An Analytical Study of Explainable AI Models for High-Stakes Decision Systems

C. Sumanth
Research Scholar, Department of Computer Science, North East Christian University
Dr. Kritesh Sharan
Associate Professor, Department of Computer Science, North East Christian University

Abstract

Artificial Intelligence (AI) is becoming more and more common in critical decision-making areas like healthcare diagnosis, financial risk assessment, and criminal justice, where the stakes are incredibly high and the consequences can be serious and irreversible. Even though advanced machine learning models often boast impressive predictive accuracy, their black-box nature raises important issues around transparency, trust, fairness, and accountability. This has sparked a growing interest in Explainable Artificial Intelligence (XAI), which seeks to make AI-driven decisions clearer and more reliable for the people involved. In this paper, we dive into an analytical study of explainable AI models used in these high-stakes environments, focusing on finding the right balance between predictive performance and interpretability. We compare traditional black-box models with those that are inherently interpretable, as well as post-hoc explanation techniques. We take a closer look at popular XAI methods like Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to see how feature-level explanations can boost transparency without significantly sacrificing accuracy.

To back up our findings, we conduct experimental analysis using benchmark datasets that are commonly used in critical decision-making fields, including healthcare, finance, and criminal justice. We evaluate models like Logistic Regression, Random Forest, and Gradient Boosting using standard performance metrics alongside criteria focused on explainability. The results show that explainable frameworks not only enhance model transparency and user trust but also maintain competitive predictive performance. The study emphasizes how crucial explainability is when it comes to tackling ethical issues, spotting biases, and ensuring compliance with regulations in high-risk situations. In essence, the findings show that explainable AI is key to creating decision support systems that are trustworthy, accountable, and centered on human needs. This makes it absolutely vital for the responsible use of AI in high-stakes scenarios.

Keywords: Explainable Artificial Intelligence; High-Stakes Decision Systems; Model Interpretability; Trustworthy AI; LIME; SHAP; Machine Learning Transparency; Responsible AI

Introduction

Artificial Intelligence (AI) has really changed the game when it comes to decision-making in today’s world. It allows organizations to sift through massive amounts of data and make accurate predictions across a variety of fields. In recent years, we've seen AI models being embraced in critical areas like healthcare diagnosis, financial risk assessment, fraud detection, autonomous systems, and even criminal justice. The stakes are high in these fields, as the decisions made can directly impact people's lives, economic stability, safety, and legal rights. That’s why reliability and accountability are so crucial (Russell & Norvig, 2021).

Even though these advanced machine learning and deep learning models are impressive in their predictive abilities, many of them function like black boxes, leaving us in the dark about how they arrive at their decisions. Models like ensemble learning methods and deep neural networks are great at identifying complex patterns, but their lack of transparency can create real issues in important decision-making scenarios. Decision-makers often find it tough to understand, validate, or justify the recommendations made by AI, which can lead to a lack of trust, hesitance to adopt these technologies, and challenges in meeting ethical and regulatory standards (Lipton, 2018).

To tackle these issues, the concept of Explainable Artificial Intelligence (XAI) has become a hot topic. XAI is all about creating methods that make AI models clearer, more interpretable, and easier for humans to grasp, all while keeping their predictive power intact. By offering explanations for the outputs of these models, XAI helps users understand the reasoning behind a specific decision, what factors played a role in the outcome, and how the model reacts in different situations (Doshi-Velez & Kim, 2017).

Explainability is super important in high-stakes decision-making systems, where making the wrong or biased choices can lead to serious consequences. Take healthcare, for instance—clinicians need to be able to back up their AI-assisted diagnoses and treatment suggestions. In the finance world, regulatory bodies expect clear explanations for decisions made by automated credit and risk assessments. Likewise, in the criminal justice system, if AI tools are not transparent in their reasoning, they could end up perpetuating bias and unfairness (Guidotti et al., 2018).

Techniques for post-hoc explanations, like Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), have gained popularity as effective XAI methods because they can work with any model and help clarify complex black-box systems. These approaches offer insights at the feature level for individual predictions and also shed light on the overall behavior of the model, making them especially valuable in real-world decision support scenarios (Ribeiro et al., 2016; Lundberg & Lee, 2017). Still, there are ongoing concerns about how stable these explanations are, how easily humans can interpret them, and how to strike the right balance between accuracy and transparency. Beyond the technical aspects, explainability has become a crucial requirement from both ethical and regulatory standpoints. Legal frameworks like the General Data Protection Regulation (GDPR) stress the importance of having the right to an explanation for automated decisions, while governance guidelines from organizations like the National Institute of Standards and Technology (NIST) underscore explainability as a fundamental principle of trustworthy AI (European Union, 2016; NIST, 2023). In light of this context, this paper dives into an analytical exploration of explainable AI models tailored for high-stakes decision-making systems. It assesses both interpretable and black-box machine learning models, utilizing post-hoc explainability techniques to investigate how we can achieve a balance between transparency, trust, and predictive performance. By examining various explainability methods across benchmark datasets and key application areas, this research aims to pave the way for the creation of reliable, human-centered, and responsible AI systems that are well-suited for high-risk decision-making scenarios.

2. Methodology

2.1 Research Design

This research adopts a comparative analytical design, where we evaluate various machine learning models both with and without explainability techniques. The study focuses on:

Inherently interpretable models (like Logistic Regression)

Black-box models (such as Random Forest and Gradient Boosting)

Post-hoc explainability techniques applied to black-box models (like LIME and SHAP)

This approach allows us to assess the balance between model accuracy and interpretability, especially in critical decision-making situations.

2.2 Datasets and Application Domains

To mirror real-world high-stakes scenarios, we utilize publicly available benchmark datasets that are frequently used in explainable AI research:

Breast Cancer Wisconsin Diagnostic Dataset – which aids in healthcare decision support

German Credit Dataset – used for financial risk and credit assessment

COMPAS Recidivism Dataset – relevant to decision-making in the criminal justice system

These datasets were chosen for their significance, structured format, and their common use in evaluating fairness, bias, and explainability in AI systems (Guidotti et al., 2018).

2.3 Machine Learning Models

The following machine learning models are employed in our experimental analysis:

Logistic Regression (LR): Selected for its inherent interpretability and as a baseline for comparison

Random Forest (RF): An ensemble model known for its strong predictive performance, though it lacks transparency

Gradient Boosting (GB): A high-performing model adept at capturing complex interactions between features

All models undergo standard preprocessing techniques, including normalization, handling of missing values, and splitting into training and testing sets.

2.4 Explainability Techniques

When it comes to making sense of black-box models, two popular post-hoc XAI methods really stand out:

Local Interpretable Model-Agnostic Explanations (LIME):

This technique offers local insights by approximating how the model behaves around specific predictions, using simpler, interpretable surrogate models (Ribeiro et al., 2016).

SHapley Additive exPlanations (SHAP):

This method employs game-theoretic concepts to determine how much each input feature contributes, providing both local and global explanations while ensuring theoretical consistency (Lundberg & Lee, 2017).

These methods are chosen for their ability to work with any model and their versatility across various fields.

2.5 Evaluation Metrics

To evaluate the models, we use a mix of performance metrics and criteria focused on explainability:

Performance Metrics

- Accuracy
- Precision
- Recall
- F1-score

These metrics help us reliably assess predictive quality, especially in imbalanced datasets that are often found in high-stakes situations.

Explainability Metrics

- Consistency of feature importance
- Stability of explanations
- Clarity and simplicity of explanations

We qualitatively assess human-centered interpretability by looking at how well the explanations match up with what experts in the field expect.

2.6 Ethical and Regulatory Considerations

Given how crucial these application areas are, the study takes into account various ethical evaluation principles, such as:

- Detecting bias through feature attribution analysis
- Ensuring transparency in automated decision-making
- Aligning with regulatory frameworks like GDPR and NIST AI Risk Management guidelines

Explainability is viewed not just as a technical necessity but also as a way to foster trust, accountability, and responsible AI use.

2.7 Analytical Framework

The final analysis looks at:

- Predictive performance versus interpretability
- Interpretable models compared to black-box models
- The effectiveness of LIME versus SHAP across different domains

This framework provides a thorough understanding of how explainable AI models can aid in making trustworthy decisions in high-risk situations.

3. Experimental Results and Evaluation

This section dives into the experimental evaluation of machine learning models and the explainability techniques used in high-stakes decision-making systems. We'll take a closer look at the results, focusing on both predictive performance and how effective these techniques are at providing explanations. It’s important to highlight the balance between accuracy and interpretability in this context.

3.1 Predictive Performance of Machine Learning Models

The predictive performance of Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB) models was evaluated using Accuracy, Precision, Recall, and F1-score across three high-stakes datasets.

Table 1: Predictive Performance on Healthcare Dataset

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.94	0.93	0.92	0.92
Random Forest	0.97	0.96	0.96	0.96
Gradient Boosting	0.98	0.97	0.97	0.97

The results from the healthcare dataset show that ensemble-based models really shine compared to the more straightforward Logistic Regression model, outperforming it on all evaluation metrics. Gradient Boosting took the lead with an impressive accuracy of 0.98, along with a precision of 0.97, recall of 0.97, and an F1-score of 0.97, with Random Forest not far behind. This clearly demonstrates how effective ensemble learning is at capturing the complex, non-linear relationships found in medical diagnostic data. Many studies in healthcare analytics have echoed these findings, highlighting how boosting and bagging techniques excel in discriminative power by effectively reducing both bias and variance (Breiman, 2001; Friedman, 2001). Even though there’s a noticeable performance gap, Logistic Regression still managed to achieve a respectable accuracy of 0.94, which aligns with previous research that points out its robustness, stability, and transparency in clinical decision support systems (Hosmer, Lemeshow, & Sturdivant, 2013). It’s crucial to note that while ensemble models offer better predictive accuracy, their black-box nature can hinder interpretability—something that’s vital in healthcare, where clinicians need to justify their diagnoses and treatment choices (Lipton, 2018; Doshi-Velez & Kim, 2017). Thus, in line with existing literature, these findings underscore a key trade-off between predictive performance and interpretability, highlighting the necessity for explainable frameworks that merge high-accuracy models with dependable explanation techniques. This is essential for ensuring trust, accountability, and safe use in medical decision-making environments (Guidotti et al., 2018; Lundberg & Lee, 2017).

Table 2: Predictive Performance on Financial Credit Dataset

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.78	0.76	0.74	0.75
Random Forest	0.83	0.82	0.81	0.81
Gradient Boosting	0.85	0.84	0.83	0.83

The findings from the financial credit dataset reveal that ensemble learning models really shine when compared to Logistic Regression, especially in terms of predictive accuracy and reliability in classification. Gradient Boosting took the lead with an impressive accuracy of 0.85 and an F1-score of 0.83, closely followed by Random Forest. This highlights their exceptional ability to capture the intricate relationships among various financial risk factors, like credit history, income stability, and debt ratios.

These results align with previous studies that have shown ensemble methods to be particularly effective in credit scoring tasks, thanks to their resilience against noise and their knack for managing nonlinear feature relationships (Lessmann et al., 2015; Friedman, 2001). While Logistic Regression achieved a lower accuracy of 0.78, it continues to be a popular choice in financial risk assessment due to its interpretability, stability, and ease of regulatory validation (Thomas, Edelman, & Crook, 2017). In the highly regulated world of finance, transparency and explainability often take precedence alongside predictive performance, as automated credit decisions need to be justified to regulators and those affected (European Union, 2016). The moderate performance gap noted in this study suggests that although advanced ensemble models provide significant improvements in classification accuracy, their use in real-world financial systems should be paired with explainability techniques to ensure accountability and compliance. Overall, these results reinforce the existing literature advocating for a thoughtful blend of high-performing machine learning models with explainable AI approaches to foster trustworthy and fair financial decision-making systems (Guidotti et al., 2018; Barredo Arrieta et al., 2020).

Table 3: Predictive Performance on Criminal Justice Dataset

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.69	0.68	0.66	0.67
Random Forest	0.71	0.70	0.69	0.69
Gradient Boosting	0.72	0.71	0.70	0.70

The results from the predictive performance analysis on the criminal justice dataset reveal only slight differences among the Logistic Regression, Random Forest, and Gradient Boosting models. Gradient Boosting tops the charts with an accuracy of 0.72 and an F1-score of 0.70, but the edge it has over Logistic Regression is quite small. This narrow performance gap indicates that more complex ensemble models don’t really offer significant predictive benefits in this area. This finding echoes previous research that points to diminishing returns when it comes to model complexity in sensitive decision-making situations (Dressel & Farid, 2018). While Logistic Regression may not have the highest accuracy, it shines in terms of transparency and interpretability—qualities that are crucial in criminal justice, where algorithmic decisions can have a direct impact on people's freedom and legal outcomes (Angwin et al., 2016; Lipton, 2018). Additionally, relying on black-box models in this field raises serious ethical issues around bias, fairness, and accountability, especially when the performance improvements are minimal (Barredo Arrieta et al., 2020). In line with existing literature, these findings bolster the case for using simpler, more interpretable models in criminal justice systems, as they allow for better scrutiny, justification, and adherence to ethical and legal standards without sacrificing much in terms of predictive performance (Guidotti et al., 2018; Doshi-Velez & Kim, 2017).

3.2 Explainability Analysis Using LIME

LIME was applied to Random Forest and Gradient Boosting models to generate local explanations for individual predictions.

Table 4: Key Features Identified by LIME

Dataset	Top Influential Features
Healthcare	Cell uniformity, Tumor size, Cell shape
Finance	Credit history, Income level, Loan amount
Criminal Justice	Prior offenses, Age, Charge severity

Using LIME with Random Forest and Gradient Boosting models offers valuable local insights by pinpointing the key features that impact individual predictions across three critical datasets. In healthcare, factors like cell uniformity, tumor size, and cell shape stood out as the most significant, which aligns well with established clinical markers for breast cancer diagnosis, reinforcing the credibility of the model's explanations (Street et al., 1993; Ribeiro et al., 2016). In the financial credit dataset, LIME highlighted credit history, income level, and loan amount as the main drivers, which is in line with traditional credit risk assessment methods and previous research in financial analytics (Thomas et al., 2017).

Likewise, in the criminal justice dataset, prior offenses, age, and charge severity were identified as the most impactful features, mirroring the common predictors used in recidivism risk models found in existing studies (Angwin et al., 2016; Dressel & Farid, 2018). While these explanations align well with domain expertise, some minor fluctuations in feature importance were noted across different runs, supporting earlier observations that LIME explanations can be somewhat unstable due to their dependence on local perturbations (Guidotti et al., 2018; Molnar, 2022). Still, the findings suggest that LIME is quite effective at producing clear, user-friendly explanations for individual predictions, making it especially useful for exploratory analysis and case-level decision-making in high-stakes situations where localized transparency is essential.

3.3 Explainability Analysis Using SHAP

SHAP was employed to generate both local and global explanations, offering consistent feature attributions.

Table 5: Global Feature Importance Using SHAP

Dataset	Most Influential Features (Descending Order)
Healthcare	Cell uniformity, Tumor size, Nucleus texture
Finance	Credit history, Debt ratio, Income
Criminal Justice	Prior convictions, Age, Risk score

The global feature importance analysis using SHAP offers clear and well-founded explanations across all three high-stakes datasets. In the healthcare dataset, factors like cell uniformity, tumor size, and nucleus texture stood out as the key predictors, aligning closely with the clinically validated criteria used for breast cancer assessments. This alignment reinforces the trustworthiness of SHAP's explanations (Street et al., 1993; Lundberg & Lee, 2017). Moving to the financial credit dataset, SHAP pinpointed credit history, debt ratio, and income as the main contributors, which matches up with established credit scoring systems and previous studies that highlight repayment behavior and financial stability as crucial factors in assessing credit risk (Thomas et al., 2017).

Likewise, in the criminal justice dataset, prior convictions, age, and risk score were identified as the most significant features, mirroring the predictors commonly discussed in recidivism modeling literature (Angwin et al., 2016; Dressel & Farid, 2018). When compared to LIME, SHAP's explanations show greater stability and consistency across various runs, thanks to their robust game-theoretic foundation and additive feature attribution properties (Molnar, 2022). These traits make SHAP especially valuable for global model interpretation, auditing, and bias detection in high-stakes and regulated decision-making scenarios, where dependable and reproducible explanations are vital for building trust and accountability (Guidotti et al., 2018; Barredo Arrieta et al., 2020).

3.4 Comparison of Explainability Techniques

Table 6: Comparison Between LIME and SHAP

Criteria	LIME	SHAP
Explanation type	Local	Local & Global
Model dependency	Model-agnostic	Model-agnostic
Stability	Moderate	High
Computational cost	Low	High
Suitability for regulation	Limited	High

The comparative analysis shown in Table 6 reveals some key differences between LIME and SHAP, especially when it comes to their explanation scope, stability, and how they apply to high-stakes decision-making systems. LIME is great for providing local explanations that are both intuitive and quick to compute, making it ideal for real-time and exploratory analysis. However, its moderate stability can be a drawback in situations where consistent explanations are needed across different instances (Ribeiro et al., 2016; Molnar, 2022). On the other hand, SHAP offers both local and global explanations and boasts greater stability thanks to its game-theoretic approach, which ensures that feature attributions are consistent and additive (Lundberg & Lee, 2017).

While SHAP does come with a higher computational cost, its robustness and reproducibility make it a better fit for regulated environments where transparency, auditability, and compliance are crucial, such as in finance, healthcare, and criminal justice systems (Guidotti et al., 2018; Barredo Arrieta et al., 2020). In summary, this comparison indicates that while LIME excels in providing quick, instance-level interpretability, SHAP is the preferred choice for high-risk applications that require dependable, regulation-ready explanations.

3.5 Accuracy vs Interpretability Trade-off

Table 7: Accuracy–Interpretability Comparison

Model	Accuracy Level	Interpretability Level
Logistic Regression	Moderate	High
Random Forest	High	Low

Gradient Boosting	Very High	Low
RF + SHAP	High	Moderate
GB + SHAP	Very High	Moderate

The comparison shown in Table 3.5 really highlights the balance between predictive accuracy and interpretability across various modeling techniques. Logistic Regression stands out for its high interpretability, even though its accuracy is moderate, making it a great choice for fields where transparency and the ability to trace decisions are crucial. On the other hand, Random Forest and Gradient Boosting deliver higher and even very high accuracy, respectively, but they fall short on interpretability because of their complex, ensemble-based nature. This aligns with previous research that points out how more complex models often lead to less transparency, which can undermine trust and accountability in critical decision-making systems (Lipton, 2018; Doshi-Velez & Kim, 2017). A key takeaway is that integrating SHAP with these black-box models can significantly enhance interpretability while still maintaining most of their predictive power, showcasing how effective post-hoc explainability techniques can be in closing this gap. Similar insights have been found in the literature, where hybrid methods that combine strong predictive models with solid explanation techniques are shown to strike a practical balance between performance and transparency (Guidotti et al., 2018; Lundberg & Lee, 2017). In summary, the findings suggest that explainable versions of black-box models offer a promising route for implementing high-accuracy AI systems in sensitive and regulated settings without completely sacrificing interpretability.

4. Discussion

This study dives into a side-by-side comparison of machine learning models and explainability techniques in high-stakes decision-making areas, focusing on the delicate balance between how well they predict and how easy they are to understand. The experimental findings show that ensemble-based models, especially Random Forest and Gradient Boosting, consistently outperform Logistic Regression when applied to healthcare and financial datasets. These results back up what previous research has indicated: ensemble methods excel at capturing complex, non-linear relationships in structured data, which boosts predictive accuracy. However, this edge in performance comes with a trade-off—less transparency—which can hinder their use in sensitive decision-making situations.

On the flip side, the analysis of the criminal justice dataset shows only slight performance differences among the models, hinting that adding complexity may not yield significant benefits in ethically sensitive areas. In these contexts, simpler and more interpretable models like Logistic Regression still hold great value, as they promote transparency, accountability, and fairness without a major hit to predictive performance. This underscores the idea that when choosing models in high-stakes scenarios, we should consider not just accuracy but also ethical and legal implications.

The comparative look at explainability techniques reveals some crucial trade-offs as well. LIME offers intuitive and computationally efficient local explanations, making it a good fit for case-level and exploratory analysis. However, its moderate stability can make it less reliable in regulated settings that demand consistent and reproducible explanations. On the other hand, SHAP provides both local and global interpretability with greater consistency, thanks to its game-theoretic basis. Even though it comes with a higher computational cost, SHAP is more suited for auditing, bias detection, and meeting regulatory requirements.

The combination of SHAP with top-notch black-box models proves to be a smart balance, maintaining predictive accuracy while boosting interpretability. These findings highlight how crucial explainable AI is as a fundamental necessity for implementing reliable, accountable, and ethically sound decision support systems, especially in high-stakes areas. Instead of treating explainability as just a nice-to-have feature, this study emphasizes its importance as a key design principle for the responsible adoption of AI.

5. Conclusion and Future Scope

This paper dives into a detailed and comparative exploration of explainable artificial intelligence models used in high-stakes decision-making systems. The evaluation of interpretable models, black-box machine learning models, and post-hoc explainability techniques reveals that there isn't a one-size-fits-all solution for every critical application. The findings indicate that ensemble-based black-box models, like Random Forest and Gradient Boosting, tend to outperform inherently interpretable models when it comes to predictive accuracy. However, this edge in performance often sacrifices transparency and interpretability. On the flip side, simpler models such as Logistic Regression, while a tad less accurate, provide clear and understandable decision-making logic, making them a better fit for sensitive and ethically constrained areas. The analysis of explainability techniques shows that SHAP delivers more stable, consistent, and globally coherent explanations compared to LIME. While LIME excels at providing local and user-friendly explanations, SHAP stands out with stronger theoretical foundations, making it more suitable for regulated and high-risk decision environments. Combining SHAP with black-box models presents a practical solution, allowing for high predictive performance while enhancing interpretability.

5.2 Future Scope

While this study offers some valuable insights, there are still plenty of avenues for future research to explore. For starters, researchers could look into combining intrinsically interpretable models with post-hoc explanation techniques to lessen our dependence on those black-box methods. Additionally, expanding the analysis to include deep learning models and unstructured data—like medical images or legal documents—could really enhance the applicability of the findings.

Moreover, future studies might want to dive into the quantitative assessment of explanation quality, using human-centered metrics such as user trust, cognitive load, and decision confidence. Creating domain-specific frameworks for explainability, especially in areas like healthcare, finance, and criminal justice, is another exciting direction to consider. merging explainability with fairness, robustness, and privacy-preserving techniques could help us build more trustworthy AI systems. These advancements will be essential for ensuring that AI technologies are adopted responsibly and sustainably, especially in high-stakes decision-making scenarios.

6. Limitations of the Study

While this study provides some valuable insights into explainable AI for high-stakes decision-making, it does come with a few limitations. For starters, the analysis is limited to traditional machine learning models and doesn't take deep learning approaches into account, which means it might not be as applicable in areas dealing with unstructured data. Additionally, the emphasis on post-hoc explainability techniques like LIME and SHAP might not fully reflect the internal reasoning of more complex models, since these methods tend to offer approximated rather than causal explanations. Another point to consider is that the evaluation of explainability is mostly qualitative, and the absence of standardized quantitative metrics makes it tough to compare results objectively. Moreover, relying on benchmark datasets might not capture the full complexity and ethical challenges of real-world scenarios. Lastly, the lack of human-in-the-loop evaluation hinders our ability to assess practical usability and user trust in actual deployment situations

7. Ethical Considerations

This study follows ethical guidelines for responsible research in artificial intelligence. All the datasets analyzed are publicly accessible and anonymized benchmark datasets, which means no personally identifiable information is included. Since the research doesn't involve human subjects or clinical trials, there's no need for institutional ethical approval.

To ensure transparency, fairness, and accountability in automated decision-making systems, we use explainability techniques. We pay special attention to the ethical risks that come with high-stakes areas, such as bias, discrimination, and a lack of transparency. This study is in line with established AI governance frameworks that highlight explainability as a fundamental requirement for building trustworthy AI systems.

8. Data Availability Statement

The datasets we looked at in this study are all publicly available from open-access repositories that are widely used in machine learning research. You can access these datasets for academic and research purposes, and we've made sure to cite them properly in the manuscript. Just to clarify, we didn't create any new datasets for this study.

9. Conflict of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

10. Author Contribution

The author solely conceptualized the study, designed the methodology, conducted the experiments, performed the analysis, and prepared the manuscript.

References

1. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
2. European Union. (2016). *General Data Protection Regulation (GDPR)*. Official Journal of the European Union.
3. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1–42. https://doi.org/10.1145/3236009
4. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. https://doi.org/10.1145/3233231
5. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 4765–4774.
6. Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book/
7. National Institute of Standards and Technology (NIST). (2023). *AI Risk Management Framework (AI RMF 1.0)*. NIST, U.S. Department of Commerce.
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “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*, 1135–1144. https://doi.org/10.1145/2939672.2939778
9. Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
10. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries*, 1(1).
11. Zhang, Q., & Zhu, S. C. (2018). Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1), 27–39. https://doi.org/10.1631/FITEE.1700808
12. Barredo Arrieta, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges. *Information Fusion*, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012