Artificial Intelligence Systems and Its Applications (AISA)
Volume 1, Issue 2, December 2025
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

Integrating SHAP Guided Feature Optimization into Gradient Boosting for Explainable Machine Learning

Muhammad Hikmal Yazid[1]

[1] Al Ihsan Education Foundation, Sidoarjo, Indonesia

Abstract: Artificial Intelligence (AI) has achieved remarkable success in predictive modeling, yet the lack of explainability in complex models remains a major challenge for adoption in high-stakes domains. This study addresses this problem by developing a machine learning pipeline that integrates explainability techniques with high-performance predictive models. The objectives are to enhance model transparency, evaluate performance on real-world datasets, and compare the proposed approach with conventional baseline models. Experimental evaluation was conducted on healthcare and finance datasets, using gradient boosting models combined with SHAP explanations to provide feature-level interpretability. The results demonstrate that the proposed approach achieves 92.5% accuracy, 91.2% precision, and 90.8% recall, outperforming baseline models while maintaining transparent decision-making. Visualization of feature contributions confirmed that the model's predictions align with domain knowledge, enhancing trust and accountability. The study highlights the feasibility of balancing predictive performance with explainability, providing a practical framework for deploying AI in critical applications. Limitations include increased computational requirements for large-scale datasets. The findings offer implications for both researchers and practitioners by demonstrating that highly accurate models can remain interpretable, promoting ethical and responsible AI deployment. Future work should explore scalability, real-time interpretability, and application to additional domains, further bridging the gap between predictive power and model transparency.

Keywords: Artificial Intelligence, Machine Learning, Explainable AI, SHAP, Gradient Boosting, Model Interpretability, Predictive Analytics

Article History:

Received: 10 October 2025 Accepted: 28 December 2025 Published: 31 December 2025

Corresponding Author: Muhammad Himal Yazid, Email: yazid.hikmal.muhammad@gmail.com

DOI: 10.65917/aisa.v1i2.34

1 Introduction

Artificial Intelligence (AI) has emerged as a transformative force in the 21st century, reshaping industries, research, and daily life through its capacity to perform tasks that traditionally required human intelligence. From autonomous vehicles to personalized healthcare recommendations, AI systems have demonstrated the ability to learn complex patterns, make predictions, and optimize processes with unprecedented efficiency. Among the most prominent subfields of AI, machine learning (ML) and deep learning (DL) have garnered significant attention due to their capacity to model highly non-linear relationships, handle vast amounts of data, and adaptively improve their performance over time. These capabilities have enabled breakthroughs in areas as diverse as natural language processing, computer vision, robotics, and finance. The growing reliance on AI systems has catalyzed societal and industrial transformations, creating new opportunities for productivity, innovation, and scientific discovery.

Despite these remarkable achievements, a critical challenge persists: the lack of interpretability and transparency in complex AI models. Modern machine learning algorithms, particularly deep neural networks and ensemble methods such as gradient boosting machines, are often treated as "black boxes" due to their intricate architectures and millions of parameters. While these models achieve high predictive accuracy, their internal decision-making processes are opaque, leaving end-users, practitioners, and regulators uncertain about how and why specific predictions are made. This lack of explainability poses significant barriers to adoption in high-stakes domains, including healthcare, finance, criminal justice, and autonomous systems, where decisions directly impact human lives, societal well-being, and ethical considerations. Modern machine learning models, particularly deep neural networks and ensemble methods, are often criticized as *black-box systems* due to their opaque decision-making processes, which hinder trust, accountability, and adoption in high-stakes domains such as healthcare and finance(Adadi & Berrada, 2020; Arrieta, 2020; Guidotti, 2018).

The lack of interpretability becomes critical in high-stakes decision-making domains such as healthcare and finance, where opaque predictions may lead to ethical, legal, and social risks. [4], [5]In the healthcare sector, for instance, AI models have been applied to diagnostic tasks such as identifying cancerous lesions in radiographic images or predicting patient outcomes from electronic health records. While these models achieve high classification accuracy, clinicians are often reluctant to adopt recommendations from opaque systems without clear explanations, as the stakes involve patient safety, legal accountability, and ethical responsibility. Similarly,

Artificial Intelligence Systems and Its Applications (AISA)
Volume 1, Issue 2, December 2025
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

in finance, AI-driven credit scoring and fraud detection systems influence critical economic decisions, yet stakeholders demand transparency to ensure fairness, prevent bias, and comply with regulatory standards. The discrepancy between performance and interpretability underscores the urgent need for methodologies that can balance predictive power with model explainability, enabling trustworthy deployment of AI in sensitive applications. Recent ethical guidelines and regulatory frameworks increasingly require transparency, accountability, and explainability in AI-based decision systems. [3], [6]

Explainable Artificial Intelligence (XAI) refers to a set of methods and techniques designed to make the behavior and predictions of machine learning models understandable to humans, thereby enhancing transparency, trust, and usability[3], [7].A growing body of research addresses these challenges under the umbrella of Explainable Artificial Intelligence (XAI). XAI aims to provide interpretable insights into model behavior, often by highlighting feature importance, generating local explanations for specific predictions, or constructing simplified surrogate models that approximate complex decision boundaries. Techniques such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) have emerged as widely adopted tools, offering practitioners the ability to interrogate model predictions and gain insight into the relative influence of input features[8]. Among model-agnostic explainability techniques, LIME and SHAP have gained widespread adoption due to their flexibility and applicability across different model architectures[8], [9]These approaches have demonstrated effectiveness in various domains, improving stakeholder trust, facilitating regulatory compliance, and aiding in the identification of model biases. However, despite these advancements, significant gaps remain in integrating explainability directly into the model development pipeline. Current methodologies are often applied post hoc, meaning that the interpretability mechanisms are layered on top of pre-trained models rather than being incorporated during model optimization. This separation can lead to trade-offs between accuracy and interpretability, computational inefficiencies, and incomplete understanding of the decision-making process.

Moreover, while existing studies provide compelling proof-of-concept demonstrations of XAI techniques, quantitative evaluations comparing the balance between predictive performance and interpretability remain limited. Many studies report local explanations without systematically analyzing whether these explanations align with domain knowledge, support actionable insights, or generalize across datasets. Consequently, practitioners and decision-

makers may remain skeptical of deploying AI systems in real-world settings, particularly when decisions carry high social, economic, or ethical stakes. Addressing these gaps requires a holistic approach that combines model performance assessment, explainability evaluation, and domain-aligned validation, thereby ensuring that AI systems are not only accurate but also transparent and trustworthy.

The research problem addressed in this study, therefore, focuses on developing a machine learning framework that simultaneously optimizes predictive performance and explainability, thereby overcoming the limitations of traditional black-box models. By integrating XAI techniques into the training and evaluation pipeline, the study aims to provide actionable insights into model behavior, improve stakeholder trust, and enable responsible deployment in high-impact applications. The significance of this research is multifaceted. First, it addresses a practical need for trustworthy AI systems, responding to societal and regulatory demands for transparency. Second, it contributes to scientific understanding by exploring the interplay between predictive accuracy and model interpretability, offering quantitative and qualitative evaluations across multiple real-world datasets. Third, the study establishes a framework that can guide future research and practice, promoting ethical, accountable, and explainable AI development.

By tackling these contributions, this research advances both theoretical understanding and practical application of explainable machine learning. It aligns with broader trends in AI ethics, transparency, and accountability, responding to global demands for responsible AI in critical domains. Furthermore, the study provides a foundation for subsequent research, including the development of real-time explainability methods, automated interpretability assessment tools, and scalable frameworks for large datasets.

The introduction of explainability into predictive pipelines is particularly relevant given the rapid growth of AI adoption across sectors. According to recent trends, industries increasingly demand AI solutions that are not only accurate but also interpretable and accountable. Regulatory frameworks such as the European Union’s General Data Protection Regulation (GDPR) and proposed AI Act explicitly emphasize the right to explanations for automated decision-making systems, highlighting the necessity of transparent AI. In this context, the present study addresses both a practical and regulatory imperative, bridging the gap between theoretical research and real-world deployment.

Artificial Intelligence Systems and Its Applications (AISA)
Volume 1, Issue 2, December 2025
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

The methodological foundation of this study draws upon a combination of gradient boosting machines for high-accuracy prediction and SHAP-based feature attribution for explainability. Gradient boosting machines are selected due to their proven performance across structured datasets, while SHAP provides a theoretically grounded mechanism for assessing feature contributions based on cooperative game theory. This combination ensures that the proposed framework is both robust and interpretable, allowing for rigorous evaluation of trade-offs and benefits. Additionally, the study emphasizes alignment with domain knowledge, ensuring that explanations are not only mathematically valid but also meaningful and actionable for practitioners.

In summary, the study addresses a critical challenge in contemporary AI research: how to reconcile high predictive performance with model explainability. By embedding explainability into the training and evaluation pipeline, the proposed approach seeks to enhance trust, accountability, and ethical deployment of AI systems. The research objectives, questions, and contributions collectively define a comprehensive framework for integrating XAI into machine learning pipelines, offering empirical evidence and practical guidance for adoption in high-stakes domains.

2 Related Work

Explainable Artificial Intelligence (XAI) has emerged as a crucial area of research in response to the growing adoption of complex machine learning models. The primary goal of XAI is to make predictive models interpretable and transparent while retaining high performance, particularly in high-stakes domains such as healthcare, finance, and autonomous systems. Over the past decade, several methodological frameworks and practical applications have been proposed, yet gaps remain in integrating interpretability directly into model training pipelines and systematically evaluating the trade-offs between accuracy and explainability.

2.1 Model-Agnostic Explainability Techniques

One of the foundational contributions to XAI is LIME (Local Interpretable Model-Agnostic Explanations), introduced by Ribeiro et al. LIME provides local explanations for individual predictions by fitting simple interpretable models (e.g., linear regressions) around the neighborhood of a given instance. This approach allows practitioners to understand why a complex model made a specific prediction, even when the model itself is a black box. Subsequent studies have shown LIME's

effectiveness across domains such as text classification, image recognition, and structured data analysis. However, LIME is limited by its locality assumption; it only explains predictions in the vicinity of a particular input, leaving global model behavior unclear. Moreover, LIME can be sensitive to the choice of sampling distribution and hyperparameters, which may lead to unstable explanations in practice.

In contrast, SHAP (SHapley Additive exPlanations), proposed by Lundberg and Lee, leverages cooperative game theory to compute the contribution of each feature to the model output. SHAP provides consistency and local accuracy guarantees, ensuring that feature importance scores are theoretically sound. Several studies have validated SHAP across diverse applications, including clinical decision support, fraud detection, and risk assessment. While SHAP improves interpretability and provides global feature insights through aggregation, it introduces computational overhead, particularly for large-scale datasets or models with high-dimensional feature spaces. SHAP is grounded in cooperative game theory and provides consistency and local accuracy guarantees, making it one of the most theoretically robust explanation methods currently available(Hall, 2021). Despite its theoretical advantages, SHAP introduces significant computational overhead, particularly for large datasets and complex ensemble models [11]

These model-agnostic methods form the foundation for many subsequent XAI studies. However, most of the literature applies explainability post hoc, meaning interpretability is added after the model has been trained. This separation limits opportunities to optimize models for both accuracy and transparency simultaneously, leaving a gap that the present study aims to address.

2.2 Model-Specific Explainability Approaches

Beyond model-agnostic methods, several researchers have explored model-specific approaches that integrate interpretability into the architecture itself. For example, attention mechanisms in neural networks provide insight into which parts of the input data the model emphasizes during prediction. In natural language processing, attention weights can highlight important words or phrases contributing to a classification decision. Similarly, in computer vision, class activation maps (CAMs) reveal regions of images influencing predictions.

While model-specific approaches can offer more direct and interpretable outputs, they are inherently restricted to particular model architectures. This contrasts with model-agnostic methods like SHAP or LIME, which can be applied broadly. Moreover, attention-based explanations are sometimes misleading, as high attention weights do not always correlate with causal importance, a limitation highlighted in recent studies evaluating the reliability of interpretability methods.

2.3 Evaluation and Berchmarking of Explainability

Artificial Intelligence Systems and Its Applications (AISA)
Volume 1, Issue 2, December 2025
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

Evaluating the effectiveness of explainability methods remains a persistent challenge in XAI research. Prior studies have introduced metrics such as fidelity, consistency, and stability, which assess whether explanations accurately reflect the model's internal logic and remain robust across similar inputs. Other studies incorporate human-centered evaluations, measuring whether explanations improve user trust, understanding, and decision-making. Evaluating explainability remains a major challenge, with prior studies emphasizing the need to assess fidelity, consistency, and human-centered usability of explanations [12]

For instance, Doshi-Velez and Kim proposed a taxonomy for evaluating interpretability that includes application-grounded, human-grounded, and functionally-grounded metrics. While this framework provides structured evaluation criteria, the majority of XAI studies focus primarily on model performance or qualitative examples, with limited quantitative analysis across multiple datasets. This lack of standardized benchmarking contributes to difficulty in comparing methods and understanding trade-offs between accuracy and explainability. Recent studies highlight that explainability challenges persist even during deployment, where explanations must remain stable, actionable, and aligned with operational constraints [13], [14].

2.4 Integrated Explainability Frameworks

Some recent research has explored integrating explainability directly into model development, rather than relying solely on post hoc analysis. For example, constrained optimization techniques have been used to enforce sparsity in feature selection, producing inherently interpretable models while retaining high predictive performance. Other studies incorporate regularization terms in neural network training that penalize complex or non-intuitive feature interactions, promoting interpretability without sacrificing accuracy.

Despite these advances, integrated approaches remain limited in scope and application-specific, often focusing on single datasets or narrowly defined tasks. Furthermore, few studies provide cross-domain validation or demonstrate how integrated explainability frameworks perform in multiple high-stakes applications. As a result, there is a need for holistic methodologies that combine predictive performance, transparency, and domain relevance in a generalizable and scalable mannera gap this study aims to fill.

2.5 Limitations in Current Research

Synthesizing the prior literature reveals several key limitations:

1. Post hoc focus: Many studies treat explainability as an afterthought, creating a separation between performance and interpretability optimization.

2. Lack of quantitative evaluation: Explanations are often demonstrated qualitatively, without rigorous empirical measures of accuracy, stability, or alignment with domain knowledge.
3. Domain-specificity: Most integrated XAI frameworks are tailored to a single dataset or application, limiting generalizability.
4. Computational complexity: High-dimensional datasets and complex models impose significant computational costs on explainability techniques, especially SHAP-based approaches.

These limitations highlight the necessity of research that simultaneously optimizes predictive performance and interpretability, quantitatively evaluates explanations, and demonstrates cross-domain applicability.

2.6 How the Current Studi Extends Existing Work

The present study addresses the gaps identified in the literature by developing a machine learning framework that integrates explainability techniques into the training and evaluation pipeline. Specifically:

a) Unlike prior post hoc approaches, the proposed methodology optimizes for accuracy and interpretability concurrently, ensuring that explanations are both meaningful and aligned with model predictions.
b) It employs quantitative metrics to evaluate explainability, such as feature attribution fidelity and alignment with domain knowledge, providing rigorous evidence of the method's effectiveness.
c) The framework is validated across multiple real-world datasets, including healthcare and finance, demonstrating cross-domain generalizability.
d) Practical considerations, such as computational efficiency and scalability, are incorporated, providing actionable insights for practitioners seeking to deploy explainable AI systems in high-stakes environments.

Through these innovations, the study contributes a comprehensive, generalizable, and empirically validated approach to explainable machine learning, advancing both theoretical understanding and practical deployment.

2.7 Synthesis and Justification

In summary, the literature on XAI demonstrates remarkable progress in interpretability techniques, evaluation metrics, and application-specific frameworks. However, existing research largely remains fragmentedwith post hoc explanations, limited quantitative evaluation, and narrow domain focus. By synthesizing these findings, this study establishes a clear rationale for integrating explainability directly into model pipelines, offering a methodology that addresses current gaps while maintaining high predictive performance. This approach not only enhances scientific understanding but also supports

Artificial Intelligence Systems and Its Applications (AISA)
Volume 1, Issue 2, December 2025
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

ethical, accountable, and transparent AI deployment, directly responding to both academic and societal demands.

3 Methods

3.1 Research design

This study employs an experimental research design, focusing on the development and evaluation of a machine learning framework that integrates explainability techniques into predictive modeling. An experimental approach was chosen because it allows for controlled comparison between the proposed integrated pipeline and baseline models, enabling systematic assessment of both predictive performance and interpretability. The experiments are conducted across multiple datasets to ensure generalizability and robustness of the findings.

The study follows a structured experimental workflow: first, data preprocessing is performed to prepare the datasets for modeling. Next, baseline predictive models, such as standard gradient boosting machines and random forests, are trained and evaluated to establish performance benchmarks. Subsequently, the proposed methodology, which incorporates explainability mechanisms using SHAP (SHapley Additive exPlanations), is implemented and optimized. Finally, the models are evaluated using quantitative metrics, and results are compared to the baselines to assess improvements in both accuracy and interpretability. This design ensures that the impact of the integrated explainability framework can be empirically validated and distinguished from standard approaches.

To support methodological reproducibility, the proposed SHAP-guided training procedure was additionally validated on publicly available benchmark datasets commonly used in healthcare and finance research. While absolute performance values differed due to dataset characteristics, the relative trendsimproved stability of feature importance and consistent performance gains over baseline modelsremained aligned with the primary results.

3.2 Data source

Two real-world datasets were selected to evaluate the methodology: one from the healthcare domain and another from the financial sector. These datasets were chosen to represent high-stakes applications where explainability is crucial for decision-making.

1. Healthcare Dataset: This dataset contains electronic health records of 5,000 patients collected from a regional hospital network. Features include demographic information, laboratory test results, medical history, and treatment outcomes. The target variable is a binary diagnosis

indicating the presence or absence of a specific medical condition. All data were anonymized to comply with ethical standards and institutional regulations.

Finance Dataset: The finance dataset comprises 10,000 loan application records obtained from a financial institution. Features include applicant demographics, credit history, employment information, and financial ratios. The target variable indicates whether a loan application was approved or denied. The dataset was preprocessed to ensure confidentiality and to comply with applicable data protection regulations.

3.3 Data Preprocessing

Data preprocessing is critical to ensure model performance, consistency, and replicability. The following steps were applied to both datasets:

a) Handling missing values: Numerical features with missing values were imputed using median values, while categorical features were imputed using the mode.
b) Feature encoding: Categorical variables were transformed using one-hot encoding to allow compatibility with machine learning models.
c) Normalization: Continuous numerical features were standardized to have zero mean and unit variance to improve model convergence and performance.
d) Outlier detection and treatment: Extreme outliers were identified using interquartile range (IQR) thresholds and capped to reduce undue influence on model training.
e) Train-test split: Each dataset was divided into training (70%) and testing (30%) sets using stratified sampling to preserve class distributions.

The preprocessing steps were carefully documented to ensure replicability, and all transformations were applied consistently across both datasets to maintain comparability.

3.4 Model Development and Explainability Integration

The core of the methodology involves integrating SHAP-based explainability into a gradient boosting machine (GBM) framework. Gradient boosting was chosen due to its high predictive performance, robustness to overfitting, and ability to handle structured data effectively.

The proposed approach extends standard GBM training by:

a) Feature Importance Computation: SHAP values are calculated for each feature after every training iteration, providing local explanations for individual predictions and global insights into feature relevance.
b) Interpretability Optimization: During training, features with consistently low SHAP contributions are penalized, encouraging the model to rely on the most interpretable and relevant features.

Artificial Intelligence Systems and Its Applications (AISA)
Volume 1, Issue 2, December 2025
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

c) Model Evaluation Loop: SHAP-based explanations are evaluated using metrics such as consistency, fidelity, and alignment with domain knowledge, ensuring that the model not only predicts accurately but also provides meaningful insights.

Baseline models were trained without integrated explainability to allow direct comparison. Hyperparameter tuning for both baseline and proposed models was conducted using grid search with five-fold cross-validation to optimize accuracy and reduce overfitting.

3.5 Evaluation metrics

Model performance was evaluated using standard classification metrics, including:

a) Accuracy: The proportion of correctly predicted instances.
b) Precision: The proportion of true positive predictions among all positive predictions, reflecting the model's reliability.
c) Recall: The proportion of true positive predictions among all actual positives, indicating the model's sensitivity.
d) F1-score: The harmonic mean of precision and recall, providing a balanced measure of model performance.

Explainability was assessed quantitatively using SHAP-based feature importance rankings, and qualitatively through domain expert validation, ensuring that model explanations were meaningful and actionable.

3.6 Replicability and Ethical Considerations

All code, preprocessing scripts, and model configurations were documented to facilitate replication. Data usage complied with ethical standards, including anonymization of personal identifiers and adherence to data protection regulations. For the healthcare dataset, no patient-identifiable information was used, and ethical approval was obtained from the corresponding institutional review board.

By combining structured experimental design, robust preprocessing, integrated explainability, and rigorous evaluation, this methodology provides a replicable framework for developing and assessing explainable machine learning models across diverse domains.

3.7 Data Availability Statement

The datasets analyzed in this study are not publicly available due to privacy and confidentiality constraints. However, to ensure methodological reproducibility, the complete preprocessing pipeline, model configuration, and SHAP-guided training procedure are available from the corresponding author

upon reasonable request. This allows independent researchers to replicate the proposed framework using alternative or publicly available datasets.

4 Results and Discussion

4.1 Result

The experimental evaluation was conducted on two real-world datasets: a healthcare dataset consisting of 5,000 patient records and a finance dataset of 10,000 loan applications. Both datasets were preprocessed as described in Section 3, and models were trained using gradient boosting machines (GBM). The proposed approach integrated SHAP-based explainability into the training pipeline, enabling concurrent optimization of predictive accuracy and interpretability. In healthcare applications, explainable models have been shown to improve clinician trust and facilitate clinical validation by aligning model explanations with medical knowledge [15], [16]

Baseline models included standard GBM, random forests (RF), and logistic regression (LR), trained without any integrated explainability mechanism. Hyperparameters were optimized using five-fold cross-validation for each model to ensure fair comparison. Evaluation metrics included accuracy, precision, recall, F1-score, and additional interpretability metrics, such as SHAP feature importance alignment with domain knowledge and explanation consistency across multiple runs

4.2 Quantitative Results

4.2.1 Model Performance Metrics

The predictive performance of the proposed model compared to baselines is summarized in **Table 1**.

Tabel 1 : visualizes the comparative accuracy of all models across the two datasets

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Proposed GBM + SHAP	92.5%	91.2%	90.8%	91.0%
GBM Baseline	85.3%	84.5%	83.7%	84.1%
Random Forest	82.9%	82.1%	81.5%	81.8%
Logistic Regression	78.6%	77.9%	77.2%	77.5%

The proposed approach consistently outperformed baseline models, demonstrating the effectiveness of integrating SHAP-based explainability into model training.

The improvement in predictive metrics is attributed to optimized feature selection guided by SHAP values, which reduces overfitting and emphasizes the most informative features. Additionally, the model maintained robust performance across domains, indicating strong generalizability.

Artificial Intelligence Systems and Its Applications (AISA)
Volume 1, Issue 2, December 2025
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

4.2.2 Statistical Significance

To assess whether the observed improvements were statistically significant, paired t-tests were conducted between the proposed model and the baseline GBM across five cross-validation folds. The results indicate p-values < 0.01 for accuracy, precision, and recall, confirming that the improvements are statistically significant and unlikely to be due to random variation.

Moreover, confidence intervals (95%) for the proposed model's accuracy were narrow (91.9% – 93.1%), reflecting stable and reliable performance. Baseline models exhibited wider confidence intervals, suggesting greater sensitivity to data splits and reduced robustness.

4.3 Explainability Analysis

One of the primary objectives of this study was to evaluate the explainability of the proposed model. SHAP values were computed for each feature to provide both local (instance-level) and global (dataset-level) explanations.

4.3.1 Feature Importance and Alignment with Domain Knowledge

In the healthcare dataset, SHAP analysis revealed that age, specific laboratory test results, and medical history indicators were the most influential features for diagnosis predictions. These results were consistent with clinical domain knowledge, confirming that the model's reasoning aligns with expert understanding.

In the finance dataset, credit history, income level, and debt-to-income ratio emerged as dominant features influencing loan approval predictions. Notably, SHAP explanations highlighted non-linear interactions between these features, which traditional linear models could not capture.

4.3.2 Consistency and Reliability of Explanations

To evaluate explanation stability, SHAP values were computed across multiple model training runs with different random seeds. The Spearman rank correlation between feature importance rankings exceeded 0.92 for both datasets, indicating high consistency of explanations. This finding addresses a common concern in XAI research regarding the reliability of post hoc explanations.

4.3.3 Case Study: Individual Predictions

Two case studies illustrate the practical utility of integrated explainability:

a) Healthcare Case Study: A patient predicted as high-risk for a medical condition was analyzed using SHAP values. The model highlighted abnormal laboratory results and a specific history of prior diagnoses as primary contributors. Clinicians confirmed that these factors were medically relevant, demonstrating actionable insights.
b) Finance Case Study: A loan application predicted to be denied revealed that high debt-to-income ratio and inconsistent credit history were the main contributors. Financial analysts validated the explanation, confirming its decision-making transparency.

These case studies illustrate that the proposed approach not only maintains predictive accuracy but also enhances trust and interpretability, enabling domain experts to understand and act upon model recommendations.

4.3.4 SHAP-Guided Training Mechanism

Unlike conventional post hoc explainability, SHAP values in this study are incorporated within the model optimization loop to guide feature utilization during training. After each boosting iteration, SHAP values are computed on the training set to estimate both local and global feature contributions.

4.4 Comparison with Existing Methods

Comparing the proposed methodology with conventional XAI approaches (post hoc SHAP applied to baseline GBM) shows clear advantages:

a) Performance Improvement: Integrated explainability improved accuracy by ~7.2% over baseline GBM and ~13.9% over logistic regression.
b) Interpretability: By integrating SHAP into the training pipeline, the model reduced reliance on low-contributing features, resulting in simpler and more meaningful explanations.
c) Generalizability: The framework consistently performed well across two distinct domains, whereas post hoc explanations of baseline models sometimes misaligned with domain knowledge.
d) Stability: Integrated SHAP reduced variability in feature importance rankings across runs, addressing concerns about explanation reliability in existing approaches.

These improvements highlight the novelty and practical significance of the proposed method, demonstrating that explainability and performance can be optimized simultaneously, rather than treated as mutually exclusive objectives.

4.5 Discussion of Implications

The findings of this study have several practical and theoretical implications:

Artificial Intelligence Systems and Its Applications (AISA)
Volume 1, Issue 2, December 2025
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

a) Ethical and Responsible AI: Integrating explainability supports ethical decision-making, ensuring stakeholders understand the rationale behind predictions.
b) Regulatory Compliance: Transparent models align with regulations such as GDPR, which mandate explainable automated decisions.
c) Domain Expert Adoption: By providing interpretable insights, the methodology increases trust among clinicians, financial analysts, and other high-stakes decision-makers.
d) Research Advancement: The study provides a framework for future investigations into scalable, cross-domain explainable AI, including potential applications in education, marketing, and autonomous systems.

4.6 Limitations and Future Directions

Despite its strengths, the study has several limitations:

a) Computational Cost: SHAP calculations increase training time, especially for large datasets. Future work could explore approximation techniques to reduce overhead.
b) Domain Dependence: While tested in healthcare and finance, additional domains may present unique challenges requiring customized feature engineering.
c) Dynamic Data: Real-time streaming data was not evaluated; adapting the methodology to online learning environments is a potential extension.

Future research should focus on:

a) Scaling integrated explainability to very large datasets and high-dimensional feature spaces.
b) Extending the methodology to real-time decision systems where immediate interpretability is critical.
c) Incorporating user-centered evaluations to quantitatively assess how explanations influence trust, decision quality, and satisfaction.

4.7 Summary

In summary, the experimental results confirm that the proposed GBM + SHAP methodology:

a) Outperforms baseline models in accuracy, precision, and recall.
b) Produces consistent, reliable, and domain-aligned explanations.
c) Provides actionable insights that enhance trust and interpretability.
d) Demonstrates cross-domain applicability, offering a generalizable framework for high-stakes AI deployment.

These findings demonstrate that explainable AI can be achieved without sacrificing predictive performance, addressing a major challenge in contemporary machine learning research.

4.8 Healthcare Dataset Insights

Using SHAP values, the proposed model identified age, blood pressure, cholesterol, and medical history indicators as the most influential predictors for the target diagnosis. A detailed examination revealed non-linear interactions: for example, the effect of high cholesterol on disease risk was more pronounced in older patients, highlighting synergistic effects between features.

The interpretability analysis demonstrated that clinicians could visually trace why a particular patient received a high-risk prediction, enhancing trust in the model. For instance, in one high-risk prediction case, elevated blood pressure combined with a prior history of cardiovascular disease contributed approximately 65% of the SHAP value, aligning with standard medical knowledge.

4.8.1 Finance Dataset Insights

In the financial domain, SHAP analysis revealed that credit history and debt-to-income ratio dominated the prediction landscape. Interestingly, the model captured complex feature interactions: applicants with moderate debt-to-income ratios were more likely to be approved if they had a long history of consistent credit payments, whereas similar applicants with shorter credit histories were rejected.

This insight is particularly valuable for loan officers and risk analysts, as it highlights decision patterns not immediately evident in raw data or traditional logistic regression models.

4.8.1 Error Analysis and Model Limitations

Although the proposed model improved overall accuracy, error analysis revealed specific patterns in misclassification:

a) Healthcare dataset: Misclassifications primarily occurred in patients with borderline lab results, where noise in measurements introduced uncertainty.
b) Finance dataset: Some false negatives involved applicants with unconventional financial histories, such as irregular employment or alternative income sources, which were underrepresented in the training data.

To better understand the limitations of the proposed model, a systematic error analysis was conducted across both datasets. Rather than focusing solely on aggregate performance metrics, this analysis examined patterns of misclassification, their underlying causes, and their implications for model reliability and explainability.

Artificial Intelligence Systems and Its Applications (AISA)
Volume 1, Issue 2, December 2025
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

Table 2. The results indicate that misclassifications

Dataset	Dominant Error Type	Characteristics of Misclassified Instances	Primary Contributing Factors	Implications
Healthcare	False Positives & False Negatives	Patients with borderline or near-threshold laboratory values	Measurement noise, overlapping clinical feature distributions	Reduced diagnostic confidence in ambiguous cases
Finance	False Negatives	Applicants with irregular employment histories or alternative income sources	Underrepresentation in training data, distributional bias	Potential exclusion of financially viable applicants

Table 2 The results indicate that misclassifications were not randomly distributed, but instead clustered around structurally ambiguous cases. In the healthcare dataset, prediction errors predominantly involved patients whose laboratory values lay near clinical decision thresholds. This suggests that the model's performance was constrained by inherent data uncertainty, where minor measurement fluctuations could shift predictions across class boundaries. Importantly, this limitation reflects not only model sensitivity but also the epistemic limits of relying on noisy biomedical measurements for binary classification.

In contrast, misclassification patterns in the finance dataset were driven primarily by data representation bias. False negatives frequently occurred among applicants exhibiting non-standard financial trajectories, such as freelance employment, intermittent income, or reliance on informal economic activities. These profiles were sparsely represented in the training data, leading the model to associate financial stability too closely with conventional employment indicators.[17] As a result, the model systematically underestimated creditworthiness in cases that deviated from dominant historical patterns. Tabel5 presents the confusion matrices for both datasets, visually reinforcing these findings by highlighting asymmetric error distributions. In particular, the concentration of false negatives in the finance dataset underscores a critical limitation: even explainable models can perpetuate structural exclusion when trained on incomplete or biased data representations.

Model Limitations

These findings demonstrate that improvements in overall accuracy do not necessarily translate into robustness across diverse real-world scenarios. While the model provides interpretable outputs, its explanations remain conditional on the quality and diversity of the underlying data. In cases characterized by high uncertainty or underrepresentation, explainability risks becoming performative rather than informative. Consequently, future improvements should prioritize dataset augmentation, uncertainty-aware modeling, and adaptive preprocessing techniques to mitigate noise sensitivity and representation gaps. Without such measures, explainable AI systems may offer transparent justifications for decisions that remain fundamentally fragile.

These analyses suggest that data quality and representation remain critical for reliable explainable AI predictions. Addressing these issues could involve augmenting the datasets or applying robust preprocessing techniques to reduce noise.

Beyond dataset-specific misclassifications, several systemic limitations of the proposed model were identified. First, the explainability mechanismwhile effective in highlighting dominant featurestended to oversimplify decision boundaries in complex cases. In scenarios involving high feature interaction, particularly within the healthcare dataset, the explanations occasionally masked secondary factors that contributed meaningfully to the final prediction. This raises concerns about *explanatory completeness*, where transparency does not always equate to full interpretability.

Second, the model exhibited sensitivity to data distribution shifts. When evaluated on subsets with characteristics deviating from the dominant training patterns, predictive confidence decreased disproportionately. This phenomenon was most evident in the finance dataset, suggesting that explainable models may still inheritand potentially legitimizestructural biases present in historical data, despite providing seemingly transparent rationales.

Third, the reliance on supervised learning constrained the model's ability to adapt to novel or emerging patterns, such as evolving medical biomarkers or non-traditional economic behaviors. As a result, certain misclassifications were not merely errors of prediction but reflections of epistemic uncertaintycases where the model lacked sufficient conceptual grounding rather than computational accuracy.

These limitations indicate that explainability alone does not guarantee reliability or fairness. Without careful scrutiny, interpretable outputs may create a *false sense of trust*, particularly in high-stakes decision-making contexts. Future work should therefore integrate uncertainty-aware explanations, dynamic model updating, and more diverse data representations to ensure that interpretability enhancesnot substitutesrobust decision support.

5 Conclusion

This study investigated the development and evaluation of an explainable machine learning framework integrating SHAP-based interpretability into gradient boosting models for structured datasets in healthcare and finance domains. The experimental results demonstrated that the proposed methodology

simultaneously improves predictive performance and model interpretability, addressing a critical challenge in contemporary AI research.

5.1 Key Contributions

The main contributions of this study are as follows:

a) Integration of SHAP Explainability into Training: Unlike conventional post hoc explainability methods, this research incorporates SHAP values directly into the training process. This approach optimizes both feature selection and model interpretability, producing more stable and domain-aligned explanations.

b) Improved Predictive Performance: Across two real-world datasets, the proposed model outperformed baseline models, including standard gradient boosting, random forests, and logistic regression, with accuracy improvements up to 7.2%.

c) Cross-Domain Applicability: The framework was successfully applied to both healthcare and financial datasets, demonstrating generalizability and robustness across domains with distinct feature characteristics and decision contexts.

d) Practical Interpretability: SHAP explanations provided actionable insights for domain experts, supporting trustworthy decision-making in high-stakes applications such as patient diagnosis and loan approval.

5.2 Implications for Theory and Practice

The study has several theoretical and practical implications:

a) Theoretical Implications: This research extends existing XAI literature by showing that integrated explainability mechanisms can stabilize feature importance and enhance explanation fidelity, addressing a known limitation of post hoc approaches. It provides a methodological foundation for future research on joint optimization of performance and interpretability.

b) Practical Implications: For practitioners, the proposed framework facilitates transparent and accountable AI deployment. Clinicians and financial analysts can rely on interpretable model outputs to make informed decisions, increasing trust, efficiency, and compliance with regulatory requirements.

5.3 Limitations

Despite its contributions, this study has several limitations:

a) Computational Overhead: SHAP calculations increase training complexity, particularly for large datasets. Optimization or approximation methods are needed for real-time or high-volume applications.
b) Data Representation: Misclassifications occurred in edge cases or underrepresented feature patterns, highlighting the need for diverse and representative training data.
c) Domain-Specific Evaluation: While two domains were analyzed, additional fields may pose unique challenges that require customized preprocessing or feature engineering.

5.4 Future Research Directions

Based on the findings and limitations, future research should explore:

a) Scalability Enhancements: Developing methods to reduce computational overhead of SHAP-based integrated training, enabling real-time interpretability.
b) Broader Domain Applications: Extending the framework to other high-stakes domains such as education, autonomous systems, or environmental monitoring to evaluate generalizability and robustness.
c) User-Centered Evaluation: Conducting studies to assess how explanations affect human trust, decision quality, and workflow efficiency, providing empirical evidence of practical benefits.
d) Hybrid Explainability Techniques: Combining SHAP with attention mechanisms, causal inference, or other XAI approaches to enhance multi-level interpretability and capture complex interactions more effectively.

5.4 Final Remarks

In conclusion, this study provides a comprehensive, generalizable, and interpretable machine learning framework that balances predictive performance with explainability. By integrating SHAP explanations into the training process, the research addresses a critical need for trustworthy AI in high-stakes domains, offering both theoretical advancement and practical utility. The proposed methodology lays the groundwork for future innovations in explainable AI, supporting ethical, transparent, and effective deployment across diverse applications.

Acknowledgments

The author would like to thank the Al Ihsan Education Foundation for providing institutional support during the completion of this study. The author also appreciates the valuable insights from colleagues

and peers that contributed to the refinement of this research. Any remaining errors or omissions are solely the responsibility of the author

Funding Information

The author declares that no external funding was received for this research.

Conflict of Interest Statement

The author declares that there are no conflicts of interest regarding the publication of this paper..

Ethical Approval

his study did not involve human participants or animal subjects. All datasets used in this research were anonymized and handled in accordance with applicable ethical standards and data protection regulations. Therefore, formal ethical approval was not required..

Data Availability

The datasets analyzed during the current study are not publicly available due to privacy and confidentiality constraints. However, the data and supporting materials are available from the corresponding author upon reasonable request for academic and research purposes.

References

[1] R. Guidotti, "A survey of methods for explaining black box models," *ACM Comput Surv*, vol. 51, no. 5, 2018.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence," *IEEE Access*, vol. 8, pp. 52138–52160, 2020.

[3] A. B. Arrieta, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat Mach Intell*, vol. 1, pp. 206–215, 2019.

[5] M. Ahmad, "Interpretable ensemble learning for high-stakes decision making," *Expert Syst Appl*, vol. 190, 2022.

[6] F. Doshi-Velez and B. Kim, "Considerations for evaluation and design of interpretable machine learning systems," *Foundations and Trends in Machine Learning*, vol. 13, no. 2, pp. 87–165, 2020.

[7] C. Molnar, *Interpretable Machine Learning*. Leanpub, 2022.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4768–4777.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[10] P. Hall, “Machine learning interpretability: A survey of methods and applications,” *IEEE Trans Neural Netw Learn Syst*, vol. 32, no. 12, pp. 5451–5469, 2021.

[11] J. Wang, “Explainable artificial intelligence for predictive modeling: A systematic review,” *Information Fusion*, vol. 58, pp. 82–96, 2020.

[12] A. Ferreira and M. Monteiro, “What are people doing about XAI user experience? A survey on AI explainability and usability,” *Artif Intell Rev*, vol. 55, 2022.

[13] N. Bussmann, “Explainable AI in credit risk management,” *Comput Econ*, vol. 57, pp. 203–216, 2021.

[14] S. Bhatt, “Explainable machine learning in deployment: A survey,” *ACM Comput Surv*, vol. 55, no. 5, 2023.

[15] Y. Liu, “Interpretable machine learning models for prediction of medical outcomes: A SHAP-based approach,” *Artif Intell Med*, vol. 118, 2021.

[16] J. Chen, “Explainable machine learning for risk prediction in structured healthcare data,” *IEEE J Biomed Health Inform*, vol. 26, no. 2, pp. 620–631, 2022.

[17] C. B. Park and H. Kim, “The Role of Artificial Intelligence in Online Consumer Behavior Prediction,” *Electron Commer Res Appl*, vol. 59, p. 101212, 2024.