Progresif: Jurnal Ilmiah Komputer
https://ojs.stmik-banjarbaru.ac.id/index.php/progresif/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877

Utilizing Explainable AI for Interpreting Machine Learning Model Results in Ceria Credit Scoring

DOI: http://dx.doi.org/10.35889/progresif.v21i2.2769

Roni Eka Setiawan[1*], Toni Wijanarko Adi Putra[2], Budi Hartono[3]
Teknik Informatika, Universitas Sains dan Teknologi Komputer, Semarang, Indonesia
[*]e-mail *Corresponding Author:* roniekasetiawan1@gmail.com

Abstract

This study aims to improve the transparency of machine learning models in credit scoring using various Explainable Artificial Intelligence (XAI) methods. The methods used include SHAP, BRCG, ALE, Anchor, and ProtoDash to explain the prediction results of machine learning models, namely logistic regression, XGBoost, and random forest. This study applies a quantitative approach with a comparative method, where Ceria loan application data from Bank Rakyat Indonesia (BRI) is analyzed using a machine learning model, then evaluated using the Explanation Consistency Framework (ECF). The results show that the XAI method can improve understanding of model decisions, with SHAP and ALE effective for global explanations, while Anchor and ProtoDash provide in-depth insights at the individual level. Evaluation using ECF shows that the post-hoc method has high consistency, although Anchor has limitations in the aspect of axiom identity. In conclusion, the XAI method can help improve trust and transparency in credit scoring at BRI.

Keywords: *Explainable Artificial Intelligence; Credit Scoring; Machine Learning; Model Interpretability; Explanation Consistency Framework*

Abstrak

Penelitian ini bertujuan untuk meningkatkan transparansi model pembelajaran mesin dalam penilaian kredit menggunakan berbagai metode *Explainable Artificial Intelligence* (XAI). Metode yang digunakan antara lain SHAP, BRCG, ALE, Anchor, dan ProtoDash untuk menjelaskan hasil prediksi model pembelajaran mesin yaitu regresi logistik, XGBoost, dan random forest. Penelitian ini menggunakan pendekatan kuantitatif dengan metode komparatif, dimana data pengajuan pinjaman Ceria dari Bank Rakyat Indonesia (BRI) dianalisis menggunakan model *machine learning*, kemudian dievaluasi menggunakan *Explanation Consistency Framework* (ECF). Hasilnya menunjukkan bahwa metode XAI dapat meningkatkan pemahaman keputusan model, dengan SHAP dan ALE efektif untuk penjelasan global, sementara Anchor dan ProtoDash memberikan wawasan mendalam pada tingkat individu. Evaluasi menggunakan ECF menunjukkan bahwa metode post-hoc memiliki konsistensi yang tinggi, meskipun Anchor memiliki keterbatasan pada aspek identitas aksioma. Kesimpulannya, metode XAI dapat membantu meningkatkan kepercayaan dan transparansi dalam credit scoring di BRI.

Kata Kunci: *Explainable Artificial Intelligence; Credit Scoring; Machine Learning; Model Interpretability; Explanation Consistency Framework*

1. Introduction

The rapid development of artificial intelligence (AI) technologies has transformed various sectors, including finance. One of the most important applications is in credit scoring, which involves assessing a borrower's creditworthiness based on historical and behavioral data. Machine learning models in credit scoring have been shown to improve the efficiency and accuracy of risk analysis [1]. However, the high complexity of these algorithms such as XGBoost and Random Forest often makes their decision-making processes a "black box", lacking transparency. This raises serious concerns related to fairness, accountability, and public trust in AI-based systems.

In practice, Bank Rakyat Indonesia (BRI), one of the largest financial institutions in Indonesia, has adopted machine learning models to accelerate credit evaluation. In 2023, BRI

disbursed over IDR 1,266 trillion in loans. Despite this achievement, the use of complex models like XGBoost faces challenges in interpretability, making it difficult for both internal decision-makers and external stakeholders (such as applicants and regulators) to understand the rationale behind credit approval or rejection. Thus, the main problem addressed in this study is: *How can interpretability and transparency of credit scoring models used at BRI be improved through Explainable Artificial Intelligence (XAI)?*

To address this challenge, this research proposes the application of five XAI methods—SHAP, BRCG, ALE, Anchor, and ProtoDash—each offering a unique approach to interpreting model outputs. These methods were selected for their complementary capabilities in providing both global (model-level) and local (instance-level) explanations. SHAP, for example, is known for its ability to measure feature contribution, while Anchor generates interpretable rule-based conditions. Additionally, the study incorporates the Explanation Consistency Framework (ECF) to evaluate the quality of each explanation method, using criteria such as identity, separability, and stability [2]. Recent literature supports the rationale that XAI methods are crucial in enhancing trust and accountability in AI-based credit systems [3] [4].

The aim of this study is to implement and compare five XAI methods in interpreting predictions made by credit scoring models at BRI and to evaluate the explanation quality using ECF. The practical contribution of this research lies in its potential to guide decision-makers in adopting more transparent, trustworthy AI systems in finance, while also offering a theoretical foundation for future research on model interpretability in real-world applications. The results of this study are expected to enhance the transparency of machine learning models in credit scoring and assist companies in making more accurate and interpretable decisions.

2. Literature Review

Examined the application of logistic regression models for credit scoring in the banking sector. Their study used a quantitative approach by applying the model to financial datasets from several banks in Jordan. The researchers evaluated the model's accuracy in distinguishing between risky and non-risky borrowers based on predictors such as income, loan size, and repayment history. The result showed that logistic regression offers a simple yet effective method for credit classification, although it lacks the ability to model non-linear relationships [1].

Conducted a comprehensive survey on machine learning interpretability. They analyzed a wide variety of explanation techniques—both intrinsic and post-hoc—used across domains such as healthcare, finance, and transportation. Their methodology involved categorizing methods based on their interpretability level (global or local), model dependence, and evaluation metrics. They emphasized that a lack of standard frameworks to assess explanation quality remains a critical gap in current research [5].

Proposed SHAP (Shapley Additive Explanations), a unified framework for interpreting predictions made by any machine learning model. Their method, inspired by cooperative game theory, computes the contribution of each feature by treating it as a "player" in the prediction outcome. SHAP was tested on models like XGBoost and deep learning classifiers, showing its effectiveness in attributing model output to input features in a consistent and theoretically sound manner [3].

Introduced Anchor, a model-agnostic explanation technique that generates high-precision IF-THEN rules. Their study used simulated datasets and real-world classifiers (including decision trees and random forest) to show how anchors can offer robust, human-interpretable explanations. The method balances precision and coverage, allowing users to understand the decision logic under specific conditions [4].

Proposed ProtoDash, a prototype selection technique that identifies the most representative data points (prototypes) based on importance weights. This method is particularly useful for instance-level explanations, especially when communicating with non-technical stakeholders. Their procedure involved training classifiers and using ProtoDash to select real samples that are closest to a given prediction target, helping to improve transparency [6].

Developed the Accumulated Local Effects (ALE) method, which addresses the shortcomings of partial dependence plots by being more robust to feature correlations. The researchers implemented ALE to visualize the marginal effect of features on the prediction outcomes of black-box models such as random forest and XGBoost. Their results showed ALE to be more stable and easier to interpret in practical machine learning applications [7].

While the aforementioned studies have provided valuable insights into interpretable AI and machine learning model analysis, they typically focus on evaluating a single explanation method in isolation or apply the method to simplified case studies. In contrast, the present study integrates five

distinct XAI methods SHAP, BRCG, ALE, Anchor, and ProtoDash and applies them to a real-world credit scoring system at Bank Rakyat Indonesia [8]. Furthermore, this study uniquely employs the Explanation Consistency Framework (ECF) to evaluate the quality of explanations across all methods using three objective criteria: identity, separability, and stability. This comprehensive and comparative approach combined with real institutional data offers a level of depth, practicality, and evaluation rigor that has not been fully explored in previous research, highlighting the novelty of this work in both academic and applied contexts [9].

3. Research methods

This study employs a quantitative comparative approach to evaluate the interpretability of various machine learning models used in credit scoring. The research stages include literature review, data acquisition from Bank Rakyat Indonesia (BRI), data preprocessing and feature selection, model training, implementation of Explainable Artificial Intelligence (XAI) methods, and explanation quality evaluation using the Explanation Consistency Framework (ECF) [10].

The following classification algorithms are used:

1) Logistic Regression

A statistical model used to predict binary outcomes based on input features. Its core function is defined as:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}} \quad (1)$$

where X represents the feature vector and β are the model coefficients.

2) Random Forest

An ensemble method that builds multiple decision trees using random subsets of features and data samples [11]. Predictions are made by aggregating the outputs of individual trees (majority voting for classification). Random forest handles overfitting better than single decision trees and is robust to noise.

3) XGBoost (Extreme Gradient Boosting)

A gradient boosting framework that builds trees sequentially, minimizing loss using a second-order Taylor expansion. Its objective function is:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \quad (2)$$

where l is the loss function and Ω is the regularization term for complexity control.

4) Dataset and Features

The dataset used in this study consists of 2,080 anonymized loan application records obtained from BRI. Each record represents a customer and includes more than 80 numerical and categorical features [12]. The target variable is a binary outcome (Accepted/Rejected) indicating credit approval status.

The key features used for modeling and interpretation include:

- feature_15: Total number of loan applications submitted on non-bank platforms in the last 180 days
- feature_19: Average nominal balance of the customer's savings account
- feature_30: Education level index
- feature_59: Time (in days) between opening the account and applying for the loan
- feature_74: Balance-to-age ratio
- feature_80: Ratio between account age and customer age
- feature_12: Maximum risk score in the past 90 days
- feature_81: Expense-to-income ratio

Feature selection was conducted based on statistical correlation and business relevance. These features were then normalized and encoded as required before training the models.

4. Reasearch Discussion

Shapley Additive Explanations (SHAP) is a post-hoc explanation method that builds an explanation model on top of a prediction model to understand how decisions are made. Shapley

values are calculated for each feature in each row of data to measure its influence on the prediction. The SHAP library provides a variety of visualization graphs to help interpret the results. In global explanation, the main focus is on identifying the most influential features and how they affect the model's decisions. The average of the absolute Shapley values is used to rank the features based on their impact on the prediction. The higher the Shapley value of a feature, the greater its influence on the model's decisions.

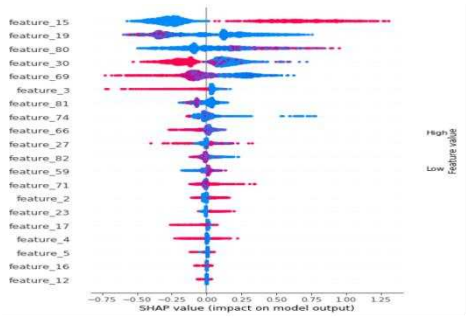

Figure 1. Feature importance graph on XGBoost

Based on the analysis in Figure 1, the features with the greatest influence in the XGBoost model are the average amount of savings (feature_19) and the number of loan applications on non-bank platforms in the last 180 days (feature_15), which significantly affect the credit scoring prediction results.

To find out how each feature affects the model in making decisions, the Shapley value can be visualized using beeswarm plots. Beeswarm plots are made using the implementation in the shap library. Figure 2, Figure 3, and Figure 4 show the level of importance of a feature and how it affects predictions in the logistic regression, XGBoost, and random forest models. In this graph, the features are sorted by their level of importance according to the Shapley value of the feature. Each point in each feature row represents the value of a sample for that feature. A red point indicates a high value for that sample, and conversely a blue point means the feature value for that sample is low.

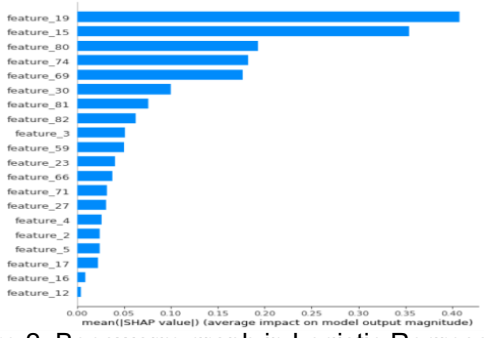

Figure 2. Beeswarm graph in Logistic Regression

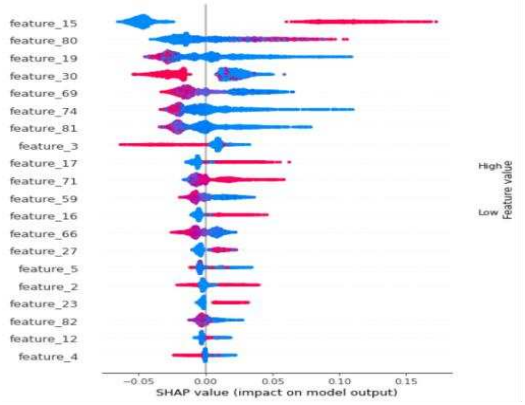

Figure 3. Beeswarm graph on XGBoost

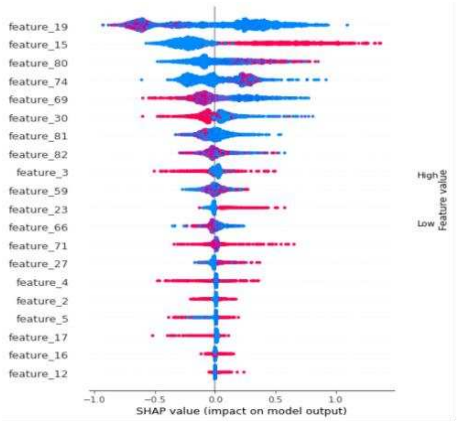

Figure 4. Beeswarm graph in Random Forest

The Shapley value shows the influence of a feature on the model prediction, where the point on the right has a positive impact on the Rejected classification, while the point on the left shows a negative impact. The higher the Shapley value of a feature, the more likely it is to encourage the model to classify an individual as Rejected. Although the features used are similar in all three models, the order of influence is different. Based on the beeswarm plot, individuals with high savings (feature_19), higher education level (feature_30), and high income-to-expense ratio are more likely to be classified as Accepted, because these features have a negative correlation with the Shapley value. This indicates that individuals with these characteristics have a low risk of default. In contrast, feature_15, which indicates the number of loan applications on non-bank platforms in the last 180 days, has a positive correlation with the Rejected decision, because frequent loan applications can indicate high risk. However, not all features have. A consistent impact, such as feature_71 (Britama savings opening time), which in some cases has a positive influence, but in other situations is negative. By understanding the features that have a significant impact, SHAP helps assess whether model decisions are in line with business expectations and intuition [13].

4.1 Global Explanation With BRCG

Boolean Rule Column Generation (BRCG) is a binary classification model that is designed to be understood without the help of explanation methods. BRCG provides explanations by showing rules in the form of disjunctive normal form (DNF) or conjunctive normal form (CNF) that the model uses to perform classification. The BRCG implementation used in this study is part of the IBM AI Explain Ability toolkit (AIX360). The number of clauses in the rule can be determined by setting the lambda0 and lambda1 parameters that regulate the complexity of the rules to be used by BRCG. lambda0 regulates the fixed cost for each clause and lambda1 regulates the cost for each additional condition in one clause.

Table 1. Rules used by BRCG in making classifications based on the values of the parameters lambda0 and lambda1

lambda0	lambda1	Akurasi	Aturan yang digunakan
0.01	0.1	0.664	’feature 15 > −999.00’
0.1	0.01	0.801	’feature 15 > −999.00 AND feature 69 <= 9.00 AND feature 80 > 0.02’
0.01	0.01	0.801	’feature 15 > −999.00 AND feature 69 <= 9.00 AND feature 80 > 0.02’

lambda0	lambda1	Akurasi	Aturan yang digunakan
0.0001	0.0001	0.897	'feature 2 <= −396.02 AND feature 2 > −999.00', 'feature 15 > −999.00 AND feature 30 <= 6.00 AND feature 30 > 3.00', 'feature 15 > −999.00 AND feature 69 <= 9.00 AND feature 69 > 2.05 AND feature 71 > 2018.00 AND feature 80 > 0.02', 'feature 5 <= 39.30 AND feature 5 > −1111.00 AND feature 15 > −999.00 AND feature 69 <= 9.00 AND feature 80 > 0.01 AND feature 81 <= 1.34'

AND feature 80 > 0.02', 'feature 5 <= 39.30 AND feature 5 > −1111.00 AND feature 15 > −999.00 AND feature 69 <= 9.00 AND feature 80 > 0.01 AND feature 81 <= 1.34'

Table 1. Comparing the rules generated by BRCG based on the combination of lambda0 and lambda1 values. The smaller the lambda value, the higher the model accuracy, but the generated rules become more specific, making them difficult to understand. Therefore, users must balance between model accuracy and explainability. BRCG allows users to understand the reasoning behind classifying an individual as Accepted or Rejected based on the generated rules. For example, with lambda0 values of 0.01 and lambda1 values of 0.01, the generated rule is 'feature 15 > −999.00 AND feature 69 ≤ 9.00 AND feature 80 > 0.02', which means individuals with a history of applying for loans on non-bank platforms in the last 180 days, applying for loans before September, and having a savings account age-to-age ratio of more than 0.02 will be classified as Rejected. This rule is interesting because many individuals who applied for loans before September 2020-2021 experienced rejections, possibly due to the peak of the Covid-19 pandemic and the PPKM policy which made banks more conservative in approving loans By understanding the rules generated by BRCG, users can evaluate the model's decisions and identify aspects that need to be improved to improve accuracy and fairness in credit scoring [14].

4.2 Global Explanation With ALE

Accumulated Local Effect (ALE) describes how a feature affects the predictions made by the model. ALE calculates the effect of a feature on the model's prediction results by averaging the changes in the predictions for each change in the feature value within a certain interval. In this study, the author tries to compare the effect of feature_19 on the prediction results of the model. feature_19 describes the average nominal savings balance of prospective creditors.

Table 2. Effect of feature_19 on model predictions based on group

fitur_19 (dalam Rp)	Ukuran	Efek		
		Logistic Regression	XGBoost	Random Forest
0,00 - 420.452,00	416	0.169231	0.130048	0.05
420.452,00 - 1.177.623,00	416	0.05625	0.021875	0.033173
1.177.623,00 - 2.615.684,00	416	-0.027885	-0.006971	-0.005288
2.615.684,00 - 6.188.443,00	416	-0.047115	-0.026202	-0.017308
6.188.443,00 - 1.054.315.000,00	416	-0.044712	-0.035817	-0.024519

Table 2 shows the effect of feature_19 value on model prediction. The data is divided into five equal groups based on the average nominal savings. The Effect column shows the effect of

the average nominal savings balance on the probability of a prospective creditor being classified as Rejected. Based on Table 4.2, all three models have the same tendency. In general, the visible pattern is that the higher the average nominal savings balance of a person, the lower the probability of a person being classified by the model as Rejected. In the two groups with the lowest average nominal savings balance, the feature_19 value increases the probability of a person being classified as Rejected. Meanwhile, in the next three groups, the feature_19 value decreases the probability of a person being classified as Rejected. The higher the value of the savings balance, the lower the probability of a prospective creditor being classified as Rejected.

4.3 Local Explanation Experiment Results

1) Local Explanation With SHAP

Untuk penjelasan lokal, informasi yang ingin diketahui adalah mengapa seorang calon kreditur diklasifikasikan sebagai Diterima atau Rejected by the Model. To explain this, the shap library provides a graph called a force plot. In a force plot, each feature is depicted as a force that influences the model's prediction. The blue features on the right are the ones that encourage the model to classify an individual as Accepted. Meanwhile, the red features on the left are the ones that encourage the model to classify an individual as Rejected.

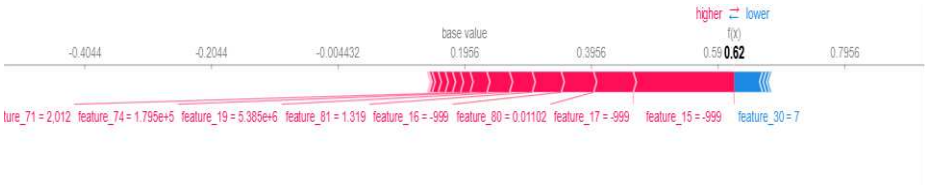

Figure 5. Force plot graph for Rejected individuals in Logistic Regression

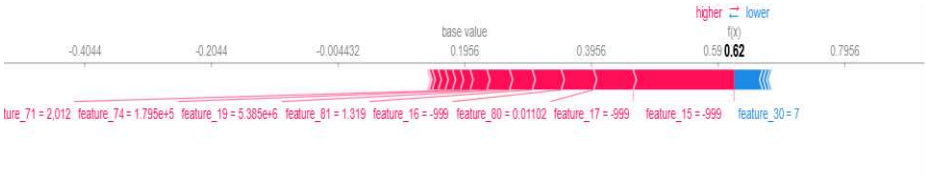

Figure 6. Force plot graph for Rejected individuals on XGBoost

Figure 5 and Figure 6 are force plots that explain how the influence of the features in each model causes the first individual to be classified as Accepted. As with the first individual, the influence of each feature on the model varies, although all classify the second individual as Rejected. In all three models, the factor that most influences the model to classify individual 2 as Rejected is feature_15 which indicates the total number of platforms where non-bank loan applications were made in the last 180 days. In general, the model classifies the second individual as Rejected because the second individual does not have a sufficient track record. This is indicated by the absence of a borrowing history in the last three months (feature_15), the difference between the age of the savings account and the time of borrowing (feature_59) is only three days, and the ratio of age to age of the savings account (feature_80) is quite high. Because the second individual does not have a sufficient track record, the model considers lending to the individual quite risky.

2) Local Explanation With Anchor

Anchor explains the prediction results of the classification model through rules called anchors, which are sufficient conditions for the model to make decisions [15]. This rule ensures that changes to other features will not affect the prediction results as long as the feature values in the anchors remain met. In this study, two precision thresholds were used, namely 0.90 and 0.95. At the threshold of 0.90, Anchor produces the same rule for all three models, where the first individual is classified as Accepted because he has an average savings balance (feature_19) > Rp2,887,740.75 and a balance to age ratio (feature_74) > 92,936.38. The precision of this rule ranges from 90-93%, with a coverage of around 23% of the sample. At the threshold of 0.95, the rules created remain the same with the addition of a maximum risk score clause in the last 90 days (feature_12) > -1034.62 and an expenditure to income ratio > 0.56 for XGBoost. Precision increases to 95-98%, but coverage is smaller, namely 6% for XGBoost and 20% for logistic regression and random forest. With this approach, Anchor helps understand the main factors that contribute to model decisions, providing more transparent insights into credit scoring.

Table 3. Comparison of explanations for Accepted individuals using Anchor

Model	Presisi	Cakupan	Anchors
Logistic Regression	0.93	0.23	feature 19 > 2887740.75 AND feature 74 >= 92936.38
	0.97	0.20	feature 19 >= 2887740.75 AND feature 74 >= 92936.38 AND feature 12 >= −1034.62
XGBoost	0.90	0.23	feature 19 >= 2887740.75 AND feature 74 >= 92936.38
	0.98	0.06	feature 19 >= 2887740.75 AND feature 74 >= 92936.38 AN feature 12 >= −1034.62 AND feature 82 >= 0.56
Random Forest	0.91	0.23	feature 19 >= 2887740.75 AND feature 74 >= 92936.38
	0.95	0.20	feature 19 >= 2887740.75 AND feature 74 >= 92936.38 AND feature 12 >= −1034.62

Table 4 shows the rules that Anchor created to explain why the three models classified the second individual as Rejected. For the second individual, Anchor provides different explanations for each model. One clause that appears in all three models is feature_15 which provides information about the total number of platforms where non-bank loan applications were made in the last 180 days, but there is still a difference in the value for this feature. The precision for each model ranges from 0.92 - 0.97 for the experiment with a threshold of 0.90 and 0.97 - 0.98 for the experiment with a threshold of 0.95. The coverage of anchors for the second individual ranges from 0.06 - 0.15 for the experiment with a threshold of 0.90 and 0.03 - 0.15 for the experiment with a threshold of 0.95.

Table 4. Comparison of explanations for Rejected individuals using Anchor

Model	**Presisi**	**Cakupan**	***Anchors***
Logistic Regression	0.97	0.15	feature 15 > −689.21 AND feature 59 <= 489.08
	0.97	0.15	feature 15 > −689.21 AND feature 59 <= 489.08
XGBoost	0.93	0.09	−999.00 < feature 15 <= 0.61 AND feature 80 <= 0.07 AND feature 17 > −999.00
	0.98	0.03	−999.00 < feature 15 <= 0.61 AND feature 80 <= 0.07 AND feature 17 > −999.00 AND feature 12 <= −1034.62
Random Forest	0.92	0.06	Anchor: −689.21 < feature 15 <= 0.61 AND feature 80 > 0.07 AND feature 81 <= 0.17
	0.98	0.03	−689.21 < feature 15 <= 0.61 AND feature 80 <= 0.07 AND 0.05 < feature 81 <= 0.17 AND feature 66 <= 5.00

Based on the explanation given by Anchor, the three models have different reasons for arriving at the same decision, but generally the differences are not too far apart. In general, the coverage of anchors for the second individual is smaller than the coverage of anchors for the first individual. This is because there are far fewer individuals classified as Rejected than individuals classified as Accepted. In using Anchor, users must balance the need for precision and coverage of the explanation. A high precision threshold will produce very specific anchors with low coverage. Anchors that are too specific can only be used to explain a small subset of the sample. Meanwhile, a low precision threshold will produce anchors with large coverage but low precision. Too low precision will make the resulting explanation unreliable. Because the explanation given by Anchor is in the form of rules that are relatively easy to understand, Anchor can be used to explain to prospective creditors why their loan application was accepted or rejected [16]. BRCG offers an interpretability approach that does not require additional processing because the model is already equipped with rules in easy-to-understand DNF or CNF formats. This makes it an ideal solution for transparent prediction models with good accuracy. However, its limitation lies in its ability to only support binary classification, limiting the scope of its use. ALE, as a feature importance method, can be used to understand the impact of features in a certain interval, even on correlated features. However, ALE is only able to analyze a maximum of two features at a time. Meanwhile, Anchor provides explanations in the form of simple rules, making it easier for prospective lenders to understand why their loan application was rejected. However, there is a trade-off between precision and coverage of the rules, which users should be aware of. ProtoDash, which explains predictions by comparing samples with similar prototypes, is useful for risk assessors but is less ideal for prospective lenders without in-depth knowledge of credit scoring [17].

Each explanation method has unique characteristics and different use cases. SHAP and ALE, while both show feature importance, have different approaches SHAP measures how much and how a feature influences the prediction, while ALE focuses on the change in prediction within a feature interval. Rule-based explanations, such as Anchor and BRCG, are more suitable for novice users because they are easier to understand. Additionally, feature importance methods such as SHAP can be used for feature selection, especially when combined with ablation studies, which evaluate the impact of a feature by removing it from the model. By combining these techniques, users can select features that are highly influential and consistent with business intuition, improving the transparency and accuracy of the model in credit scoring.

This study provides a significant contribution to the growing body of literature on Explainable Artificial Intelligence (XAI) in credit scoring by offering a broader comparative perspective than previous works. Unlike earlier studies that often focused on a single explanation method or simplified case studies, this research integrates five distinct XAI techniques SHAP, ALE, BRCG, Anchor, and Proto Dash and applies them simultaneously to a real-world credit scoring system at Bank Rakyat Indonesia (BRI). This comprehensive integration strengthens prior findings such as those of Barredo Arrieta et al. (2020) [18], who emphasized the importance of transparency in AI for financial decision-making, and Zhou & Hooker (2023) [19], who argue that no single method can fully capture the complexities of modern credit models.

In addition, this study enhances the evaluation of explanation quality by incorporating the Explanation Consistency Framework (ECF), as proposed by Doshi-Velez and Kim (2017) [20], which has not been widely applied in empirical XAI studies within the credit scoring domain. The ECF allows for a structured, multi-dimensional evaluation based on identity, separability, and stability addressing a major gap identified in previous interpretability research [21] .

Furthermore, the use of actual institutional data from BRI adds substantial practical relevance, moving beyond the synthetic or small-scale datasets used in earlier works, such as the simulated examples in Ribeiro et al.'s (2018) development of Anchor [4]. This real-world deployment bridges a significant research-to-practice gap and demonstrates how XAI can be operationalized effectively in financial institutions in emerging markets.

Altogether, the findings of this study not only reinforce key outcomes from prior research such as the effectiveness of SHAP and ALE in global explanation tasks but also highlight underexplored limitations, such as Anchor's lower identity consistency, offering new insights into the comparative robustness of these methods. Therefore, this work contributes to the consolidation of XAI knowledge by integrating empirical evidence with theoretical evaluation, and it provides a methodological and practical reference point for future development of interpretable AI in finance.

5. Conclusion

This study shows how explainability methods can improve the transparency of machine learning models in credit scoring. The models used include logistic regression, XGBoost, and random forest, with explanations divided into global and local. Global explanations, using SHAP and ALE, highlight features that are influential in the model's decisions. An alternative is a predictive model that can explain its own decisions, such as BRCG, which provides simple rules for classification. Local explanations use SHAP, Anchor, and ProtoDash to explain the model's decisions for a given individual, each with a different approach: SHAP shows the impact of each feature, Anchor provides rules, and ProtoDash compares with other samples.

Evaluation using the Explanation Consistency Framework (ECF) shows that the post-hoc explanation method has high consistency, although Anchor scores lower on the identity axiom due to the effect of randomization. In conclusion, the explanation method can help overcome the complexity of the black box model in credit scoring. BRI can adopt this method to explore more complex models with better performance. The choice of method must be adjusted to the purpose of the explanation and the intended audience, ensuring that credit decisions are more transparent and reliable.

Daftar Referensi

[1] M. Bekhet and S. Eletter, "Credit scoring model using logistic regression technique: A study of banks in Jordan," *Eur. Sci. J*, vol. 10, no. 10, pp. 271–281, 2014.

[2] M. R. Honegger and S. Blanc, "Shedding light on black box machine learning algorithms," Master's," *thesis, Karlsruhe Inst. Techno*, 2018.

[3] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *in Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[4] C. G. M. T. Ribeiro, S. Singh, "Anchors: High-precision model-agnostic explanations," in *in Proc. AAAI Conf. Artificial Intelligence*, 2018, pp. 1527–1535.

[5] and J. S. C. D. V. Carvalho, E. M. Pereira, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, pp. 832-841, 2019.

[6] and C. A. K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, "Efficient data representation by selecting prototypes with importance weights," in *in Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2019, pp. 260–269.

[7] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *J. R. Stat. Soc. Ser. B*, vol. 82, no. 4, pp. 1059–1086, 2020.

[8] H. Hilmin, "Internasionalisasi nilai moderasi beragama dalam kurikulum merdeka," *Muaddib*, vol. 7, no. 1, pp. 37–45, 2024.

[9] M. Mukhlis, "Signifikansi dan kontribusi guru PAI dalam pembentukan karakter siswa," *Integr. Educ. J*, vol. 1, no. 1, pp. 22–42, 2024.

[10] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box," *Harvard J. Law Technol*, vol. 31, no. 2, pp. 841–887, 2018.

[11] A. D. Saputra and A. Tunnafia, "Penguatan pendidikan karakter pada anak sekolah dasar," *Phenomenon*, vol. 2, no. 2, pp. 69–92, 2024.

[12] N. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," in *arXiv preprint arXiv:1702.08608*, 2017.

[13] I. Rahmawati, "Kesehatan mental siswa korban bullying," *Undergrad. thesis Univ. Ponorogo*, 2023.

[14] A. P. Setiani and L. N. Hidayah, "Dampak bullying terhadap kesehatan psikologis siswa," *Liberosis*, vol. 2, no. 1, pp. 41–50, 2024.

[15] R. V. Astifionita, "Memahami dampak bullying pada siswa sekolah menengah," *Iebah*, vol. 18, no. 1, pp. 36–46, 2024.

[16] B. Waltl and R. Vogl, "Increasing transparency in algorithmic-decision-making with explainable AI," *Datenschutz und Datensicherheit*, vol. 42, no. 10, pp. 613–617, 2018.

[17] and D. W. S. Dash, O. Günlük, "Boolean decision rules via column generation," *arXiv Prepr. arXiv1805.09901*.

[18] Barredo Arrieta, A., Díaz-Rodríguez, N., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[19] R. Zhou, J., & Hooker, "Explainable AI for credit scoring: A survey," *J. Financ. Technol.*, vol. 5, no. 2, pp. 123–139, 2023.

[20] B. Doshi-Velez, F., & Kim, "Doshi-Velez, F., & Kim, B," *arXiv Prepr. arXi 1702.08608.*, 2017.

[21] M. Chen, L., Liu, P., & Zhang, “Rule-based explanations for tree ensembles,” in *In Proc. IJCAI 2020*, 2020, pp. 5432–5438.