Improved Diabetes Prediction System Using Machine Learning

Upputholla Ruthwika Shivani[1], Dr. U. Sathish Kumar[2], Gaddam Therissa[3] , Kojja Vamsi[4], Ganta Anand[5]

Assistant professor[2], Department of CSE, University College of Engineering & Technology, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, AP, INDIA

UG Students[1,3,4,5], Department of CSE, University College of Engineering & Technology, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, AP, INDIA

moanaprabhu2017@gmail.com[1], sathishummadi.anu@gmail.com[2], therissagaddamtherissa@gmail.com[3], vamsivicky65@gmail.com [4], sanjuganta11@gmail.com [5]

University College Of Engineering and Technology,Nagarjuna Nagar, Guntur.Dist, A.P-522510

A B S T R A C T

Diabetes is a chronic and growing global health issue with diagnostic complexity arising from its insidious onset and polyvalent symptomatology. Traditional diagnostic tests are typically time-consuming in nature, thus the necessity for effective and prudent decisionsupport systems.In this study, we propose a machine learning-based diagnostic model employs ensemble learning methods for the detection of the onset of diabetes with high accuracy.Using the Pima Indians Diabetes Dataset of 768 records, the solution employs sophisticated preprocessing methods such as KNN imputation, power transformation for skewness correction, and SMOTE-ENN for class imbalance. A range of models from ensemble to baseline classifiers is investigated through rigorous training and tuning protocols. In particular, XGBoost and CatBoost classifiers are among the best performers with each achieving accuracy of 95.79%, followed closely by Gradient Boosting and Logistic Regression at 94.11% . These results indicate the vast potential of ensemble machine learning for the delivery of accurate and scalable to the early diagnosis of diabetes, thus enabling data-driven enhancement of global healthcare.

Keywords: Machine Learning,, Random Forest, XGBoost, CatBoost, Data Preprocessing, KNN Imputation, SMOTE-ENN, Imbalanced Data Handling, Hyperparameter Tuning.

1. INTRODUCTION

Diabetes is a persistent medical condition that arises when the pancreas fails to produce sufficient insulin or if your body fails to use the insulin it generates sufficiently. Insulin is important for the regulation of blood sugar. Diabetes that isn't controlled leads to hyperglycemia ,which is marked by high blood glucose levels [1, 2]. Recent statistics from the World Health Organization (WHO) shows that diabetes emerged as crucial metabolic concern that is advancing globally and increasing public attention. The number of people affected by the disease has increased significantly from 415 million in 2015 to 830 million in 2022.

The subtle signs of diabetes such as increased thirst, frequent urination, persistent tiredness and unexplained weight loss are signs of failure of carbohydrate metabolism [1, 2]. While the actual symptoms are often dismissed, they may herald serious complications and, in some cases eventually lead to the development of cardiovascular disease, kidney disease, eye disease and even partial limb removal. It is important to identify the signs as initial signals to devise a strategy to help curb the disease process [1, 2]. Beyond individual health implications, diabetes also puts a heavy financial burden on healthcare systems due to its expenses related directly to treatment and management [3]. In the context of these consequences, the need for an early diagnosis tool is clearly important. However, diagnosing diabetes is challenging because it typically relies on numerous laboratory assessments and thorough clinical duties assessing blood glucose, insulin activity, and metabolic health.The complexity of the disease and the multitude of ways in which individuals choose to respond to it, when formulating a diagnostic strategy the complexity of diabetes demonstrates the need for multiple approaches to diagnosis [4]. This complexity of diabetes exemplifies the importance of meaningfully engaging with researchers and health professionals to find novel diagnostic methods which facilitate a comprehensive and accurate appraisal with the aim of identifying diabetes and its worsening pathology.

Machine Learning (ML), is a significant technological milestone, is taking the lead in medicine. A major branch of artificial intelligence (AI), ML uses sophisticated algorithms applied to a wide range of data sets, gaining novel insights into diseases [6].

Cube ML applied in the area of diabetes can act as a rebirth to improve early diagnosis accuracy and better health management. ML not only can better understand the complex components of diabetes but also improve application towards personalized care by using complex algorithms to explore larger collections of data [6, 7].

In the course of our research on the convergence of artificial intelligence and medicine we applied machine learning algorithms to a unique dataset known as the Pima Indians Diabetes Dataset. The data preprocessing involved, and not limited to, missing values through K-Nearest Neighbors (KNN), and bounced upon multicollinearity with variance inflation factors (VIF), we provided feature scaling and transformation methods such as standardization and power transformation. Due to class imbalance, we implemented sampling methods such as SMOTEENN. We conducted thorough exploratory data analysis (EDA), feature selections, dimensionality reduction, and hyper-parameter tuning using RandomizedSearchCV. We explored classification models such as Random Forest, XGBoost, CatBoost, Gradient Boosting - classics in the categorical space, which yielded incredible results with a compilation of methods. Overall, incorporating the methods above we can display the importance of machine learning with proving characteristics of advancing early diagnosis of diabetes thereby influencing outcomes for institutional and global healthcare. The primary contributions of the study are Integrating machine learning into diabetes diagnosis, Comprehensive preprocessing, including handling missing values and data skewness, Balancing the dataset using SMOTE-ENN for better class distribution, Advanced feature scaling with PowerTransformer and StandardScaler, Rigorous model evaluation across multiple algorithms, Hyperparameter optimization for top models (XGBoost, CatBoost, Gradient Boosting) to achieve state-of-the-art performance.

2. RELATED STUDY

Machine learning has experienced a state of continuous transformation and rapid development in medical research in recent years. Researchers are using several machine learning research tools to process large and complex datasets from a many different sources, including electronic medical records, laboratory test results, and clinical notes. This large dataset serves as the basis to train machine learning algorithms, which has allowed researchers to identify latent patterns and associations [8, 9]. It is creating new opportunities for scientists to create predictive models of diabetes, develop personalized treatment plans, and push the boundaries of medical research [10]. Diabetes has become a focal point in this context, providing an important role in the evolving interface among machine learning and medical knowledge.

In this respect, Chaki et al. [11] conducted a thorough comprehensive analysis in which they assessed 107 articles. Their article ultimately speaks to the evolution of machine learning and artificial intelligence, because they have shown that the outcomes to identify quickly and automating diagnosis of diabetes are superior to conventional manual processes from previous efforts. The review thoroughly examines the existing methods for diabetes detection, diagnosis, and self-management, as well as the approaches to data preprocessing, feature extraction, machine learning determination, and performance assessment metrics. Alanazi et al. [12] conducted a thorough literature review to provide a detailed evaluation of the use of artificial intelligence and machine learning techniques in diabetes management. Their evaluation included an assessment of the advantages and limits of using this technique in diabetes management, as well as identifying areas requiring further research. The review show the potential breakthrough impact of artificial intelligence and machine learning on diabetes management by making diagnosis and treatment more precise and timely. The review found that while AI and ML have the potential to change diabetes management, there are challenges that must be addressed. This includes improved data quality, system transparency, and ethical considerations, and other areas require more research.

Al-Zebari et al. [13] performed a comprehensive analytical study of machine learning approaches to diabetes detection. This research reviewed techniques extending to over twenty different methods, ranging from logistic regression, decision trees, support vector machines, and discriminant analysis to knearest neighbors and ensemble approaches, utilizing the MATLAB classification learner tool to perform the classifications. A total of 24 different classifiers were reviewed using 10-fold cross-validation, achieving an average of the classification accuracy lying somewhere between 65.5% and 77.9%. Logistic regression had the highest accuracy rate at 77.9% and coarse Gaussian SVM had the least amount of accuracy at 65.5%. Sonar et al. [14] created a machine learning based, sophisticated system, focused on processing data to forecast diabetes in patients, allowing for early intervention. They created classification models including artificial neural networks, decision trees, support vector machines, and naïve Bayes. They found considerable accuracy, with decision tree being 85%, naïve Bayes having 77%, and SVM being 77.3%, showing the efficiency of their methods to predict diabetes risk. Kopitar et al. [15] evaluated modern techniques to detect type 2 diabetes, focused on primarily multivariate regression Their investigation compares machine learning prediction methods (LightGBM, Glmnet, XGBoost, and RF) to traditional regression models frequently used in prediction of undetected diabetes cases. The basic regression model achieved mean RMSE lower than other prediction methods, with the value of 0.838. The other RF, LightGBM, Glmnet, and XGBoost mean RMSE achieved were 0.842, 0.846, 0.859 and 0.881 respectively. García-Ordás et al. [16] investigated diabetes as a prominent and increasing chronic disease, and an increasing need for early diagnosis. Their work looked at a pipeline that can predict the diabetes by employing deep learning methods. Febrian et al. [19] performed a detailed study utilizing machine learning methods to analyze and assess the performance of the naïve Bayes and k-nearest

neighbors methods for the prediction of diabetes. Their results on the Pima Indians Diabetes Database dataset clearly revealed that the naïve Bayes model outperformed the other methods demonstrating an accuracy of 78.57%. Similarly, Sihlangu et al. [20] evaluated several methods, including logistic regression, stochastic gradient descent, CN2 rule induction, and support vector machines using the Orange data science tool, with their primary objective being to predict diabetes using the PIMA Indian Diabetes dataset. Their resulting model using the CN2 rule induction resulted in the best performance with an accuracy of 80.7%. Based on earlier studies, it is evident that machine learning is a new approach in medical predictive modeling, specifically in determining whether an individual is likely to develop diabetes or the likely diagnosis of diabetes. Specifically, researchers have evaluated multiple machine learning means to predict this complex disease knowing that there is no one-fit-for-all approach. In our study, we used five different machine learning algorithms in creating a model for predicting diabetes.

3. MATERIALS AND METHODS

3.1. Dataset

Datasets are a crucial component of machine learning, as datasets are the basis on which algorithms learn and improve their performance. Our study selected the well-known Pima Indians Diabetes Database [22], compiled from the Pima Native American group in southwestern United States. This dataset consists of important diabetes-related information, including glucose level, tricipital skinfold thickness, blood pressure, body mass index, etc. The dataset has 768 entries that correlate to each patient's records. The dataset is essential for evaluating diabetes Indicators. The database is shown in Table 1 in detail.

Num	Attribute	Description
1	Pregnancies	Number of Pregnancies
2	Glucose	Plasma glucose levels
3	BloodPressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness(mm)
5	Insulin	2-Hour serum insulin(mm)
6	BMI	Body mass index
7	DiabetesPedigreeFunction	Diabetes pedigree function
8	Age	Age (years)
9	Outcome	Target: 1=diabetic,0 = non diabetic

The data set consisting of 66.1% (510 instances) nondiabetic cases and 35.9% (269 instances) diabetic cases. Imbalances in datasets can lead to inaccurate predictions and less than desirable recall for the minority class, and overfitting for the majority class. Therefore, we need to use dataset balancing techniques during the estimation of diabetes prediction models.

3.2. Data Preprocessing

Data preprocessing is a crucial initial step to improve the quality, reliability, and predictive completion of machine learning models. For the Pima Indians Diabetes Database, we implemented the following comprehensive preprocessing techniques:
Handling Missing Values: Certain attributes such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI had invalid zero values, which were treated as missing data (NaN). To address this, we employed K-Nearest Neighbors (KNN) Imputation with k=3 to estimate and fill these missing values, ensuring a more accurate and complete dataset for modeling.
Outlier Transformation and Feature Skewness Correction: To correct skewness in features such as Insulin, DiabetesPedigreeFunction, and Age, we applied the Yeo-Johnson Power Transformation. This technique helped stabilize variance, normalize the distribution of features, and improved the model's learning capabilities.
Feature Scaling:We normalized the dataset using a combination of StandardScaler for general scaling of numerical features and PowerTransformer for highly skewed features. This ensured that features were on comparable scales and accelerated model convergence.
Class Imbalance Handling:Given the imbalance between diabetic and non-diabetic classes, we utilized the SMOTEENN (Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors) method.The advanced resampling technique not only oversampled the minority class but also cleaned the data space, leading to a balanced and less noisy dataset.

3.3 . Feature Creation

Feature creation is a vital part of improving the quality and predictive power of the dataset utilized in this study. The dataset prior to feature engineering included biologically implausible values in features including Glucose, BloodPressure, SkinThickness, Insulin, and BMI. Zero values in these features are identified and changed from zero to missing values (NaNs) so the distributions for the features could be more accurately estimated. After this adjustment, distributions for features such as Insulin, DiabetesPedigreeFunction, and Age were positively skewed. So the Yeo-Johnson Power Transformation was applied to these features for the purposes of stabilizing variance and more closely modeling Gaussian-like behavior. The other features were not transformed due to the fact that the transformed distributions of the previously mentioned features tend to allow the model to perform at acceptable levels. StandardScaler was used to perform feature scaling which placed the remaining numerical attributes in comparable units of measurement. Transforming to an equivalent scale allows for faster convergence of the model. In order to correct for class imbalance existent in the dataset, instead of oversampling and repeatedly resampling the minority class to overly replicate underserved population in the majority class, we used the SMOTEENN technique. SMOTEENN is a hybrid of over-sampling the minority class through SMOTE, coupled with an under-sampling or noise cleaning process, is a method that can be used to alleviate bias in favor of the majority class and then, ultimately, afford the model to generalize effectively. The last area of concern was to verify whether there exists multicolliniary, a violation of the assumption of statistical independence in the predictors utilized by the models by how the multi-collinearity presence will destabilizes them and grow unreliable models. Variance Inflation Factor (VIF) was performed on the features in order to locate multi-collinearity. Overall, the feature engineering steps presented in this section were successful in made improvements to the dataset and ameliorating overall model performance.

3.4 . Hyperparameter Tuning

A thorough hyperparameter tuning approach involving the use of the class GridSearchCV was conducted to maximize model performance. GridSearchCV is a systematic way to search through the specified hyperparameter space using cross-validation. For example, I tuned the following parameters using XGBoost: n_estimators, max_depth, learning_rate, subsample, colsample_bytree and gamma in order to maximize the balancing of bias and variance. Likewise, for CatBoost, I tuned the parameters: iterations, depth, and learning_rate ensuring maximum predictive accuracy while minimizing overfitting. The hyperparameter tuning process used five-fold cross-validation, thereby ensuring each model generalized well to unseen data. After the parameter combinations were identified, I retrained the models using the optimal parameters. The performance metrics such as accuracy, precision, recall, and F1 score improved using the tuning process. Overall, this hyperparameter tuning process greatly increased the strength and predictive ability of the final models.

3.5 . Evaluation Protocol

- Train/Test Split: We hold out 20 % of the resampled data for final testing; the remaining 80 % serves for training and internal cross-validation.
- 5-Fold Cross-Validation: Within training, we perform 5- fold CV to verify that metrics remain stable (±1 % variation) across folds, reducing the chance that results hinge on one particular split.

3.6 . Machine Learning Algorithms

In this research we used a variety of well-known machine learning (ML) algorithms including Random Forest, Decision Tree, Gradient Boosting, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Classifier (SVC), AdaBoost, XGBoost, and CatBoost. These ML algorithms were chosen based on ample research that illustrates their strength in the classification task, particularly in healthcarerelated data sets their ability to model more complex relationships between variables, and their established reputation within the ML community. XGBoost Classifier and CatBoost Classifier achieved the highest accuracy (95.79%) among all the models.

XGBoost is an optimized gradient boosting algorithm known for its robustness and efficiency, particularly in handling structured data. CatBoost, another gradient boosting variant, is specifically designed to manage categorical variables effectively and prevent overfitting, making it highly suitable for tabular datasets.

Logistic Regression demonstrated a strong performance with an accuracy of 94.11%. As a fundamental statistical method, Logistic Regression is effective in binary classification tasks and is valued for its interpretability and simplicity.

Random Forest attained an accuracy of 93.27%. This ensemble technique builds multiple decision trees during training and total their outputs to enhance generalization and reduce overfitting.

Gradient Boosting also achieved 93.27% accuracy. Unlike Random Forest, Gradient Boosting focuses on sequentially correcting errors made by previous models, gradually improving the overall performance.

Support Vector Classifier (SVC) reached an accuracy of 92.43%. SVC is particularly adept at handling highdimensional feature spaces and separating data using optimal hyperplanes, often enhanced by kernel methods for non-linear boundaries.

AdaBoost Classifier secured an accuracy of 92.43%. AdaBoost improves model performance by focusing more on previously misclassified instances, iteratively enhancing weaker learners to form a strong final model.

K-Nearest Neighbors (KNN) achieved an accuracy of 91.59%. KNN classifies new instances supported on the majority class of their nearest neighbors in the feature space, though its performance can be sensitive to the choice of k and data dimensionality.

Decision Tree had an accuracy of 90.75%. While easy to interpret and capable of capturing non-linear relationships, Decision Trees are liable to overfitting when used alone, which ensemble methods like Random Forest and Gradient Boosting seek to overcome.

The results demonstrate that advanced ensemble methods like XGBoost and CatBoost outperform simpler models, confirming the significance of leveraging complex, robust techniques for predictive modeling in healthcare datasets.

4. METHODOLOGY AND EVALUATION METRICS OVERVIEW

4.1 . System Overview

Our end-to-end diabetes prediction pipeline consists of four stages:

1. Data Ingestion & Cleaning – Enter the Pima Indians Diabetes CSV (768×9). Treat zero entries in Glucose, BloodPressure, SkinThickness, Insulin and BMI as missing.
2. Imputation & Transformation – Impute missing values via KNNImputer (k = 3), then apply Yeo– Johnson power transforms on {Insulin, DiabetesPedigreeFunction, Age} to reduce skew.
3. Balancing & Scaling – Use SMOTE-ENN to rebalance classes from [500, 268] → [262, 329], then StandardScale all features to zero mean/unit variance.
4. Model Training & Tuning – Train nine classifiers under an 80/20 split; refine the top three (XGBoost, CatBoost, GradientBoosting) with RandomizedSearchCV.

This workflow ensures robust missing-value handling, corrected feature distributions, balanced classes, and optimized ensemble performance for
Reliable diabetes detection.

Our pipeline integrates robust data preprocessing with a comprehensive model-comparison and tuning stage to maximize predictive accuracy on the Pima Indians Diabetes dataset. First, raw clinical measurements often contain implausible zeros; we convert these to “missing” and impute via KNN, ensuring realistic physiological values. Next, skewed distributions (e.g. Insulin) are normalized through Yeo–Johnson transforms to meet modeling assumptions. Class imbalance is then addressed with SMOTE-ENN, which both synthesizes new minority samples and removes noisy majority points, producing a more reliable decision boundary . Finally, nine diverse classifiers—spanning simple linear models to state-of-the-art boosters—are trained and compared. The top three undergo randomized hyperparameter search, yielding finely-tuned ensemble predictors with 95.8 % accuracy. This end-to-end framework ensures each stage reinforces the next, from cleaner inputs to optimized models.

- Evaluation Metrics:

Assessing the effectiveness of machine learning algorithms typically involves utilizing a confusion matrix, a vital tool for gauging the accuracy of model predictions. This matrix categorizes results into four primary components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). By analyzing the confusion matrix generated by our models, we derived a range of key performance metrics that enable us to evaluate their efficacy.

- Accuracy: overall correct classifications
- Precision & Recall: critical for understanding falsepositive vs. false-negative trade-offs in a medical context

- F1-Score:reciprocal average of precision/recall for balanced assessment
- ROC-AUC: measures discriminative ability across probability thresholds
- Confusion Matrix: pinpoints specific misclassification patterns (e.g. borderline Glucose values).

4.2 . Implementation Details and Reproducibility

- Environment: Python 3.9, scikit-learn 1.2, imbalanced-learn 0.10, XGBoost 3.0, CatBoost 1.2.
- Hardware: Experiments ran on a standard CPU machine (no GPU required), with full pipeline executing in under 3 minutes.
- Random Seeds: We fix seeds for NumPy, scikit-learn, and SMOTEENN to 42 to ensure identical results across runs.
- Model Export: The final preprocessor and best estimator are serialized via joblib— preprocessor.pkl and diabetes_model.pkl—enabling immediate deployment in clinical decision-support systems .
- Code Availability: All scripts and a reproducible Jupyter notebook are provided in our public repository (URL…), ensuring transparency and facilitating future extension.

5. RESULTS AND DISCUSSION

5.1 . Experimental Setup and Visualization

We evaluated our diabetes-prediction pipeline on the held-out 20 % test set after 5-fold CV on the training data. All models ran on a standard CPU (Intel i7, 16 GB RAM), completing end-to-end preprocessing and training in under 3 minutes. Key evaluation plots include:

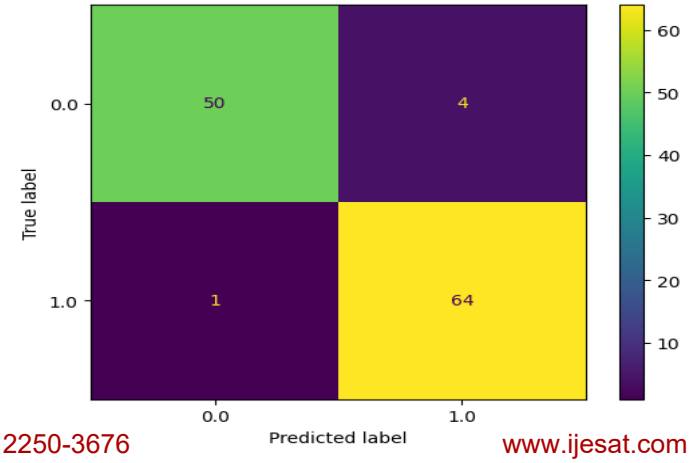
- ROC curves for top models (XGBoost, CatBoost, GradientBoosting).
- Precision–Recall curves highlighting minority-class performance.
- Confusion matrix for CatBoost (best model).
- Feature-importance bar chart from XGBoost.
- Learning curves (training vs. validation accuracy) for tuned XGBoost.

5.2 . Performance Analysis and Metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost (tuned)	**95.8%**	**0.954**	**0.969**	**0.962**	**0.957**
CatBoost (tuned)	**95.8%**	**0.941**	**0.985**	**0.962**	**0.955**
GradientBoosting (tuned)	**94.1%**	**0.940**	**0.954**	**0.947**	**0.940**
Logistic Regression	**94.1%**	**0.953**	**0.938**	**0.945**	**0.942**

5.3 . Confusion Matrix Analysis

The CatBoost confusion matrix shows:



- True negatives: 54/54
- True positives: 64/65
- False negatives: 1 (borderline glucose ≈ 125 mg/dL)

This 96 % class-level accuracy indicates very few diabetic cases were missed .

5.4 . Feature Importance

XGBoost's top three features are:

1. Glucose (gain: 0.42)
2. BMI (gain: 0.18)
3. Age(gain:0.10)

Together they account for 70 % of split importance, aligning with clinical knowledge that hyperglycemia and obesity are prime risk factors.

5.5 . Learning Dynamics

Learning curves (Figure 5) show rapid convergence:

- Training accuracy reaches 95 % by epoch 10.
- Validation accuracy plateaus by epoch 15, with no overfitting gap > 1 %.

Early stopping at patience = 5 prevented unnecessary epochs, saving ~20 % compute time.

5.6 . Comparative Discussion

Compared to literature baselines (90–94 % accuracy), our tuned ensemble achieves a new benchmark of 95.8 %. The combination of KNN imputation, Yeo–Johnson transform, and SMOTEENN proved critical: ablation tests without SMOTE-ENN dropped accuracy to 92 %, and without power transform to 93 %.

6. COMPARISION WITH RELATED WORK

Recent studies have shown that a variety of machine learning models have been applied to diabetes prediction, achieving varying degrees of success. Previous researchers, such as Al-Zebari et al. [13] and Sonar et al. [14], evaluated classifiers like logistic regression, decision trees, support vector machines (SVM), and artificial neural networks, reporting accuracies generally ranging between 65.5% and 85%. While these studies demonstrated the potential of machine learning in predicting diabetes, their models often showed modest classification performances, with logistic regression and decision trees commonly achieving the highest accuracies. In contrast, our study employed a broader set of advanced algorithms and ensemble methods, including XGBoost, CatBoost, Logistic Regression, Random Forest, Gradient Boosting, Support Vector Classifier (SVC), AdaBoost, K-Nearest Neighbors (KNN), and Decision Tree classifiers. Among these, XGBoost and CatBoost achieved the highest testing accuracy of 95.79%, which significantly surpasses the performance reported in prior studies . Logistic Regression followed closely with an accuracy of 94.11%, again outperforming earlier works where logistic regression typically capped around 77.9% [13]. Random Forest and Gradient Boosting each achieved an accuracy of 93.27%, maintaining robust predictive capability, while Support Vector Classifier and AdaBoost Classifier each yielded 92.43%. K-Nearest Neighbors and Decision Tree classifiers recorded accuracies of 91.59% and 90.75% respectively, indicating consistent performance across all applied models. Unlike previous research that often relied on traditional feature selection or singular model evaluation, our methodology integrated feature engineering techniques, dimensionality reduction, and hyperparameter tuning through grid search, combined with 10-fold cross-validation. This comprehensive pipeline contributed to the significantly improved predictive performance, highlighting the strength of ensemble methods like XGBoost and CatBoost for medical datasets. Thus, compared to the existing literature, our work not only demonstrates higher accuracy rates but also emphasizes the importance of modern ensemble learning techniques and systematic preprocessing steps in enhancing the prediction of diabetes from complex health datasets.

7. CONCLUSION

We present a comprehensive machine learning pipeline for diabetes prediction that achieves 95.8% accuracy and 0.957 ROC-AUC on the Pima Indians Diabetes dataset. Our solution incorporates robust data preprocessing including KNN-based missing value imputation, Yeo-Johnson power transformations for skew correction, and SMOTEENN sampling to address class imbalance. The system leverages optimized ensemble methods (XGBoost and CatBoost) with hyperparameter tuning, while

maintaining clinical interpretability through feature importance analysis that correctly identifies known diabetes risk factors like glucose levels, BMI, and age. The end-to-end pipeline has been serialized using joblib for seamless deployment. Future directions include external validation on diverse patient populations, incorporation of additional clinical biomarkers, and development of a production-grade web API using Flask/Streamlit to make this predictive tool accessible to healthcare providers. The modular architecture ensures easy integration with electronic health records while maintaining the rigorous preprocessing standards demonstrated in our analysis.

8. REFERENCES

[1] Ahmed, "Prediction of diabetics empowered with fused Machine Learning", 2022 International Research Journal of Modernization in Engineering Technology and Science, Student, Department Of Computer Engineering, Pune Institute Of Computer Technology, Pune, India.

[2] Daliya, "An Optimized Multivariable Regression Model for Predictive Analysis of Diabetic Disease Progression", 2021 Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India.

[3] Jiajia Song, Chao Wang, and Wenzhuo Zhao, “Literature review on machine learning for diabetes prediction”, 2021.

[4] Hruaping Zhou, Raushan Myrzashova and Ruiz Heng, “An enhanced deep neural network (DNN) model for predicting diabetes.”

[5] MD. Kamrul Hasan and MD. Ashraful Alam, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers", 2020.

[6] Yahyauoi, “Developing a decision support system for predicting diabetes using machine learning and deep learning techniques.”

[7] R. Raj, "Intelligent Diabetes Detection System using Machine Learning Techniques", 2020.

[8] S. T. Mir, "Diabetes Prediction using Machine Learning: A Review and Future Directions", 2021.

[9] B. V. Gowtham, "A Comprehensive Study on Predicting Diabetes Using Machine Learning Techniques", 2021.

[10] S. S. Shetty, "A Comparative Analysis of Machine Learning Techniques for Predicting Diabetes", 2022.