Original Article

Check for updates

Available online at
www.heca-analitika.com/ijma

Indatu Journal of Management and Accounting

Vol. 3, No. 1, 2025

Credit Card Fraud Detection Through Explainable Artificial Intelligence for Managerial Oversight

Muksalmina Muksalmina [1,*], Ahmad Syahyana [1], Ferdy Hidayatullah [1], Ghalieb Mutig Idroes [2] and Teuku Rizky Noviandy [3]

[1] Department of Management, Faculty of Social Sciences and Education, Universitas Ubudiyah Indonesia, Banda Aceh 23114, Indonesia; muksalmina@uui.ac.id (M.M.); ahmadsyahyana@uui.ac.id (A.S.); ferdyhidayatullah@uui.ac.id (F.H.)
[2] Interdisciplinary Innovation Research Unit, Graha Primera Saintifika, Aceh Besar 23371, Indonesia; ghaliebidroes@outlook.com (G.M.I.)
[3] Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia; rizky_si@abulyatama.ac.id (T.R.N.)

* Correspondence: muksalmina@uui.ac.id

Article History

Received 19 March 2025
Revised 22 May 2025
Accepted 1 June 2025
Available Online 8 June 2025

Keywords:
Transparency
Accountability
Compliance
Interpretability
Decision support

Abstract

As digital payment systems grow in volume and complexity, credit card fraud continues to be a significant threat to financial institutions. While machine learning (ML) has emerged as a powerful tool for detecting fraudulent activity, its adoption in managerial settings is hindered by a lack of transparency and interpretability. This study examines how explainable artificial intelligence (XAI) can enhance managerial oversight in the deployment of ML based fraud detection systems. Using a publicly available, simulated dataset of credit card transactions, we developed and evaluated four ML models: Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest. Performance was assessed using standard metrics, including accuracy, precision, recall, and F1-score. The Random Forest model demonstrated superior classification performance but also presented significant interpretability challenges due to its complexity. To fill this gap, we applied SHAP (SHapley Additive exPlanations), a leading method for explaining the outputs of the Random Forest model. SHAP analysis revealed that transaction amount and merchant category were the most influential features in determining the risk of fraud. SHAP plots were used to make these insights accessible to non-technical stakeholders. The findings underscore the importance of XAI in promoting transparency, facilitating regulatory compliance, and fostering trust in AI-driven decisions. This study offers practical guidance for managers, auditors, and policymakers seeking to integrate explainable ML tools into financial risk management processes, ensuring that technological advancements are balanced with accountability and informed human oversight.

1. Introduction

Credit card fraud continues to pose a significant and growing threat to financial institutions, merchants, and consumers worldwide [1–4]. As digital transactions become more common and complex, fraudsters are adopting increasingly sophisticated tactics to exploit system vulnerabilities [5–7]. According to industry estimates, billions of dollars are lost annually due to fraudulent activity, making fraud detection a critical operational and strategic priority [8–12]. Traditional rule-

based systems, which rely on static if-then logic, are increasingly insufficient in identifying novel or subtle patterns of fraud [13]. These systems often produce high false-positive rates, fail to adapt to evolving threats, and lack the speed required for real-time decision-making [14]. Therefore, there is a clear and pressing need for more intelligent, adaptive, and interpretable fraud detection systems that can keep pace with emerging risks.

In response, many organizations have turned to machine learning (ML) algorithms to enhance the accuracy and efficiency of fraud detection [4]. These data-driven systems are capable of learning from vast transactional datasets to identify subtle patterns and anomalies that might indicate fraudulent behavior. However, despite their superior predictive power, ML models often operate as "black boxes." Their decision-making processes are opaque, making it difficult for managers, auditors, and regulators to understand or justify why a given transaction was flagged as fraudulent [15].

This lack of interpretability creates significant barriers to the organizational adoption of artificial intelligence (AI) in fraud detection. From a managerial perspective, trust, accountability, and regulatory compliance are as important as model accuracy [16]. Managers must be able to explain automated decisions to stakeholders, investigate edge cases, ensure the ethical use of data, and comply with increasingly stringent financial regulations, such as GDPR, PSD2, and other transparency mandates. Consequently, the challenge is not only technical, but also managerial: making these models understandable, justifiable, and actionable within real-world decision-making environments [17].

To address this challenge, the field of Explainable Artificial Intelligence (XAI) has emerged. XAI refers to a suite of methods and tools designed to make AI models more interpretable and transparent to human users [18]. In fraud detection, XAI can provide post-hoc explanations for why a transaction was flagged, enabling decision-makers to evaluate the reasoning behind algorithmic predictions. These explanations are crucial for establishing trust in AI systems, informing operational decisions, and fulfilling compliance obligations.

Among various XAI techniques available, this study selects SHAP (Shapley Additive exPlanations) due to its strong theoretical foundation, model-agnostic design, and ability to provide both global and local interpretability [19–22]. Unlike other methods, such as LIME, which generate local approximations that may vary across runs, SHAP offers consistent and additive explanations based on cooperative game theory. These properties make SHAP particularly suitable for fraud detection, where transparency, reproducibility, and trust are essential for managerial oversight and regulatory compliance.

While prior research has demonstrated the effectiveness of ML techniques in detecting fraudulent transactions, many studies focus primarily on technical performance metrics such as accuracy and precision, with limited attention to how these models can be made understandable or actionable in managerial [23–25]. Specifically, there is a lack of empirical work examining how explainable AI tools can support operational decision-making, regulatory compliance, and stakeholder trust in real-world financial environments. This study addresses that gap by applying SHAP to interpret a high-performing fraud detection model and evaluating not only the predictive accuracy but also the practical utility of the explanations for managers, auditors, and compliance officers.

This paper aims to fill that gap by exploring how explainable AI can support managerial oversight in credit card fraud detection. Using a simulated credit card transaction dataset, we evaluate the performance of multiple ML models and apply SHAP to interpret the outputs of the best-performing algorithm. The analysis focuses not only on technical metrics but also on how the explanations produced can support managerial decision-making, risk governance, and regulatory compliance. By integrating predictive accuracy with interpretability, this study offers practical insights for managers, auditors, and policymakers seeking to implement AI in a responsible, transparent, and effective manner.

2. Materials and Methods

To provide a clear overview of the research process, Figure 1 illustrates the workflow adopted in this study. It begins with the acquisition of the dataset, followed by data preprocessing to ensure analytical relevance and compliance with privacy regulations. The process continues with data augmentation, ML model development, performance evaluation, and classification interpretation using explainable AI techniques.

2.1. Dataset

The dataset utilized in this study was retrieved from Kaggle and published by Shenoy [26]. It is a simulated credit card transaction dataset generated using Spark's data generation techniques. The dataset comprises transactions conducted over a two-year period, from January 1, 2019, to December 31, 2020.

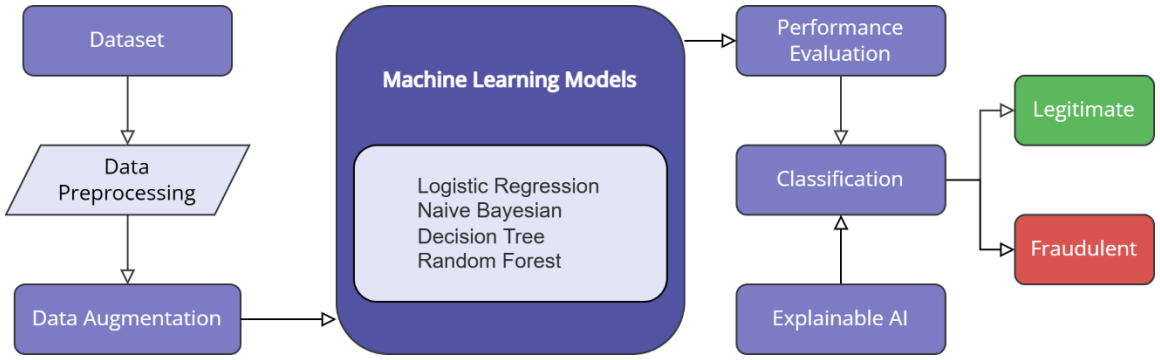

Figure 1. Workflow for building and interpreting machine learning models in credit card fraud detection.

Table 1. Dataset feature description.

Feature Name	Description
trans_date_trans_time	Timestamp of the transaction
cc_num	Anonymized credit card number
merchant	Merchant name involved in the transaction
category	The category of the merchant
amt	Transaction amount
first	First name of the cardholder
last	Last name of the cardholder
gender	Gender of the cardholder
street	Street address of the cardholder
city	City of the cardholder
state	State of the cardholder
zip	Zip code of the cardholder
lat	Latitude coordinate of the cardholder's address
long	Longitude coordinate of the cardholder's address
city_pop	Population of the cardholder's city
job	Occupation of the cardholder
dob	Date of birth of the cardholder
trans_num	Unique identifier for the transaction
unix_time	Unix timestamp of the transaction
merch_lat	Latitude coordinate of the merchant's location
merch_long	Longitude coordinate of the merchant's location
is_fraud	Binary label indicating fraud status (1 = fraud, 0 = not fraud)

The data comprises 1,852,394 transactions involving 1,000 unique customers and a pool of 800 merchants. Of these, 1,842,743 transactions are labeled as non-fraudulent, while 9,651 are marked as fraudulent, reflecting the typical class imbalance found in real-world fraud detection problems. Each transaction entry contains a rich set of features ranging from personal demographics and location data to transaction metadata. The complete list of features is provided in Table 1.

2.2. Data Preprocessing

To ensure the dataset was both privacy-compliant and analytically efficient, several data cleaning and transformation steps were performed. Initially, several columns containing personally identifiable information (PII) or irrelevant metadata were removed. These included credit card numbers, names, addresses, and transaction identifiers such as cc_num, first, last, street, city, zip, Unnamed: 0, trans_num, and unix_time. In addition, columns containing raw date-time and geolocation information, such as trans_date_trans_time, dob, lat, long, merch_lat, and merch_long, were excluded to protect customer privacy and reduce redundancy in later analysis.

For categorical features, encoding techniques were applied to make the data compatible with ML algorithms [27]. Specifically, the gender column was converted into a numerical format, mapping female ('F') to 0 and male ('M') to 1. This transformation helps eliminate bias and ensures the feature can be used effectively in model training.

To enhance the dataset with spatial context, a new feature, distance_to_merchant, was engineered using the Haversine formula. This formula calculates the shortest distance between two points on a sphere, based on their latitude and longitude coordinates. By computing the geodesic distance between the customer's and the

Table 2. Final dataset features after preprocessing.

Feature Name	Description
merchant	Merchant name involved in the transaction
category	The category of the merchant
amt	Transaction amount
gender	Encoded gender of the cardholder (0 = female, 1 = male)
state	State of the cardholder
city_pop	Population of the cardholder's city
job	Occupation of the cardholder
is_fraud	Binary label indicating fraud status (1 = fraud, 0 = not fraud)
distance_to_merchant	Geodesic distance between the cardholder and the merchant location

merchant's locations, this feature can offer critical insight into potentially suspicious transactions, such as a purchase occurring far from the customer's usual location. After preprocessing, the final set of features used in the model is summarized in Table 2.

2.3. Data Augmentation

One of the critical challenges in fraud detection is the imbalance in class distribution; fraudulent transactions typically represent only a small fraction of the total data. This imbalance can lead ML models to become biased toward predicting legitimate transactions, thereby reducing their ability to identify fraud correctly [28].

To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied as a data augmentation strategy. SMOTE was selected due to its effectiveness in addressing class imbalance without discarding valuable data from the majority class. Unlike random downsampling, which reduces the number of legitimate transactions and risks information loss, SMOTE generates synthetic samples by interpolating between existing instances of the minority class [29].

This approach has the added benefit of maintaining the overall dataset size, which is important for retaining statistical power and ensuring model robustness. Additionally, the generated synthetic samples contribute to improved pattern recognition by enhancing the feature diversity of the minority class. Although other techniques, such as cost-sensitive learning and ensemble-based resampling, were considered for future investigation, they were not implemented in this initial study. The primary reason for this decision was to maintain model interpretability and reduce computational complexity, particularly in early-stage model evaluation and benchmarking.

2.4. Machine Learning Models

In this study, four ML algorithms were selected based on their complementary strengths and their suitability for common challenges in fraud detection. These challenges include class imbalance, complex feature relationships, and the need for model interpretability. Although SMOTE was used to balance the training data, class imbalance remains a central issue in real-world fraud detection environments. Therefore, it was crucial to select models that are known to perform well in dealing with imbalanced data. The selected models are Logistic Regression, Naive Bayes, Decision Tree, and Random Forest. Together, they provide a range of approaches that support both performance evaluation and practical considerations.

Prior to model training, the dataset was stratified and split into 80% for training and 20% for testing to ensure that the proportion of fraudulent and non-fraudulent transactions remained consistent across both sets [30]. Stratified sampling is particularly important in the imbalanced classification problems, such as fraud detection, where the minority class (fraudulent transactions) comprises only a small percentage of the total dataset. Without stratification, there is a risk that the test set may not adequately represent the minority class, leading to unreliable performance evaluation and potential overestimation of model accuracy.

Logistic Regression was chosen due to its simplicity, computational efficiency, and high level of interpretability. It estimates the probability of fraud based on a linear combination of input features, making it easy to understand and implement. This level of transparency is especially valuable in financial applications, where stakeholders often require clear explanations for automated decisions [31].

Naive Bayes was included for its speed and effectiveness in handling high-dimensional data. The model applies Bayes' Theorem under the assumption of feature independence, which allows it to scale well and perform reliably in large datasets. Although this independence assumption is often violated in practice, Naive Bayes can still produce strong classification results, particularly in domains like fraud detection, where certain features may individually signal suspicious behavior. Its probabilistic outputs also provide a straightforward measure of

decision confidence, which can be useful for prioritizing alerts [32].

Decision Tree was selected for its intuitive, rule-based structure that closely mirrors human decision-making. Each decision path can be visualized and explained, making it easier for managers, auditors, and compliance officers to trace how a transaction was classified and understood. This interpretability is particularly important in regulated sectors such as finance, where accountability and transparency are essential [33].

Random Forest was included for its ability to improve accuracy and generalization by aggregating predictions from multiple decision trees. It is well-suited to capturing complex, non-linear relationships in transactional data and tends to perform well even in noisy or imbalanced datasets. While less interpretable than a single decision tree, Random Forest still offers insights through feature importance scores, helping bridge the gap between predictive performance and explainability [34].

All machine learning models in this study were trained using the default hyperparameters provided by the Scikit-learn library. This decision aligns with a common practice in initial modeling phases, where baseline performance is established before engaging in more computationally intensive hyperparameter tuning. Using default settings also promotes reproducibility and enables fair comparison across different algorithms, as it ensures that each model is evaluated under standardized conditions. While further performance improvements could potentially be achieved through techniques such as grid search or randomized search, the focus in this phase was on comparing fundamental model capabilities and assessing their suitability for fraud detection tasks under practical constraints.

2.5. Performance Evaluation

To assess the effectiveness of each ML model in detecting fraudulent transactions, several standard performance metrics were utilized. These included Accuracy, Precision, Sensitivity (Also Known as Recall), Specificity, and the F1-Score. Each of these metrics provides a distinct perspective on model performance, which is particularly important in the context of fraud detection, where the costs of false positives and false negatives can vary significantly.

- Accuracy reflects the overall proportion of correctly predicted transactions (both fraudulent and legitimate) but can be misleading in imbalanced datasets [35].
- Precision indicates the proportion of correctly identified fraud cases out of all transactions the model flagged as fraudulent, making it crucial for minimizing false alarms [36].
- Sensitivity measures the model's ability to accurately detect actual fraud cases, ensuring that high-risk transactions are not overlooked [37].
- Specificity represents the model's capacity to correctly identify legitimate transactions, which helps prevent unnecessary disruptions for genuine customers [38].
- F1-Score provides a balanced measure that considers both precision and recall, making it particularly useful when managing the trade-off between catching fraud and avoiding false accusations [39].

In addition to these metrics, Receiver Operating Characteristic (ROC) curves were plotted for each model. The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) across various threshold settings. The Area Under the Curve (AUC) was also considered as a summary indicator of model discrimination ability; the higher the AUC, the better the model is at distinguishing between fraudulent and legitimate transactions [40].

2.6. Explainable AI

To enhance the transparency and interpretability of ML predictions in fraud detection, SHAP was used in this study [41]. SHAP is a model-agnostic explanation method based on concepts from cooperative game theory. It calculates the contribution of each input feature to a specific prediction by assigning a value that reflects how much that feature increased or decreased the model's output [42]. This is done by evaluating all possible combinations of feature inputs, similar to how Shapley values are used to fairly divide gains among players in a coalition. SHAP is particularly useful in high-stakes settings, such as fraud detection, because it provides both global insights into how the model behaves overall and local explanations for individual predictions. This level of interpretability supports managerial accountability, regulatory compliance, and trust in automated decision systems.

SHAP was used to identify the features that had the greatest influence on the model's classification decisions for transactions as either fraudulent or legitimate. High transaction amounts, significant geographic distance between the customer and merchant, and abnormal

Table 3. Performance comparison of machine learning models.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Logistic Regression	84.86	80.03	92.91	76.81	85.99
Naïve Bayesian	83.55	76.07	97.91	69.20	85.62
Decision Tree	98.79	99.16	98.40	99.17	98.78
Random Forest	99.52	99.59	99.44	99.59	99.52

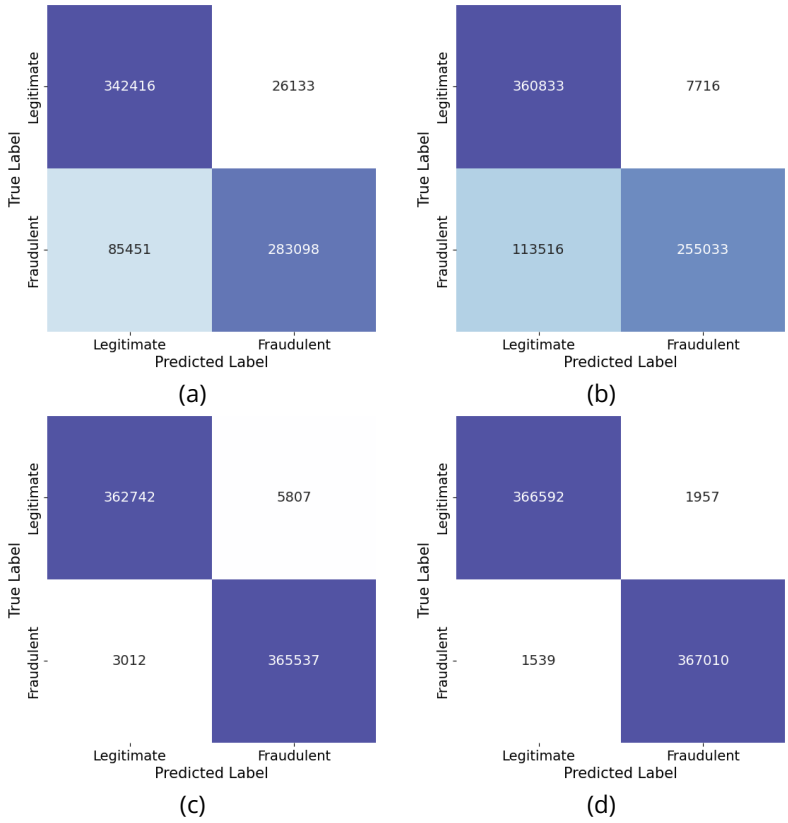

Figure 2. Confusion matrices for (a) Logistic Regression, (b) Naïve Bayesian, (c) Decision Tree, and (d) Random Forest models.

transaction timing were among the top contributors to fraud predictions across models. Additionally, SHAP highlighted instances where transactions were borderline or incorrectly flagged, providing valuable insights into the model's limitations and areas for potential refinement.

SHAP visualizations, such as summary plots, force plots, and dependence plots, were instrumental in making these insights accessible to both technical and non-technical stakeholders. These visuals effectively illustrated how individual features interacted and contributed to outcomes, thereby supporting auditability and compliance in high-stakes financial environments [43].

3. Results and Discussion

To evaluate the effectiveness of various machine learning algorithms in detecting credit card fraud, multiple performance metrics were calculated. These results are presented in Table 3.

The performance metrics demonstrate that both Decision Tree and Random Forest models significantly outperform Logistic Regression and Naive Bayes across all evaluation criteria. Random Forest exhibited the highest overall performance, achieving an accuracy of 99.52% and an F1-score of 99.52%, indicating a strong balance between precision and recall. Its high specificity (99.59%) also suggests that it is highly effective at minimizing false positives, a critical factor in fraud detection, as unnecessary flagging of legitimate transactions can lead to customer dissatisfaction.

Decision Tree also performed exceptionally well, with an accuracy of 98.79% and an F1-score of 98.78%. While slightly lower than Random Forest, it maintains high interpretability due to its rule-based structure, which is useful in regulated environments that require decision traceability. Logistic Regression and Naive Bayes showed moderate performance, with Logistic Regression performing slightly better in terms of precision and F1-Score. Logistic Regression's simplicity and ease of

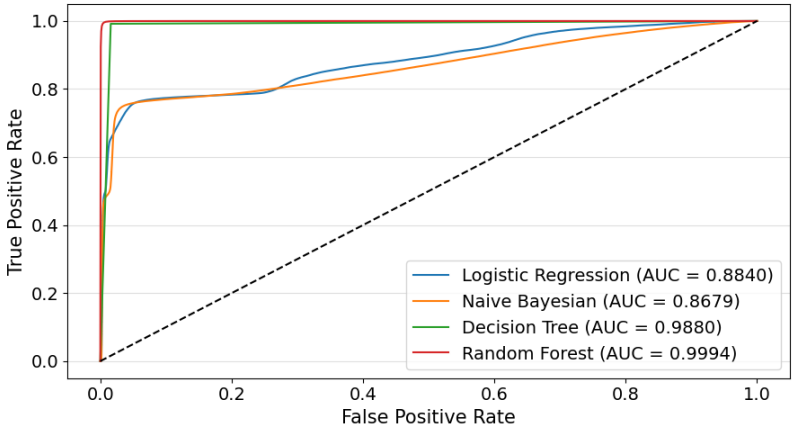

Figure 3. ROC curves and AUC scores for all four models, highlighting Random Forest's superior classification performance.

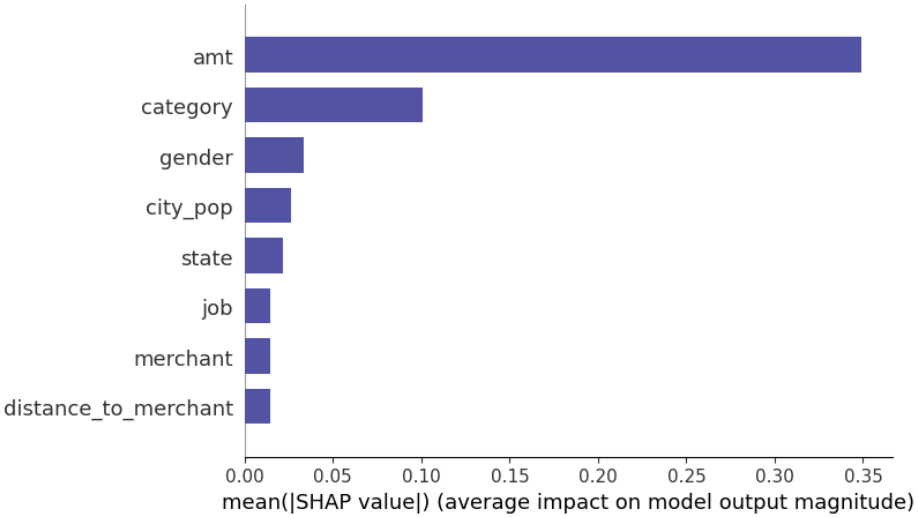

Figure 4. SHAP summary plot showing the average impact of each feature on Random Forest model predictions.

interpretation make it a suitable baseline model, particularly in contexts requiring model transparency. Naïve Bayesian achieved a notably high sensitivity of 97.91%, making it effective in identifying actual fraud cases; however, its lower specificity (69.20%) resulted in a relatively higher number of false positives.

Further insight into the performance of each ML model can be gained from the confusion matrices shown in Figure 2. Logistic Regression (Figure 2a) correctly classified 283,098 fraudulent transactions (true positives) but also missed 85,451 (false negatives), indicating a substantial number of fraud cases were overlooked. It also generated 26,133 false positives, incorrectly flagging legitimate transactions as fraudulent. Naïve Bayes (Figure 2b) showed an even higher false negative count (113,516), despite a relatively low false positive rate of 7,716. This suggests that while Naïve Bayes was cautious about labeling transactions as fraudulent, it often failed to catch actual fraud. The Decision Tree model (Figure 2c) dramatically improved both aspects, with only 3,012 false negatives and 5,807 false positives, demonstrating its capacity to make highly accurate and interpretable classifications. Random Forest (Figure 2d) achieved the best results, with only 1,539 false negatives and 1,957 false positives among hundreds of thousands of transactions.

To complement the evaluation of classification metrics, Receiver Operating Characteristic (ROC) curves were plotted for each model, as illustrated in Figure 3. The ROC curve visualizes the trade-off between the true positive rate and the false positive rate across various threshold settings. The Area Under the Curve (AUC) provides a single scalar value that summarizes the model's ability to discriminate between fraudulent and legitimate transactions. Figure 3 reveals clear differences in model performance. Random Forest achieved an outstanding AUC of 0.9994, indicating near-perfect discrimination capability. The Decision Tree model also performed exceptionally well with an AUC of 0.9880, reflecting its effectiveness in capturing fraud patterns while

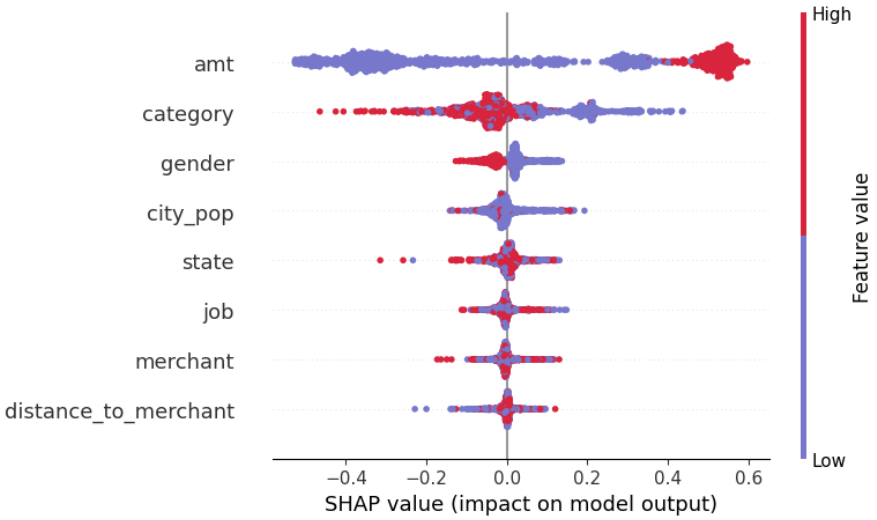

Figure 5. SHAP beeswarm plot showing the distribution and direction of feature impacts on individual Random Forest model predictions.

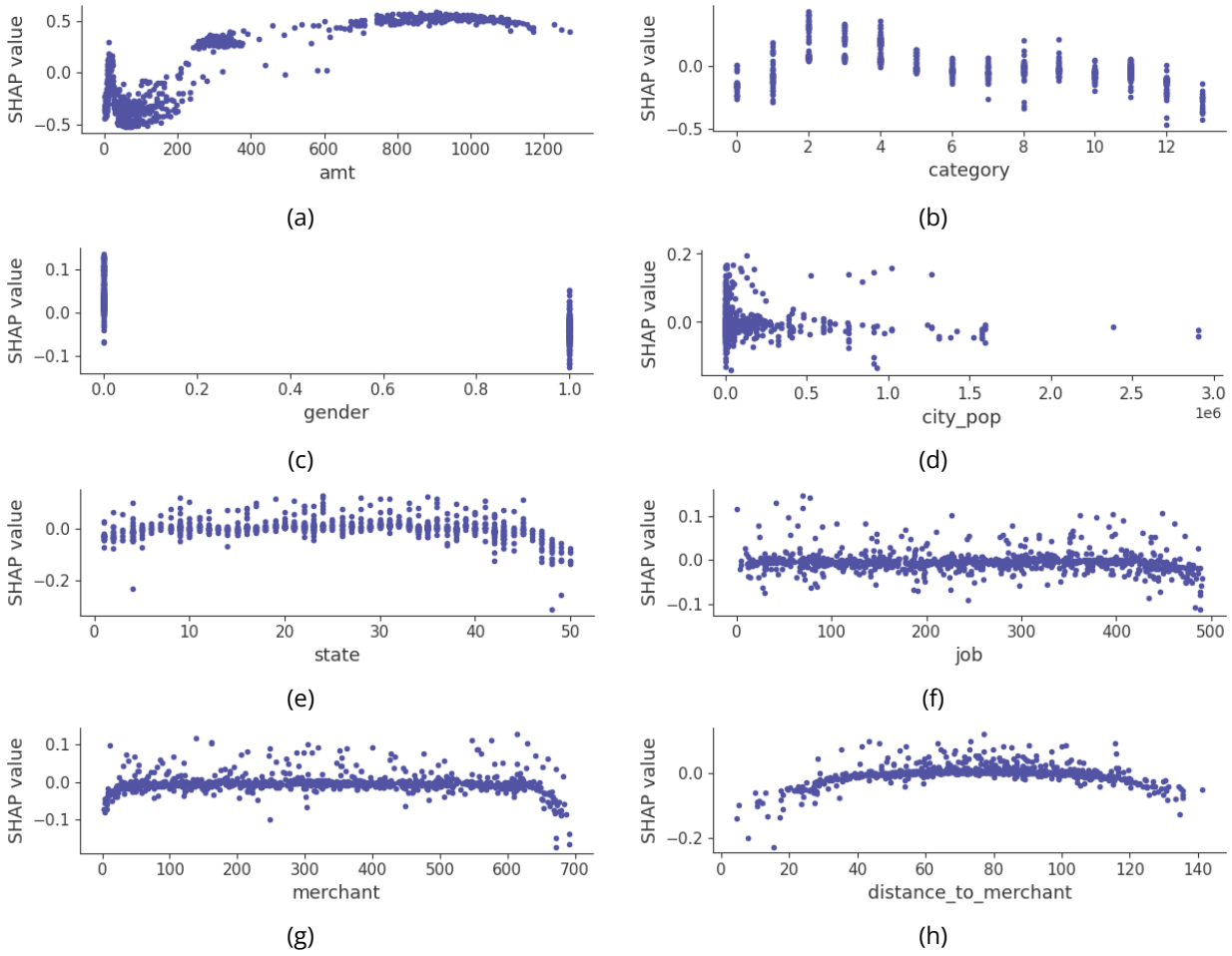

Figure 6. SHAP dependence plots for each input feature: (a) amt, (b) category, (c) gender, (d) city_pop, (e) state, (f) job, (g) merchant, and (h) distance_to_merchant.

maintaining a low false alarm rate. Logistic Regression and Naïve Bayes recorded AUC values of 0.8840 and 0.8679, respectively, which are acceptable but notably lower than those of tree-based models. These results confirm the superior classification capabilities of ensemble methods, such as Random Forest, particularly when applied to imbalanced datasets characteristic of fraud detection tasks.

These results confirm that more complex models, such as Random Forest, can deliver superior predictive accuracy, as evidenced by their near-perfect scores across

performance metrics, confusion matrix outcomes, and ROC analysis. However, this high performance comes at the cost of interpretability, a crucial concern in financial applications where transparency, accountability, and regulatory compliance are paramount. This is where XAI methods, such as SHAP, become essential, as they provide the tools needed to demystify black-box models and help stakeholders understand the reasoning behind each prediction. To solve the gap between accuracy and explainability, the next section examines how SHAP was applied to interpret the Random Forest model, the algorithm that performed best in this study. It reveals the key factors that drive its fraud detection decisions.

To enhance the interpretability of the high-performing Random Forest model, SHAP was applied to understand the contribution of each feature to the model's predictions. SHAP values are grounded in cooperative game theory, allowing us to decompose a prediction into the sum of contributions from each feature. This helps both technical and non-technical stakeholders understand not just what the model predicted, but why.

Figure 4 presents a SHAP summary plot, showing the average absolute SHAP value for each feature, which reflects its overall impact on the model's output. It is evident that amt has the highest influence on the model's predictions, followed by the category of the merchant. These two features significantly dominate the decision-making process. Other features, such as gender, city_pop, state, and job, also contribute, but to a much lesser extent. Surprisingly, distance_to_merchant, which was hypothesized to signal suspicious behavior, showed minimal impact, suggesting either limited variation in this feature or its lesser relevance in the specific dataset used.

To gain deeper insight into how each feature influences the Random Forest model's predictions, the SHAP beeswarm plot in Figure 5 was analyzed. Unlike the summary bar chart in Figure 4, this plot visualizes the distribution of SHAP values across all samples, colored by the feature value (low to high). Each point represents a single prediction, enabling us to understand not only the importance of a feature but also how its value influences the prediction toward fraud or non-fraud.

The plot reinforces the dominance of the amt feature; high transaction amounts (shown in red) tend to strongly increase the probability of a transaction being classified as fraudulent, as indicated by positive SHAP values. Similarly, certain merchant category values also drive predictions toward fraud. For features such as gender and state, the effect is more variable, with both high and low values contributing positively or negatively, depending on the context. Interestingly, features such as distance_to_merchant and merchant appear to have less consistent or weaker effects, often clustering around zero SHAP values, indicating limited influence on model output.

This visualization offers essential interpretability, particularly in environments that require accountability for model decisions. It allows practitioners to verify that model behavior aligns with domain expectations and to identify potential biases or unexpected dependencies that may require mitigation. For example, if SHAP plots consistently show that high transaction amounts and specific merchant categories are strong indicators of fraud, a risk manager could use this insight to adjust manual review thresholds or prioritize certain types of transactions for real-time flagging. Compliance officers might also use this information to document the rationale behind automated alerts, ensuring that decisions are traceable and aligned with regulatory expectations. These visualizations can thus directly inform fraud response protocols, risk scoring models, and internal audit processes.

To better understand how individual features affect the prediction output of the Random Forest model, SHAP dependence plots were generated for all key variables, as shown in Figure 6. Each subplot illustrates the relationship between the actual value of a feature and its corresponding SHAP value, highlighting how changes in that feature impact the likelihood of a transaction being flagged as fraudulent.

In Figure 6a, the amt plot shows a clear positive correlation between transaction amount and SHAP value, indicating that higher transaction amounts significantly increase the predicted fraud probability. Similarly, in Figure 6b, certain category codes are associated with higher SHAP values, suggesting that specific merchant types are more commonly linked to fraudulent behavior.

The gender plot (Figure 6c) shows a noticeable separation between male and female cardholders, though its impact is minor compared to other features. city_pop (Figure 6d) shows that transactions originating from low-population areas have slightly higher fraud risk, as reflected by higher SHAP values. The state and job plots (Figures 6e and 6f) exhibit a more uniform SHAP distribution, indicating limited variability in model influence across these categories. Interestingly, the merchant (Figure 6g) and distance_to_merchant (Figure 6h) plots reveal low SHAP values across their respective ranges, reinforcing earlier findings that these features have minimal influence on the model's output. However, some mild non-linear patterns can be observed, suggesting possible localized effects or interactions with other features.

The findings of this study highlight the potential of ML, particularly ensemble methods such as Random Forest, for enhancing fraud detection systems in the financial services sector. However, as fraud detection becomes increasingly automated and reliant on complex algorithms, the need for interpretability becomes paramount from a managerial oversight perspective. While traditional rule-based systems are easy to audit and explain, high-performing ML models often operate as black boxes, posing challenges for compliance, stakeholder trust, and decision accountability.

Using SHAP, we demonstrated how managers can gain insight into the inner workings of a high-performing Random Forest model. SHAP enabled us to identify the most influential predictors of fraud, primarily transaction amount and merchant category, and to visualize how specific values of these features affect the likelihood of a transaction being flagged for fraud. These explanations support transparency in model decisions, which is crucial for risk managers, auditors, and compliance officers who must justify automated actions to internal stakeholders and external regulators.

Moreover, SHAP outputs can aid in strategy formulation and operational policy. For instance, knowing that higher transaction amounts disproportionately influence fraud predictions may prompt risk teams to reassess thresholds for manual review or escalation. Similarly, understanding the modest influence of features such as geographic distance or cardholder occupation may encourage a reevaluation of data collection practices, thereby reducing unnecessary data handling and associated privacy concerns.

From an organizational change perspective, integrating explainable AI allows firms to fill the gap between data science teams and management. Managers without deep technical expertise can still engage with model behavior using SHAP visualizations, fostering cross-functional collaboration in fraud strategy design. This also aligns with broader ESG (Environmental, Social, and Governance) and AI ethics frameworks that advocate for the responsible and interpretable deployment of AI in decision-making processes that impact individuals. Importantly, the use of XAI supports regulatory compliance. Financial institutions are subject to strict regulations that require transparency in the approval or decline of transactions, particularly when fraud is suspected. Explainable models mitigate legal and reputational risks by providing transparent, justifiable rationales for algorithmic decisions. As regulatory bodies worldwide begin to scrutinize AI-driven processes more closely, the implementation of XAI can serve as a critical risk mitigation tool.

While the models demonstrated strong performance on the test set, it is important to acknowledge that generalization to real-world fraud detection environments is not guaranteed and was not directly evaluated in this study. The dataset used, although simulated, was designed to reflect realistic patterns in transaction behavior, including class imbalance, merchant diversity, and temporal distribution. However, real-world deployment involves additional complexities such as evolving fraud tactics, data drift, and integration with live transaction systems. As such, further validation using live or institution-specific data, as well as longitudinal testing, would be necessary to confirm the robustness and adaptability of these models in operational settings.

4. Conclusions, Implications and Limitations

This study examined the integration of XAI into credit card fraud detection systems, with a focus on its role in enhancing managerial oversight and decision-making. By applying SHAP to the best-performing Random Forest model, we demonstrated that complex ML algorithms, while highly accurate, require interpretability tools to be effectively and responsibly deployed in real-world financial environments. The SHAP analysis provided both global and local explanations, enabling managers to understand the logic behind fraud predictions and to align ML systems with internal policies and external regulatory expectations.

The implications of this research extend beyond technical model optimization into the strategic and operational realms of fraud risk management. Explainable AI enables transparent decision-making, allowing non-technical stakeholders such as compliance officers, auditors, and senior managers to interpret and validate fraud alerts. This fosters better collaboration between data science teams and business units. Additionally, explainability supports regulatory compliance. As financial institutions face increasing pressure to justify automated decisions, tools like SHAP provide the transparency and auditability needed to reduce legal risk and demonstrate accountability. Interpretability also contributes to the ethical deployment of AI, particularly in sensitive areas such as fraud detection, where false positives can harm customer relationships or damage brand reputation. XAI allows managers to examine edge cases and assess whether specific features or thresholds are introducing operational inefficiencies or unintended bias.

From a practical standpoint, the findings of this study suggest that SHAP-enhanced Random Forest models can be directly applied to improve real-world fraud detection systems. Feature importance insights, for instance, can

guide the adjustment of manual review thresholds, inform the design of fraud escalation procedures, and support the documentation of automated decisions for compliance reporting. By integrating SHAP explanations into fraud management workflows, institutions cannot only make their AI systems more accurate but also more transparent, trustworthy, and aligned with organizational goals and regulatory standards.

Despite these contributions, the study has several limitations. First, the dataset used was simulated and may not fully reflect the complexity of real-world fraud scenarios. While useful for prototyping and controlled analysis, future work should validate findings using live transactional data from financial institutions to ensure accuracy and reliability. Second, the models were trained with default parameters. While this reflects standard practice in many initial deployments, further performance gains might be achieved through hyperparameter tuning or more advanced model architectures. Third, SHAP is only one of several explainability techniques. Although it offers robust theoretical grounding and intuitive visuals, future studies could compare its effectiveness with other XAI methods. Finally, this study focused on the managerial interpretation of model outputs. Still, future research could explore organizational adoption processes, such as training, change management, and integration into existing fraud operations workflows. Understanding how explainability influences trust, uptake, and performance at the team level remains a critical next step.

Author Contributions: Conceptualization, M.M. and T.R.N.; methodology, M.M.; software, T.R.N.; validation, M.M., T.R.N., and G.M.I.; formal analysis, M.M.; investigation, M.M.; resources, M.M.; data curation, T.R.N.; writing—original draft preparation, M.M.; writing—review and editing, M.M.; visualization, T.R.N.; supervision, M.M.; project administration, M.M.; funding acquisition, A.S. and F.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Data Availability Statement: The dataset used in this study is publicly available on Kaggle and can be accessed at https://www.kaggle.com/datasets/kartik2112/fraud-detection. It contains transaction data used for fraud detection analysis.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Fedotova, G. V., Chugumbaev, R. R., Chugumbaeva, N. N., Sukhinin, A. V., and Kuzmina, E. V. (2019). Increase of Economic Security of Internet Systems of Credit Organizations, 922–931. doi:10.1007/978-3-030-00102-5_98.
2. Karpoff, J. M. (2021). The Future of Financial Fraud, *Journal of Corporate Finance*, Vol. 66, 101694. doi:10.1016/j.jcorpfin.2020.101694.
3. Laxman, V., Ramesh, N., Jaya Prakash, S. K., and Aluvala, R. (2024). Emerging Threats in Digital Payment and Financial Crime: A Bibliometric Review, *Journal of Digital Economy*, Vol. 3, 205–222. doi:10.1016/j.jdec.2025.04.002.
4. Hafez, I. Y., Hafez, A. Y., Saleh, A., Abd El-Mageed, A. A., and Abohany, A. A. (2025). A Systematic Review of AI-Enhanced Techniques in Credit Card Fraud Detection, *Journal of Big Data*, Vol. 12, No. 1, 6. doi:10.1186/s40537-024-01048-8.
5. Beju, D.-G., and Făt, C.-M. (2023). Frauds in Banking System: Frauds with Cards and Their Associated Services, 31–52. doi:10.1007/978-3-031-34082-6_2.
6. Hardi, I., Idroes, G. M., Márquez-Ramos, L., Noviandy, T. R., and Idroes, R. (2025). Inclusive Innovation and Green Growth in Advanced Economies, *Sustainable Futures*, Vol. 9, 100540. doi:10.1016/j.sftr.2025.100540.
7. Hardi, I., Afjal, M., Khan, M., Idroes, G. M., Noviandy, T. R., and Utami, R. T. (2024). Economic Freedom and Growth Dynamics in Indonesia: An Empirical Analysis of Indicators Driving Sustainable Development, *Cogent Economics & Finance*, Vol. 12, No. 1. doi:10.1080/23322039.2024.2433023.
8. Hasham, S., Joshi, S., and Mikkelsen, D. (2019). Financial Crime and Fraud in the Age of Cybersecurity, *McKinsey & Company*, Vol. 2019.
9. Hardi, I., Ray, S., Duwal, N., Idroes, G. M., and Mardayanti, U. (2024). Consumer Confidence and Economic Indicators: A Macro Perspective, *Indatu Journal of Management and Accounting*, Vol. 2, No. 2, 81–95. doi:10.60084/ijma.v2i2.241.
10. Idroes, G. M., Maulidar, P., Marsellindo, R., Afjal, M., and Hardi, I. (2024). The Impact of Credit Access on Economic Growth in SEA Countries, *Indatu Journal of Management and Accounting*, Vol. 2, No. 2, 96–104. doi:10.60084/ijma.v2i2.256.
11. Hardi, I., Nghiem, X.-H., Suwal, S., Ringga, E. S., Marsellindo, R., and Idroes, G. M. (2024). Starting a Business: A Focus on Construction Permits, Electricity Access, and Property Registration, *Indatu Journal of Management and Accounting*, Vol. 2, No. 2, 105–117. doi:10.60084/ijma.v2i2.245.
12. Hardi, I., Idroes, G. M., Hamaguchi, Y., Can, M., Noviandy, T. R., and Idroes, R. (2025). Business Confidence in the Shift to Renewable Energy: A Country-Specific Assessment in Major Asian Economies, *Journal of Economy and Technology*, Vol. 3, 44–68. doi:10.1016/j.ject.2024.08.002.
13. Vashishth, T. K., Chaudhary, A., Sharma, V., Chaudhary, S., Sharma, N., Sharma, R., Kaushik, V., and Sharma, S. (2025). Adaptive AI Systems for Financial Fraud Detection and Risk Management, 431–454. doi:10.4018/979-8-3373-1200-2.ch021.
14. Khanum, A., K S, C., Singh, B., and Gomathi, C. (2024). Fraud Detection in Financial Transactions: A Machine Learning Approach vs. Rule-Based Systems, *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, IEEE, 1–5. doi:10.1109/IITCEE59897.2024.10467759.
15. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. (2024). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence, *Cognitive Computation*, Vol. 16, No. 1, 45–74. doi:10.1007/s12559-023-10179-8.
16. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., and Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions, *ACM Computing Surveys*, Vol. 55, No. 9, 1–33. doi:10.1145/3561048.
17. Awosika, T., Shukla, R. M., and Pranggono, B. (2024). Transparency and Privacy: The Role of Explainable AI and Federated Learning in Financial Fraud Detection, *IEEE Access*, Vol. 12, 64551–64560. doi:10.1109/ACCESS.2024.3394528.
18. Ponzoni, I., Páez Prosper, J. A., and Campillo, N. E. (2023). Explainable Artificial Intelligence: A Taxonomy and Guidelines for Its Application to Drug Discovery, *WIREs Computational Molecular Science*, Vol. 13, No. 6. doi:10.1002/wcms.1681.

19. Lundberg, S. M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems*, Vol. 30.

20. Noviandy, T. R., Idroes, G. M., Hardi, I., Afjal, M., and Ray, S. (2024). A Model-Agnostic Interpretability Approach to Predicting Customer Churn in the Telecommunications Industry, *Infolitika Journal of Data Science*, Vol. 2, No. 1, 34–44. doi:10.60084/ijds.v2i1.199.

21. Idroes, G. M., Noviandy, T. R., Idroes, G. M., Hardi, I., Duta, T. F., Hamoud, L. M., and Al-Gunaid, H. T. (2024). Prognostication of Differentiated Thyroid Cancer Recurrence: An Explainable Machine Learning Approach, *Narra X*, Vol. 2, No. 3, e183. doi:10.52225/narrax.v2i3.183.

22. Noviandy, T. R., Nisa, K., Idroes, G. M., Hardi, I., and Sasmita, N. R. (2024). Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer's Drug Discovery with LightGBM, *Journal of Computing Theories and Applications*, Vol. 1, No. 4, 358–367. doi:10.62411/jcta.10129.

23. Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., and Adejoh, J. (2024). Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach, *Big Data and Cognitive Computing*, Vol. 8, No. 1, 6. doi:10.3390/bdcc8010006.

24. Feng, X., and Kim, S.-K. (2024). Novel Machine Learning Based Credit Card Fraud Detection Systems, *Mathematics*, Vol. 12, No. 12, 1869. doi:10.3390/math12121869.

25. Noviandy, T. R., Hardi, I., and Idroes, G. M. (2024). Forecasting Bank Stock Trends Using Artificial Intelligence: A Deep Dive into the Neural Prophet Approach, *The International Journal of Financial Systems*, Vol. 2, No. 1, 29–56. doi:10.61459/ijfs.v2i1.41.

26. Shenoy, K. (2020). Credit Card Transactions Fraud Detection Dataset, Kaggle.

27. Srikanth Yadav., M., and Kalpana., R. (2019). Data Preprocessing for Intrusion Detection System Using Encoding and Normalization Approaches, *2019 11th International Conference on Advanced Computing (ICoAC)*, IEEE, 265–269. doi:10.1109/ICoAC48765.2019.246851.

28. Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., and Idroes, R. (2023). Credit Card Fraud Detection for Contemporary Financial Management Using XGBoost-Driven Machine Learning and Data Augmentation Techniques, *Indatu Journal of Management and Accounting*, Vol. 1, No. 1, 29–35. doi:10.60084/ijma.v1i1.78.

29. Chawla, N. V, Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique, *Journal of Artificial Intelligence Research*, Vol. 16, 321–357.

30. Joseph, V. R. (2022). Optimal Ratio for Data Splitting, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Vol. 15, No. 4, 531–538. doi:10.1002/sam.11583.

31. Matuszelański, K., and Kopczewska, K. (2022). Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach, *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 17, No. 1, 165–198. doi:10.3390/jtaer17010009.

32. Chen, H., Hu, S., Hua, R., and Zhao, X. (2021). Improved Naive Bayes Classification Algorithm for Traffic Risk Management, *EURASIP Journal on Advances in Signal Processing*, Vol. 2021, No. 1, 30. doi:10.1186/s13634-021-00742-6.

33. Noviandy, T. R., Idroes, G. M., and Hardi, I. (2024). Enhancing Loan Approval Decision-Making: An Interpretable Machine Learning Approach Using LightGBM for Digital Economy Development, *Malaysian Journal of Computing (MJOC)*, Vol. 9, No. 1, 1734–1745. doi:10.24191/mjoc.v9i1.25691.

34. Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., and Idroes, R. (2023). QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms, *Heca Journal of Applied Sciences*, Vol. 1, No. 1, 1–7. doi:10.60084/hjas.v1i1.12.

35. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Afidh, R. P. F., and Idroes, R. (2024). An Explainable Multi-Model Stacked Classifier Approach for Predicting Hepatitis C Drug Candidates, *Sci*, Vol. 6, No. 4, 81. doi:10.3390/sci6040081.

36. Tharwat, A. (2021). Classification Assessment Methods, *Applied Computing and Informatics*, Vol. 17, No. 1, 168–192. doi:10.1016/j.aci.2018.08.003.

37. Noviandy, T. R., Idroes, G. M., and Hardi, I. (2025). Integrating Explainable Artificial Intelligence and Light Gradient Boosting Machine for Glioma Grading, *Informatics and Health*, Vol. 2, No. 1, 1–8. doi:10.1016/j.infoh.2024.12.001.

38. Noviandy, T. R., Idroes, G. M., and Hardi, I. (2024). An Interpretable Machine Learning Strategy for Antimalarial Drug Discovery with LightGBM and SHAP, *Journal of Future Artificial Intelligence and Technologies*, Vol. 1, No. 2, 84–95. doi:10.62411/faith.2024-16.

39. Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., and Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery, *Malacca Pharmaceutics*, Vol. 1, No. 2, 48–54. doi:10.60084/mp.v1i2.60.

40. Noviandy, T. R., Idroes, G. M., and Hardi, I. (2024). Machine Learning Approach to Predict AXL Kinase Inhibitor Activity for Cancer Drug Discovery Using XGBoost and Bayesian Optimization, *Journal of Soft Computing and Data Mining*, Vol. 5, No. 1, 46–56. doi:10.30880/jscdm.2024.05.01.004.

41. Gramegna, A., and Giudici, P. (2021). SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk, *Frontiers in Artificial Intelligence*, Vol. 4. doi:10.3389/frai.2021.752558.

42. Noviandy, T. R., Maulana, A., Irvanizam, I., Idroes, G. M., Maulydia, N. B., Tallei, T. E., Subianto, M., and Idroes, R. (2025). Interpretable Machine Learning Approach to Predict Hepatitis C Virus NS5B Inhibitor Activity Using Voting-Based LightGBM and SHAP, *Intelligent Systems with Applications*, Vol. 25, 200481. doi:10.1016/j.iswa.2025.200481.

43. Le, T.-T.-H., Kim, H., Kang, H., and Kim, H. (2022). Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method, *Sensors*, Vol. 22, No. 3, 1154. doi:10.3390/s22031154.