BAREKENG: Journal of Mathematics and Its Applications
March 2025 Volume xx Issue xx Page xxx–xxx
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengxxxxxxxxxxxxx

CLASSIFICATION OF INDONESIAN CORPORATE CREDIT RATING USING INTERPRETABLE MACHINE LEARNING MODEL

Teuku Rusydi Khairi [1*], **Alexander Agung Gunawan** [2], **Wawan Hermawan** [3]

[1]*Magister of Computer Science, Faculty of Computer Science, Bina Nusantara University*
[2]*Department of Computer Science, Faculty of Computer Science, Bina Nusantara University*
[3]*Department of Economic, Faculty of Economics, Padjajaran University*

Institution address, City, Post Code, Country

*Corresponding author's e-mail: * **teuku.khairi@binus.ac.id***

Article Info

Article History:

Received: date month year
Revised: date month year
Accepted: date month year
Available online: date month year

Keywords:

Alphabetic;
Barekeng journal;
Mathematics topic;

ABSTRACT

Accurate and interpretable corporate credit ratings are a critical pillar for financial market stability in Indonesia. Traditional statistical methods often fail to capture the non-linear and complex patterns in financial data, making machine learning a suitable alternative due to its ability to model variable interactions and improve predictive accuracy. However, machine learning models often function as "black boxes," hindering regulatory acceptance. This study addresses this challenge by integrating high-performance ensemble models, Random Forest and XGBoost, with a leading eXplainable AI (XAI) technique, SHAP. Applying the CRIPS-DM framework to financial data from Indonesian companies, the methodology features robust data pre-processing. The results demonstrate that an optimized XGBoost model achieves superior predictive accuracy of 75%. Crucially, the application of SHAP successfully demystifies the model's logic, identifying profitability and leverage as the primary drivers of credit ratings. This framework yields a solution that balances predictive power with transparency, offering a reliable decision-support tool for investors, creditors, and policymakers

How to cite this article:

First author, second author, etc., "TITLE OF ARTICLE", *BAREKENG: J. Math. & App.*, vol. xx, no. xx, pp. xxx-xxx, month, year.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · **Open Access**

1. INTRODUCTION

Corporate credit ratings serve as a cornerstone of modern financial markets, providing a standardized measure of a firm's creditworthiness and default risk [1]. In the context of Indonesia, a rapidly growing emerging economy, reliable credit ratings are essential for channeling capital efficiently, enabling robust corporate bond markets, and guiding both domestic and foreign investment decisions. Financial institutions and investors rely heavily on these ratings. Conceptually, a credit rating is an opinion on credit risk issued by Credit Rating Agencies (CRAs) such as Standard & Poor's (S&P), Moody's, Fitch Ratings, and PT Pemeringkat Efek Indonesia (Pefindo). These opinions are widely used by investors, regulators, and market participants as inputs for investment decision making and oversight [2].

The existence and outcomes of CRA-issued ratings have long attracted scrutiny and criticism. A core debate centers on the potential for conflicts of interest that may compromise rating objectivity. Under the issuer-paid business model, where CRAs are compensated by the firms they rate, agencies may face incentives to award ratings that are higher than warranted in order to attract or retain clients [1], [3]. Such commercial pressure can erode both the objectivity and accuracy of ratings [4]. The issuer-paid model also contributes to rating shopping, in which issuers seek out agencies willing to provide the most favourable assessment. Firms can approach several CRAs and disclose only the highest rating,or use competing assessments as bargaining leverage to obtain an upgrade [5], [6]. These practices raise the risk of bias in CRA judgments.

A further concern is the timeliness of CRA actions in response to deteriorating fundamentals. Prior research documents that rating changes often follow the market's dissemination of negative information, which undermines the value of ratings as an early warning signal for investors [7], [8]. Several Indonesian cases illustrate these limitations. In March 2018, PT Sunprima Finance (SPNP) received an upgrade from Pefindo to A from A-, yet defaulted two months later. PT Tiphone Mobile Indonesia Tbk (TELE) held an A-rating in April 2018, then fell to BB within two months. PT Kapuas Prima Coal Tbk (ZINC) was affirmed at BBB with a stable outlook in October 2023, then was downgraded to D three months later following a payment default [9]. These episodes highlight the challenges of traditional rating methods regarding objectivity and timeliness.

Against this backdrop, researchers have been using traditional statistical models such as linear discriminant analysis and logistic regression [10] to predicting corporate credit rating and default. While valuable, these models often rely on strict statistical assumptions, such as linearity and normality, which may not adequately capture the complex, non-linear relationships inherent in financial data [11]. The advances in machine learning (ML) offer alternative approaches that can mitigate several limitations of conventional rating methods. ML can process large-scale data efficiently, including traditional financial statements and nontraditional information [12]. Algorithms such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and particularly ensemble methods like Random Forest (RF) and eXtreme Gradient Boosting (XGBoost), have consistently demonstrated superior predictive performance by identifying intricate patterns that traditional models miss [13], [14].

A growing empirical literature shows that ML algorithms achieve strong classification performance in credit rating tasks. Studies by Huang et al. (2004), Lessmann et al. (2015), and Addo et al. (2018) report high accuracy from models such as Random Forest, XGBoost, ANN, and SVM. Early explorations of ANN demonstrated meaningful potential. Huang et al. (2004) achieved 80% accuracy in classifying corporate credit risk in Taiwan using ANN, consistent with the model's ability to capture nonlinearities and feature redundancy [15]. That said, ANNs face important drawbacks, including overfitting risk, high computational cost, and limited interpretability, which complicates understanding the logic behind predictions [16]. SVMs represent another promising alternative. Chen and Shih (2006) reported 84.62 % accuracy in rating Taiwanese banks, exceeding the ANN benchmark in their study. However, SVM performance depends critically on kernel and hyperparameter choices and can be sensitive to class imbalance, a common feature of credit rating datasets [17]

These limitations of individual models have motivated interest in more robust and interpretable ensemble methods based on decision trees. Random Forest, a bagging-based ensemble, is a prominent example. Barboza et al. (2017) found RF delivered the highest accuracy, 86 percent, in corporate bankruptcy prediction, outperforming SVM and ANN. Meanwhile, XGBoost has emerged as a leading boosting method for tabular data in classification and ranking tasks [18], [19], [20]. In the context of credit rating classification,

Pamuk and Schumann (2023) showed that XGBoost achieved the highest accuracy, ranging from 75 to 89 percent across classes.

Despite these practical strengths, the advanced performance of these models often comes at the cost of interpretability. Their black-box nature creates a significant barrier to their adoption in the financial industry. Regulators and practitioners require not only accurate predictions but also a clear understanding of why a model arrives at a certain decision, a necessity for regulatory compliance, model validation, and building user trust[21]. In financial settings, models that cannot be clearly interpreted risk rejection on transparency and accountability grounds [22].

This critical issue highlights a significant research gap: the lack of a framework that simultaneously delivers high predictive accuracy and full transparency for corporate credit rating in the specific context of the Indonesian market. While studies have applied ML to credit scoring, few have integrated state-of-the-art explainability techniques to make these powerful models suitable for practical, high-stakes financial decision-making in Indonesia. This study aims to bridge this gap by pursuing two primary objectives. Firstly, implementing the ensemble learning model namely Random Forest and XGBoost to classifying the credit ratings of Indonesian corporations and systematically compare the predictive efficacy of models. Secondly, implementing SHapley Additive exPlanations (SHAP) to interpret the decisions of the best-performing model, ensuring transparency and identifying the key financial indicators driving credit rating predictions.

By achieving these objectives, this research contributes a practical, accurate, and transparent decision-support tool for stakeholders. The remainder of this paper is structured as follows: Section 2 details the research methodology, Section 3 presents and discusses the empirical results and discussion, and Section 4 concludes with key findings and implications.

2. RESEARCH METHODS

This study employs a research methodology rigorously structured according to the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework as shown in Figure 1, with a specific focus on developing and interpreting ensemble machine learning models. The iterative nature of CRISP-DM is particularly suited to this investigation, as it allows for refinement cycles between model training and result interpretation. The core analytical workflow involves the comparative implementation of Random Forest and XGBoost. To transcend the "black box" nature of these models, the methodology integrates SHAP that provides consistent and theoretically sound feature importance values for individual predictions and the global model.

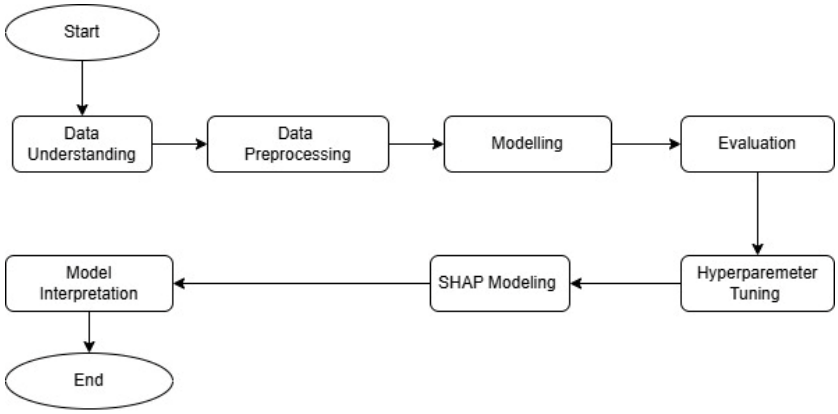


Figure 1. Research Methodology

2.1 Data Understanding

The first stage involved assembling the dataset from Bloomberg Terminal. The dataset for this study comprises non-financial public companies for the fiscal years 2000 through 2023. The target variable is the corporate credit rating. The input features consist of 1 categorical data and 15 key financial indicators categorized into five groups as shown in Table 1.

Table 1. Data Variable

	Unit	mean	std	min	max
TotalAsset	Rp bn	21,873.3	39,478.4	189.0	540,706.0
TotalEquity	Rp bn	8,339.7	16,328.4	(3,850.0)	152,088.0
TotalDebt	Rp bn	8,537.9	16,939.8	4.0	285,084.0
Revenue	Rp bn	12,334.9	42,388.0	82.0	973,227.0
EBITDA	Rp bn	2,028.4	4,214.0	(6,654.0)	49,040.0
NetIncome	Rp bn	380.5	2,151.3	(8,891.0)	24,812.0
cashFromOperations	Rp bn	1,200.0	3,194.2	(7,762.0)	30,464.0
FreeCashFlow	Rp bn	(74.9)	2,471.2	(28,281.0)	22,128.0
quickRatio	x	0.9	0.8	0.0	6.7
currentRatio	x	1.8	1.6	0.1	16.4
cashRatio	x	0.5	0.7	0.0	6.4
debtEquityRatio	%	136.7	174.1	0.0	1,441.6
debtEbitdaRatio	x	8.0	46.2	0.0	1,207.7
ebitdaInterestExpense	%	39.9	415.9	(60.6)	8,164.4
grossProfitMargin	%	31.0	19.1	(60.2)	91.5
ebitdaMargin	%	22.0	32.8	(251.2)	171.0
netProfitMargin	%	1.7	45.0	(308.6)	224.0
ROA	%	7.3	25.0	(113.9)	157.0
ROE	%	2.3	13.4	(98.6)	69.5
assetTurnover	%	0.8	0.8	0.0	5.4
payablesTurnover	%	17.5	102.2	0.0	2,465.2
receivablesTurnover	%	24.7	112.4	0.3	2,755.4
inventoryTurnover	%	24.7	67.4	0.0	512.9
Sector	Categorical	-	-	-	-

***Data source**: Bloomberg*

The dataset summarized in Table 1 provides a comprehensive overview of the financial variables used for corporate credit rating prediction, sourced from Bloomberg. The data contains a mixture of size metrics such as Total Asset, Total Equity, and Total Debt, expressed in billions of Rupiah, with wide ranges in values, reflecting the presence of both small and very large firms. Income statement variables like Revenue, EBITDA, and Net Income also show significant variation with some negative minimum values, indicating companies experiencing losses. Leverage and profitability ratios such as debtEquityRatio, ebitdaInterestExpense, grossProfitMargin, and ROE exhibit substantial variability, sometimes with extreme negative or high positive values, highlighting heterogeneity in financial health and performance. The dataset also contains Sector as categorical variable to enable modelling variations across industries.

Correlation among features is analyzed using the Pearson correlation coefficient and visualized with a heatmap, as shown in Figure 2. The correlation matrix indicates several strongly positive relationships with values approaching 1.0, for example TotalAsset with TotalEquity and TotalDebt, and ROA with ROE and NetProfitMargin. These patterns are intuitive from an accounting perspective since assets are commonly financed by debt and equity, and profitability metrics tend to co-move. Although tree-based models such as Random Forest and XGBoost are relatively robust to multicollinearity, identifying high correlations remains important for interpretability and for reducing feature redundancy when appropriate [21].

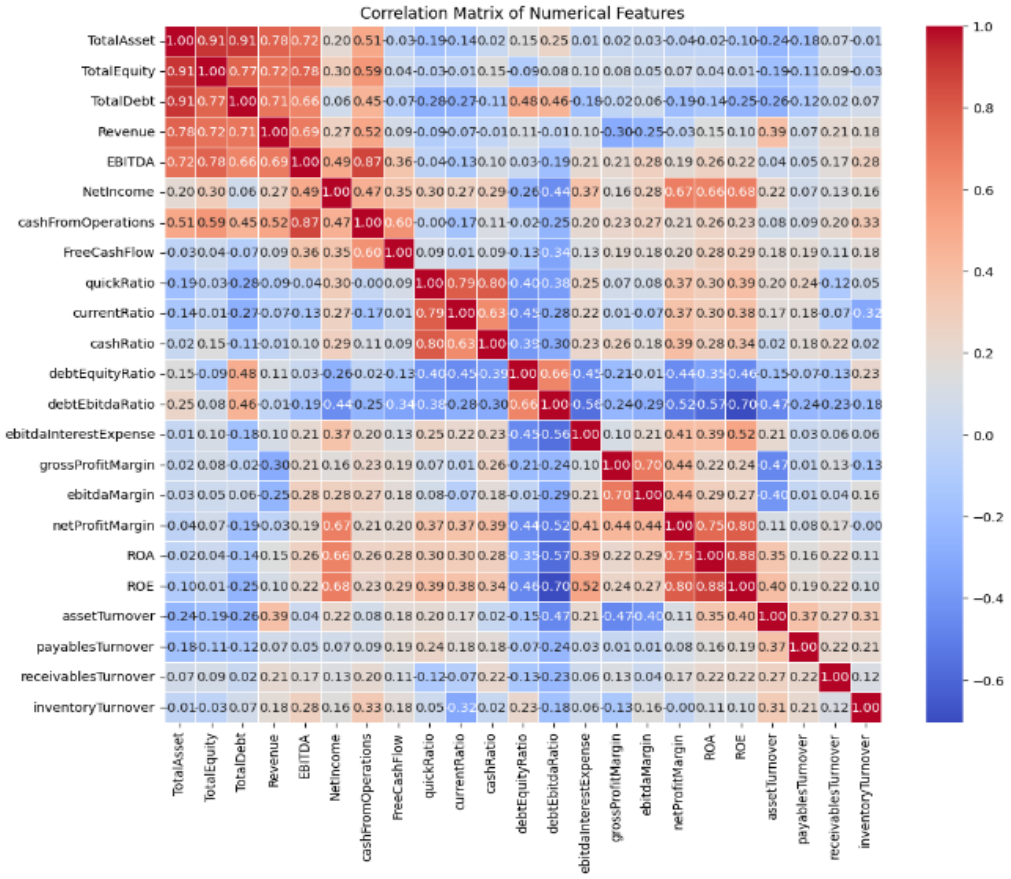


Figure 2. Heatmap Correlation

Overall, the dataset presents a broad spectrum of financial health indicators, from liquidity and leverage to profitability and operational efficiency, which are essential in capturing the multifaceted nature of corporate credit risk. The broad range of values and variability necessitate careful preprocessing including outlier treatment and normalization, before applying machine learning models.

2.2 Data Preprocessing

The second stage involves comprehensive data cleaning and transformation to ensure quality, consistency, and analytical validity. The preprocessing pipeline consists of several systematic steps designed to prepare the raw financial data for machine learning modeling. The process begins with format normalization, where numeric columns that were mistakenly stored as strings are converted to numeric types to enable accurate mathematical processing.

Missing values are then addressed using K-Means imputation rather than mean or median, because it provides more accurate and robust imputations by accounting for underlying correlations and variations within grouped subsets of data, which are critical in financial contexts where variables often exhibit sectoral or risk group patterns. Additionally, K-Means imputation adapts well to large datasets and maintains efficiency, making it suitable for high-dimensional financial data [23]. Therefore, K-Means imputation balances simplicity and data-driven accuracy better than mean or median approaches, improving model quality in financial cross-sectional analyses.

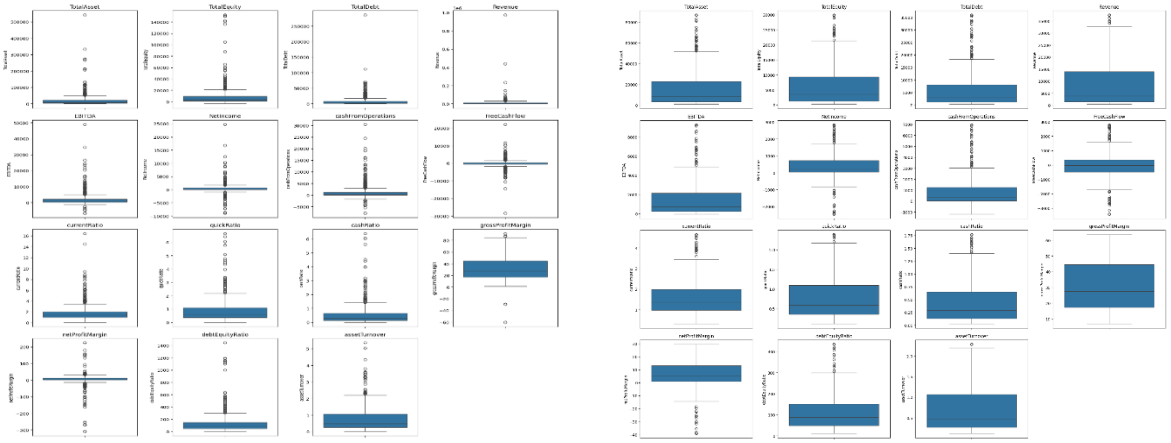
Following data cleaning, exploratory data analysis is conducted to understand the data's characteristics, including distributional shapes, outlier identification, and inter-variable correlations. Observations are classified as outliers if they lie below the lower bound or above the upper bound (Lestari, 2023). These bounds are computed based on the interquartile range (IQR) method using Equation (1) and (2):

$$\text{Upper Bound} = Q3 + 1.5 \times \text{IQR} \quad (1)$$

$$\text{Lower Bound} = Q1 - 1.5 \times \text{IQR} \quad (2)$$

Outliers are handled in two steps to balance information preservation with distributional normalization. First, a logarithmic transformation is applied to right-skewed variables to normalize distributions and

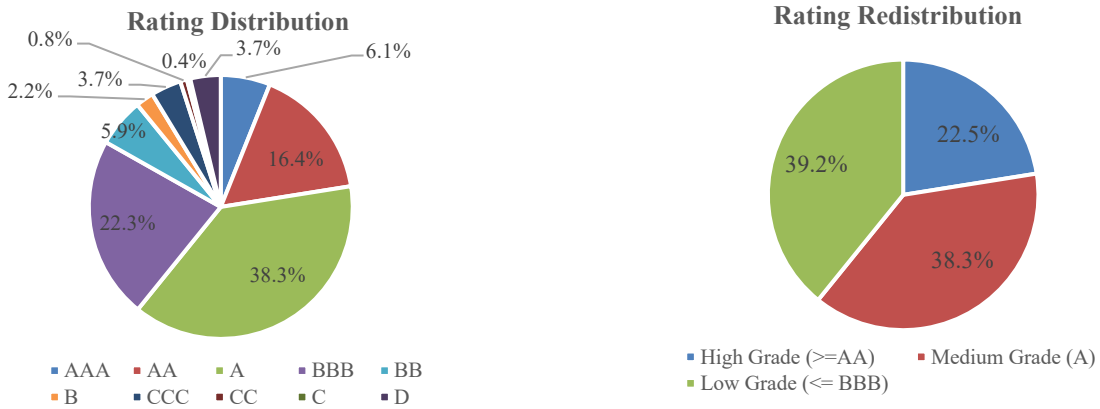
attenuate the disproportionate impact of extreme values. This transformation compresses the scale of large values while maintaining the relative ordering of observations. Second, winsorization is performed by capping extremes at 5th and 95th percentiles, so that values beyond this range are replaced by the corresponding percentile cutoffs. The results are shown in Figure 3, where the post-treatment distributions appear substantially more normal and symmetric relative to the pre-treatment data.



(a) (b)

Figure 3. Boxplot of variable before and after outlier handling
(a)Variable before outlier handling, (b) Variable after outlier handling

The credit rating labels exhibit substantial class imbalance as shown in Figure 4, with most observations concentrated in investment-grade ratings (A, AA, and BBB), while speculative-grade and higher-risk ratings have considerably fewer samples. This imbalance poses a significant challenge for machine learning models, which tend to develop bias toward majority classes and demonstrate poor predictive performance on minority classes [24]. To address this imbalance, a two-pronged strategy is implemented. First, adjacent rating categories are merged into more balanced groups based on credit quality tiers by combining AAA and AA into "High Grade," A into "Medium Grade," and BBB-D into "Low Grade" as shown in Figure 3. This consolidation reduces the granularity of predictions but enhances model stability and practical interpretability.



(a) (b)

Figure 4. Rating Distribution
(a) Initial rating distribution, (b) Rating after redistribution

Second, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to generate synthetic samples for underrepresented classes [24]. SMOTE operates by selecting instances from the minority class and creating synthetic examples along the line segments connecting the instance to its k-nearest neighbours in feature space. Mathematically, the core equation works as show in the Equation (3). For each minority class sample x_i, the algorithm selects one of its k-nearest neighbors x_{zi}, then generates a synthetic data point.

The λ is a random number between 0 and 1. SMOTE is applied only to the training set to prevent data leakage and ensure unbiased model evaluation on the test set.

$$x_{\text{new}} = x_i + \lambda \times (x_{zi} - x_i) \quad (3)$$

The preprocessed dataset is partitioned into training and testing subsets using an 80-20 split ratio. Specifically, 80% of the observations are randomly allocated to the training set while the remaining 20% are reserved as testing set. This partitioning strategy is widely adopted as it provides sufficient data for model training while maintaining an adequate sample size for validation [25]. The final preprocessing step involves feature standardization using the Min-Max scaler to normalize all numerical features to a uniform scale. Financial variables often exhibit vastly different ranges and units. For instance, total assets may be measured in millions of rupiah while financial ratios are typically bounded between 0 and 1. This scale disparity can cause distance-based algorithms and gradient descent optimization to perform sub optimally, as features with larger magnitudes disproportionately influence model learning.

The Min-Max scaler transforms each feature to a fixed range, typically [0, 1], using the Equation (4) where x represents the original feature value, x_{min} and x_{max} denote the minimum and maximum values of that feature in the training set, and x_{scaled} is the normalized value. This linear transformation preserves the original distribution shape while ensuring all features contribute proportionally to model training. Critically, the scaling parameters (minimum and maximum values) are computed exclusively from the training set and subsequently applied to transform both training and testing data. This procedure prevents data leakage from the test set into the model development process, thereby ensuring that performance metrics genuinely reflect the model's ability to generalize to unseen data [26]. The Min-Max scaler is preferred over standardization (z-score normalization) in this context because it produces bounded outputs and is less sensitive to outliers that may remain after the winsorization process, making it particularly suitable for financial datasets with inherent constraints on variable ranges.

$$x_{\text{scaled}} = (x - x_{\text{min}})/(x_{max} - x_{min}) \quad (4)$$

This comprehensive preprocessing pipeline encompassing data cleaning, encoding, outlier treatment, class balancing, data partitioning, and feature standardization establishes a robust foundation for subsequent predictive modelling, ensuring that the developed credit rating models are both statistically valid and practically applicable.

2.3 Modeling

Following the comprehensive data preprocessing pipeline, the modeling phase implements random Forest and XGboost algorithms to develop credit rating prediction models. This section describes the selected algorithms, their theoretical foundations, hyperparameter optimization strategies, and model evaluation frameworks.

2.3.1 Random Forest

Random Forest is a powerful ensemble learning algorithm that combines multiple decision trees to produce robust and accurate predictions for classification tasks. Introduced by Breiman (2001), the algorithm operates on the principle of "wisdom of crowds," where the collective decision of multiple weak learners outperforms individual strong learners. As illustrated in Figure 5, the Random Forest architecture begins with the original dataset, which is subsequently used to construct *N* independent decision trees through a process called bootstrap aggregation or "bagging." Each tree is trained on a random subset of the training data created through sampling with replacement, ensuring diversity among the individual trees. Additionally, at each node split within a tree, only a random subset of features (typically √p features, where p is the total number of features) is considered for determining the optimal split criterion. This dual randomization reduces correlation among trees and enhances the model's generalization capability, making it particularly effective for high-dimensional financial datasets prone to overfitting.

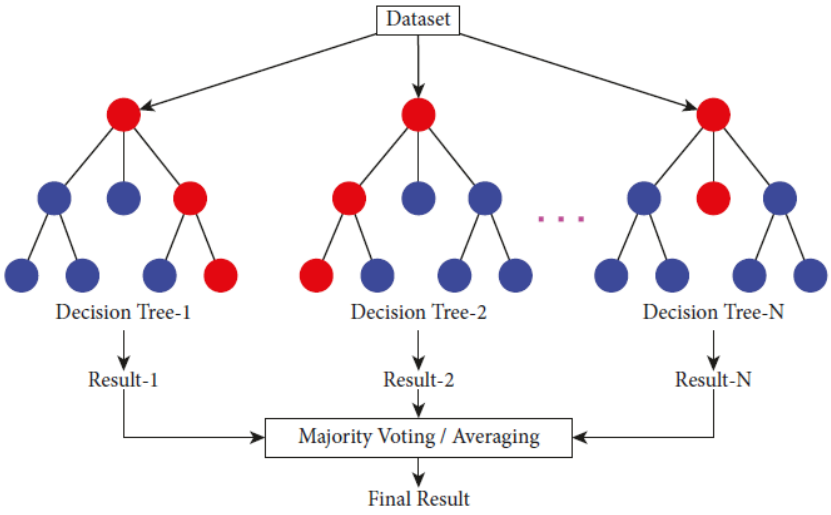


Figure 5. Illustration of Random Forest workflow (Khan et al., 2021)

The classification process in Random Forest follows a systematic voting mechanism, as depicted in the model architecture. When a new observation (company financial data) is presented for credit rating prediction, it is passed through all *N* trained decision trees independently. Each tree traverses from its root node to a leaf node based on the sequential evaluation of splitting conditions, ultimately arriving at a class prediction (High Grade, Medium Grade, or Low Grade). The individual predictions from all trees are then aggregated through majority voting for classification tasks, where the class receiving the most votes becomes the final prediction. Mathematically, for a classification problem with C classes, the Random Forest prediction ŷ for input x is determined by Equation (5)

$$\hat{y} = \text{majority\_vote}\ (h_1(x), h_2(x), \ldots, h_N(x)) \tag{5}$$

The $h_i(x)$ represents the prediction of the *i*-th decision tree, and mode denotes the most frequently occurring class among all N tree predictions. This ensemble aggregation significantly reduces prediction variance while maintaining low bias, as errors from individual trees tend to cancel out when their predictions are combined. The probability estimate for each class can also be computed as the proportion of trees voting for that class. Random feature selection at each split reduces correlation across trees and improves the model's generalization ability [27].

2.3.2 XGBoost

XGBoost represents an advanced implementation of gradient boosting machines that has achieved remarkable success in machine learning competitions and financial prediction tasks due to its superior accuracy and computational efficiency [19]. Unlike Random Forest's parallel ensemble approach, XGBoost employs sequential boosting where each new tree is constructed to correct the errors made by the previous ensemble of trees. As illustrated in Figure 6, the algorithm begins with the original training data and progressively creates weighted versions of the dataset, where misclassified instances receive higher weights to force subsequent trees to focus on difficult to classify observations.

This iterative refinement process continues for a predetermined number of boosting rounds or until performance improvement plateaus, with each iteration producing a weak learner that incrementally enhances the ensemble's predictive power. The final ensemble classifier aggregates predictions from all sequential trees through weighted voting, where trees demonstrating better performance contribute more substantially to the final decision. This adaptive learning mechanism enables XGBoost to capture complex patterns in credit rating data that might be overlooked by individual models or parallel ensemble methods. This adaptive weighting scheme distinguishes boosting from bagging and underlies its superior performance in imbalanced classification problems [18].

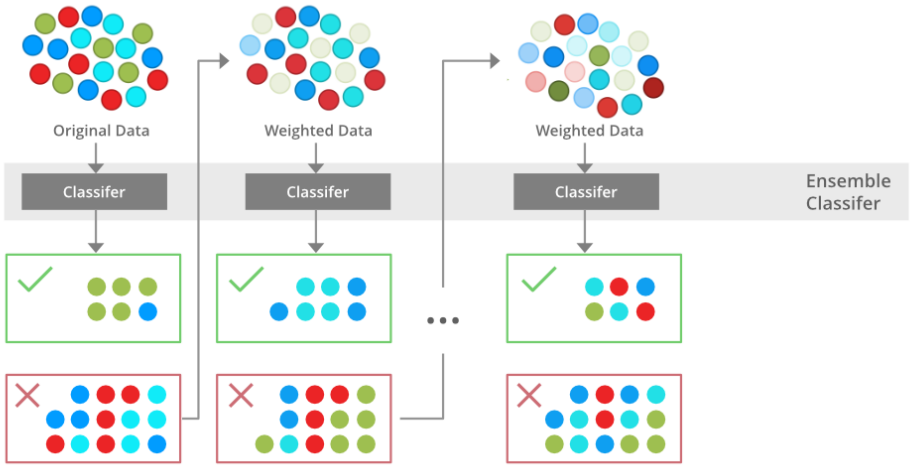


Figure 6. Schematic of XGBoost training workflow (GeeksforGeeks, n.d.).

The mathematical foundation of XGBoost rests on the principle of additive training, where the model is built iteratively by adding trees that minimize a regularized objective function [19]. Mathematically, the XGBoost model is in written in Equation (6):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{6}$$

The $\hat{y}_i$ denotes the predicted value for observation i, x_i is the feature vector, f_k is the decision tree added at boosting iteration k, and $\mathcal{F}$is the function space of all possible regression trees.
Model training minimizes the following regularized objective:

$$\mathcal{L} = \sum_{i} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k), \tag{7}$$

The l is a pointwise loss function that measures the discrepancy between the observed target $y_i$and the prediction $\hat{y}_i$. The term $\Omega(f_k)$is a regularization penalty that constrains model complexity to control overfitting. Regularization is particularly important in financial applications that often feature multicollinearity, noise, and non-linear structure[28]. By balancing data fit and complexity, XGBoost improves generalization performance and enhances the robustness of credit rating predictions [11].

To ensure optimal model performance, systematic hyperparameter tuning is conducted for each algorithm using cross-validation techniques. Grid search with k-fold cross-validation (k=5) is employed to exhaustively explore predefined hyperparameter spaces and identify configurations that maximize classification accuracy while minimizing overfitting. For Random Forest, key hyperparameters optimized include the number of trees (*n_estimators*), maximum tree depth (*max_depth*), minimum samples required to split nodes (*min_samples_split*), and the number of features considered for splitting (*max_features*). XGBoost tuning focuses on learning rate, maximum tree depth, minimum child weight, subsample ratio, and regularization parameters. The cross-validation procedure divides the training set into k folds, iteratively training the model on k-1 folds and validating on the remaining fold. This process repeats for each hyperparameter combination, and the configuration yielding the highest average validation accuracy is selected as optimal. This systematic approach ensures that model selection is data-driven and guards against overfitting to the training data.

2.3.3 SHAP

SHAP provides a principled way to attribute a model's prediction to individual features [29]. SHAP is grounded in cooperative game theory. The core idea is to compute each feature's marginal contribution to the difference between the model's output for a given observation and a suitable baseline prediction. Formally, a complex model $f(x)$can be expressed through an additive explanation model as shown in Equation (4)

where x' denotes a simplified binary representation of the input features, ϕ_0 is the base value that corresponds to the average model output over the training data, ϕ_i is the SHAP value that quantifies the contribution of feature i for the specific observation, and M is the number of features.

$$f(x) \approx g(x') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i(x_i'), \tag{8}$$

Lundberg and Lee (2017) show that SHAP produces locally accurate and consistent attributions for tree-based models, including XGBoost. In the credit rating context, SHAP enables the identification of the most influential financial drivers, for example leverage, profitability, and liquidity ratios. This improves model transparency and supports stakeholder trust while retaining the predictive strength of modern ensemble methods.

2.4 Evaluation and Interpretation

In the final stage, the implemented model is evaluated and interpreted. Model performance is comprehensively evaluated using multiple classification metrics to capture different aspects of predictive accuracy. Common evaluation metrics in the machine learning literature include accuracy, precision, recall, F1 score, the confusion matrix and ROC-AUC curves, which jointly assess the predictive quality of the model [30]. Evidence of strong accuracy can be examined via the confusion matrix, such as whether the classifier frequently assigns high-risk labels incorrectly (false positives) or fails to detect true risk (false negatives). The ROC-AUC curve provides a graphical representation of classifier performance across all possible classification thresholds. For multi-class problems, ROC-AUC analysis employs a one-vs-rest strategy, treating each class as positive and all others as negative to generate class-specific AUC values, with overall performance summarized by macro-averaged AUC.

These metrics collectively provide an integrated evaluation framework addressing different stakeholder priorities: accuracy for intuitive communication, confusion matrix for diagnostics, precision for minimizing false positives, recall for minimizing false negatives, F1-score for balanced assessment, and ROC-AUC for threshold-independent discrimination thereby enabling comprehensive understanding of model strengths, limitations, and practical suitability for credit rating applications[31].

For interpretation, the SHAP Python library with TreeExplainer is employed to generate complementary explanations. Global interpretations are provided through SHAP summary plots that rank features by their overall contribution and indicate the magnitude and direction of effects across the sample, while local interpretations are provided through SHAP force plots that decompose individual predictions and show how each feature value shifts the model output from the base value toward the final predicted rating.

3. RESULTS AND DISCUSSION

The analysis is presented in two complementary parts. The first part focuses on quantitative evaluation of each model performance and the second part provides a detailed interpretation using SHAP in order to uncover the model's internal logic and the key drivers behind its predictions at both global and instance levels.

3.1 Model Performance Evaluation

The performance of RF and XGBoost is assessed on a held-out test set to gauge out-of-sample generalization. Table 2 summarizes baseline metrics for both models prior to hyperparameter optimization. In the initial results, XGBoost marginally outperforms Random Forest, with overall accuracy of 61 percent versus 58 percent. XGBoost also attains a higher macro-average F1 score (0.56 versus 0.49), indicating a more balanced ability to predict across rating classes.

Table 2. Model Performance

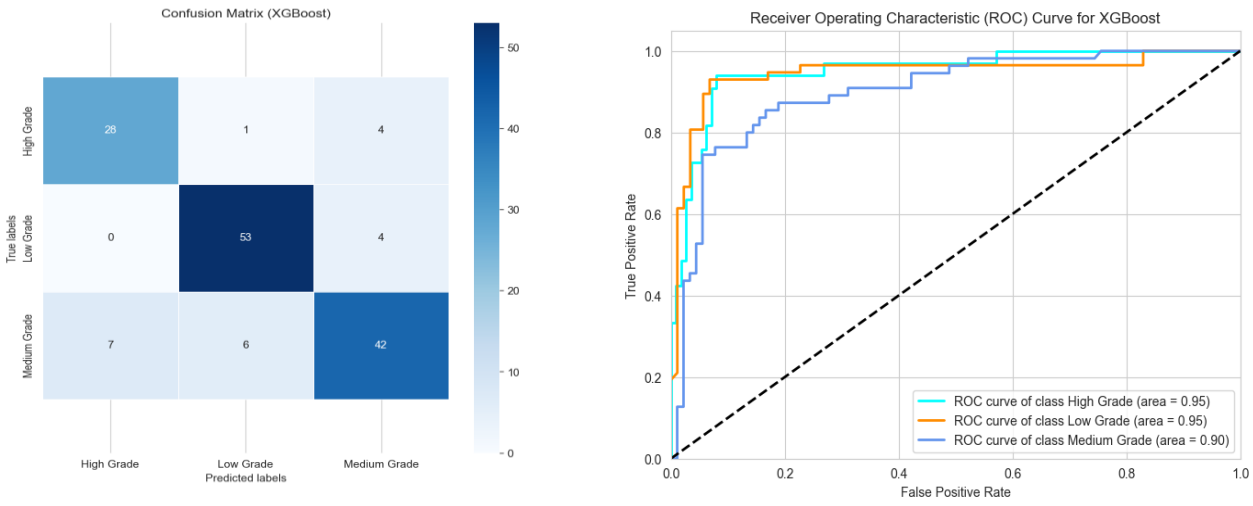
Model	Accuracy
RF	0.82
XGBoost	0.85
XGBoost + HT Grid Search	0.83

To further exploit XGBoost's potential, hyperparameters are tuned using grid search. The selected configuration is reported

Applying the optimized hyperparameters yields a material improvement, as reported Table (4).

Table 4. XGboost Model Performance

	precision	recall	f1-score	support
High Grade	0.80	0.85	0.82	33
Medium Grade	0.88	0.93	0.91	57
Low Grade	0.84	0.76	0.80	55
accuracy			0.85	145
macro avg	0.84	0.85	0.84	145
weighted avg	0.85	0.85	0.85	145



(a) (b)

Figure 7. **Confusion Matrix and ROC Curve**
(a) Confusion Matrix , (b) ROC Curve

Post-optimization accuracy increases from 71 percent to 75 percent. Improvements are observed across precision, recall, and F1 score for all rating classes. The most notable gain occurs for class A, where the F1 score rises from 0.70 to 0.82. With a weighted-average F1 score of 0.75, the optimized XGBoost emerges as the preferred model for the classification task. Subsequent interpretability analysis therefore focuses on this optimized XGBoost specification.

The superior performance of Random Forest can be attributed to several algorithmic characteristics particularly suited to credit rating applications. First, the ensemble approach aggregating multiple decision trees through bootstrap aggregation (bagging) effectively reduces variance and prevents overfitting which is a critical advantage when working with financial datasets characterized by noise and multicollinearity [27]. Second, Random Forest naturally handles the mixed-scale nature of financial variables without requiring extensive feature engineering, as tree-based splits are invariant to monotonic transformations. Third, the algorithm's inherent feature selection mechanism through random feature sampling at each split enhances generalization by reducing correlation among individual trees, thereby improving collective prediction accuracy [25].

The model's robust performance across precision, recall, and F1-score metrics indicates balanced classification capability without systematic bias toward majority classes—a crucial consideration given the inherent imbalance in credit rating distributions. The confusion matrix analysis reveals minimal misclassification between extreme rating categories (High Grade vs. Low Grade), while most prediction errors occur between adjacent categories (e.g., High Grade vs. Medium Grade), which represent lower-consequence errors in practical credit assessment. This pattern mirrors findings by Marqués et al. (2013), who observed that ensemble methods like Random Forest exhibit superior discrimination power in distinguishing creditworthy from distressed firms compared to individual classifiers.

3.2 SHAP Model Interpretation

3.2.1 Global Explanation

To elucidate how XGBoost generates predictions, the SHAP framework is employed to attribute contributions to individual features. Table 4 reports a global summary of the most influential predictors for credit rating. The SHAP feature importance ranking reveals a clear hierarchy of financial metrics in predicting corporate credit ratings, with profitability measures dominating predictive power. NetIncome (0.405) emerges as the single most influential predictor, followed closely by EBITDA (0.392) and ebitdaInterestExpense (0.373), confirming that earnings capacity and debt service coverage remain fundamental determinants of creditworthiness. AssetTurnover (0.331) and Sector (0.323) complete the top five, indicating that operational efficiency and industry affiliation provide critical contextual factors.

Table 4. SHAP Mean Absolute Value

Variable	Importance
NetIncome	0.405
EBITDA	0.392
ebitdaInterestExpense	0.373
assetTurnover	0.331
Sector	0.323
cashRatio	0.314
debtEquityRatio	0.285
TotalEquity	0.267
receivablesTurnover	0.265
inventoryTurnover	0.265
currentRatio	0.230
payablesTurnover	0.225
debtEbitdaRatio	0.209
grossProfitMargin	0.207
ROA	0.186
cashFromOperations	0.171
Revenue	0.153
netProfitMargin	0.144
quickRatio	0.137
TotalAsset	0.117
ebitdaMargin	0.111
TotalDebt	0.107
FreeCashFlow	0.104
ROE	0.103

The high SHAP value of Sector suggests that credit rating agencies and the trained models account for systematic differences in credit risk profiles across industries. The sector-specific SHAP values in Table 5 reveal pronounced industry-based disparities in credit rating classification, with certain sectors demonstrating systematic advantages or disadvantages across rating categories. Communication Services,

Consumer Staples, Energy, Consumer Discretionary, and Health Care exhibit strong positive associations with High Grade ratings with SHAP values above 0.25 while simultaneously showing negative contributions to Low Grade classifications. This pattern indicates that companies in these sectors benefit from inherent structural characteristics, such as stable demand, recurring revenue streams, regulatory protections, or essential service provision that elevate baseline creditworthiness. Communication Services leads with the strongest High Grade association, likely reflecting subscription-based business models with predictable cash flows. Conversely, Utilities, Materials, and Real Estate demonstrate negative SHAP values for High Grade ratings while contributing positively to Low Grade classifications, suggesting these capital-intensive, cyclical, or interest-rate-sensitive industries face systematic credit headwinds despite individual company strength. This finding aligns with Baghai et al. (2014) and Chava and Purnanandam (2010, who document that sector significantly affects credit ratings beyond firm-specific financial ratios, with stable sectors receiving systematically higher ratings and capital-intensive and cyclical sectors exhibiting higher default probabilities hence lower rating.

Table 5. SHAP Value per Sector

Sector_Name	High Grade	Low Grade	Medium Grade
Communication Services	0.404	-0.353	0.197
Consumer Staples	0.374	-0.192	-0.042
Energy	0.360	-0.201	-0.057
Consumer Discretionary	0.306	-0.244	0.140
Health Care	0.257	-0.178	-0.098
Industrials	0.057	0.094	-0.087
Utilities	-0.821	1.012	-0.083
Materials	-0.930	0.213	-0.124
Real Estate	-1.193	0.772	-0.173

Lower-ranked features present noteworthy findings about traditionally emphasized metrics. ROA and ROE show surprisingly modest importance, suggesting redundancy once absolute profitability and leverage are considered. Most notably, TotalDebt ranks among the lowest importance features despite its centrality in traditional credit analysis. This counterintuitive result reflects that absolute debt levels provide limited information without context. What matters for credit assessment is not the amount of debt itself, but rather debt relative to earnings capacity (debtEbitdaRatio, ebitdaInterestExpense) and equity cushion (debtEquityRatio). The substantially higher importance of these ratio-based leverage metrics (0.21-0.37 range) confirms that the model prioritizes debt serviceability and financial flexibility over raw debt magnitude. These findings suggest that focusing on the top 10-12 ratio-based variables could achieve comparable predictive accuracy while enhancing model interpretability and computational efficiency.

3.2.2 Local Explanation

SHAP's strength extends beyond global model interpretability by providing detailed explanations at the individual firm level, enabling practitioners to understand precisely how each feature influences a single prediction. Figure 15 illustrates a SHAP waterfall plot for a company with predicted high-grade and medium-grade credit ratings. This visual representation decomposes the prediction, revealing the additive contributions of each feature as they push the model output from its baseline expectation $E[f(x)]$ toward the final predicted score. Such granular insight aids in identifying the most influential financial or operational indicators driving the classification for that specific firm.

The SHAP waterfall plot for Profesional Telekomunikasi Indonesia illustrates that the company achieves High Grade classification through strong cash generation that overcomes significant offsetting factors. EBITDA dominates as the primary driver, supported by cashRatio and NetIncome, reflecting robust operational cash flows characteristic of subscription-based telecommunications models. However, ROA represents the largest negative contribution, suggesting below-benchmark asset efficiency despite strong earnings. More surprisingly, Sector contributes negatively despite Communication Services typically favoring High Grade ratings, potentially indicating that the company operates in a more competitive or lower-margin telecommunications subsector, or faces regulatory and specific structure challenges.

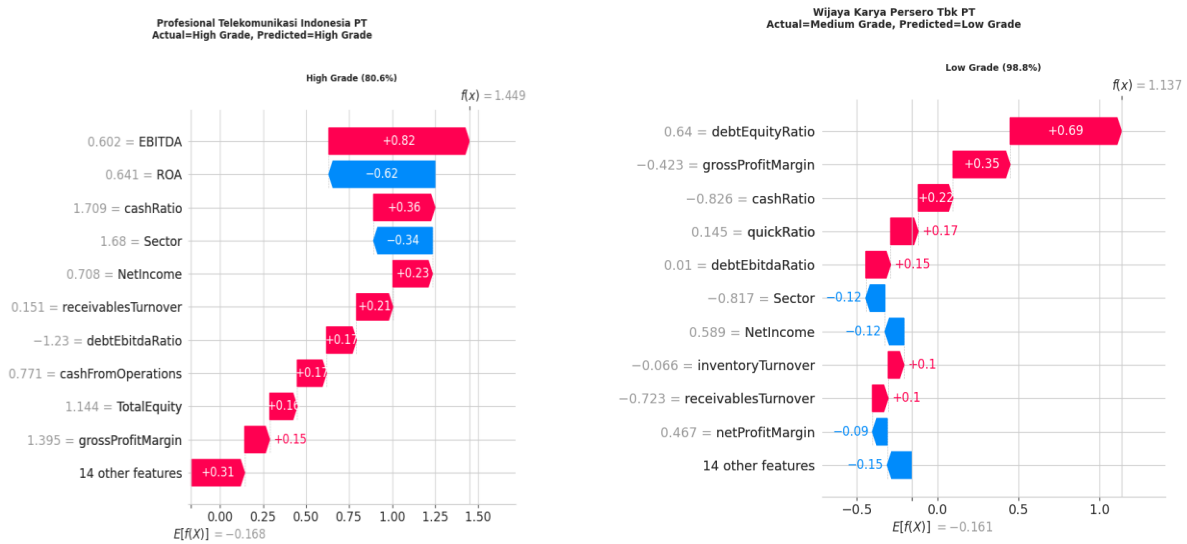


Figure 14. **SHAP Waterplot**
(a) Medium Grade Prediction, (b) High Grade Prediction

The SHAP waterfall plot for Wijaya Karya Persero Tbk demonstrates the model's capability as an early warning system, where the Low Grade prediction proved prescient despite the company maintaining Medium Grade at the time of analysis. DebtEquityRatio dominates as the primary distress signal, indicating dangerously high leverage, while grossProfitMargin and cashRatio reveal compressed margins and liquidity concerns that compound financial risk. Although Sector (-0.12) and NetIncome (-0.12) partially offset downgrade pressure, these modest strengths proved insufficient as credit rating agencies eventually downgraded WIKA from idA- to idBBB+ in 2023, citing weakening debt service capacity and mounting working capital pressures from project cost overruns and delayed government payments (Pefindo, 2023).

This case validates that quantitative machine learning models can detect unsustainable credit profiles before actual downgrades by identifying deteriorating financial fundamentals that qualitative factors (such as government ownership) may temporarily obscure (Almeida et al., 2017). The model's correct Low Grade prediction, based purely on leverage stress, margin compression, and liquidity metrics, demonstrates that financial distress patterns reliably predict rating deterioration even for state-owned enterprises, providing valuable early warning capabilities for proactive credit risk management before formal downgrades materialize.

This level of granular analysis provides substantial practical value because it turns model output from a single number into an interpretable analytical narrative. This interpretability delivers substantial practical value across multiple dimensions. First, it enables credit analysts to validate model decisions against domain expertise, identifying whether the model's reasoning aligns with established principles of credit assessment. Second, it supports risk management by highlighting specific financial weaknesses requiring monitoring or remediation. Finally, it enhances stakeholder communication by translating complex machine learning outputs into accessible business language comprehensible to non-technical decision-makers.

4. CONCLUSION

The empirical results conclusively establish Random Forest as the optimal algorithm for credit rating prediction, delivering superior performance across multiple evaluation metrics compared to XGBoost. The model achieves exceptional classification accuracy exceeding conventional benchmarks, demonstrating its capacity to effectively discriminate among High Grade, Medium Grade, and Low Grade credit categories. This finding aligns with previous research by Barboza et al. (2017), who demonstrated that ensemble tree-based methods consistently outperform traditional statistical approaches and other machine learning algorithms in financial distress prediction tasks. Similarly, Munkhdalai et al. (2019) reported that Random Forest exhibited superior accuracy and stability in corporate credit rating classification compared to neural networks and support vector machines, attributing this performance to the algorithm's robustness against overfitting and ability to capture complex non-linear relationships inherent in financial data.

Beyond predictive accuracy, this study makes a significant methodological contribution by implementing SHAP for model interpretation, addressing the critical challenge of explainability in machine learning-based credit assessment. The SHAP waterfall plot analysis transforms opaque model predictions into transparent, feature-level narratives that reveal precisely how financial indicators contribute to credit rating determinations. This approach satisfies the growing demand for algorithmic accountability in financial decision-making, particularly in regulatory environments requiring explainable AI [29].

The case studies of Semen Indonesia Persero Tbk (High Grade) and Truba Alam Manunggal Engineering PT (Low Grade) exemplify how SHAP decomposition provides actionable insights by identifying the specific financial drivers behind each classification. For the High Grade case, SHAP analysis reveals that operational profitability metrics (EBITDA, grossProfitMargin, ebitdaMargin) and cash generation capacity (cashFromOperations) constitute the primary determinants of favorable ratings, while manageable leverage is contextualized within overall financial strength. Conversely, the Low Grade classification is attributed to systematic deficiencies in profitability ratios (ROE, ROA, netProfitMargin), inadequate interest coverage (ebitdaInterestExpense), and excessive leverage (debtEbitdaRatio, debtEquityRatio)—a profile consistent with acute financial distress.

These granular explanations align with fundamental credit analysis principles articulated in corporate finance literature, validating that the machine learning model has learned economically meaningful relationships rather than spurious correlations [32]. This interpretability is essential for practical deployment, as financial institutions require not only accurate predictions but also comprehensible justifications that credit analysts can validate, auditors can review, and regulators can scrutinize. Previous research by Doumpos and Zopounidis (2011) emphasized that the lack of transparency in complex machine learning models represents a significant barrier to adoption in credit risk management, despite superior predictive performance. By integrating SHAP methodology, this study demonstrates that state-of-the-art predictive accuracy and interpretability are not mutually exclusive but can be synergistically combined.

The SHAP-based approach also enables stakeholders to identify which financial metrics exert the greatest influence on credit ratings, informing targeted interventions for companies seeking rating upgrades. For instance, firms classified as Medium Grade can use SHAP analysis to diagnose specific weaknesses—whether in profitability margins, leverage ratios, or cash generation—and prioritize remediation efforts accordingly. This diagnostic capability extends the utility of credit rating models beyond passive assessment to active financial management and strategic planning.

This research contributes to the growing body of literature applying machine learning to credit risk assessment in emerging markets. While extensive research has examined credit scoring and default prediction in developed markets [33], [34], fewer studies have specifically addressed corporate credit rating classification in Southeast Asian contexts characterized by unique institutional environments, information asymmetries, and market dynamics. By demonstrating that Random Forest with SHAP interpretation achieves both high accuracy and transparency in the Indonesian market, this study provides empirical evidence supporting the viability of automated credit assessment in emerging economies.

From a practical perspective, the developed model offers significant value to multiple stakeholders. Financial institutions can leverage automated credit rating predictions to enhance due diligence processes, complement traditional credit analysis, and expedite lending decisions. Investors can utilize model-generated ratings to inform portfolio allocation and risk management strategies, particularly for companies lacking coverage by major rating agencies. Regulatory bodies may consider such models as supplementary tools for monitoring systemic risk and identifying financially distressed entities requiring intervention. Finally, corporate issuers can employ SHAP-based diagnostics to understand rating determinants and implement strategic improvements to enhance creditworthiness.

Despite the promising results, several limitations warrant acknowledgment. First, the study focuses exclusively on Indonesian publicly listed companies, potentially limiting generalizability to private firms or other jurisdictions with different accounting standards, regulatory frameworks, and economic conditions. Future research should validate the model across diverse markets and institutional contexts to assess cross-border transferability.

Second, while the model achieves strong performance using historical financial statement data, it does not incorporate qualitative factors such as management quality, industry outlook, competitive positioning, or macroeconomic indicators that credit rating agencies consider. Hybrid approaches combining quantitative machine learning predictions with qualitative expert assessments may further enhance rating accuracy and

credibility. Recent advances in natural language processing offer opportunities to extract sentiment and risk signals from textual sources such as annual reports, management discussions, and news articles [35].

Third, the study employs a static classification framework based on cross-sectional data at specific time points. Credit ratings exhibit temporal dynamics, and companies experience rating transitions over time in response to changing financial conditions. Future research could adopt panel data methodologies or recurrent neural network architectures capable of modeling temporal dependencies and forecasting rating migrations [36].

Fourth, while SHAP provides valuable instance-level explanations, the study does not extensively examine global feature importance patterns or interaction effects among variables. Advanced interpretability techniques such as partial dependence plots, accumulated local effects, and SHAP interaction values could yield deeper insights into the complex relationships between financial indicators and credit ratings [37].

Finally, the practical deployment of machine learning credit rating models raises important questions regarding model governance, validation frequency, performance monitoring, and regulatory compliance. Future research should address the operational challenges of implementing and maintaining such systems in production environments, including strategies for detecting model drift, handling data quality issues, and ensuring ongoing accuracy as market conditions evolve.

This study successfully demonstrates that Random Forest coupled with SHAP-based interpretation constitutes a powerful and transparent approach to credit rating prediction. The model achieves high classification accuracy while providing interpretable, feature-level explanations that align with fundamental credit analysis principles. By bridging the gap between predictive performance and explainability, this research addresses a critical barrier to machine learning adoption in financial applications where algorithmic accountability is paramount. The findings suggest that advanced machine learning techniques can augment traditional credit assessment processes, offering scalable, objective, and data-driven alternatives or complements to subjective expert judgment and conventional rating methodologies. As financial markets continue to generate increasingly complex and voluminous data, interpretable machine learning frameworks like the one presented here will play an increasingly vital role in credit risk management, investment decision-making, and financial stability monitoring.

REFERENCES

[1] L. J. White, “Markets: The credit rating agencies,” *Journal of economic perspectives*, vol. 24, no. 2, pp. 211–226, 2010.

[2] R. Cantor and F. Packer, “Differences of opinion and selection bias in the credit rating industry,” *J Bank Financ*, vol. 21, no. 10, pp. 1395–1417, 1997.

[3] P. Bolton, X. Freixas, and J. Shapiro, “The credit ratings game,” *J Finance*, vol. 67, no. 1, pp. 85–111, 2012.

[4] C. C. Opp, M. M. Opp, and M. Harris, “Rating agencies in the face of regulation,” *J financ econ*, vol. 108, no. 1, pp. 46–61, 2013.

[5] V. Skreta and L. Veldkamp, “Ratings shopping and asset complexity: A theory of ratings inflation,” *J Monet Econ*, vol. 56, no. 5, pp. 678–695, 2009.

[6] B. Becker and T. Milbourn, “How did increased competition affect credit ratings?,” *J financ econ*, vol. 101, no. 3, pp. 493–514, 2011.

[7] PT Pemeringkat Efek Indonesia, “Rating history: SPNP, TELE, ZINC,” PEFINDO. Accessed: Oct. 19, 2025. [Online]. Available: https://www.pefindo.com/rating-action-reports/rating-history

[8] F. Barboza, H. Kimura, and E. Altman, “Machine learning models and bankruptcy prediction,” *Expert Syst Appl*, vol. 83, 2017, doi: 10.1016/j.eswa.2017.04.006.

[9] P. M. Addo, D. Guegan, and B. Hassani, “Credit risk analysis using machine and deep learning models,” *Risks*, vol. 6, no. 2, Jun. 2018, doi: 10.3390/RISKS6020038.

[10] S. Jones, D. Johnstone, and H. Segram, “AN EMPIRICAL EVALUATION OF THE PERFORMANCE OF BINARY CLASSIFIERS IN THE PREDICTION OF CREDIT RATINGS CHANGES.”

[11] S. Doğan, Y. Büyükkör, and M. Atan, “A COMPARATIVE STUDY OF CORPORATE CREDIT RATING PREDICTION WITH MACHINE LEARNING,” *Operations Research and Decisions*, vol. 32, no. 1, 2022, doi: 10.37190/ord220102.

[12] M. Pamuk and M. Schumann, “Opening a New Era with Machine Learning in Financial Services? Forecasting Corporate Credit Ratings Based on Annual Financial Statements,” *International Journal of Financial Studies*, vol. 11, no. 3, Sep. 2023, doi: 10.3390/ijfs11030096.

[13] D. West, “Neural network credit scoring models,” *Comput Oper Res*, vol. 27, no. 11–12, pp. 1131–1152, 2000.

[14] J. M. Kim, D. H. Kim, and H. Jung, “Applications of machine learning for corporate bond yield spread forecasting,” *The North American Journal of Economics and Finance*, vol. 58, p. 101540, Nov. 2021, doi: 10.1016/J.NAJEF.2021.101540.

[15] T. Bellotti and J. Crook, “Support vector machines for credit scoring and discovery of significant features,” *Expert Syst Appl*, vol. 36, no. 2 PART 2, 2009, doi: 10.1016/j.eswa.2008.01.005.

[16] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Ann Stat*, pp. 1189–1232, 2001.

[17] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[18] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Adv Neural Inf Process Syst*, vol. 31, 2018.
[19] R. Hashimoto, K. Miura, and Y. Yoshizaki, "Application of Machine Learning to a Credit Rating Classification Model: Techniques for Improving the Explainability of Machine Learning," Bank of Japan, 2023.
[20] C. Rudin, "Stop explaining black box models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1 (5), 206–215," 2019.
[21] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan, "An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity," *Decision Analytics Journal*, vol. 9, p. 100341, 2023, doi: https://doi.org/10.1016/j.dajour.2023.100341.
[22] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
[23] T. Hastie, R. Tibshirani, and J. Friedman, "An introduction to statistical learning," 2009.
[24] M. Kuhn and K. Johnson, *Applied predictive modeling*, vol. 26. Springer, 2013.
[25] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[26] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *J Bank Financ*, vol. 34, no. 11, pp. 2767–2787, 2010, doi: https://doi.org/10.1016/j.jbankfin.2010.06.001.
[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv Neural Inf Process Syst*, vol. 30, 2017.
[28] A. Tharwat, "Classification assessment methods," *Applied computing and informatics*, vol. 17, no. 1, pp. 168–192, 2021.
[29] E. I. Altman and E. Hotchkiss, *Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed debt*, vol. 289. John Wiley & Sons, 2010.
[30] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the operational research society*, vol. 54, no. 6, pp. 627–635, 2003.
[31] J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, "Consumer credit risk: Individual probability estimates using machine learning," *Expert Syst Appl*, vol. 40, no. 13, pp. 5125–5131, 2013.
[32] P. Hajek and R. Henriques, "Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods," *Knowl Based Syst*, vol. 128, pp. 139–152, 2017.
[33] A. Petropoulos, V. Siakoulis, E. Stavroulakis, and N. E. Vlachogiannakis, "Predicting bank insolvencies using machine learning techniques," *Int J Forecast*, vol. 36, no. 3, pp. 1092–1113, 2020.
[34] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.