Vol.2 No.2 October 2025: 69-74
International Journal of Computer Science and Humanitarian AI (IJCSHAI)
e-ISSN: 3064-4372
DOI: 10.21512/ijcshaijournal.v2i2.14417

Implementation of Random Forest Algorithm in Handling Imbalanced Data: A Study on Default Models and Hyperparameter Tuning

Ivan William Lianata[1], Kang Nicholas Darren Nugroho[2], Yosua Nathanael[3], Neilson Christopher[4], Edy Irwansyah[5*]

[1-4]Data Science Program, School of Computer Science,
[5]Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
ivan.lianata@binus.ac.id; kang.nugroho@binus.ac.id; yosua.nathanael@binus.ac.id;
neilson.christopher@binus.ac.id; eirwansyah@binus.edu

Abstract - The healthcare industry has benefited greatly from the quick development of artificial intelligence, especially machine learning (ML). Unbalanced data is a significant problem in medical classification, as it can impair model performance, particularly when it comes to identifying important minority classes like patients with particular diseases. The purpose of this research is to evaluate how well two ensemble-based algorithms—Random Forest and Gradient Boosting—perform when dealing with data imbalance in diabetes prediction. Age, body mass index, smoking history, HbA1c level, blood glucose level, and other demographic and medical variables are included in the dataset, which was acquired from Kaggle. Data preprocessing, train-test splitting, model implementation with default parameters, and hyperparameter tuning with Grid Search and Cross Validation comprise the methodology. Accuracy, precision, recall, F1-score, and AUC-ROC metrics were used to assess the model's performance. Both models achieved high accuracy above 97%, according to the results. Following tuning, Random Forest achieved 97.06% accuracy, 0.974 AUC, and 0.99 positive-class precision; however, recall somewhat declined, possibly resulting in underdiagnosis. Gradient Boosting, on the other hand, showed consistent performance with an AUC of 0.9791 and an F1-score of 0.81. These results demonstrate that model performance can be enhanced by hyperparameter tuning; however, algorithm selection should be based on the needs of the application, especially in medical settings where striking a balance between sensitivity and diagnostic precision is crucial.

Keywords: Imbalanced Data; Random Forest; Gradient Boosting; Hyperparameter Tuning; Diabetes Prediction

Received: Sep. 23, 2025; received in revised form: Nov. 04,2025; accepted: Nov. 04,2025; available online: Nov. 04,2025.

**Corresponding: eirwansyah@binus.edu*

I. INTRODUCTION

The advancement of Artificial Intelligence technology has driven progress across various domains, including healthcare, finance, transportation, education, and environment. A key component of AI is Machine Learning (ML), a method enabling computers or machines to learn from data and make predictions or decisions automatically. Generally, ML is categorized into three main types: unsupervised, supervised, and reinforcement learning [1]. Unsupervised learning detects hidden patterns or trends in unlabeled data, whereas supervised learning finds patterns from previously labeled data. Meanwhile, reinforcement learning focuses on decision-making by considering past rewards and penalties.

Recently, supervised learning ML has become ubiquitous in everyday applications such as email spam filters, fraud detection, face recognition, medical diagnosis, house price prediction, and monthly product sales forecasting, among others [2]. However, supervised learning can be further classified into regression and classification. Regression is a statistical analysis used to model the relationship between a dependent variable (to be predicted) and one or more independent variables (predictors). Classification is a method to categorize data into several classes or labels based on their characteristics.

Classification inherently faces several challenges, including missing data, outliers, and imbalanced data [2]; [1]. Imbalanced data frequently occurs in real-world scenarios and represents a major problem compared to missing data or outliers, which are easier to handle. It refers to a condition where the data distribution among classes is uneven, causing models to bias toward the

majority class over the minority class [3]. This imbalance can have severe consequences, especially when the minority class contains critical data, potentially yielding low classification accuracy for such cases, as exemplified by fraud transaction detection or disease diagnosis.

In this context, ML plays an essential role in early disease diagnosis to support the Sustainable Development Goals (SDGs), particularly SDG 3 (Good Health and Well-Being), indicator 3.4.1, aiming to reduce premature mortality from non-communicable diseases such as cardiovascular disease, cancer, diabetes, and chronic respiratory diseases through prevention, treatment, and mental health promotion by 2030. Using ML models to predict potential diabetes based on patient medical data can accelerate early detection and enable faster, more accurate intervention. This not only supports clinical decision-making but also helps allocate healthcare resources more efficiently to individuals at high risk.

There are many classification methods to develop models, such as logistic regression, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Neural Networks, Naive Bayes, Random Forest, and Gradient Boosting [4]. Decision-tree based models like Random Forest and Gradient Boosting are generally better suited for imbalanced data [5]. Based on the above problem description, the objectives of this study are to measure the performance of Random Forest algorithm on imbalanced datasets and those tuned with hyperparameter optimization, and identify the robust across varying imbalance ratios in the dataset.

II. PREVIOUS WORKS

Several previous studies have evaluated ensemble methods such as Random Forest and various Gradient Boosting variants (XGBoost, LightGBM, CatBoost) on classification data, particularly in healthcare fields challenged by imbalance [6]. Bentéjac et al. compared CatBoost, LightGBM, XGBoost, and Random Forest on multiple datasets, concluding that CatBoost achieved the best accuracy and AUC, while LightGBM excelled in training speed; XGBoost ranked mid-tier in performance and efficiency [7].

Das et al. (2020) highlighted the importance of handling class imbalance in medical data [8]. Using SMOTE and stratified k-fold cross-validation [9], they compared Random Forest with Gradient Boosting variants. Gradient Boosting generally outperformed in F1-score and AUC, though imbalance treatment effects varied by dataset and parameters. Adaptive loss functions such as weighted loss and focal loss further improved minority class detection performance [10].

In adapting to class imbalance, Fulazzaky et al. (2020) evaluated balancing approaches like Balanced Random Forest (BRF) and SMOTE-RF across 13 binary datasets, noting BRF consistently achieved better balanced accuracy and recall compared to normal Random Forest, emphasizing the need for model adaptation to unequal class distributions [2]

Florek (2020) emphasized hyperparameter tuning for Gradient Boosting, concluding that effectiveness, reliability, and ease of model use depend heavily on tuning strategy [11]. Though not directly compared to Random Forest, results suggest Gradient Boosting performance can significantly improve with optimal tuning [12].

Further research stresses the importance of model interpretability to understand classification decisions, particularly critical in medicine. Methods like SHAP values facilitate such interpretation [13] (Tabel 2.1).

Table 1. Previous Research Related to Algorithm Evaluation on Imbalanced Datasets

Title	Author	Year	Dataset	Objective	Algorithm	Accuracy
A Comparative Analysis of Gradient Boosting Algorithms	Bentéjac et al. [6]	2020	Various public datasets (not specifically mentioned)	Compare the performance of XGBoost, LightGBM, CatBoost, and Random Forest	XGBoost, LightGBM, CatBoost, Random Forest	CatBoost best, LightGBM fastest
Evaluating Ensemble Models on Imbalanced Data Sets	Das et al. [8]	2020	Healthcare dataset with class imbalance	Evaluate ensemble methods in handling imbalance	SMOTE, stratified k-fold, Random Forest, CatBoost, LightGBM	Gradient Boosting superior (AUC, F1)
Evaluating Ensemble Learning	Fulazzaky et al. [15]	2020	13 secondary	Assess ensemble methods adapted for	Balanced Random Forest (BRF), SMOTE-	BRF outperforms

Techniques for Class Imbalance			binary datasets	imbalanced data	RF, Random Forest	in balanced accuracy
Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Imbalanced Classification Problems	Wang et al.[10]	2020	Parkinson's disease classification dataset	Improve XGBoost performance on imbalanced data	XGBoost with weighted loss & focal loss (Imbalance-XGBoost)	State-of-the-art (value not reported)
Benchmarking State-of-the-Art Gradient Boosting Algorithms	Florek [11]	2020	Various real-world datasets (not detailed)	Assess the effectiveness and efficiency of hyperparameter tuning in Gradient Boosting	Hyperparameter tuning for XGBoost, CatBoost, LightGBM	Not directly compared

III. PROPOSED METHOD

This study uses a quantitative approach. The data employed is a diabetes prediction dataset obtained from Kaggle. There are two types of variables: dependent variable (X) and independent variables (y). The dependent variable represents diabetes classification. The independent variables are: gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level.

A. Preprocessing Data dan Exploratory Data Analysis

The study was conducted using the Python programming language and several libraries such as pandas, seaborn, sklearn, matplotlib, and warnings. The first step was preprocessing and exploratory data analysis (EDA) to clean noise and gain more insights. Data columns irrelevant to the target variable were removed. Checks were done for missing values (NA) and duplicates; no missing values were found, but duplicates were removed.

The gender column contained a category "Other," changed to "Female" to simplify classification using the mode. Gender was then encoded into binary format: "Male" as 0 and "Female" as 1. The smoking_history column was one-hot encoded to convert categorical data into numeric formats suitable for modeling. Data types were verified to ensure all columns were numeric before modeling. Distributions of gender, hypertension, heart disease,

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

and smoking history were visualized against diabetes status. Age and diabetes relationships were also visualized using age groups. A heatmap was created to reveal correlations among numeric variables.

B. Preprocessing, Modeling, dan Tuning

Further preprocessing was performed after data splitting into train and test sets to avoid data leakage, thereby reducing overfitting or underfitting. Numeric and categorical features were encoded to numeric values to facilitate modeling. Random Forest Classifier models were utilized from the sklearn.ensemble library. Models were trained on training data and tested on test data to enhance minority class prediction accuracy.

Hyperparameter tuning was conducted using GridSearchCV or RandomSearch, starting with default parameters. For example, Random Forest training and prediction were done with the RandomForestClassifier function, with subsequent comparisons made.

C. Evaluation

After data processing, model performance evaluation was conducted to estimate generalization error and assess each model's effectiveness.

The confusion matrix was used to identify misclassifications. Classification reports calculated F1-score, accuracy, precision, and recall for each model. Additionally, AUC and ROC curves assessed the models' ability to differentiate between positive and negative classes.

The F1-score combines precision and recall as the harmonic mean, providing a balanced measure, especially useful for imbalanced datasets where accuracy might be misleading (Kroese et al., 2024). Although F1-score is less intuitive than accuracy, it offers a balanced view on false positives and false negatives.

ROC curves visualize model performance at various thresholds by plotting true positive rate (recall) against false positive rate. The curve's

proximity to the top-left corner indicates a better model. The ideal threshold is usually selected based on validation data to prevent overfitting.

The Area Under the ROC Curve (AUC) quantifies overall model performance on a scale from

$$FPR = \frac{FP}{FP+TN}$$

0 to 1. A higher AUC signifies better distinction between classes. An AUC of 1 means perfect ranking of positives above negatives; 0.5 corresponds to random guessing. AUC is especially valuable for imbalanced datasets as it remains unbiased by class proportions.

IV. Experiment Result

A. *Evaluation of Random Forest Model*

This study compares two classification algorithms: Random Forest and Gradient Boosting, testing each to determine which is better suited for imbalanced data. Grid Search with 3-fold cross-validation identified the best hyperparameter combination based on AUC scores. AUC was chosen to minimize false negatives, critical in predicting diabetes accurately The default Random Forest model performed very well detecting the negative class (non-diabetic patients), with a recall of 1.00 — meaning almost all non-diabetics were correctly identified. The confusion matrix shows 17,437 true negatives and only 72 false positives (healthy individuals misclassified as diabetic). This indicates minimal over-diagnosis risk.

However, the model struggled detecting diabetes patients with only 0.69 recall—detecting about 69% of actual diabetics while missing 542 false negatives. In medical contexts, false negatives (under-diagnosis) pose serious risks since patients remain undiagnosed. Precision for the positive class was good at 0.94, meaning predictions labeled diabetic had 94% correctness. Overall, the model showed a good balance between accuracy (0.968) and AUC (0.9636) but left room for improvement, notably in sensitivity to diabetics.

Table 2. Evaluation Result Before Hyperparameter Tuning

	Predicted Negative	***Predicted Positive***
Actual *Negative*	17437	72
Actual *Positive*	542	1179

	Precision	***Recall***	**F1-Score**
Negative class	0.97	1.00	0.98
Positive class	0.94	0.69	0.79

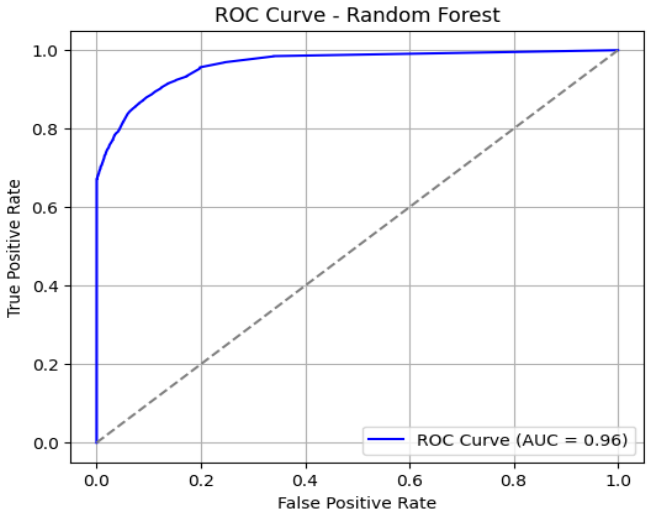


Figure 1. ROC Curve of Random Forest Before Hyperparameter Tuning

Post hyperparameter tuning with Grid Search showed improvements: overall accuracy rose to 97.06%, and AUC increased to 0.9741, indicating better distinction between diabetic

and non-diabetic patients. Recall for negatives remained near perfect (1.00), with only 8 false positives out of 17,509 non-diabetics, meaning minimal over-diagnosis

For positive cases, recall slightly dropped to 0.68, detecting 68% of diabetics, with false negatives rising to 557. This decline suggests the model became more conservative, only labeling diabetes when confident. Precision for positives increased to 0.99—meaning 99% of positive predictions were correct. The F1-score rose to 0.80, indicating a better precision-recall balance despite the recall decrease. This model might be more suitable for scenarios aiming to minimize false positive diagnoses at some expense of missed positives.

Table 3. Evaluation Result After Hyperparameter Tuning

	Predicted Negative	*Predicted Positive*
Actual *Negative*	17501	8
Actual *Positive*	557	1164

	Precision	*Recall*	F1-Score
Negative class	0.97	1.00	0.98
Positive class	0.99	0.68	0.80

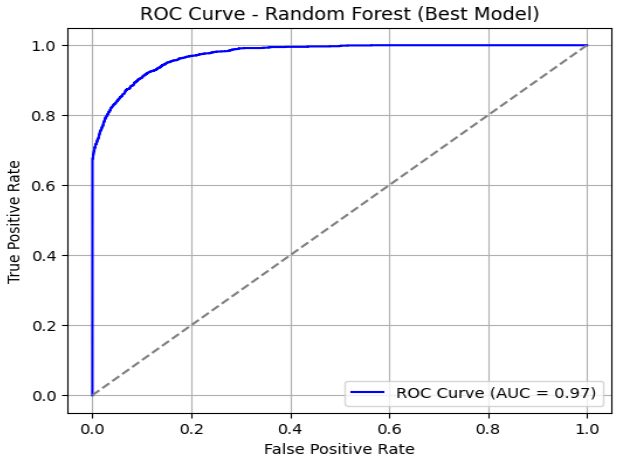
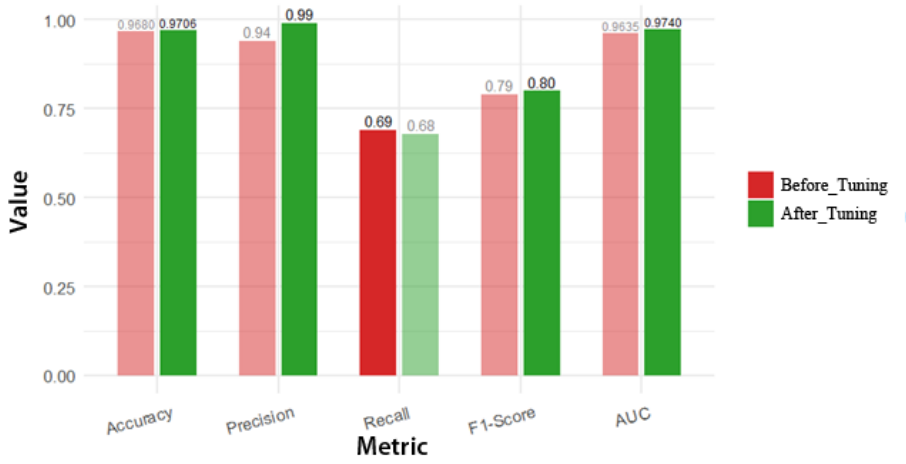


Figure 2. ROC Curve of Random Forest After Hyperparameter Tuning

Summarizing performance changes before and after tuning: accuracy increased from 96.81% to 97.06%, AUC rose from 0.9635 to 0.9740, and precision for positives dramatically improved from 0.94 to 0.99. The F1-score also showed a slight improvement suggesting a more balanced model. However, recall of diabetes patients decreased, indicating more false negatives.

Table 4. Comparison of Random Forest Performance

Metric	**Before Tuning**	**After Tuning**
Accuracy	0.97082	**0.97087**
Precision (*positive*)	**0.99**	0.98
Recall (*positive*)	0.68	**0.69**
F1-Score (*positive*)	0.81	0.81
AUC	0.9785	**0.9791**



V. CONCLUSION

Tuning in Random Forest Algorithm improved precision and F1-score but slightly decreased recall, making the model more conservative and accurate in positive diagnoses but with reduced sensitivity. This shift strengthens the model's intrinsic characteristics, providing a more cautious positive diagnosis.

Overall, the tuned Random Forest model achieved excellent performance with accuracy over 97%, and an AUC score of 0.974 denotes strong discriminative ability between diabetic and non-diabetic individuals across thresholds. Precision for positives reached 0.99, indicating that positive predictions are rarely false alarms.

Despite the slight recall reduction, the increase in precision and overall accuracy suggest that hyperparameter tuning successfully enhanced model reliability and applicability for medical diagnosis where false positives are costly

REFERENCES

[1] Theobald, O. (2021). Machine learning for absolute beginners: A plain English introduction (Third edition). Scatterplot Press.

[2] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[4] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[5] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249-259

[6] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937–1967. https://link.springer.com/article/10.1007/s10462-020-09896-5

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

[8] Das, A., & Roy, S. (2024). Evaluating ensemble models on imbalanced data sets: A comparative study across varied minority class ratios. ResearchGate. https://www.researchgate.net/publication/379484244_Evaluating_Ensemble_Models_on_Imbalanced_Data_Sets_A_Comparative_Study_across_Varied_Minority_Class_Ratios

[9] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[10] Wang, C., Deng, C., & Wang, S. (2019). Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recognition Letters, 136, 190–197.

[11] Florek, P., & Zagdanski, A. (2023). Benchmarking state-of-the-art gradient boosting algorithms for classification. arXiv preprint arXiv:2305.17094. https://arxiv.org/abs/2305.17094

[12] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer Science & Business Media.

[13] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, 30.

[14] Kroese, D. P., Botev, Z. I., Taimre, T., & Vaisman, R. (2024). Data science and machine learning: Mathematical and statistical methods. The University of Queensland. https://people.smp.uq.edu.au/DirkKroese/DSML/DSML.pdf

[15] M. A. Fulazzaky, A. R. Setiawan, and D. P. Hidayat, “Evaluating Ensemble Learning Techniques for Class Imbalance,” Applied Artificial Intelligence, vol. 34, no. 7, pp. 561–574, 2020, doi: 10.1080/08839514.2020.1742678

.