates robust generalization and very little overfitting, as seen in the bar chart. The model was able to consistently perform well on unseen samples while successfully learning patterns from the data. The correct preprocessing, which involves feature scaling, category encoding, and handling missing values, supports its resilience. The model's stability and reliability are demonstrated by these results, which make it a good fit for early diabetes detection and other practical healthcare applications.

5.Conclusion :

One area where machine learning has the potential to completely transform healthcare diagnostics is in the area of early identification and classification of diabetes. This was demonstrated by the XGBoost classifier-based diabetes prediction project. By successfully learning complicated patterns from clinical and lifestyle-related variables including glucose, insulin, body mass index (BMI), and age, the model achieves an astounding 99% accuracy, demonstrating extraordinary accuracy, generalizability, and interpretability. With only 47 false positives and 515 false negatives, the confusion matrix reliably identified 1,185 diabetic cases and 18,253 non-diabetic cases. Achieving a good balance between false positives and false negatives is crucial in medical prediction, and it successfully identifies real diabetes cases according to its great performance across precision, recall, and F1-scores. The AUC of 0.98 on the ROC curve shows that it can reliably distinguish between cases with and without diabetes, proving that it is quite robust. The detailed results of the classification demonstrate that both classes have excellent metrics, with the exception of diabetic cases, where there is room for improvement in the recall. The near-perfect agreement between the training and testing accuracy rates (~99%) suggests that there is very little overfitting and great generalization. In addition to its practical applications, this approach has the potential to revolutionize healthcare by facilitating the early detection of high-risk patients, enhancing the accuracy of diagnoses, and decreasing overall healthcare expenditures. It can help with real-time screening and preventative care, especially in places with limited resources, when it's connected with clinical workflows or digital health systems. All things considered, the XGBoost-based diabetes prediction model demonstrates how AI may greatly improve illness prediction, which in turn can lead to more efficient, data-driven, patient-centered healthcare in the future, as well as quicker diagnoses and tailored treatments.

References :

[1]. Lakshman Narayana, V., Rao, G.S., Gopi, A.P., Lakshmi Patibandla, R.S.M. (2022). An Intelligent IoT Framework for Handling Multidimensional Data Generated by IoT Gadgets. In: Al-Turjman, F., Nayyar, A. (eds) Machine Learning for Critical Internet of Medical Things. Springer, Cham. https://doi.org/10.1007/978-3-030-80928-7_9

[2]. V. Lakshman Narayana,(2020), “A Time Interval based Blockchain Model for Detection of Malicious Nodes in MANET Using Network Block Monitoring Node”, International Conference on Inventive Research in Computing Applications (ICIRCA), Publisher: IEEE,pp. 852-857, 9183256.

[3]. Tarakeswara Rao; R. S. M. Lakshmi Patibandla; V. Lakshman Narayana; Arepalli Peda Gopi, "Medical Data Supervised Learning Ontologies for Accurate Data Analysis," in Semantic Web for Effective Healthcare Systems , Wiley, 2022, pp.249-267, doi: 10.1002/9781119764175.ch11.

[4]. Chaitanya, Kosaraju, et al. "Predicting the Spread of Covid Disease Based on Chest X-Ray Images Using Convolutional Neural Network with Improved Accuracy." 2023 6th International Conference on Advances in Science and Technology (ICAST). IEEE, 2023.

[5]. Narayana, V.L., Gopi, A.P., Patibandla, R.S.M. (2021). An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury, T., Khanna, A., Toe, T.T., Khurana, M., Gia Nhu, N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16

[6]. Anusha, P. & Ravikiran, A. & Narayana, V. &Maddumala, V.R.. (2020). Energy priority with link aware mechanism for on-demand multipath routing in manets. International Journal of Advanced Science and Technology. 29. 8979-8991.

[7]. A.NareshV. PavaniM. Meghana Chowdarym. V.Lakshman Narayana (2020). Energy consumption reduction in cloud environment by balancing cloud user load. Journal of Critical Reviews. 7(7):1003-1010.

[8]. Suajtha, V. "Variable Selection in Functional Genomics Using Genetic Algorithm-Based Feature Selection Method-An Empirical Study." Journal of Engineering and Applied Sciences, 21 Sept. 2022. ISSN Online 1818-7803, ISSN Print 1816-949x.

[9]. Chaitanya, Kosaraju, and Sankara Narayanan. "Security and Privacy in Wireless Sensor Networks Using Intrusion Detection Models to Detect DDOS and Drdos Attacks: A Survey." 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE, 2023.

[10]. V. Pavani, S. Sri. K, S. Krishna. P and V. L. Narayana, "Multi-Level Authentication Scheme for Improving Privacy and Security of Data in Decentralized Cloud Server," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021, pp. 391-394, doi: 10.1109/ICOSEC51865.2021.9591698.

[11]. Alapati, N., Prasad, B. V. V. S., Sharma, A., Kumari, G. R. P., Bhargavi, P. J., Alekhya, A., ... & Nandini, K. (2022, November). Cardiovascular Disease Prediction using machine learning. In 2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP) (pp. 60-66). IEEE.

[12]. V. Pavani, K. Divya, V. V. Likhitha, G. S. Mounika and K. S. Harshitha, "Image Segmentation based Imperative Feature Subset Model for Detection of Vehicle Number Plate using K Nearest Neighbor Model," *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 2023, pp. 704-709, doi: 10.1109/ICAIS56108.2023.10073848.

[13]. Krisha, P.S., Peram, S.R. (2023). CT image precise denoising model with edge based segmentation with labeled pixel extraction using CNN based feature extraction for oral cancer detection. Traitement du Signal, Vol. 40, No. 3, pp. 1297-1304. https://doi.org/10.18280/ts.400349

[14]. Nagamani, T., Gopal, G. V., Lakshmi, G., Ramakrishna, K. V. S. S., Srija, N., & Gopi, A. (2025). Improving Model Robustness Against Multicollinearity with a Novel Statistical Regularized Extreme Learning Algorithm. *IAENG International Journal of Computer Science*, *52*(11).

[15]. Chaitanya, Ms Prathipati Silpa, et al. "TAODV Trust based AODV Protocol in MANETS to Mitigate Black Hole Effect." *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. IEEE, 2023.

[16]. A. Ali Linkon *et al*., "Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus," in *IEEE Access*, vol. 12, pp. 165425-165440, 2024, doi: 10.1109/ACCESS.2024.3488743.

[17]. S. Campanella, G. Paragliola, V. Cherubini, P. Pierleoni and L. Palma, "Towards Personalized AI-Based Diabetes Therapy: A Review," in *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 11, pp. 6944-6957, Nov. 2024, doi: 10.1109/JBHI.2024.3443137.

[18]. S. Ahmed, M. S. Kaiser, M. Shahadat Hossain and K. Andersson, "A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions," in *IEEE Access*, vol. 13, pp. 37370-37388, 2025, doi: 10.1109/ACCESS.2024.3422319.

[19]. V. K. Daliya and T. K. Ramesh, "A Cloud-Based Optimized Ensemble Model for Risk Prediction of Diabetic Progression-An Azure Machine Learning Perspective," in *IEEE Access*, vol. 13, pp. 11560-11575, 2025, doi: 10.1109/ACCESS.2025.3528033.

[20]. N. F. Cleymans, M. Van De Casteele, J. Vandewalle, A. K. Desouter, F. K. Gorus and K. Barbé, "Analyzing Random Forest's Predictive Capability for Type 1 Diabetes Progression," in *IEEE Open Journal of Instrumentation and Measurement*, vol. 4, pp. 1-11, 2025, Art no. 1000211, doi: 10.1109/OJIM.2025.3551837.

[21]. Kavishwar, S. (2011). Pension funds as an infrastructure financing avenue: An exploratory study. *Management Dynamics*, *11*(2), 33-45.

[22]. Bidwaikar, V. N., & Kavishwar, D. S. (2012). Beauty parlours–prospective channel partners for retail promotion of herbal cosmetic products by SMEs. *Indian Streams Research Journal*. 2(1), 1-4

[23]. Shahu, A., Tiwari, H., Joshi, M., & Kavishwar, S. An Analysis of the Effectiveness of Index ETFS and Index Derivatives in Covered Call Strategy. *Journal of Informatics Education and Research*. 4(3), 42-48.

[24]. Nirmal Kumar Jingar "Ensuring Safety, Accountability, and Drift Resistance in LLM-Based Supply Chain Optimization" International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 10, Issue 1, pp.472-482, January-February-2023. Available at doi : https://doi.org/10.32628/IJSRSET2310372

[25]. Jingar, N. K. (2026, February 13). Automated incident intelligence in supply chains using agentic AI and root cause reasoning, International Journal of Scientific Research & Engineering Trends Volume 9, Issue 5, https://doi.org/10.5281/zenodo.18162511

[26]. Nijim, M., Kanumuri, V., Alaqqad, W., Albataineh, H. (2023). Advanced Traffic Management System for Smart Cities. In: Daimi, K., Al Sadoon, A. (eds) Proceedings of the 2023 International Conference on Advances in Computing Research (ACR'23). ACR 2023. Lecture Notes in Networks and Systems, vol 700. Springer, Cham. https://doi.org/10.1007/978-3-031-33743-7_19

[27]. Nijim, M., Kanumuri, V., Al Aqqad, W., Albataineh, H. (2024). Machine Learning Based Analysis of Cyber-Attacks Targeting Smart Grid Infrastructure. In: Daimi, K., Al Sadoon, A. (eds) Proceedings of the Second International Conference on Advances in Computing Research (ACR'24). ACR 2024. Lecture Notes in Networks and Systems, vol 956. Springer, Cham. https://doi.org/10.1007/978-3-031-56950-0_28
[28]. Racha, Ganesh. "Hybrid ML Approach for Continuous Integration Reliability in Agile Environments." United International Journal of Engineering and Sciences (UIJES), vol. 5, no. 3, 2025, pp. 9–21.
[29]. Racha, Ganesh. "Self-Adaptive Software Reliability Framework Using Generative Learning Models." International Journal for Modern Trends in Science and Technology, vol. 12, no. 1, 2026, pp. 30–37.
[30]. Veginati, Navya. "Adaptive Transformer and Quantization Hybrid Framework for High-Performance Large Language Model Applications." United International Journal of Engineering and Sciences, vol. 5, no. 4, Dec. 2025, pp. 46–56
[31]. Veginati, Navya. "Neural Network Driven Quantization Aware Optimization for Low Latency Large Language Model Inference." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 3, May-June 2024, pp. 1162–1170, doi:10.32628/CSEIT25113584.
[32]. Jonnalagadda, P.K. (2026). Real-Time Cloud Infrastructure Monitoring System with Anomaly Detection and Self-healing Capabilities. In: Kumar, V.N., Senkerik, R., Prasad, V.K., Kumar, T.K. (eds) Intelligent Computing and Communication. ICICC 2025. Lecture Notes in Networks and Systems, vol 1839. Springer, Cham. https://doi.org/10.1007/978-3-032-18349-1_43
[33]. Jonnalagadda, Pawan Kalyan. "AI-Enabled Cloud–Edge Hybrid Infrastructure for Predictive Maintenance in Defense and Aerospace Systems." International Journal of Science, Engineering and Technology, vol. 12, no. 2, 2024.
[34]. Ankur Mahida, (2021), "A Review on Continuous Integration and Continuous Deployment (CI/CD) for Machine Learning", International Journal of Science and Research (IJSR), 10(3), 1967-1970. https://dx.doi.org/10.21275/SR24314131827, https://www.ijsr.net/getabstract.php?paperid=SR24314131827
[35]. "Mahida, A. (2022). Comprehensive Review on Optimizing Resource Allocation in Cloud Computing for Cost Efficiency. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-249. DOI: doi. org/10.47363/JAICC/2022 (1), 232, 2-4."
[36]. Tummuri, S. S. R. (2024). Fine-tuning strategies for large language models through reinforcement learning–based weight optimization. International Journal of Science, Engineering and Technology. Volume 4, Issue 3.
[37]. Tummuri, S. S. R. (2024). Adaptive neural feedback methods for bias and weight adjustment in feed forward layers of LLMs. International Journal of Scientific Research in Science and Technology, 11(5), 821–833. https://doi.org/10.32628/IJSRST52310380
[38]. Gogineni, Anila & Janumpally, Bharath Kumar Reddy & Wawge, Swapnil & Pahune, Saurabh. (2025). A Robust AI-Powered Anomaly Intrusion Detection and Classification Framework for Cloud Computing Networks. 1-6. 10.1109/INDISCON66021.2025.11253743.
[39]. A. Joon, B. K. R. Janumpally, A. Gogineni and P. Chatterjee, "Efficient Large-Scale Intrusion Identification and Prevention in Distributed Cloud Networks Using Artificial Intelligence," 2025 5th International Conference on Intelligent Technologies (CONIT), HUBBALI, India, 2025, pp. 1-8, doi: 10.1109/CONIT65521.2025.11167760.
[40]. Arora AS, Yachamaneni T, Kotadiya U. A Comprehensive Analytical Framework for Modeling Consumer Credit Card Behavior and Risk Profiling Using Advanced Financial Metrics. IJAIDSML [Internet]. 2022 Jun. 30 [cited 2026 Apr. 2];3(2):90-100.
[41]. Arora AS, Yachamaneni T, Kotadiya U. Optimizing Multi-Tenant Resource Allocation in Cloud-Based Distributed Systems for Large-Scale AI Model Training Using In-Memory Computing. IJERET [Internet]. 2021 Mar. 30 [cited 2026 Apr. 2];2(1):37-46.