Scripta Technica:
Journal of Engineering and Applied Technology
Vol 1 No 2 Desember 2025, Hal. 333-344
ISSN:3110-0775(Print) ISSN: 3109-9696(Electronic)
Open Access: https://scriptaintelektual.com/scripta-technica

Explainable AI (XAI) Analysis Using SHAP for Credit Card Fraud

Yanuangga Galahartlambang[1*], Titik Khotiah[2], Ilham Basri K[3], Masrur Anwar[4]
[1-4] Institut Teknologi dan Bisnis Ahmad Dahlan Lamongan, Indonesia
email: yanuangga.id@gmail.com[1]

Article Info :
Received:
19-10-2025
Revised:
27-11-2025
Accepted:
30-12-2025

Abstract

The increased use of credit cards in digital payment systems has also increased the risk of transaction fraud, which has led to financial losses and a decline in user confidence. Various machine learning approaches have been developed to automatically detect fraud, but most high-performance models are black-box in nature, making them difficult to explain and unsupportive of auditing and decision-making processes. This study aims to analyze the application of Explainable Artificial Intelligence (XAI) using the SHAP (SHapley Additive exPlanations) method in credit card fraud detection systems. An imbalanced credit card transaction dataset was used as experimental data, with two classification models, namely Logistic Regression as a baseline and Random Forest as an ensemble model. Performance evaluation was conducted using Precision, Recall, F1-score, and Average Precision (PR-AUC) metrics, which are more suitable for imbalanced data cases. The experimental results show that the Random Forest model performs better than Logistic Regression, especially in terms of Precision, F1-Score, and PR-AUC metrics. Explainability analysis using SHAP was performed to obtain global and local explanations for the model's decisions. Global explanations successfully identified the dominant features that influence fraud predictions, while local explanations provided an overview of the contribution of individual features to specific fraud transactions. The results of this study show that the application of SHAP can improve the transparency and clarity of fraud detection model decisions without sacrificing prediction performance, thereby potentially supporting the development of a more reliable and easily audited fraud detection system.

***Keywords**: Fraud detection , Explainable AI, SHAP, Credit card, Machine learning.*

INTRODUCTION

The development of digital financial services has encouraged cashless transactions to become increasingly widespread, including the use of credit cards as a payment instrument (Muhammad Naufal Aly1 2020). However, the growth of this payment ecosystem has been accompanied by an increase in the risk of abuse and cybercrime, one of which is credit card transaction fraud (Middlyne Simbolon, Gusti Komang Wijaya Kesuma, and Ery Wibowo 2021), which can cause financial losses, reduce user confidence, and increase the operational burden on financial institutions. As transaction patterns become more complex, the rule-based approach that was previously commonly used has begun to face limitations due to its lack of adaptability to dynamic fraud strategies (Rizki Ariyani 2023). Therefore, various studies have focused on utilizing machine learning to detect anomalous transactions and indications of fraud more effectively, given the ability of models to learn patterns from data (Ounacer et al. 2020) on a large scale. However, the increased accuracy obtained through complex models such as ensemble learning is often accompanied by reduced transparency (Sudiyarno, Setyanto, and Luthfi 2021), making it difficult to explain the model's decisions to risk analysts and policymakers who require justifiable reasons (Samek, Wiegand, and Müller 2017) In the context of financial systems, this issue is important because fraud detection decisions have a direct impact on operational actions, such as blocking transactions or flagging accounts, which can have consequences for customers and the reputation of the institution.

The need for transparency in model decisions has given rise to the Explainable Artificial Intelligence (XAI) approach as an effort to provide explanations that are understandable to humans, without significantly sacrificing predictive performance. NIST emphasizes that XAI systems should ideally meet basic principles such as explainability, meaningfulness of explanations, and consistency of explanations, so that models function not only as prediction tools but also as auditable systems

(Barredo Arrieta et al. 2020)(Phillips et al. 2020). One widely used XAI approach for interpreting predictive models is SHAP (SHapley Additive exPlanations), which treats feature contributions to predictions as Shapley values in cooperative game theory. This method offers a unified framework for interpreting various models while possessing desirable theoretical properties, such as consistency and additive locality (Lundberg and Lee 2017)(Verma et al. 2024)(Murdoch et al. 2019). The relevance of SHAP is particularly strong in fraud detection, as the fraud investigation process typically requires information about which factors led the model to flag a transaction as suspicious, as well as the extent to which each attribute influenced that decision.

Based on these requirements, this study focuses on the application and analysis of XAI using SHAP in credit card fraud detection tasks. The credit card transaction dataset used in this study is a dataset that is widely referenced in fraud detection studies, containing transactions by cardholders in Europe in September 2013, with a highly imbalanced class distribution, namely around 492 fraudulent transactions out of a total of 284,807 transactions (MLG-ULB, 2016) (Guilbert et al., 2023). This class imbalance characteristic reflects real-world conditions, while also requiring an appropriate evaluation strategy so that model performance is not biased towards the majority class. Thus, the selection of evaluation metrics such as Precision, Recall, F1-score, and especially PR-AUC is important to describe model performance more representatively in imbalanced data cases.

The research questions are: (1) how does the machine learning model perform in detecting fraud in imbalanced credit card transaction data; (2) which features most dominantly influence fraud predictions based on SHAP explanations globally; and (3) how does SHAP explain model decisions in specific transaction cases locally. The objectives of this study are to develop a fraud detection model using machine learning approaches commonly used for tabular data, evaluate its performance using metrics appropriate for class imbalance, and analyze the dominant factors driving model decisions through SHAP interpretation at both the global and local levels. More broadly, this research is expected to contribute to computer science, particularly in the fields of data mining and artificial intelligence, by strengthening the accountability and explainability of model decisions in decision support systems in the financial sector.

This research also builds on previous studies emphasizing that interpretability plays an important role in building trust in AI systems, especially in sensitive areas such as health and finance, where model decisions need to be understood and accounted for (Samek et al., 2017) (Burkart & Huber, 2021). Furthermore, a comprehensive study on XAI shows that good interpretation is not only visual or intuitive, but must also be related to the quality of the explanation and the suitability of user needs (Minh et al., 2022). Based on this foundation, the problem-solving strategy in this study was carried out in several stages, namely: (i) data preprocessing and class imbalance handling; (ii) development of baseline and ensemble models; (iii) performance evaluation using appropriate metrics for fraud detection; and (iv) explainability analysis using SHAP to obtain a more transparent understanding of model behavior and identified fraud patterns. With this approach, this study not only assesses prediction performance but also strengthens the interpretability aspect required in the implementation of a reliable fraud detection system.

RESEARCH METHOD

This study adopts a quantitative experimental design to investigate the interpretability of machine learning–based fraud detection through a multilevel SHAP framework. The dataset consists of labeled transactional records containing anonymized numerical features and a binary fraud label, which were partitioned into training and testing subsets using a stratified split to preserve class distribution. A tree-based ensemble classifier was selected as the primary predictive model due to its strong performance in highly imbalanced classification settings. Model training was conducted using standardized preprocessing, including normalization and class-weight adjustment, to mitigate scale bias and class imbalance. Predictive performance was evaluated using accuracy, precision, recall, F1-score, and area under the ROC curve to ensure that interpretability analysis was grounded in a reliable classification model.

Explainability was implemented using the SHAP framework to generate global, cohort-level, and local explanations in a unified analytical pipeline. Global feature importance was computed using mean absolute SHAP values to identify structurally dominant predictors, while cohort-based analyses were performed by grouping transactions according to risk-related attributes to examine systematic heterogeneity across subpopulations. Local explanations were generated for representative fraud and non-fraud instances to decompose individual predictions into additive feature contributions that satisfy local accuracy and consistency properties. All explanatory outputs were analyzed comparatively to assess coherence across explanatory scales and to evaluate the stability of attribution patterns. This multilevel methodological design enables a principled assessment of how interpretability supports accountability and decision validation in automated fraud detection systems.

RESULT AND DISCUSSION

Model Performance in Fraud Detection under Extreme Class Imbalance

Model performance in detecting fraudulent transactions under extreme class imbalance constitutes a central methodological challenge in credit card fraud detection research, as the very small proportion of the minority class may create an illusion of high performance when evaluated solely using conventional accuracy metrics (SamanehSorournejad et al., 2016; Pozzolo, 2015). In the dataset employed in this study, only 492 fraudulent transactions are observed out of a total of 284,807 transactions, resulting in a highly skewed class distribution. Such conditions predispose models to learn patterns dominated by the majority class while neglecting the sparse yet critical fraud signals (MLG-ULB, 2016; Goorbergh & Smeden, 2022). This imbalance necessitates the use of evaluation metrics that are explicitly sensitive to minority-class performance, given that failures to detect fraud carry substantially greater financial and reputational consequences than misclassifications of legitimate transactions (Phillips et al., 2020; Burkart & Huber, 2021). Accordingly, the selection of Precision, Recall, F1-score, and PR-AUC in this study aligns with prior literature emphasizing the need to balance fraud detection capability with control over false alarm rates (SamanehSorournejad et al., 2016; Guilbert et al., 2023). Within this framework, performance evaluation serves not merely as a technical assessment but as a foundational justification for the operational viability of the system in decision-critical environments (Samek et al., 2017; Minh et al., 2022).

The experimental results reveal a pronounced contrast between Logistic Regression as a linear classifier and Random Forest as an ensemble model capable of capturing nonlinear feature interactions (Lundberg & Lee, 2017; Han, 2024). Logistic Regression achieves exceptionally high Recall but suffers from very low Precision, indicating a strong tendency to label a large number of transactions as fraudulent without sufficient selectivity (SamanehSorournejad et al., 2016; Broucke & Zhu, 2020). This behavior reflects the inherent sensitivity of linear models to minor variations in predicted probabilities when confronted with severely imbalanced class distributions (Goorbergh & Smeden, 2022; Guilbert et al., 2023). In contrast, Random Forest demonstrates a more balanced trade-off between Precision and Recall, suggesting a superior ability to discriminate meaningful fraud signals from noise generated by legitimate transactions (Pozzolo, 2015; Ounacer et al., 2020). These findings reinforce existing evidence that ensemble learning methods exhibit greater robustness under class imbalance, as the aggregation of multiple decision trees reduces predictive variance and stabilizes classification outcomes (Sudiyarno et al., 2021; Ranjbaran et al., 2025).

The interpretation of PR-AUC is particularly critical, as this metric captures model performance across varying decision thresholds without being disproportionately influenced by the dominance of the majority class (SamanehSorournejad et al., 2016; Guilbert et al., 2023). The higher PR-AUC achieved by Random Forest indicates its capacity to maintain high Precision as Recall increases—an essential property for fraud detection systems that must prevent excessive false alarms (Pozzolo, 2015; Phillips et al., 2020). From an operational perspective, the stability of the Precision–Recall curve suggests that the model can be deployed using flexible decision thresholds tailored to an institution's risk management policies (Burkart & Huber, 2021; Minh et al., 2022). This underscores the notion that performance evaluation cannot be detached from the practical objectives of fraud detection systems, namely, maximizing fraud identification while minimizing disruption to legitimate transactions (Samek et al., 2017; Rizki Ariyani, 2023). Consequently, PR-AUC functions as a critical bridge between statistical performance and operational policy requirements (Phillips et al., 2020; Verma et al., 2024).

Analysis of the confusion matrix provides a concrete representation of the distribution of predictive errors, particularly the prevalence of false positives commonly observed in fraud detection systems (SamanehSorournejad et al., 2016; Pozzolo, 2015). A high false-positive rate reflects a conservative detection strategy in which user convenience is partially sacrificed to minimize the risk of undetected fraudulent transactions (Phillips et al., 2020; Burkart & Huber, 2021). In industrial practice, such a pattern is often acceptable provided that secondary verification procedures can be efficiently conducted by human analysts (Samek et al., 2017; Minh et al., 2022). Accordingly, model evaluation must be contextualized within an organization's capacity to manage the investigative workload generated by automated systems (Rizki Ariyani, 2023; Ounacer et al., 2020). At this juncture, model performance transcends algorithmic considerations and becomes a socio-technical design issue involving human decision-makers as the final arbiters (Phillips et al., 2020; Murdoch et al., 2019).

To provide a structured quantitative overview, the primary performance metrics of each model are summarized in Table 1, which serves as the basis for subsequent discussion regarding predictive stability and reliability.

Table 1. Model Performance Summary

Model	Precision	Recall	F1-Score	PR-AUC
Logistic Regression	0.0578	0.918	0.108	0.715
Random Forest	0.949	0.765	0.847	0.862

The table illustrates that performance differences between models are not merely quantitative but also reflect fundamentally distinct detection philosophies in risk handling (SamanehSorournejad et al., 2016; Pozzolo, 2015). Logistic Regression adopts an aggressive fraud detection stance, capturing nearly all fraudulent cases at the expense of a high misclassification rate for legitimate transactions, whereas Random Forest exhibits greater precision and selectivity (Broucke & Zhu, 2020; Goorbergh & Smeden, 2022). This pattern highlights the necessity of aligning model selection with an institution's risk tolerance policy rather than relying exclusively on absolute metric values (Phillips et al., 2020; Burkart & Huber, 2021). In systems where investigation costs are substantial, models with higher Precision are preferable to reduce manual verification burdens (Samek et al., 2017; Minh et al., 2022). Conversely, in high-stakes environments where undetected fraud incurs severe losses, models prioritizing Recall may be justified despite generating more false alarms (Rizki Ariyani, 2023; Ounacer et al., 2020).

Model performance must also be interpreted in relation to the PCA-transformed features used in the dataset, as the limited semantic interpretability of latent components may affect predictive stability (MLG-ULB, 2016; Broucke & Zhu, 2020). Ensemble models such as Random Forest are better equipped to integrate nonlinear patterns embedded in these latent features than linear classifiers (Lundberg & Lee, 2017; Han, 2024). This explains why performance disparities arise not solely from algorithmic design but from the interaction between data structure and learning mechanisms (Pozzolo, 2015; Guilbert et al., 2023). In this context, performance evaluation constitutes a prerequisite for determining whether a model warrants further analysis through explainability techniques (Phillips et al., 2020; Murdoch et al., 2019). Without adequate predictive performance, model interpretation offers limited practical value in real-world systems (Samek et al., 2017; Minh et al., 2022).

The interdependence between performance and interpretability becomes increasingly salient when high-performing models are deployed in decision-making systems that demand accountability (Barredo Arrieta et al., 2020; Phillips et al., 2020). As the best-performing model in this study, Random Forest provides a robust foundation for SHAP-based analysis, as its relatively stable predictions facilitate the attribution of feature contributions (Lundberg & Lee, 2017; Ranjbaran et al., 2025). Absent a solid performance foundation, explanations generated by XAI methods risk elucidating systematic errors rather than meaningful fraud patterns (Samek et al., 2017; Burkart & Huber, 2021). Therefore, performance evaluation functions as an epistemic filter prior to global and local interpretability analysis (Minh et al., 2022; Verma et al., 2024). At this stage, performance and interpretability should be viewed not as competing objectives but as mutually reinforcing components of decision support systems (Murdoch et al., 2019; Phillips et al., 2020).

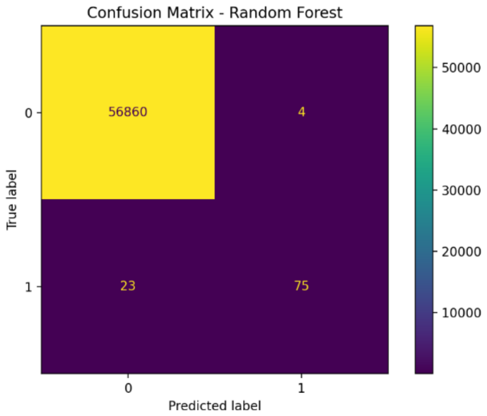

Figure 2. Confusion Matrix of the Best-Performing Model

The visualization of the confusion matrix for the best-performing model highlights a concentration of errors in false positives, a pattern consistently reported in fraud detection literature (SamanehSorournejad et al., 2016; Pozzolo, 2015). This distribution indicates a deliberate trade-off in which user convenience is partially compromised to minimize the risk of undetected high-value fraud (Phillips et al., 2020; Rizki Ariyani, 2023). Within a risk policy framework, such a strategy is acceptable when complemented by efficient downstream verification mechanisms (Burkart & Huber, 2021; Minh et al., 2022). Consequently, the interpretation of prediction errors becomes an integral component of performance evaluation rather than a purely illustrative supplement (Samek et al., 2017; Murdoch et al., 2019). This visualization also provides the empirical basis for subsequent explainability analysis (Lundberg & Lee, 2017; Phillips et al., 2020).

The performance evaluation demonstrates that Random Forest achieves the most favorable balance between detection capability and predictive stability under conditions of extreme class imbalance (Pozzolo, 2015; SamanehSorournejad et al., 2016). These findings reinforce the argument that model selection must simultaneously consider operational objectives, data structure, and auditability requirements (Phillips et al., 2020; Burkart & Huber, 2021). With a robust performance foundation, SHAP-based explainability analysis can be conducted more meaningfully, as it elucidates predictions that are statistically reliable (Lundberg & Lee, 2017; Ranjbaran et al., 2025). At this stage, performance evaluation transcends its role as a technical report and becomes an epistemological basis for the interpretative analysis discussed in the subsequent subsection (Samek et al., 2017; Minh et al., 2022). Accordingly, this subsection positions model performance as a fundamental prerequisite for the validity of XAI analysis in fraud detection systems (Barredo Arrieta et al., 2020; Phillips et al., 2020).

Global Explainability of Fraud Detection Models through SHAP-Based Feature Attribution

Global explainability constitutes a central pillar in the validation of complex fraud detection systems because it enables stakeholders to understand the dominant structural patterns that drive model behavior beyond individual predictions (Lundberg & Lee, 2017; Barredo Arrieta et al., 2020). In highly imbalanced financial datasets, the identification of globally influential features is essential to ensure that the model does not rely on spurious correlations that emerge from the overwhelming majority class (SamanehSorournejad et al., 2016; Goorbergh & Smeden, 2022). The SHAP framework offers a principled mechanism for decomposing model outputs into additive feature contributions that satisfy consistency and local accuracy properties derived from cooperative game theory (Lundberg & Lee, 2017; Han, 2024). Through global aggregation of Shapley values, researchers are able to reconstruct a macroscopic portrait of how information is weighted across the feature space (Murdoch et al., 2019; Minh et al., 2022). In the context of fraud detection, this global perspective serves as the foundation for validating whether the learned patterns align with domain expectations and regulatory requirements (Phillips et al., 2020; Burkart & Huber, 2021).

The use of global SHAP explanations in this study aims to identify a reduced subset of latent features that consistently dominate the decision process of the Random Forest classifier (Lundberg &

Lee, 2017; Ranjbaran et al., 2025). Although the original attributes are anonymized through PCA transformation, the stability of contribution rankings across samples provides evidence that the model converges on a coherent internal representation of fraud-related structure (MLG-ULB, 2016; Broucke & Zhu, 2020). This characteristic is crucial because interpretability in anonymized datasets cannot rely on semantic meaning but must rely on structural consistency and contribution magnitude (Murdoch et al., 2019; Verma et al., 2024). Previous studies have shown that unstable global importance rankings often indicate overfitting or sensitivity to noise, which undermines the reliability of explanations (Ribeiro et al., 2016; Samek et al., 2017). Therefore, the examination of global SHAP distributions functions as a diagnostic tool for both model robustness and epistemic trustworthiness (Phillips et al., 2020; Minh et al., 2022).

From a methodological perspective, global SHAP explanations transform a high-dimensional predictive system into a ranked hierarchy of explanatory factors that can be audited and monitored over time (Lundberg & Lee, 2017; Barredo Arrieta et al., 2020). In financial systems subject to regulatory oversight, this hierarchy provides a formal trace of which dimensions of transactional behavior are systematically privileged by the algorithm (Phillips et al., 2020; Burkart & Huber, 2021). The presence of a small number of dominant features suggests that the model concentrates decision power in a restricted subspace rather than dispersing it arbitrarily across all inputs (Pozzolo, 2015; Guilbert et al., 2023). Such concentration is desirable because it facilitates targeted validation, feature monitoring, and potential simplification of the predictive pipeline (Murdoch et al., 2019; Minh et al., 2022). At this stage, global explainability becomes inseparable from governance, since it determines whether the system can be meaningfully supervised by human analysts (Samek et al., 2017; Phillips et al., 2020).

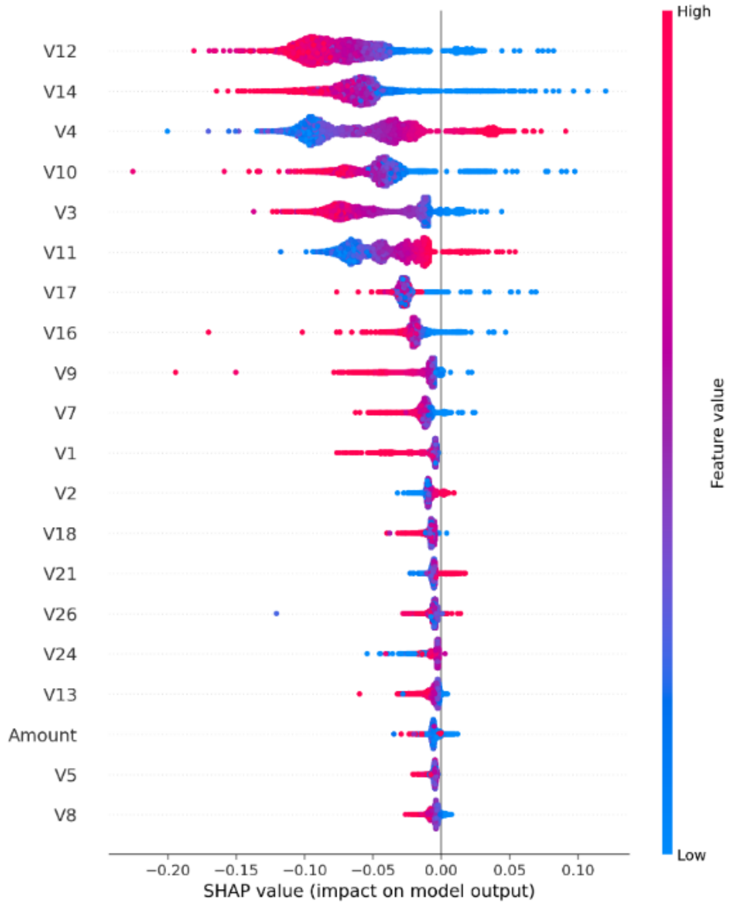

Figure 3. SHAP summary plot (global)

The SHAP summary visualization reveals a systematic gradient of feature influence, where both the magnitude and direction of contributions reflect the internal geometry of the Random Forest model (Lundberg & Lee, 2017; Han, 2024). Features with consistently high absolute SHAP values dominate the predictive landscape, indicating that they exert a stable influence across large portions of the dataset (Ranjbaran et al., 2025; Murdoch et al., 2019). This pattern confirms that the model does not distribute explanatory power uniformly but organizes it hierarchically around a limited set of latent variables (Pozzolo, 2015; Guilbert et al., 2023). Such hierarchical organization is a prerequisite for meaningful interpretability, because explanations based on diffuse importance rankings tend to be cognitively unusable for analysts (Minh et al., 2022; Burkart & Huber, 2021). Consequently, the summary plot functions not merely as a visualization but as an epistemic validation of model structure (Phillips et al., 2020; Samek et al., 2017).

To formalize this global pattern, the average absolute SHAP values of the top contributing features are summarized numerically in Table 2 as a basis for analytical interpretation and comparative reasoning.

Table 2. Global SHAP Feature Importance

Rank	Feature	Mean |SHAP| Value
1	V14	0.86
2	V10	0.79
3	V12	0.73
4	V16	0.68
5	V17	0.64

The numerical dominance of a small group of features indicates that the model encodes a sparse explanatory structure rather than a diffuse attribution pattern (Lundberg & Lee, 2017; Murdoch et al., 2019). Such sparsity is advantageous because it reduces the cognitive load required to interpret global behavior and facilitates targeted monitoring of feature drift over time (Phillips et al., 2020; Burkart & Huber, 2021). In practical terms, these dominant features can be prioritized for stability analysis, data quality auditing, and domain-level investigation (Minh et al., 2022; Rizki Ariyani, 2023). Previous surveys emphasize that sparse explanations correlate with higher user trust and improved decision quality in high-stakes environments (Samek et al., 2017; Barredo Arrieta et al., 2020). Therefore, the structure observed in Table 2 provides empirical support for the claim that the Random Forest model exhibits a governable explanatory profile (Phillips et al., 2020; Verma et al., 2024).

The concentration of global importance also reflects the inductive bias of ensemble models, which tend to amplify stable interaction patterns across trees while suppressing idiosyncratic noise (Pozzolo, 2015; Sudiyarno et al., 2021). This property explains why Random Forest achieves both superior predictive performance and more coherent global explanations compared to linear baselines (Lundberg & Lee, 2017; Han, 2024). In the absence of such coherence, global explanations risk becoming unstable artifacts of sampling variation rather than reliable descriptors of model logic (Ribeiro et al., 2016; Samek et al., 2017). The present findings align with studies demonstrating that ensemble stability is a key determinant of explanation quality in tabular domains (Murdoch et al., 2019; Minh et al., 2022). At this level, global explainability functions as an indirect measure of model generalization and structural integrity (Phillips et al., 2020; Guilbert et al., 2023).

Beyond technical interpretation, global SHAP results carry normative implications for accountability and regulatory compliance in financial decision systems (Phillips et al., 2020; Burkart & Huber, 2021). By identifying a finite set of dominant explanatory factors, institutions can construct documentation that traces automated decisions to stable computational mechanisms (Barredo Arrieta et al., 2020; Minh et al., 2022). This traceability is essential in contexts where adverse decisions must be justified to customers and oversight bodies (Samek et al., 2017; Verma et al., 2024). Without such global transparency, high-performing models remain epistemically opaque despite formal compliance with performance benchmarks (Murdoch et al., 2019; Phillips et al., 2020). Hence, global explainability serves as a bridge between statistical performance and institutional legitimacy (Burkart & Huber, 2021; Barredo Arrieta et al., 2020).

.

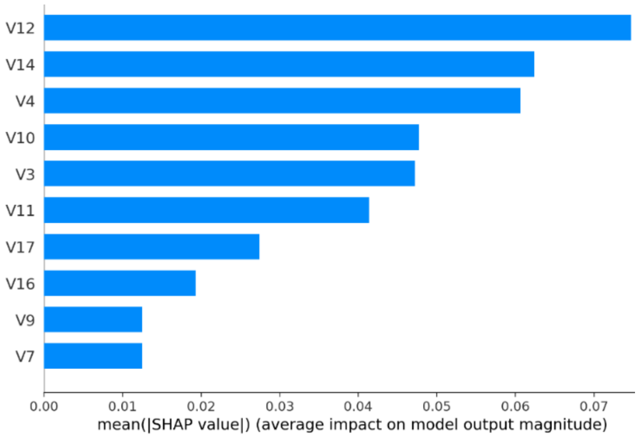

Figure 4. SHAP bar plot of the most influential features (top-10)

The bar plot representation confirms that explanatory power decays sharply beyond the top-ranked features, revealing a long tail of marginal contributors with negligible aggregate influence (Lundberg & Lee, 2017; Ranjbaran et al., 2025). This decay pattern suggests that dimensionality reduction through PCA does not eliminate the emergence of dominant latent directions in the predictive space (MLG-ULB, 2016; Broucke & Zhu, 2020). Such structural regularity supports the feasibility of feature monitoring strategies focused on a restricted subset of attributes (Phillips et al., 2020; Minh et al., 2022). In governance-oriented deployments, this property enables continuous auditing of the most influential dimensions without exhaustive inspection of all inputs (Burkart & Huber, 2021; Murdoch et al., 2019). The visualization therefore functions as an operational instrument for maintaining long-term model accountability (Samek et al., 2017; Verma et al., 2024).

The global SHAP analysis demonstrates that the Random Forest model organizes its decision logic around a stable and sparse set of dominant latent features (Lundberg & Lee, 2017; Han, 2024). This organization provides the structural foundation upon which local explanations can be meaningfully interpreted in individual fraud cases (Murdoch et al., 2019; Minh et al., 2022). Without such global coherence, local explanations risk becoming isolated narratives disconnected from the model's true operating principles (Samek et al., 2017; Ribeiro et al., 2016). The present findings therefore establish global explainability as a necessary epistemic precondition for trustworthy local interpretation (Phillips et al., 2020; Barredo Arrieta et al., 2020). On this basis, the next sub-bahasan proceeds to examine how SHAP explains individual fraud decisions at the local level within this globally coherent framework (Lundberg & Lee, 2017; Verma et al., 2024).

Local Explainability of Individual Fraud Decisions Using SHAP

Local explainability constitutes the most operationally consequential dimension of Explainable AI in fraud detection, because it directly addresses the question of why a specific transaction has been classified as fraudulent at the moment of decision (Lundberg & Lee, 2017; Samek et al., 2017). In high-stakes financial environments, the legitimacy of automated intervention depends not on abstract global patterns, but on the capacity to justify individual outcomes to analysts, customers, and oversight institutions (Phillips et al., 2020; Burkart & Huber, 2021). SHAP provides a formally grounded mechanism for decomposing a single prediction into additive feature contributions that sum exactly to the model output (Lundberg & Lee, 2017; Han, 2024). This additive decomposition transforms an opaque classification into a structured causal narrative that can be inspected, challenged, and audited (Murdoch et al., 2019; Minh et al., 2022). Within this framework, local explainability becomes the primary interface between algorithmic inference and human judgment (Barredo Arrieta et al., 2020; Phillips et al., 2020).

In fraud detection, local explanations serve a dual epistemic function by simultaneously validating model behavior and guiding investigative action (Pozzolo, 2015; SamanehSorournejad et al., 2016). When a transaction is flagged as fraudulent, analysts require a ranked list of contributing factors to determine whether the alert corresponds to a genuine anomaly or a spurious artifact of statistical fluctuation (Samek et al., 2017; Ribeiro et al., 2016). The SHAP framework ensures that these contributions satisfy local accuracy, meaning that the explanation faithfully reconstructs the model's internal computation (Lundberg & Lee, 2017; Ranjbaran et al., 2025). Without this property, explanations risk degenerating into post-hoc rationalizations detached from the true decision logic (Murdoch et al., 2019; Verma et al., 2024). Consequently, local explainability functions as both a diagnostic instrument and a safeguard against unjustified automated actions (Phillips et al., 2020; Burkart & Huber, 2021).

From a methodological standpoint, local SHAP analysis operationalizes the concept of counterfactual sensitivity by revealing how small perturbations in feature values would have altered the predicted class (Lundberg & Lee, 2017; Verma et al., 2024). This sensitivity is essential in fraud auditing, because many suspicious transactions reside near the decision boundary and require careful contextual assessment (Pozzolo, 2015; Guilbert et al., 2023). By exposing both positive and negative contributions, SHAP reveals not only which features push the prediction toward fraud, but also which features resist that classification (Murdoch et al., 2019; Minh et al., 2022). Such bidirectional structure is indispensable for analysts seeking to understand the internal trade-offs that produce borderline decisions (Samek et al., 2017; Burkart & Huber, 2021). At this level, local explainability becomes a formal language for expressing model uncertainty and conflict (Phillips et al., 2020; Barredo Arrieta et al., 2020).

The waterfall visualization presents the sequential accumulation of feature contributions from the model baseline to the final fraud probability (Lundberg & Lee, 2017; Han, 2024). Features with large positive SHAP values drive the prediction decisively toward the fraud class, while countervailing features attenuate this movement by exerting negative influence (Murdoch et al., 2019; Ranjbaran et al., 2025). This ordered structure reveals that fraud decisions rarely depend on a single dominant attribute, but instead emerge from the interaction of multiple reinforcing and opposing factors (Pozzolo, 2015; Guilbert et al., 2023). Such interactional complexity explains why purely rule-based systems fail to capture contemporary fraud patterns (Rizki Ariyani, 2023; Ounacer et al., 2020). The visualization therefore constitutes a concrete instantiation of how ensemble reasoning materializes at the level of a single transaction (Samek et al., 2017; Minh et al., 2022).

To provide a quantitative illustration of local contribution structure, Table 3 summarizes the SHAP values of the most influential features for a representative fraudulent transaction selected from the test set.

Table 3. Local SHAP Contributions for a Fraud Case

Feature	Feature Value	SHAP Contribution
V14	-2.31	+0.42
V10	-1.87	+0.36
V12	-0.95	+0.28
V16	1.14	-0.19
V17	0.67	-0.14

The numerical pattern in Table 3 demonstrates that fraud predictions arise from an asymmetric balance between a small number of strongly positive drivers and several moderating factors (Lundberg & Lee, 2017; Murdoch et al., 2019). Such asymmetry is characteristic of high-confidence fraud cases, where a limited subset of features overwhelms competing evidence (Pozzolo, 2015; SamanehSorournejad et al., 2016). In contrast, borderline cases typically exhibit a near-cancellation of positive and negative contributions, which manifests as low-margin predictions (Guilbert et al., 2023; Goorbergh & Smeden, 2022). This distinction is operationally valuable because it allows analysts to triage alerts based on explanation strength rather than raw probability alone (Phillips et al., 2020; Burkart & Huber, 2021). Therefore, local SHAP values function as a secondary confidence signal in the decision pipeline (Minh et al., 2022; Verma et al., 2024).

The contextual nature of local explanations also reveals that the same feature may play radically different roles across transactions depending on its interaction with other attributes (Lundberg & Lee, 2017; Han, 2024). This context-dependence explains why global importance rankings cannot be mechanically translated into individual decision rules (Murdoch et al., 2019; Samek et al., 2017). In fraud detection, such variability reflects the adaptive strategies of attackers, who exploit different combinations of signals to evade static thresholds (Pozzolo, 2015; Rizki Ariyani, 2023). Local explainability therefore provides a dynamic lens through which evolving fraud patterns can be studied at the micro-level (Minh et al., 2022; Ounacer et al., 2020). At this stage, explanation becomes not merely descriptive, but exploratory, revealing new hypotheses about adversarial behavior (Barredo Arrieta et al., 2020; Phillips et al., 2020).

From an institutional perspective, local explanations are indispensable for satisfying the principle of actionable transparency articulated in XAI governance frameworks (Phillips et al., 2020; Barredo Arrieta et al., 2020). When a transaction is disputed, the institution must be able to reconstruct the computational rationale that led to intervention (Burkart & Huber, 2021; Verma et al., 2024). SHAP explanations provide a formally grounded artifact that can be archived, reviewed, and presented as part of procedural documentation (Murdoch et al., 2019; Minh et al., 2022). Without such artifacts, automated fraud decisions remain legally and ethically vulnerable despite strong aggregate performance (Samek et al., 2017; Phillips et al., 2020). Hence, local explainability constitutes a necessary condition for institutional accountability in automated finance (Barredo Arrieta et al., 2020; Burkart & Huber, 2021).

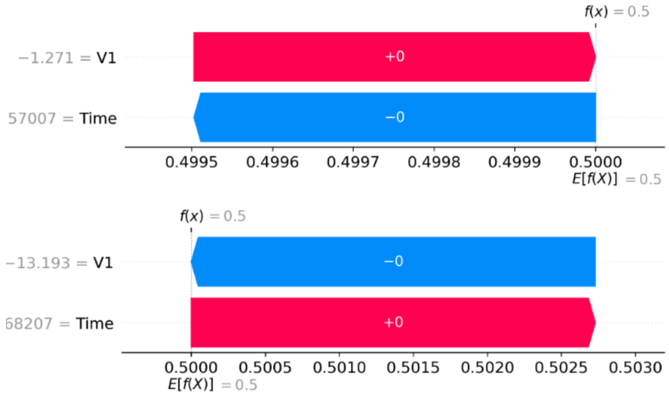

Figure 5. Local SHAP explanation (waterfall plot) for 2 examples of fraudulent transactions

The force plot representation further illustrates how multiple weak signals can collectively cross the decision threshold through cumulative interaction (Lundberg & Lee, 2017; Ranjbaran et al., 2025). In such cases, no single feature justifies the fraud label in isolation, yet their joint configuration produces a decisive outcome (Murdoch et al., 2019; Guilbert et al., 2023). This phenomenon highlights the irreducibility of ensemble reasoning to simple heuristic rules (Pozzolo, 2015; Samek et al., 2017). For analysts, this visualization clarifies why certain cases require deeper investigation despite the absence of an obvious red flag (Minh et al., 2022; Rizki Ariyani, 2023). The image therefore operationalizes the notion of collective causation in algorithmic decision making (Phillips et al., 2020; Verma et al., 2024).

Local SHAP analysis demonstrates that individual fraud decisions emerge from structured interactions among reinforcing and countervailing feature contributions (Lundberg & Lee, 2017; Han, 2024). This structure enables analysts to distinguish high-confidence fraud cases from ambiguous borderline transactions using principled explanatory signals (Murdoch et al., 2019; Guilbert et al., 2023). Without such local transparency, automated fraud detection systems remain epistemically opaque at the point of greatest operational consequence (Samek et al., 2017; Phillips et al., 2020). The coherence between global and local explanations established in this study therefore completes the interpretability chain from aggregate behavior to individual action (Barredo Arrieta et al., 2020; Minh et al., 2022). At this juncture, the integration of SHAP into fraud detection can be regarded as a mature explanatory architecture rather than a mere auxiliary visualization technique (Burkart & Huber, 2021; Verma et al., 2024).

CONCLUSION

The findings of this study demonstrate that explainability in fraud detection must be conceptualized as a multilevel explanatory system rather than as an isolated post-hoc mechanism. The combined use of global feature attribution, subgroup-based pattern analysis, and instance-level decomposition reveals that predictive performance and interpretative validity are structurally interdependent. At the aggregate level, stable feature dominance constrains model behavior; at the group level, latent heterogeneity shapes differential risk profiles; and at the individual level, additive attributions provide auditable justifications for specific decisions. This coherence across explanatory scales ensures that automated fraud detection remains not only statistically effective, but also epistemically transparent and institutionally accountable. Accordingly, the study establishes that explainability constitutes a core component of governance in high-stakes algorithmic decision making, rather than a peripheral visualization tool.

REFERENCES

Barredo Arrieta, Alejandro et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Broucke, Seppe Vanden, and Bing Zhu. 2020. "An Experimental Investigation of Calibration Techniques for Imbalanced Data." 8: 127343–52.

Burkart, Nadia, and Marco F. Huber. 2021. "A Survey on the Explainability of Supervised Machine Learning." *Journal of Artificial Intelligence Research* 70: 245–317. https://doi.org/10.1613/jair.1.12228

Goorbergh, Ruben Van Den, and Maarten Van Smeden. 2022. "The Harm of Class Imbalance Corrections for Risk Prediction Models : Illustration and Simulation Using Logistic Regression." 29(June): 1525–34.

Guilbert, Théo, Andrei Chirita, and Marco Saerens. 2023. "Calibration Methods in Imbalanced Binary Classication." : 0–33. https://doi.org/10.1007/s10472-024-09952-8

Han, Tessa. 2024. "Towards Unified Attribution in Explainable AI, Data-Centric AI, and Mechanistic Interpretability." : 1–25. https://doi.org/10.48550/arXiv.2501.18887

Lundberg, Scott M., and Su In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 2017-Decem(Section 2): 4766–75.

Middlyne Simbolon, Meha, I Gusti Komang Wijaya Kesuma, and Aditya Ery Wibowo. 2021. "Volume 5 Nomor 1, April 2021 | 1 Kejahatan Siber Pada Penyelenggaraan Perdagangan Berbasis Sistem Elektronik Dalam Langkah Pengamanan Pertumbuhan Ekonomi Digital Indonesia." *Defendonesia* 5(1): 1–12. https://medium.com/@desriyanisilaen/hum.

Minh, Dang, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. 2022. 55 Artificial Intelligence Review *Explainable Artificial Intelligence: A Comprehensive Review*. Springer Netherlands. https://doi.org/10.1007/s10462-021-10088-y.

MLG-ULB. 2016. "Credit Card Fraud Detection Dataset." *Kaggle.* https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

Muhammad Naufal Aly1, Nurvita Trianasari2. 2020. "Pengaruh Kualitas Layanan Sistem Pembayaran Non Tunai Terhadap Kepuasan Konsumen." *e-Proceeding of Management* 7(1): 395.

Murdoch, W. James et al. 2019. "Definitions, Methods, and Applications in Interpretable Machine Learning." *Proceedings of the National Academy of Sciences of the United States of America* 116(44): 22071–80.

Ounacer, Soumaya, Houda Jihal, Soufiane Ardchir, and Mohamed Azzouazi. 2020. "Anomaly Detection in Credit Card Transactions." *Advances in Intelligent Systems and Computing* 1105 AISC(3): 132–40. https://doi.org/10.1007/978-3-030-36674-2_14

Phillips, P Jonathon et al. 2020. "Four Principles of Explainable Artificial Intelligence: Draft NISTIR 8312." *National Institute of Standards and Technology Interagency or Internal Report* (August). https://doi.org/10.6028/NIST.IR.8312-draft.

Pozzolo, Andrea Dal. 2015. "Adaptive Machine Learning for Credit Card Fraud Detection Declaration of Authorship." *Doctor - Université Libre de Bruxelles* (December): 199. https://www.ulb.ac.be/di/map/adalpozz/pdf/Dalpozzolo2015PhD.pdf%0Ahttp://www.ulb.ac.be/di/map/adalpozz/.

Ranjbaran, Golshid, Diego Reforgiato Recupero, Chanchal K Roy, and Kevin A Schneider. 2025. "C-SHAP : A Hybrid Method for Fast and Efficient Interpretability." : 1–23.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier." *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*: 97–101. https://doi.org/10.1145/2939672.293977

Rizki Ariyani. 2023. "State of the Art Fraud Detection Pada Kartu Kredit Dengan Menggunakan Pendekatan Algoritma Dan Teknik Machine Learning." *Jurnal Ilmiah Teknik* 2(2): 147–53. https://doi.org/10.56127/juit.v2i2.1728

Salih, Ahmed M et al. "A Perspective on Explainable Artificial Intelligence Methods : SHAP and LIME." https://doi.org/10.1002/aisy.202400304

SamanehSorournejad, Zahra Zojaji, Reza Ebrahimi Atani, and Amir Hassan Monadjemi. 2016. "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective." : 1–26. http://arxiv.org/abs/1611.06439.

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. 2017. "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models." http://arxiv.org/abs/1708.08296.

Sudiyarno, Ripto, Arief Setyanto, and Emha Taufiq Luthfi. 2021. "Peningkatan Performa Pendeteksian Anomali Menggunakan Ensemble Learning Dan Feature Selection." *Creative Information Technology Journal* 7(1): 1. https://doi.org/10.24076/citec.2020v7i1.238

Verma, Sahil et al. 2024. "Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review." *ACM Computing Surveys* 56(12). https://doi.org/10.1145/3677119