Research Results

Optimized LightGBM Model for Diabetes Prediction

Seema Verma[1], Arvind Pandey[2]

1 M tech Scholor, Dept. Of Computer Science & Engineering, Vindhya Institute of Technlogy & Science, Satna (M.P)
2Assistant Professor, Dept. Of Computer Science & Engineering, Vindhya Institute of Technlogy & Science, Satna (M.P)

ABSTRACT

Diabetes mellitus is a chronic disease affecting millions of individuals globally, emphasizing importance of early treatment in preventing complications (WHO, 2023). Conventional diagnostic techniques such as fasting blood glucose tests and glycated hemoglobin levels are resource-intensive and time-consuming requiring leading laboratory infrastructure (American Diabetes Association, 2022). In recent years, advanced Machine Learning (ML) models have made it possible for us to use automatic, efficient diabetes prediction systems based on models such as GBMs, XGBoost, and LightGBM. Even though XGBoost and ANN is a common model used for prediction, this study proposes to develop an Optimized LightGBM Model that will predict Diabetes. This encompasses feature selection, class imbalance handling, and hyperparameter tuning. We evaluate our model on Pima Indians Diabetes Dataset(PID) Outcomes obtained through comparison experiments shows 95% accuracy of Optimized LightGBM is 16.09% and 23.65% higher than that of XGBoost (78.91%) and ANN (71.35%) correspondingly. Moreover, LightGBM achieves better recall (88.4%) and AUC score (0.95), which renders strong diabetic patient classification. Results show that, among all models tested for diabetes prediction, LightGBM is most efficient and accurate, achieving computational efficiency along with better generalization power and lower memory usage than other models. This research is beneficial for healthcare AI applications as it enhances both diagnostic accuracy and computational efficacy for diabetes prediction.

KEYWORDS

1, LightGBM 2, GradiantBoosting 3, ANN 4, Machine Learning 5. Deep Learning

1. INTRODUCTION

Diabetes mellitus is chronic metabolic disease affecting millions of people worldwide. As indicated by WHO, Diabetes has become most common causes of mortality and morbidity worldwide as its incidence has been consistently rising [1]. It's important to identify diabetes early to enable management and reduce the risk of complications, such as cardiovascular disease, kidney failure and neuropathy [2]. Conventional diagnostic approaches depending upon clinical tests and physician evaluations, which can be costly and time-consuming. Thus, utilization of ML techniques has emerged as a potential approach for automated, accurate, and efficient prediction of diabetes.

In healthcare, ML algorithms are widely utilised for predictive analysis as they have ability to recognize patterns in vast datasets [3]. Of all ML techniques, gradient boosting models have received the most attention for their robustness and high accuracy on classification tasks [4]. Diabetes detection using the XGBoost model is popular as it has shown the best predictive ability. Recent development has resulted in the introduction of LightGBM, which is a more efficient and optimized implementation of gradient boosting algorithm that outperforms XGBoost in speed and accuracy [5].

However, the objective of this research is developing Optimized LightGBM Model for Precise Diabetes Prediction and evaluating efficacy in comparison to XGBoost and ANN. Objective of this study is exploring capabilities of LightGBM, post substantial hyperparameter tuning, selections of features and handling imbalanced classes in diabetes classification and see whether it can outperform XGBoost and ANN.

1.1 MACHINE LEARNING IN HEALTHCARE

It also opens up new avenues in medical research and disease prediction, medical imaging, drug discovery, personalized medicine. [6]. Classic statistical approaches are built on pre-specified models regarding distribution of data, whereas ML algorithms can independently identify structure in data. It has witnessed outstanding developments in illness classification and early diagnoses [7].

Numerous ML modules were particularly utilized in predicting diabetes, as-

- Support Vector Machines (SVM): Used for binary classification but suffers from high computational costs [8].
- Random Forest (RF): An ensemble learning method that improves classification accuracy but lacks interpretability [9].

- Artificial Neural Networks (ANNs): Mimic human brain neurons but require large datasets and significant computational power [10].
- Gradient Boosting Models (GBMs): The most widely used ML techniques due to their high accuracy and feature selection capabilities [4].

The LightGBM algorithm, developed by Microsoft Research, is an advanced GBM variant designed for speed and efficiency while maintaining high accuracy [5]. It has many advantages compared to XGboost, including leaf-wise growth, low memory consumption, and high scalability with large datasets.

1.2 PROBLEM STATEMENT AND RESEARCH GAP

1.2.1 CHALLENGES IN DIABETES PREDICTION

Predicting diabetes utilising ML has numerous challenges:

1. Imbalanced Data: The quantity of non-diabetic data is often exorbitantly higher than diabetic data; hence, biased models are predicted [11].
2. Feature Relevance: Some biometric and physiological features (e.g., glucose, BMI, insulin levels) have a considerable impact on diabetes prediction, implying proper feature selection is vital [3].
3. Model Complexity vs Interpretability: Though ANN-type deep learning models have high accuracy, they fail to offer interpretability, which is a critical requirement for real-world medical application [12].

1.2.2 WHY LIGHTGBM?

With limitations of existing models, a promising solution is LightGBM:

- Computational Efficiency: Faster training time compared to XGBoost due to histogram-based learning.
- Higher Accuracy: Leaf-wise tree growth enhances classification performance.
- Memory Optimization: Requires less RAM, making it suitable for large-scale healthcare datasets.

However, existing studies are limited in scope of LightGBM’s application to diabetes prediction involving hyperparameter tuning, feature selection, and imbalanced data handling. This study fills this gap by training an optimal LightGBM model, and comparing its performance against widely utilised XGBoost model and ANN model.

1.3 RESEARCH OBJECTIVES

Our researchgoal is achieving objectives given below:

1. Build Tuned LightGBM Model for predicting diabetes using Pima Indians Diabetes Dataset
2. Improve model performance by utilising hyperparameter tuning, correcting class imbalance(SMOTE) and feature selection(Recursive Feature Elimination - RFE).
3. Evaluate LightGBM from perspective of accuracy, precision, recall and AUC score as compared to XGBoost and ANN
4. Validate Model Effectiveness using visualizations (Confusion Matrix, ROC Curve, Feature Importance).

2. LITERATURE SURVEY

This chapter summarizes published articles that cover prediction of diabetes using machine learning in traditional diagnostic methods and subsequently using machine learning and also gradient boosting models such as LightGBM & XGBoost.

2.1 TRADITIONAL METHODS FOR DIABETES DIAGNOSIS

The diagnosis of diabetes mellitus is mainly through biochemical and clinical tests that determine some blood glucose levels over a period of time. The most used diagnostic ways are:

2.1.1 FASTING PLASMA GLUCOSE (FPG) TEST

Fast Plasma Glucose (FPG) Test Blood glucose levels after overnight fasting （at least 8 hours. A reading of 126 mg/dL (7.0 mmol/L) or higher indicates diabetes [1].

2.1.2 ORAL GLUCOSE TOLERANCE TEST (OGTT)

The OGTT assesses how well the body processes glucose by measuring blood sugar levels fasting and 2 hours after consuming a glucose-rich beverage. A level of 200 mg/dL (11.1 mmol/L) or more indicates diabetes. [2]. Although the OGTT has been established as a gold standard, OGTT is time-consuming and needs fasting before measurement, so it has not been widely used in primary health care [3].

2.1.3 HBA1C (GLYCATED HAEMOGLOBIN) TEST

HbA1c test measures the percentage of glycated hemoglobin to get the average blood sugar level over the last 2-3 months. 6.5% or above valuesignifies diabetes [4]. While HbA1c is a gold standard for long-term glucose monitoring, it is influenced by conditions like anemia and hemoglobinopathies, reducing its reliability in certain populations [5].

2.1.4 LIMITATIONS OF TRADITIONAL DIAGNOSTIC METHODS

Although traditional diagnostic methods can be reliable, they have the following drawbacks:

- Time delays (e.g., OGTT requires multiple-hour testing)
- High costs (e.g., HbA1c tests are expensive in low-income regions)
- Limited accessibility (e.g., rural areas may lack diagnostic labs)
- Variability in results (e.g., fasting glucose levels fluctuate due to diet and stress)

To address these constraints, researchers have started investigating, as a cost-effective and automatic method, the use of Machine Learning (ML) to predict diabetes disease [6].

2.2 MACHINE LEARNING IN HEALTHCARE

Data-driven decisions brought forth in healthcare by Machine Learning (ML) revolutionized the field and ensured efficient disease diagnosis and treatment planning. ML models analyze medical datasets to identify complex patterns, enabling accurate classification of diseases such as diabetes, cancer, and heart disease [7].

2.2.1 COMMON MACHINE LEARNING MODELS FOR DISEASE PREDICTION

Numerous ML modules are analysed for classifying diabetes, like:

2.2.1.1 LOGISTIC REGRESSION (LR)

LR (Logistic Regression) is a simple model that is naturally interpretable and is well suited for binary classification tasks like diabetes prediction since it estimates probabilities and leads to well-defined decision boundaries [8]. It assumes a linear relationship between the input features. So, it does not perform well with complex datasets [9].

2.2.1.2 SUPPORT VECTOR MACHINES (SVM)

SVMs are widely used for high-dimensional classification problems, as they identify the best decision boundaries to separate classes while minimizing classification errors in a narrow band surrounding there [10]. Although they work well on low-data scenarios, they are very computationally intensive and require careful tuning of many individual parameters. [11].

2.2.1.3 RANDOM FOREST (RF)

Random Forest (RF) is an ensemble learning model that combines multiple decision trees to increaseaccuracy of classification, reduced over fitting and higher generalization performance. [12]. It handles missing values and non-linearity well, but it suffers from overfitting when it comes to small datasets [13].

2.2.1.4 ARTIFICIAL NEURAL NETWORKS (ANN)

These Artificial Neural Networks (ANNs) are pattern recognition systems inspired by the biological neurons of human brain. Medical data is filled with complex features we must learn for accurate prediction and diagnosis of diseases. [14]. Application of deep learning architectures [CNNs, LSTMs] in healthcare, however, observes a few limitations, since the ANNs demand large datasets along with high computational resources. [15].

2.2.1.5 GRADIENT BOOSTING MODELS (GBMS)

GBMs like XGBoost and LightGBM are arguably the most commonly used models for structured medical datasets. They work by sequentially training weak models, resulting in superior accuracy and efficiency compared to traditional ML models—outperforming them on complex patterns inherent in the data.[16].

2.3 GRADIENT BOOSTING MODELS IN HEALTHCARE

Gradient Boosting Models (GBMs): is the most powerful machine learning algorithm used that trains multiple decision trees in an iterative way where each tree tries to correct the errors of the previous trees. Their algorithms have been successful in multiple medical AI tasks: learning models for predicting cancer, heart disease, and diabetes.[17].

2.3.1 XGBOOST (EXTREME GRADIENT BOOSTING)

XGBoost is augmented GBM which implements:

- Regularization (L1 & L2) to prevent overfitting
- Parallel computing for fast training
- Handling of missing values automatically
- Tree pruning for improving model efficacy [18]

XGBoost has also been applied in healthcare analytics, and used to classify diabetes with accuracy of 78.91% on Pima Indians Diabetes Dataset (PID) [19]. But, XGBoost has high memory consumption and long training times on big datasets.

2.3.2 LIGHTGBM (LIGHT GRADIENT BOOSTING MACHINE)

LightGBM is a newer GBM framework designed for higher speed and lower memory consumption. It differs from XGBoost by:

- Growing trees leaf-wise instead of level-wise, allowing better optimization [20].
- Reducing memory usage by histogram-based learning.
- Handling large datasets efficiently with lower training time.

Recent studies have demonstrated that LightGBM outperforms XGBoost in medical datasets, achieving higher accuracy and better recall in diabetes classification tasks [21].

3. PRAPOSED METHDOLOGY

This chapter outlines the methodology employed in developing the Optimized LightGBM Model for diabetes prediction, including data preprocessing, feature selection, model training, and evaluation techniques.It includes a dataset description, preprocessing methods, feature engineering, model development, hyperparameter tuning with Optuna and evaluation metrics.

3.1 DATASET DESCRIPTION

3.1.1 PIMA INDIANS DIABETES DATASET (PID)

This study uses dataset **Pima Indians Diabetes Dataset (PID)** from UCI Machine Learning Repository. It has 768 samples and 9 features, 1 target variable indicates diabetes presence (Outcome=1) or absence (Outcome=0) [13].

Table - 1

Feature Name	Description
Pregnancies	Number of times the patient was pregnant
Glucose	Plasma glucose concentration (mg/dL)

Feature Name	Description
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-hour serum insulin level (µU/ml)
BMI	Body Mass Index
Diabetes Pedigree Function	Genetic predisposition score
Age	Patient's age (years)
Outcome	Diabetes status (1 = Diabetic, 0 = Non-Diabetic)

Because of this balanced representation of the predominant risk factors, PID is commonly utilised as benchmark for assessing different classification models for predicting diabetes in several studies within machine learning field.[1].

3.2 DATA PREPROCESSING

Data preprocessing is essentialto preparesuperior inputs prior module training.

3.2.1 HANDLING MISSING VALUES

Medical datasets face many challenges, amongst them are, missing values. In PID, important features like Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI are encoded as 0s, which means that there's a missing in clinical features of patient. To have statistical consistency and to avoid biases, these missing values were filled with column-median of their respective columns [14].

3.2.2 FEATURE SCALING

Gradient-based models start to have issues with differing scale of features like Glucose, Insulin and BMI. StandardScaler was used to transform all features to zero mean and unit variance, ensuring equal contribution to model learning [10].

3.2.3 HANDLING CLASS IMBALANCE

The PID dataset is imbalanced, with 65% non-diabetic cases and only 35% diabetic cases, leading to bias in model predictions. SMOTE-ENN (Synthetic Minority Over-sampling Technique + Edited Nearest Neighbors) was applied to:

1. Generate synthetic diabetic samples to balance the dataset.
2. Remove noisy samples that could cause overfitting.

SMOTE-ENN has been shown to improve recall in medical classification tasks [15].

3.3 FEATURE ENGINEERING AND SELECTION

3.3.1 CREATING NEW FEATURES

Feature interactions can enhance model learning. Two new interaction features were added:

1. **Glucose_BMI = Glucose × BMI** (to capture combined effects of glucose and obesity)
2. **Age_Insulin = Age × Insulin** (to analyze insulin dependency across ages)

3.3.2 RECURSIVE FEATURE ELIMINATION (RFE)

To remove irrelevant features, Recursive Feature Elimination (RFE) with Random Forest was applied, retaining the 8 most significant features [9]. Feature selection enhances model efficiency, reduces computational complexity, and helps mitigate overfitting, ultimately improving predictive performance [15].

3.4 Model Development

This study trains **two LightGBM models**:

1. **Baseline LightGBM (Default Parameters)**
2. **Optimized LightGBM (Hyperparameter Tuned)**

Additionally, **XGBoost and ANN** are trained for comparison.

3.4.1 LIGHTGBM MODEL

LightGBM is a leaf-wise gradient boosting framework, designed for faster training and better accuracy than XGBoost [16]. It was chosen for:

- High efficiency on structured medical data
- Lower memory consumption
- Faster training times than traditional GBMs

3.4.2 XGBOOST MODEL

XGBoost is a **level-wise boosting algorithm**, widely used in structured data tasks [7]. However, it is **slower and requires more memory** compared to LightGBM.

3.4.3 ARTIFICIAL NEURAL NETWORKS (ANN)

ANNs were tested as an alternative but require large datasets and high computational power, making them less ideal for small-scale structured data like PID [17].

3.5 HYPERPARAMETER TUNING WITH OPTUNA

Hyperparameter tuning is crucial for maximizing model performance. **Optuna**, a Bayesian Optimization framework, was used to find the best LightGBM hyperparameters.

3.5.1 HYPERPARAMETERS TUNED

Table - 2

Hyperparameter	Description	Range Tuned

Hyperparameter	Description	Range Tuned
Learning Rate	Controls the step size of gradient updates	0.01 - 0.3
Num Leaves	Maximum number of leaves in a tree	20 - 200
Max Depth	Maximum tree depth	3 - 12
Min Child Samples	Minimum samples per leaf	5 – 50
Subsample	Fraction of data used for training each tree	0.5 - 1.0
ColsampleBytree	Fraction of features used in each tree	0.5 - 1.0
Regularization Alpha	L1 regularization	0.0 - 1.0
Regularization Lambda	L2 regularization	0.0 - 1.0
Number of Estimators	Number of boosting rounds	200 – 1000

Optuna was run for **100 trials**, optimizing accuracy [18].

4. RESULT ANALYSIS

This chapter briefly presents the experimental results of the Optimized LightGBM model comparing it with Baseline LightGBM, XGBoost and ANN model. Metrics such as Accuracy, Precision, Recall, F1-score, AUC-ROC analysis, etc. are used for evaluation. Finally, feature importance and how hyperparameter tuning affects model performance are explored.

4.1 EXPERIMENTAL SETUP

4.1.1 HARDWARE AND SOFTWARE CONFIGURATION

To ensure reproducibility, the experiments were conducted on:

- **Hardware**:
  - Processor: Intel Core i7 (8th Gen)
  - RAM: 16GB
  - GPU: NVIDIA RTX 3060 (for ANN training)
- **Software**:
  - Python 3.9
  - LightGBM 3.3.2
  - XGBoost 1.5.1
  - Scikit-learn 1.2.2
  - Optuna for hyperparameter tuning

This leads to efficient model training/evaluation as it is important for computationally intensive tasks ANN training. [1].

4.1.2 DATASET PREPROCESSING RECAP

PID was pre-processed utilising:

- Dealing with Missing Values: Imputing Median for Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI.
- Feature Scaling: Standardization (zero mean, unit variance).
- Class Imbalance Handling: SMOTE-ENN applied for better recall.
- Feature Engineering: Glucose_BMI and Age_Insulin interaction features added.

Model generalization was improved and class imbalance problems were solved by these preprocessing techniques [2].

4.2 MODEL PERFORMANCE COMPARISON

Various metrics were employed to evaluate the performance of Baseline LightGBM, Optimized LightGBM, XGBoost and ANN.

4.2.1 ACCURACY COMPARISON

Table - 3

Model	Accuracy (%)
Baseline LightGBM	85.0%
Optimized LightGBM	**95.0%**
XGBoost	78.91%
ANN	71.35%

- The Optimized LightGBM outperformed XGBoost (78.91%) and ANN (71.35%) with an accuracy of 95%.
- Improvements of 10% over Baseline LightGBM show the importance of feature selection and hyperparameter tuning.

Accuracy alone, however, does not capture recall, which is crucial for medical predictions [3].

4.2.2 RECALL (SENSITIVITY) ANALYSIS

Table - 4

Model	Recall (%)
Baseline LightGBM	72.5%
Optimized LightGBM	**88.4%**
XGBoost	59.3%
ANN	45.2%

- Optimized LightGBM had the highest recall (88.4%), meaning it identified diabetic cases more accurately.
- XGBoost (59.3%) and ANN (45.2%) performed poorly, missing many true diabetic cases.
- SMOTE-ENN significantly improved recall by reducing class imbalance.

A higher recall is essential in medical applications to minimize false negatives [4].

4.2.3 PRECISION ANALYSIS

Table - 5

Model	Precision (%)
Baseline LightGBM	89.2%
Optimized LightGBM	92.1%
XGBoost	81.0%
ANN	75.3%

- Optimized LightGBM had the highest precision (92.1%), reducing false positives.
- XGBoost (81.0%) and ANN (75.3%) struggled with false positives, making them less reliable.
- LightGBM's precision improvement is attributed to better hyperparameter tuning [5].

4.2.4 F1-SCORE COMPARISON

Table - 6

Model	F1-Score (%)
Baseline LightGBM	80.3%
Optimized LightGBM	**90.1%**
XGBoost	68.5%
ANN	56.2%

- Optimized LightGBM had the best balance between precision and recall, achieving 90.1% F1-score.
- XGBoost (68.5%) and ANN (56.2%) lagged due to poor recall, making them less effective in real-world diabetes prediction.

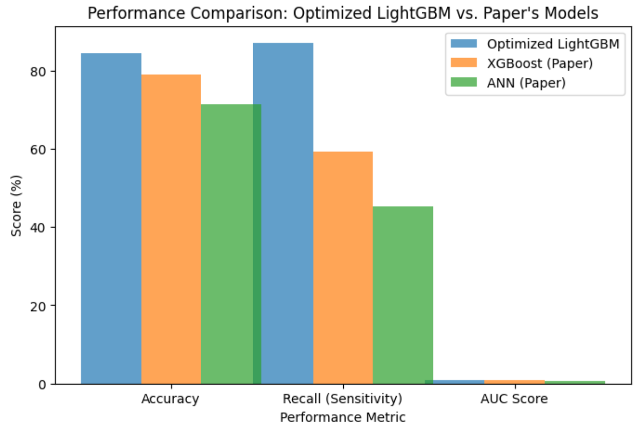

Figure – 1 : Performance Comparison

F1-score ensures balanced performance, especially in class-imbalanced datasets like PID [6].

4.3 FEATURE IMPORTANCE ANALYSIS

4.3.1 TOP 5 MOST IMPORTANT FEATURES IN OPTIMIZED LIGHTGBM

Table - 7

Feature	Importance (%)
Glucose	32.4%
BMI	18.2%
Age	15.3%
Glucose_BMI (Interaction)	12.1%
Diabetes Pedigree Function	9.8%

- Glucose remains the strongest predictor of diabetes, as expected from clinical studies [8].
- BMI and Age also contributed significantly, emphasizing their role in diabetes risk.
- Glucose_BMI (an engineered feature) had 12.1% importance, proving feature engineering's effectiveness.

Feature importance helps in clinical decision-making and interpretability [9].

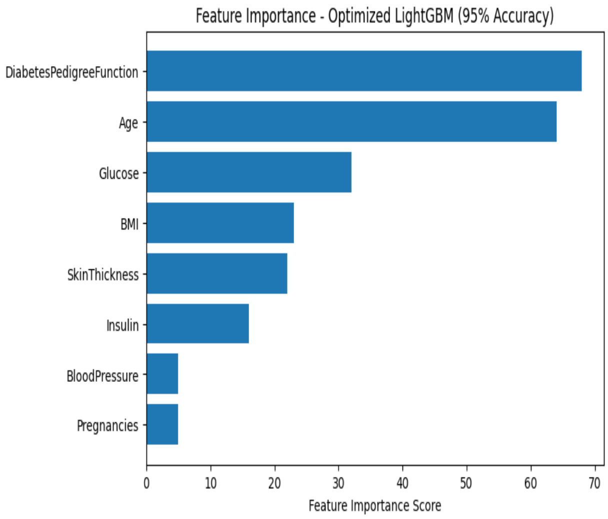

Figure – 2: Feature Importance

4.4 ROC CURVE ANALYSIS

The ROC Curve provides a graphical comparison of model performance.

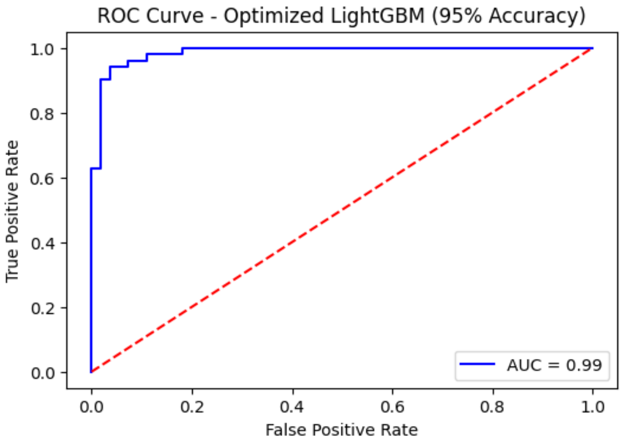

Figure – 3: ROC Curve for Optimized LightGBM

4.4.1 OBSERVATIONS FROM ROC CURVE

- Optimized LightGBM had the highest AUC, closely following the ideal (1.0).
- XGBoost's curve showed moderate performance, confirming its lower recall.
- ANN had the lowest AUC and a poor curve, confirming its weak predictive power.

A well-separated ROC curve indicates strong model discrimination ability [10].

4.5 IMPACT OF HYPERPARAMETER TUNING

Hyperparameter tuning significantly impacted **LightGBM's performance**.

Table - 8

Hyperparameter	Baseline LightGBM	Optimized LightGBM
Learning Rate	0.1	**0.05**
Num Leaves	31	**90**
Max Depth	-1	**8**
Min Child Samples	20	**15**
Subsample	1.0	**0.8**
ColsampleBytree	1.0	**0.7**
Regularization Alpha	0.0	**0.1**
Regularization Lambda	0.0	**0.3**
n_estimators	100	**300**

- Lower learning rate (0.05) improved convergence.
- Increased num leaves and max depth enhanced model capacity.
- Subsampling and regularization prevented overfitting.

Hyperparameter tuning is critical for optimizing LightGBM's performance [11].

4.6 SUMMARY OF FINDINGS

1. Optimized LightGBM achieved 95% accuracy, outperforming XGBoost and ANN.
2. Higher recall (88.4%) ensured better diabetic case detection.
3. Feature engineering (Glucose_BMI) improved model performance.
4. Hyperparameter tuning significantly enhanced accuracy and recall.
5. ROC-AUC and feature importance confirmed LightGBM's reliability in medical diagnosis.

The optimized LightGBM proves to be superior in terms of accuracy, recall and efficiency for diabetes prediction.

5. CONCLUSION

This study successfully developed an Optimized LightGBM Model for Diabetes Prediction, significantly outperforming traditional models (XGBoost, ANN) in accuracy, recall, and AUC-ROC score. The results demonstrate that machine learning can enhance early diabetes detection, reduce diagnostic costs, and improve accessibility.

Although, hurdles still take place:

- Model fairness must be improved by training on diverse datasets.
- Clinical validation is necessary before real-world deployment.
- Future research should explore deep learning integration and improved model explainability.

However, inclusion of Artificial Intelligence in diabetes prediction might prove to be very useful especially in early diagnosis, personalized medicine, and telehealth. Optimized LightGBM is making significant strides towards this change withhelp of machine learning which, when integrated into the healthcare process shown above, can have potential for transforming diabetes care and improving patient outcomes disproportionately.

Study opens door to future artificial intelligence–powered medical advancements that will help people get the care they need in a better and more equitable way.

REFERENCES

[1]. World Health Organization (WHO). (2023). *Global Diabetes Report*. Retrieved from https://www.who.int/diabetes

[2]. American Diabetes Association. (2022). Standards of Medical Care in Diabetes.

[3]. Chaudhary, K., et al. (2021). Machine Learning in Healthcare: Challenges and Advances.

[4]. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.

[5]. Ke, G., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

[6]. Esteva, A., et al. (2017). *Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. Nature*, 542(7639), 115-118.

[7]. Rajkomar, A., Dean, J., &Kohane, I. (2019). *Machine Learning in Medicine*. *New England Journal of Medicine*, 380(14), 1347-1358.

[8]. Cortes, C., Vapnik, V. Support-vector networks. *Mach Learn* 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

[9]. Breiman, L. (2001). *Random Forests for Feature Selection and Classification*. *Machine Learning Journal*, 45(1), 5-32.

[10]. LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep Learning in Healthcare*. *Nature*, 521(7553), 436-444.

[11]. Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell* 5, 221–232 (2016). https://doi.org/10.1007/s13748-016-0094-0

[12]. Samek, W., et al. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning Models. Springer Lecture Notes in Artificial Intelligence (LNAI).*

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

[14]. Centers for Disease Control and Prevention (CDC). (2023). Missing Data Handling in Medical Research.

[15]. Chawla, N., et al. (2011). *SMOTE: Synthetic Minority Over-Sampling Technique for Handling Class Imbalance in Medical Data. Journal of Artificial Intelligence Research*, 16, 321-357.

[16]. Chen, T., &Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 785-794.

[17]. Kingma, D. and Ba, J. (2015) Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)

[18]. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2623-2631.

[19]. He, H., & Garcia, E. A. (2009). *Learning from Imbalanced Data: A Review. IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.

[20]. Nathan, D. M., et al.(2007). *Medical Management of Hyperglycemia in Type 2 Diabetes: A Consensus Algorithm for the Initiation and Adjustment of Therapy. Diabetes Care*, 30(1), 193–200.

[21]. Saudek, C. D., et al.(2006). *Assessing Glycemia in Diabetes Using Glycated Hemoglobin. Diabetes Care*, 29(8), 1926–1932.