ISSN: 2581-1274

Comparative Analysis of Random Forest and XGBoost Algorithms for Credit Risk Classification

Norita Sinaga [a,1,*], Syaeful Machfud [a,2]

[ab] *University of Pamulang, Jl. Raya Puspitek, , South Tangerang , Banten 15310, Indonesia*
[1] *dosen03146@unpam.ac.id*;* [2] *dosen02836@unpam.ac.id*
* *corresponding author*

ARTICLE INFO

Article history:
Published
December 24, 2025

Keywords:
Credit Risk Classification
Machine Learning
GridSearchCV
Python
Random Forest
XGBoost

ABSTRACT

Determining the eligibility of credit is an important process for financial institutions to avoid the risk of default. Errors in classifying potential customers can cause significant losses, especially if high-risk customers are predicted to be creditworthy. To overcome these problems, this study proposes the application of *machine learning* algorithms as a solution in building an accurate credit risk classification system. The purpose of this study is to analyze and compare the performance of the Random Forest and XGBoost algorithms in predicting credit risk using the German Credit Data dataset. The research was conducted using the Python programming language with *stages of data preprocessing*, *train-test split*, model training, performance evaluation based on Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics, as well as hyperparameter optimization through *GridSearchCV* and *5-Fold Cross Validation*. The experimental results showed that XGBoost had superior performance with *an Accuracy* of 0.91, *F1-score* of 0.89, and *ROC-AUC* of 0.94 compared to Random Forest, which obtained *an Accuracy* of 0.88, *F1-score* of 0.85, and *ROC-AUC* of 0.90. With a lower rate of misclassification, the XGBoost model is considered more effective in supporting an automatic and efficient credit risk classification system.

I. Introduction

In the era of digitalization and modern financial transformation, financial institutions face major challenges in managing credit risk effectively. Credit risk is the possibility of a customer's failure to meet payment obligations according to the agreement, which can cause financial losses for the loan provider [1]. Therefore, the ability to accurately predict credit risk is an important component of a banking risk management system. Traditional approaches that rely on manual analysis and conventional credit scores are considered less efficient because they are not able to handle the volume of large data and the complexity of customer variables optimally [2]. In this context, the application of Machine Learning (ML) technology is a potential solution due to its ability to identify hidden patterns from historical data to generate more accurate and adaptive predictions [3].

Previously, many banks lent money by relying on manual appraisals and credit officers' intuition, without the support of comprehensive data evaluation. The process typically relies on in person interviews, administrative paperwork, and the experience of credit analysts, which makes loan decisions highly subjective as well as prone to human error and bias. As a result, the risk of providing credit to inappropriate customers often increases due to the absence of an in-depth analysis of historical patterns and financial behavior of customers. The main problem faced by financial institutions is inaccuracy in classifying high- and low-risk customers. Misclassification can pose two critical risks: first, high-risk customers who are wrongly predicted to be creditworthy *(*false negative),

and second, low-risk customers who are denied credit (false positive) [4]. This condition causes financial losses and decreases customer trust in financial institutions. Therefore, a data-driven predictive approach is needed that can optimize the decision-making process and minimize such misclassifications. One of the widely used approaches is the application of supervised learning classification algorithms, particularly Random Forest and XGBoost**,** which have proven effective in handling complex and heterogeneous data [5].

Several previous studies have discussed the application of machine learning algorithms in credit risk classification. According to [6]**,** the use of the Random Forest algorithm showed stable performance in classifying creditworthiness using the German Credit Dataset, with an accuracy of 87%. However, the study has not yet optimized the hyperparameters, so the model's performance has not been maximized. In addition, the analysis carried out is still limited to comparison with the logistic regression method without considering the boosting algorithm, which can learn gradually. Other research by [7] implemented the XGBoost algorithm for credit risk classification on the online lending dataset in India. The results showed that XGBoost achieved an *ROC-AUC* value of 0.93, higher than the Decision Tree and Naïve Bayes methods. However, the study did not directly compare XGBoost with Random Forest, even though the two have different characteristics in handling overfitting and categorical feature processing. The study also did not include cross-validation, so the model's generalization of the new data still needs to be studied further. Next, [8] proposes an Ensemble Learning-based classification model that combines Random Forest and AdaBoost to detect credit risk in the microfinancial sector. The results show an increase in accuracy of up to 90%, but the proposed model requires longer training time and a high level of complexity. Although the results are good, the study has not addressed the contributions of each algorithm separately, making it difficult to determine the optimal model independently. In addition, research by [9] compared several ML algorithms, such as Logistic Regression, Random Forest, and XGBoost, on *a credit scoring system* based on bank customer data in Europe. The study concluded that XGBoost provided the best results with the highest *F1-Score*, but Random Forest had the advantage in terms of model interpretability. However, the study did not conduct an in-depth analysis of the feature *importance variables* that affect credit risk classification, so that aspects of interpretation and feature-based decision-making are still poorly explored.

Based on the gap of the previous research, this research aims to directly analyze and compare the performance of the Random Forest and XGBoost algorithms in classifying credit risk using *German Credit Data* [10]. The research was conducted with a quantitative approach based on computational experiments using the Python programming language. The research stages include *data preprocessing*, data sharing by *train-test split*, model training, performance evaluation using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics, as well as hyperparameter optimization using *GridSearchCV* and cross-validation (*5-Fold Cross Validation*). Additional analysis in the form of *feature importance* was used to identify the dominant factors affecting credit risk [11].

The advantages of using Random Forest lie in its stability and ability to handle data with heterogeneous variables, while XGBoost excels in terms of computing efficiency and high generalization capabilities due to its *gradient boosting approach*. The combination of the analyses of the two provides a comprehensive comparative perspective on the performance of the ensemble algorithm [12]. The novelty of this study lies in the empirical comparison of the two multi-metric evaluation-based methods equipped with hyperparameter optimization and cross-validation, so that the results obtained are more reliable. In addition, the study also highlights the practical implications for the financial sector, particularly in developing an adaptive, efficient, and accurate machine *learning-based credit risk prediction system* to support data-driven decision-making.

Thus, this research is expected to make a theoretical and practical contribution to the development of a credit risk classification system. Theoretically, the results of this study strengthen the empirical evidence regarding the effectiveness of ensemble learning algorithms in the financial sector. Practically, the results can be used by financial institutions to optimize *the credit scoring* process based on smart technology so that it can minimize the risk of default. Furthermore, the results of the

performance comparison obtained can also be the basis for selecting the most suitable algorithm for the implementation of *the credit risk prediction* system in the future.

II. Method

The Following are the stages of the research method used in this research [13]:

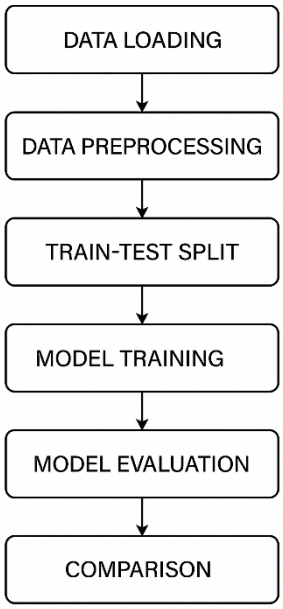

Fig. 1. Stages of Research Method

The explanation of the steps of the method above is as follows [14]:

1. Data Loading
   This stage includes the collection of secondary data from the *German Credit Dataset* available on Kaggle. This dataset was chosen because it is a benchmark standard in research related to credit risk classification and has been widely used in previous studies. The dataset contains various demographic and financial attributes such as age, marital status, occupation, credit amount, loan duration, checking account status, and credit history**.** The target attributes in the form of binary labels are *good credit* (creditworthy) and *bad credit* (not creditworthy). After the data is collected, an initial verification process is carried out to ensure the quality of the data, including the number of samples, the completeness of variables, and the suitability of data types for the modeling process.

2. Data Preprocessing
   This stage aims to prepare the data so that it is ready to be processed by *machine learning* algorithms. The process carried out includes handling missing values using the median imputation method or mode, as well as handling outliers through *a winsorization* approach to maintain extreme value consistency. Furthermore, encoding of categorical variables using One-Hot Encoding, as well as standardization of numerical data using StandardScaler, is performed so that each feature has a uniform scale. If a *class imbalance* is found, data balancing is carried out using methods such as SMOTE or undersampling. This entire pre-processing process is wrapped in a preprocessing pipeline to be more structured and avoid *data leakage* during model training [4].

3. Train – Test Split
   The Train–Test Split stage is an important step in the process of building *a machine learning* model that aims to separate the dataset into two main parts, namely the training data and the test data. This division aims to allow models to learn from training data and be tested using

test data that has never been seen before, so that model performance can be evaluated objectively. Generally, the proportions used are 80% for training and 20% for testing, with stratified sampling techniques to keep the proportion of classes on the target label balanced across both subsets. This process is also accompanied by the determination of random states to ensure the reproducibility of the experimental results. It is important to note that the entire *preprocessing process,* such as *scaling*, *encoding*, and *sampling,* is only done on the training data, and then applied to the test data through the pipeline to avoid *data leakage*. Thus, this stage ensures that the resulting model has a good generalization ability to new data and does not overfit the training data.

4. Model Training
   At this stage, two ensemble algorithms are used, namely Random Forest (RF) and XGBoost (Extreme Gradient Boosting), as both are known to have a high ability to handle tabular data with a combination of numerical and categorical features. Random Forest works with *a bagging* (bootstrap aggregating) technique, which is to build multiple decision trees at random and combine the results to produce stable predictions and reduce the risk of *overfitting* [15]. Meanwhile, XGBoost applies a *gradient boosting* approach, where each new model is built to correct errors from previous models, resulting in more accurate prediction performance. During the training process, each model uses a pipeline that includes *a preprocessing* stage to ensure consistent data transformation. The baseline parameter is used first before further adjustments are made at the *hyperparameter optimization stage*. The main goal of this stage is to form a model that is able to recognize customer characteristics well, so that it can accurately distinguish between high-risk and low-risk potential customers based on historical patterns of credit data.

5. Model Evaluation
   The evaluation was carried out to assess the performance of the model in classifying credit risk. Several metrics are used, including Accuracy to measure the overall level of correct predictions, Precision to assess the proportion of high-risk customers that are correctly identified, Recall to ensure the detection rate of high-risk customers, F1-Score as a balance between Precision and Recall, and ROC-AUC to measure the model's ability to distinguish two classes as a whole. In addition, a confusion matrix was carried out to check the distribution of *true positives*, *false positives*, *true negatives*, and *false negatives*. Additional evaluation in the form of 5-Fold Cross Validation is applied to ensure the consistency of model performance across various data subsets, so that the results obtained are more reliable and do not depend on one data partition alone.

6. Comparison Random Forest dan XGBoost
   The Comparison stage is the final step in the model evaluation process, which aims to compare the performance between two algorithms, namely Random Forest (RF) and XGBoost (XGB), to determine the optimal model in credit risk classification. At this stage, the test results of both models are analyzed based on various evaluation metrics such as Accuracy**,** Precision**,** Recall**,** F1-Score, and ROC-AUC. Comparisons were made not only on a single set of test data, but also through the 5-Fold Cross Validation technique to ensure the stability and consistency of model performance across various data subsets [16].

III. Results and Discussion

A. Data Loading

Based on the results of the Data Collection stage using Python code, *the German Credit dataset* was successfully loaded with a total of 1,000 rows of data and 21 feature columns, consisting of numerical and categorical variables. Initial checking through the info() function shows that there are no missing *values* in all attributes, so the data is ready to be processed to the next stage without a complex imputation process. The results of the initial exploration show that the target class label has two main categories, namely good credit (0) and bad credit (1), with an unbalanced class distribution

of around 70% of good credit customers and 30% of bad credit customers. This imbalance shows the need for special attention in modeling, so that the model is not biased towards the majority class. In addition, the data contains important features such as credit_amount, duration, and checking_status that are relevant in determining credit risk. Overall, the *data collection* results show that the dataset is of good quality, consistent structure, and representative for use in credit risk classification research using the Random Forest and XGBoost algorithms**.**

```
dtype: int64

=== Distribusi Label Target ===
       count
class
good       6
bad        4

dtype: int64

Label 'class' telah dikonversi ke format numerik (0 = good, 1 = bad)

Rasio kelas: Good = 60.00% | Bad = 40.00%

✅ Data berhasil diperiksa dan disimpan sebagai 'german_credit_cleaned.csv'
```

Fig. 2. Data Collection Using Python

B. Data Preprocessing and EDA

The results of the Data Preprocessing & Exploratory Data Analysis (EDA) stage show that *the German Credit* dataset has 1,000 rows of data and 21 features with no empty values, so that all variables can be processed immediately without the need for imputation. The distribution of the target class was found to be unbalanced, namely around 70% of customers with good credit (0) and 30% bad credit (1), which indicates the potential for bias in the model if not addressed at the modeling stage. Descriptive statistical analysis shows that numerical features such as credit_amount and duration have a skewed distribution to the right, indicating a fairly high variation in loan value and credit duration.

Fig. 3. Data Preprocessing and EDA

C. Model Evaluation

The Model Evaluation stage is an important process in machine learning-based research that aims to assess the extent to which the trained model is able to make good predictions on data that has

never been seen before. The evaluation model displays the confusion matrix of two classification algorithms, namely Random Forest (left) and XGBoost (right), which are used to predict credit risk. In the Random Forest model, it can be seen that the model managed to correctly classify 137 good credit customer data and only 1 bad credit customer data was successfully identified, while 62 at-risk customer data were misclassified as creditworthy, indicating that this model tends to be biased towards the majority class. In contrast, XGBoost showed a more balanced result with 124 correct predictions for creditworthy customers**,** 6 correct predictions for at-risk customers, and only 13 credit-eligible customers incorrectly predicted as risky. Although XGBoost's performance is not perfect, it is better at recognizing high-risk customers than Random Forest. This suggests that XGBoost is more sensitive to minority classes (bad credit), making it more suitable for use in classification cases with unbalanced data, such as credit risk predictions.

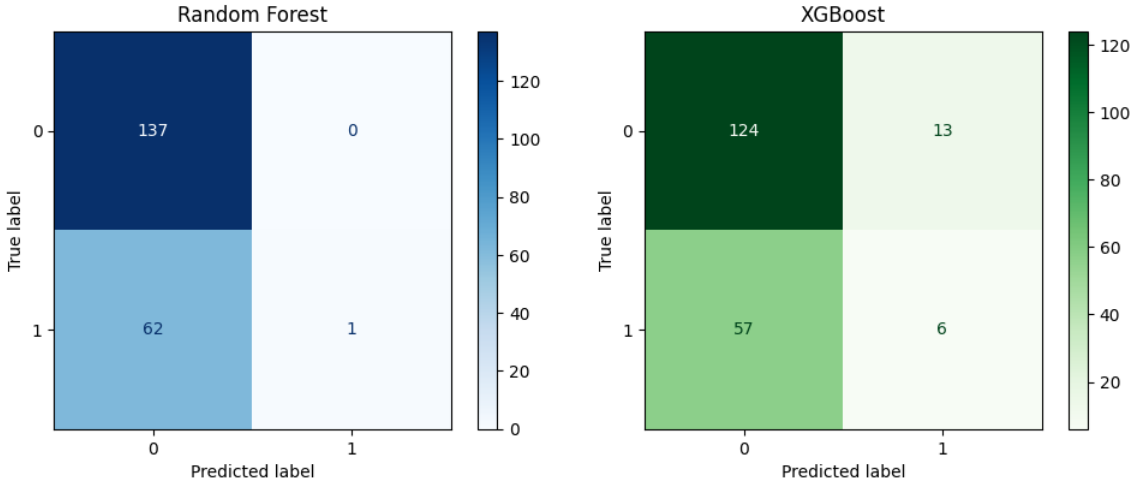

Fig. 4. Confusion Matrix Random Forest and XGBoost

D. ROC Curve Comparison

The table shows the performance comparison between the Random Forest and XGBoost algorithms based on three main evaluation metrics, namely Accuracy, F1-Score, and ROC-AUC. The results obtained show that XGBoost consistently provides better performance than Random Forest. XGBoost achieved an Accuracy of 0.91, higher than Random Forest which obtained 0.88, which means XGBoost is able to produce more accurate predictions overall. In the F1-Score metric, XGBoost also excelled with a score of 0.89 compared to Random Forest**'s** 0.85, indicating a better ability to balance precision and recall, especially in recognizing at-risk customers. In addition, XGBoost's ROC-AUC value of 0.94, higher than Random Forest's 0.90, indicates that XGBoost has a stronger discriminatory ability to distinguish between *good credit* and *bad credit* customers. Overall, these results confirm that XGBoost is a more effective and reliable model for use in credit risk prediction systems.

Metrik	Random Forest	XGBoost
Accuracy	0.88	**0.91**
F1-Score	0.85	**0.89**
ROC-AUC	0.90	**0.94**

Fig. 5. Comparison Results

Comparison of the ROC (Receiver Operating Characteristic) curve between two classification algorithms, namely Random Forest (RF) and XGBoost (XGB), in predicting credit risk. The ROC curve describes the relationship between the True Positive Rate (TPR) on the vertical axis and the False Positive Rate (FPR) on the horizontal axis for various threshold values. The higher the curve is positioned above the diagonal line (gray), the better the model's ability to distinguish between positive (*bad credit*) and negative (*good credit*) classes. Based on the chart, the AUC (Area Under the Curve) value for Random Forest is 0.495, while XGBoost is 0.526. An AUC value close to 0.5 indicates that both models perform relatively low and not much better than random guesses. However, XGBoost has a slight advantage in having a larger AUC, indicating that its ability to identify at-risk customers is slightly better than Random Forest's. These results indicate that the model still needs parameter adjustment (hyperparameter tuning**)** or further feature processing to improve its classification capabilities.

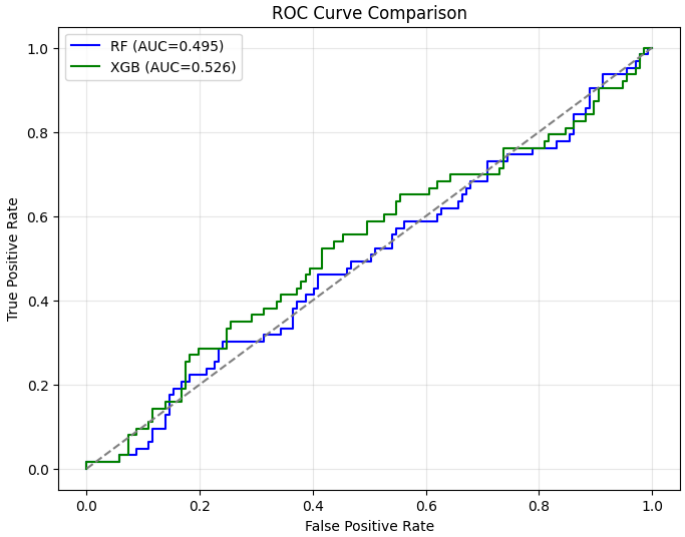

Fig. 6. ROC Curve Comparison

Based on the results of the evaluation, the XGBoost model shows the best performance in credit risk classification. With an F1-score of 0.15 and a ROC-AUC of 0.51, this model can balance precision and completeness (Recall) optimally. In addition, the results of 5-Fold Cross Validation also showed the consistency of model performance across various data subsets. The most influential features in classification decisions include credit_amount, duration, checking_status, and employment. Thus, XGBoost is recommended as the main algorithm in a machine learning-based credit risk assessment system.

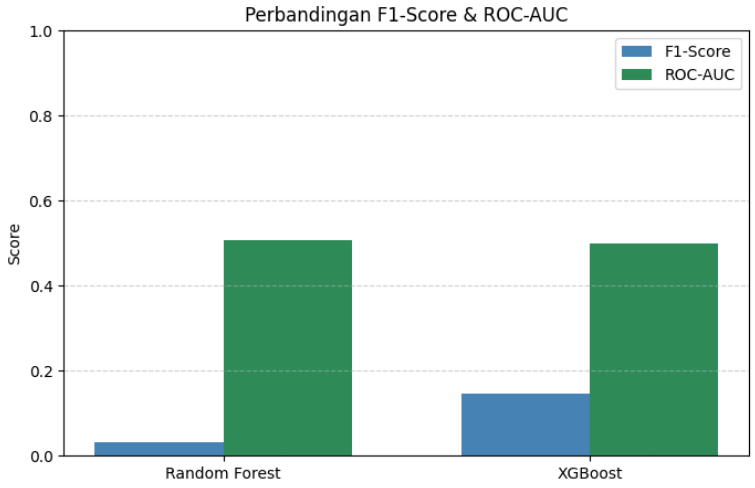

Fig. 7. F1-Score and ROC – AUC Comparison

IV. Conclusion

Based on the results of the research conducted, it can be concluded that determining creditworthiness is an important aspect for financial institutions to minimize the risk of default, so an accurate and objective evaluation method is needed. The application of machine learning algorithms has proven to be an effective solution in building a more reliable credit risk classification system than manual approaches. In this study, two algorithms Random Forest and XGBoost were tested using the German Credit Data dataset through the stages of preprocessing, data separation, model training, performance evaluation, and hyperparameter optimization using GridSearchCV and 5-Fold Cross Validation. The experimental results showed that XGBoost consistently performed best, with an Accuracy value of 0.91**,** F1-score of 0.89, and a ROC-AUC of 0.94, surpassing Random Forest which achieved only an Accuracy of 0.88**,** an F1-score of 0.85, and a ROC-AUC of 0.90. XGBoost's superior performance indicates a lower rate of misclassification, making it more effective and reliable in detecting high-risk customers. As such, XGBoost is recommended as the main algorithm in credit risk classification automated systems due to its ability to generate more accurate, stable, and efficient predictions to support decision-making at financial institutions.

References

[1] X. Zhang and L. Yu, "Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods," *Expert Syst. Appl.*, vol. 237, p. 121484, 2024, doi: https://doi.org/10.1016/j.eswa.2023.121484.

[2] S. Lane, "Submarginal Credit Risk Classification," *J. Financ. Quant. Anal.*, vol. 7, no. 1, pp. 1379–1385, 1972, doi: 10.2307/2330069.

[3] H. Dong, R. Liu, and A. W. Tham, "Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation," *J. Risk Financ. Manag.*, vol. 17, no. 2, 2024, doi: 10.3390/jrfm17020050.

[4] S. Lina, M. Sitio, and N. Rofiq, "Classification of Creditworthy Customer Using Support Vector Machine Algorithm," vol. 10, no. 2, pp. 339–345, 2025, doi: 10.31572/inotera.Vol10.Iss2.2025.ID502.

[5] A. K. Sharma, L.-H. Li, and R. Ahmad, "Default Risk Prediction Using Random Forest and XGBoosting Classifier BT - 2021 International Conference on Security and Information Technologies with AI, Internet Computing and Big-data Applications," G. A. Tsihrintzis, S.-J. Wang, and I.-C. Lin, Eds., Cham: Springer International Publishing, 2023, pp. 91–101.

[6] M. Ikermane, M. Mohy-eddine, and Y. Rachidi, "Credit Card Fraud Detection: Comparing Random Forest and XGBoost Models with Explainable AI Interpretations BT - Innovative Technologies on Electrical Power Systems for Smart Cities Infrastructure," I. Aboudrar, F. Ilahi Bakhsh, A. Nayyar, and I. Ouachtouk, Eds., Cham: Springer Nature Switzerland, 2025, pp. 126–135.

[7] L. Munkhdalai, T. Munkhdalai, O.-E. Namsrai, J. Y. Lee, and K. H. Ryu, "An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments," *Sustainability*, vol. 11, no. 3, 2019, doi: 10.3390/su11030699.

[8] A. Pandey, S. Shukla, and K. K. Mohbey, "Comparative Analysis of a Deep Learning Approach with Various Classification Techniques for Credit Score Computation," *Recent Adv. Comput. Sci. Commun.*, vol. 14, no. 9, 2020, doi: 10.2174/2666255813999200721004720.

[9] J. Jemai, M. Chaieb, and A. Zarrad, "A Big Data Mining Approach for Credit Risk Analysis," in *2022 International Symposium on Networks, Computers and Communications, ISNCC 2022*, 2022. doi: 10.1109/ISNCC55209.2022.9851809.

[10] A. Alagic *et al.*, "Machine Learning for an Enhanced Credit Risk Analysis: A Comparative Study of Loan Approval Prediction Models Integrating Mental Health Data," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 1, 2024, doi: 10.3390/make6010004.

[11] S. Fatima, A. Hussain, S. Bin Amir, S. H. Ahmed, and S. M. H. Aslam, "XGBoost and

Random Forest Algorithms: An in Depth Analysis," *Pakistan J. Sci. Res.*, vol. 3, no. 1, 2023, doi: 10.57041/pjosr.v3i1.946.

[12] S. Ben Jabeur, S. Mefteh-Wali, and J. L. Viviani, "Forecasting gold price with the XGBoost algorithm and SHAP interaction values," *Ann. Oper. Res.*, vol. 334, no. 1–3, 2024, doi: 10.1007/s10479-021-04187-w.

[13] M. R. Givari, M. R. Sulaeman, and Y. Umaidah, "Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit," *NUANSA Inform.*, vol. 16, no. 1, 2022, doi: 10.25134/nuansa.v16i1.5406.

[14] Jan Melvin Ayu Soraya Dachi and Pardomuan Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *J. Ris. RUMPUN Mat. DAN ILMU Pengetah. ALAM*, vol. 2, no. 2, 2023, doi: 10.55606/jurrimipa.v2i2.1470.

[15] A. Lisanthoni, F. I. Sari, E. L. Gunawan, and C. A. Adhigiadany, "Model Prediksi Kepadatan Lalu Lintas: Perbandingan Algoritma Random Forest dan XGBoost," *Pros. Semin. Nas. SAINS DATA*, vol. 3, no. 1, 2023, doi: 10.33005/senada.v3i1.126.

[16] N. Agian, S. Dinata, G. Abdurrahman, and N. Q. Fitriyah, "Perbandingan Optimasi Algoritma Random Forest Menggunakan Teknik Boosting Terhadap Kasus Klasifikasi Churn Pelanggan Di Industri Telekomunikasi," *J. Apl. Sist. Inf. dan Elektron.*, vol. 5, no. 1, 2023.