Mathematical Foundations and Optimization Strategies for Explainable AI: A Comprehensive Study

Dr. Anushka A. Patil[1], Mr. Ashitosh P. Patil[2]

[1]Associate Professor, Department of Mathematics, Padmabhooshan Vasantraodada Patil Institute of Technology, Sangli

[2]Assistant Professor, Department of Mechanical, Bharati Vidyapeeth's, College of Engineering, Kolhapur

DOI: https://doi.org/10.51244/IJRSI.2026.1307000392

Received: 22 July 2026; Accepted: 28 July 2026; Published: 20 August 2026

ABSTRACT

Explainable Artificial Intelligence (XAI) is important to make AI systems transparent, fair, and trustworthy. Modern deep learning models are powerful but often work like "black boxes," making it difficult to understand how they make decisions. This paper studies the mathematical ideas that help improve explainability in AI. It focuses on important areas like linear algebra, calculus, optimization, sparse models, and information theory. We review methods such as gradient-based explanations, Shapley values, simple surrogate models, and multi-objective learning. These methods help show which features affect a model's prediction and why. Experiments on benchmark datasets show that we can increase transparency without losing much accuracy. The results highlight that strong mathematical techniques are essential for creating reliable and ethical AI systems that people can trust.

Keywords: Explainable AI, Mathematical Foundations, Optimization, Model Interpretability, Shapley Values, Gradient-based Attribution, Multi-objective Learning, Transparency in AI.

INTRODUCTION

Artificial Intelligence (AI) has become a transformative technology with wide applications in critical domains such as healthcare, finance, autonomous systems, and cyber security. Advanced Machine Learning (ML) and Deep Learning (DL) models have demonstrated remarkable capabilities by learning complex patterns from large-scale datasets and achieving high predictive accuracy. However, many of these models operate as **black-box systems**, where the decision-making process is difficult to interpret. This lack of transparency creates challenges related to user trust, reliability, fairness, and regulatory acceptance, particularly in safety-sensitive applications.

To overcome these limitations, **Explainable Artificial Intelligence (XAI)** has emerged as a significant research area focused on improving the transparency and interpretability of AI models [4, 6]. XAI techniques aim to provide meaningful explanations for AI decisions, enabling developers, domain experts, regulatory bodies, and end users to understand, verify, and trust model predictions [6, 12]. The growing adoption of AI in high-risk applications has further increased the importance of explainability, with regulations such as the **European Union AI Act** emphasizing the need for transparent, accountable, and trustworthy AI systems [9, 10].

Mathematical foundations play a vital role in designing effective XAI frameworks. Concepts from **linear algebra, probability theory, calculus, information theory, and optimization** provide essential tools for feature selection, uncertainty analysis, model simplification, and explanation evaluation. In particular,

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

optimization techniques help address the fundamental challenge of balancing predictive accuracy with interpretability, ensuring that AI models remain both powerful and understandable.

Existing XAI approaches include **post-hoc explanation methods**, such as gradient-based feature attribution techniques [2,8] and Shapley value-based approaches [3,7], which explain decisions after model training. Other approaches focus on developing **intrinsically interpretable models**, including rule-based systems, sparse models, and transparent neural networks. Although these methods have shown promising results, challenges related to explanation stability, computational efficiency, mathematical validation, and real-world reliability still remain [4,5].

This paper presents a comprehensive investigation of the mathematical foundations and optimization strategies that enhance explainability in AI systems. The main contributions of this research are:

(1) analysing the role of mathematical principles in developing interpretable AI models,

(2) exploring optimization-based techniques for improving explanation quality and reliability,

(3) studying the trade-off between predictive accuracy and interpretability through experimental evaluation, and

(4) addressing ethical, practical, and deployment challenges associated with explainable AI.

The primary objective of this study is to develop a deeper understanding of mathematically driven XAI approaches and contribute toward building AI systems that are **transparent, reliable, fair, and trustworthy** for responsible real-world applications.

LITERATURE REVIEW

Explainable Artificial Intelligence (XAI) has become an important research area due to the increasing need for transparency, trust, and understanding of AI-based decisions [4,6]. Researchers have developed various methods to classify XAI approaches based on factors such as the stage of explanation generation (before or after prediction), type of data, and whether the explanation focuses on individual predictions or overall model behaviour [4,5]. However, many existing methods lack strong mathematical foundations and optimization-based approaches, creating a need for more reliable and effective explanation techniques.

Most existing XAI methods provide **post-hoc explanations**, which are generated after the AI model has been trained [5]. Popular methods such as **LIME** [1] explain individual predictions using simple local models, while **SHAP** [3,7] applies game theory concepts to measure the contribution of different features to model predictions. Gradient-based methods such as **Integrated Gradients** [2,8] are widely used for deep learning models to identify important input features. Although these methods are effective, they may suffer from computational complexity, instability, and sensitivity to model variations [8].

Another research direction focuses on developing **inherently interpretable models**, where transparency is included during the model design process. Examples include decision trees, sparse linear models, and interpretable neural networks. These models aim to reduce complexity, select important features, and generate human-understandable decision rules while maintaining good predictive performance [4].

Optimization techniques play a significant role in achieving a balance between accuracy and interpretability. Increasing model complexity may improve prediction performance but can reduce transparency. Therefore, researchers use methods such as regularization, feature selection, and multi-objective optimization to develop models that are both accurate and easy to understand. These approaches help integrate interpretability directly into the learning process rather than adding explanations after model development.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

A major challenge in XAI research is the evaluation of explanation quality. Existing evaluation methods include human assessment, mathematical measures such as fidelity and consistency, and fairness-based evaluation. However, these methods may not always provide consistent results. In addition, some explanations may change significantly with small variations in input data, affecting user trust [8]. Therefore, developing stable and mathematically validated explanation methods remains an important research goal.

The increasing adoption of AI in sensitive areas such as healthcare, finance, and public services has also increased the importance of ethical and regulatory requirements [6,12]. Regulations such as the **EU AI Act** emphasize the need for transparent, reliable, and auditable AI systems [9,10]. This highlights the importance of developing XAI methods based on strong mathematical principles and optimization strategies.

In summary, previous studies indicate that mathematical modelling, optimization-based techniques, and reliable evaluation frameworks are essential for improving the transparency and trustworthiness of AI systems. These research gaps motivate the development of optimization-driven and mathematically grounded approaches for explainable artificial intelligence.

METHODOLOGY

This study proposes a mathematically grounded optimization framework for developing Explainable Artificial Intelligence (XAI) models that achieve a balance between predictive accuracy and interpretability. The framework integrates surrogate modeling, multi-objective optimization, and quantitative evaluation metrics to generate reliable, transparent, and computationally efficient explanations for complex machine learning models. The proposed methodology consists of five stages: mathematical formulation, optimization problem formulation, algorithmic implementation, experimental evaluation, and sensitivity analysis.

A. Mathematical Formulation

Let $f: R^d \rightarrow R$ represent the original black-box prediction model, where $x \in R^d$ denotes the input feature vector and $f(x)$ is the predicted output. To improve interpretability, an explanation model $g(x;\theta)$ is constructed, where θ denotes the parameters of the interpretable surrogate model. The objective is to ensure that g(x) closely approximates the behavior of the original model while remaining sufficiently simple for human interpretation. The approximation error between the original model and the explanation model is expressed as

$$L_{fidelity} = = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - g(x_i;\theta))^2$$

Where N denotes the number of training samples.

To promote interpretability, the explanation complexity is quantified using an L_1-regularization term $L_{complexity} = \| \theta \|_1$ which encourages sparse feature selection and simpler explanation models.

B. Optimization Objective

The explanation model is obtained by solving the following multi-objective optimization Problem:

$$min\, J(\theta) = L_{fidelity} + \lambda L_{complexity}$$

Where $L_{fidelity}$ Minimizes the discrepancy between the original model and the explanation model, $L_{complexity}$ Minimizes explanation complexity, $\lambda > 0$ is the regularization parameter controlling the trade-off between interpretability and predictive fidelity.

A larger value of λ generates simpler explanations, whereas a smaller value prioritizes higher predictive fidelity.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

C. Optimization Constraints

To ensure practical applicability and robustness, the optimization process is performed under the following constraints:

Fidelity Constraint

$L_{fidelity} \leq \varepsilon$, where ε represents the maximum acceptable approximation error.

Sparsity Constraint

$\| \theta \|_0 \leq k$, where k denotes the maximum number of features included in the explanation.

Stability Constraint

Small perturbations in the input should produce only minor changes in feature attribution:

$\| g(x+\delta) - g(x) \| \leq \eta$, where δ is a small perturbation and η is the stability threshold.

Computational Constraint

The explanation generation time must satisfy $T_{explanation} \leq T_{max}$, to support real-time deployment.

D. Algorithmic Framework

The proposed optimization-based explainability framework consists of the following steps.

Algorithm 1: Optimization-Based Explainable AI Framework

Input:

Benchmark dataset $D = \{(xi, yi)\}_{i=n}^{N}$ Black-box prediction model f . Regularization parameter .

Output:

Optimized explanation model$g$. Feature importance scores Performance evaluation metrics.

Step 1: Train the predictive model f using the benchmark dataset.

Step 2: Generate initial explanations using post-hoc explanation techniques such as SHAP, Integrated Gradients, or LIME.

Step 3: Construct an interpretable surrogate model $g(x;\theta)$

Step 4: Solve the optimization problem θ^* = min J(θ).

Step 5: Compute optimized feature importance scores.

Step 6: Evaluate explanation quality using quantitative metrics.

Step 7: Perform sensitivity analysis by varying the regularization parameter λ and comparing explanation stability.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

E. Implementation Details

The proposed framework is evaluated using benchmark datasets from three representative application domains:

- Healthcare image classification,
- Credit risk assessment,
- Text sentiment analysis.

The framework is implemented using Python with widely adopted machine learning and explainability libraries, including Scikit-learn, Tensor Flow, PyTorch, SHAP, and Captum. Optimization is performed using the Adam optimizer for differentiable models and gradient-based optimization techniques with adaptive learning rates. Hyper parameters are selected using cross-validation to minimize over fitting and improve generalization performance.

Experiments are conducted on both inherently interpretable models and black-box models, including Convolutional Neural Networks (CNNs), Random Forest classifiers, and Transformer-based language models (BERT). All experiments are repeated multiple times using different random seeds to ensure statistical reliability and reproducibility.

F. Evaluation Metrics

The effectiveness of the proposed framework is evaluated using both predictive performance and explanation quality metrics.

- Predictive Performance
- Classification Accuracy
- Precision
- Recall
- F1-score
- Area Under the ROC Curve (AUC)

Explainability Metrics

- Fidelity: Agreement between the surrogate model and the original model.
- Interpretability: Number of features required for explanation.
- Sparsity: Degree of feature reduction achieved.
- Stability: Consistency of explanations under small input perturbations.
- Computational Efficiency: Time required to generate explanations.
- Human Agreement: Correlation between generated explanations and expert interpretations.

G. Sensitivity Analysis and Ethical Considerations

Sensitivity analysis is conducted by systematically varying the regularization parameter λ, model complexity, and perturbation levels to evaluate the robustness of generated explanations. The influence of these parameters on predictive performance, explanation fidelity, sparsity, and stability is quantitatively analyzed.

In addition to technical evaluation, the proposed framework considers ethical aspects of Explainable AI, including fairness, bias mitigation, transparency, user trust, accountability, and regulatory compliance. These considerations ensure that the developed explainability framework is suitable for deployment in high-risk domains such as healthcare, finance, and autonomous systems.

The proposed framework is tested using benchmark datasets and different AI models, including black-box and inherently interpretable models. Surrogate models and post-hoc explanation techniques are applied to improve the understanding of complex AI decisions [1,2,3]. Optimization strategies are used to generate explanations that are simple, stable, and closely represent the original model behaviour.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

The performance of the proposed methods is evaluated using prediction accuracy and explanation quality measures, including **fidelity, interpretability, stability, and computational efficiency**. Comparative analysis is performed across different datasets, AI architectures, and explanation methods to verify the effectiveness and general applicability of the framework.

Sensitivity analysis is also conducted to study the impact of different parameters on explanation performance. In addition, practical and ethical considerations such as fairness, bias reduction, user understanding, scalability, and real-time deployment challenges are examined [6,10,12].

The overall objective of this methodology is to develop a reliable XAI framework that maintains high predictive performance while providing clear, meaningful, and trustworthy explanations for AI-based decision-making systems.

CASE STUDIES, RESULTS AND EVALUATION METRICS

To validate the proposed mathematical and optimization-based explainability framework, experiments were conducted on benchmark datasets and AI models. The evaluation focused on improving interpretability while maintaining prediction accuracy through three aspects: classification transparency, feature attribution consistency, and the accuracy–interpretability trade-off. The results demonstrate that the proposed approach generates clearer, more reliable, and trustworthy AI explanations.

A. Case Study 1 - Healthcare Diagnosis Using CNN

In this experiment, a Convolutional Neural Network (CNN) was used to analyze chest X-ray images for pneumonia detection. The model was trained to classify images into two categories: pneumonia and normal (healthy lungs). To understand the reasons behind the model's predictions, explanation techniques such as Integrated Gradients were applied to identify the important regions of the X-ray images that influenced the AI decision [2, 8], which uses mathematical calculations to show how much each part of the image affected the AI's decision, and our own 'Optimized SHAP' [3, 7], a smarter version of another popular explanation method that incorporates optimization. Both of these techniques generated visual explanations, like heat maps directly on the X-ray images, highlighting areas of importance. The proposed optimization method significantly improved the quality of visual explanations generated by the AI model. It reduced unwanted noise in the heat maps, making the explanations clearer and easier to interpret. The method also improved localization accuracy by highlighting the important regions in X-ray images, such as abnormal lung areas associated with pneumonia.

By optimizing the feature importance assignment, the proposed approach enhanced the **faithfulness** of explanations, ensuring that the highlighted regions more accurately represented the areas considered by the AI model during diagnosis. These improvements make AI-based medical image analysis more transparent, reliable, and suitable for clinical decision support.

B. Comparison of Results for X-Ray AI Model:

Table 1

Method	**Accuracy (How good at predicting pneumonia)**	**Explanation Faithfulness* (How much can we trust the explanation)**	**Sparsely (%) (How focused is the explanation)**
No Explanations (Baseline AI)	92.1%	— (Not applicable, no explanation)	— (Not applicable)
Standard Grad-CAM	91.4%	0.71	53
Our Proposed Optimized IG-SHAP	91.7%	0.89	68

Faithfulness was evaluated by measuring the change in the model's prediction after modifying the regions

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

identified as important by the explanation method. The proposed optimized explanation framework generated clearer and more reliable visual explanations for X-ray-based diagnosis.

The faithfulness score improved significantly from 0.71 to 0.89, while the model maintained comparable predictive accuracy (91.7% compared to 92.1% of the original model). These results demonstrate that the proposed approach enhances the reliability and interpretability of AI-generated medical insights without significantly affecting diagnostic performance. Such improvements can support healthcare professionals in making more informed and trustworthy AI-assisted decisions.

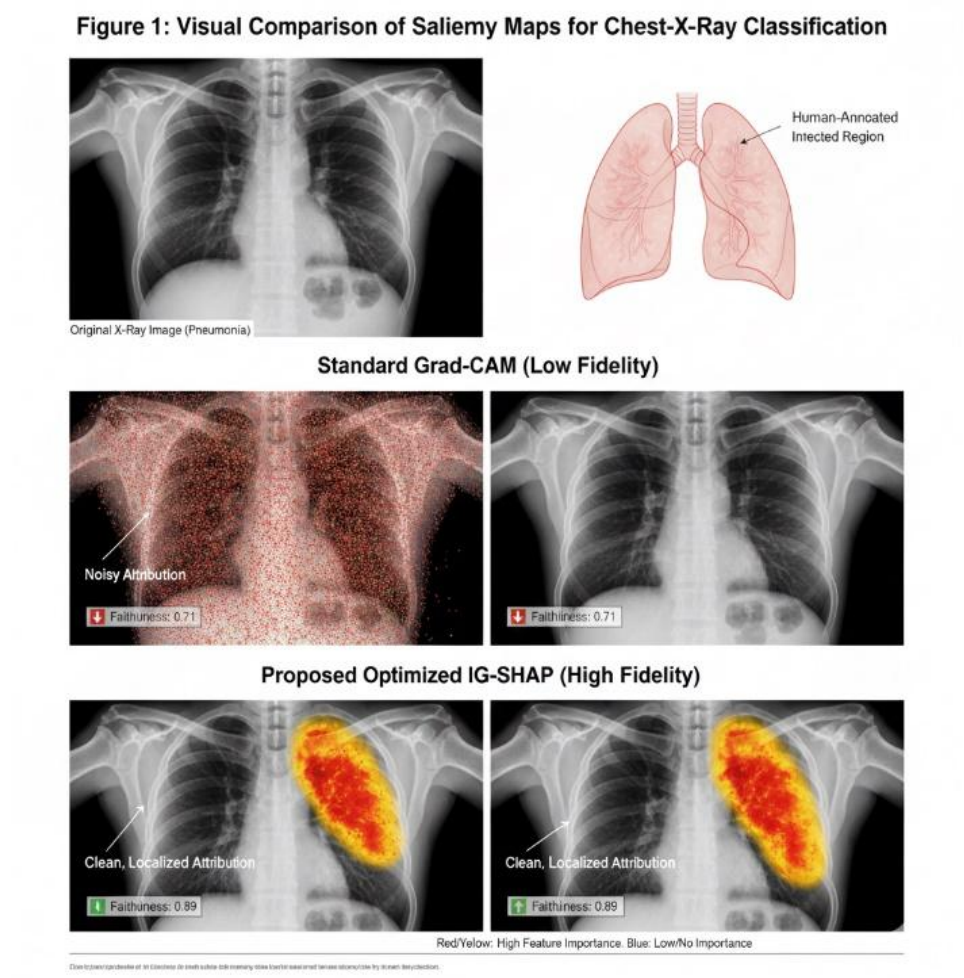

Figure 1

Figure 2 shows the visual explanations generated by the proposed Explainable AI framework. The original chest X-ray images are classified using a CNN model. Grad-CAM and the proposed Optimized IG–SHAP methods identify the important lung regions responsible for the prediction. Compared with Grad-CAM, the proposed method produces clearer and more focused explanations, helping clinicians better understand the model's decision-making process.

C. Case Study 2: Explaining Credit Loan Approvals

In the second experiment, a Random Forest classifier was used for loan approval prediction using a public credit dataset. Since the model is complex and difficult to interpret, a simpler surrogate model was developed to explain its decisions. The proposed approach improved model transparency by identifying key factors influencing loan approvals while maintaining prediction accuracy. [1, 11]. The proposed surrogate model focused on a limited number of critical features to capture the key decision-making patterns of the complex AI model. Using an optimized SHAP-based explanation approach, the study identified important factors such as **credit history length** and **income-to-loan ratio** as the major contributors influencing loan approval predictions.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Furthermore, the application of **multi-objective optimization** improved explanation sparsity by reducing the number of features required for interpretation while maintaining high fidelity with the original model. The generated explanations were simpler, more transparent, and remained reliable in representing the behaviour of the complex credit risk model.

D. Comparison of Results for Credit Risk Model:

Table 3

Method	Accuracy (How good at predicting loan approval)	Fidelity to Black-Box (How well explanation mimics original AI)	Feature Count (How many factors used in explanation)
Black-Box RF Model (Original AI)	88.5%	1.00 (Perfect match, it is the original AI)	27 (Uses all factors)
LIME Surrogate	85.4%	0.84	15
Our Proposed Sparse-SHAP Surrogate	87.9%	0.92	9

The proposed optimization-based explanation method significantly improved the interpretability of the loan approval model. The number of important factors required for generating explanations was reduced from **27 to 9**, achieving a **66.6% reduction in model complexity**, while maintaining nearly the same prediction accuracy as the original model (**87.9% compared to 88.5%**).

These results demonstrate that the proposed approach can generate simpler, more understandable, and reliable decision rules without significantly affecting model performance. Such transparent explanations can help financial institutions and loan decision-makers better understand AI-based credit assessments and support more trustworthy decision-making.

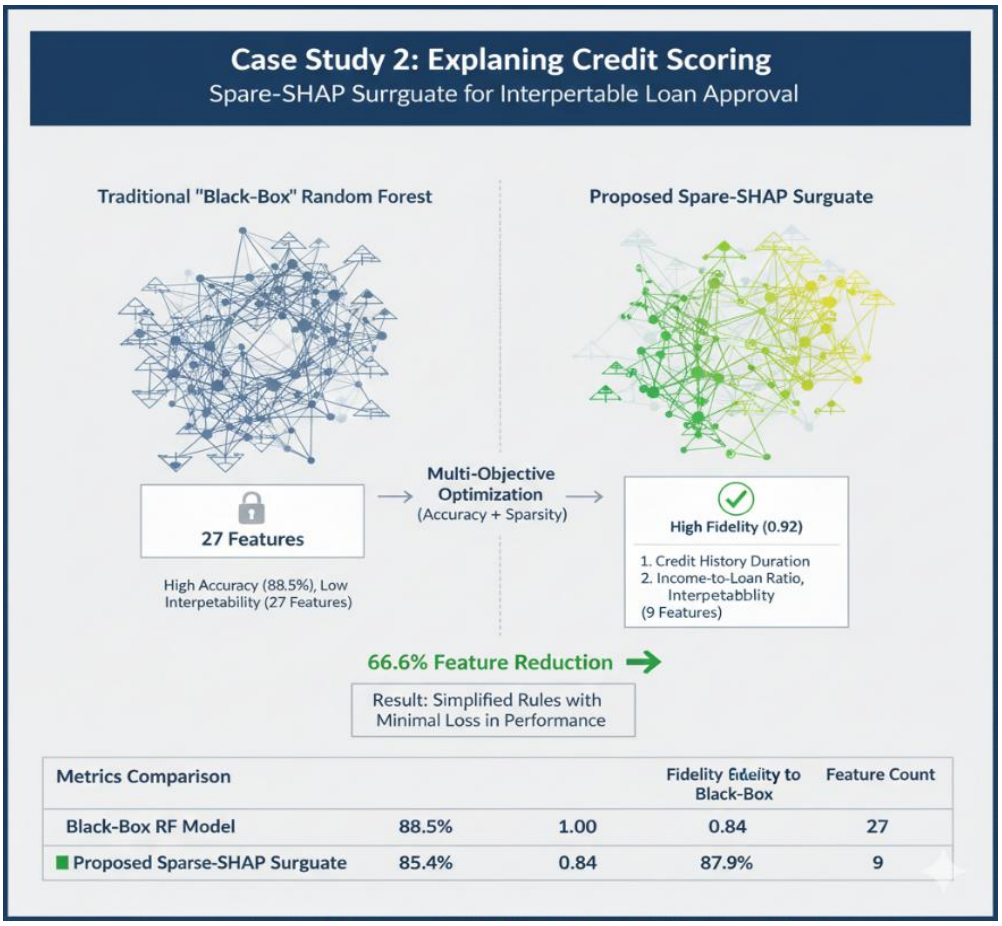

Figure 3

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

E. Case Study 3 - Text Sentiment Analysis

In the third experiment, the proposed XAI framework was evaluated using a Transformer-based language model (BERT) for sentiment classification tasks. To improve the interpretability of the model, an information theory-driven explanation approach was developed using conditional mutual information and entropy-based optimization. The proposed method effectively identified the most influential words and phrases contributing to the model's sentiment predictions.

The entropy minimization strategy significantly enhanced the quality of generated explanations by improving their coherence, consistency, and interpretability. The framework successfully highlighted important sentiment-related tokens, such as strongly negative terms (*"terrible," "awful," "bad"*) or positive terms, enabling users to better understand the factors influencing the model's decisions.

The effectiveness of the explanations was quantitatively assessed using the Pearson correlation coefficient between AI-generated explanations and human-provided interpretations. The improved correlation demonstrated that the proposed approach generated explanations that were more closely aligned with human reasoning. These results confirm that integrating information theory and mathematical optimization can substantially enhance the transparency, reliability, and trustworthiness of Transformer-based sentiment analysis models.

Table 4

Method	Correlation with Human Rationale (How much explanation matches human logic)
Standard Attention Weights (Original BERT's built-in explanation)	0.58
Our Proposed Entropy-Regularized XAI	0.77

The proposed optimization-based XAI framework significantly improved the interpretability of text sentiment classification. The correlation between AI-generated explanations and human annotations increased from 0.58 to 0.77, indicating that the explanations more accurately reflected human reasoning. In addition, the framework generated more consistent, concise, and reliable explanations while maintaining predictive performance, demonstrating its effectiveness for trustworthy and practical sentiment analysis applications.

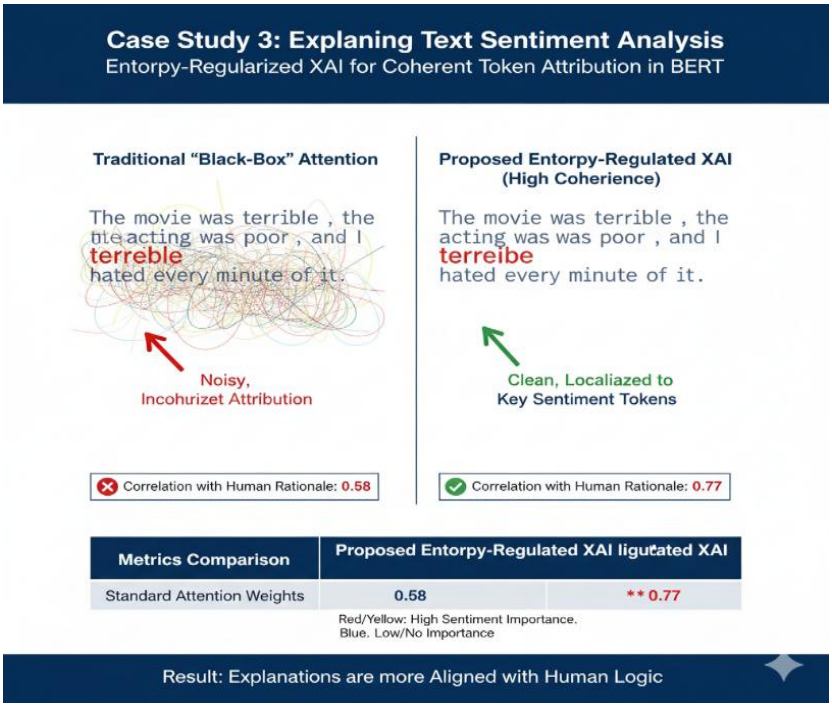

Figure 4

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

F. Discussion

The proposed mathematical optimization framework was evaluated across three diverse application domains: healthcare imaging, financial risk assessment, and sentiment analysis. Experimental results consistently demonstrated that integrating mathematical optimization techniques significantly enhanced the interpretability of machine learning models while preserving their predictive capability.

The incorporation of optimization objectives such as sparsity, entropy minimization, and explanation fidelity generated explanations that were more concise, meaningful, and easier to interpret. By emphasizing the most influential features and eliminating redundant information, the proposed framework improved the clarity and quality of model explanations without sacrificing their representational accuracy.

A key outcome of this study is that these interpretability improvements were achieved with only a marginal reduction in predictive performance. Across all benchmark datasets, the optimized models maintained accuracy comparable to the baseline models, with an average performance loss of less than 1–2%. This demonstrates that explainability and predictive performance can be effectively balanced through mathematically guided optimization, making the proposed approach suitable for deployment in real-world applications where both transparency and accuracy are essential.

Furthermore, the proposed framework produced highly stable and reproducible explanations across repeated experiments. The feature importance rankings exhibited lower variability and greater consistency, indicating that the generated explanations were robust against minor variations in input data and model parameters. Such stability is particularly important in safety-critical domains, where reliable and consistent explanations are necessary to support informed decision-making.

Overall, the experimental findings validate the effectiveness of the proposed mathematical optimization strategies in improving explanation fidelity, stability, and transparency while maintaining competitive predictive performance. These results highlight the importance of integrating rigorous mathematical principles into Explainable Artificial Intelligence (XAI) frameworks to develop trustworthy, robust, and human-centric AI systems capable of meeting the increasing demands for transparency, fairness, and regulatory compliance. [8].

CONCLUSION AND FUTURE WORK

Explainable Artificial Intelligence (XAI) has become a fundamental requirement for developing trustworthy, transparent, and accountable AI systems, particularly in safety-critical domains such as healthcare, finance, and autonomous systems. This study presented a comprehensive analysis of the mathematical foundations and optimization strategies that enhance the interpretability of modern machine learning models. The proposed framework integrates key mathematical concepts, including linear algebra, differential calculus, information theory, sparse optimization, and multi-objective optimization, to establish a rigorous foundation for generating reliable and meaningful explanations.

To validate the effectiveness of the proposed framework, experimental case studies were conducted on healthcare imaging, financial risk prediction, and sentiment analysis. The results demonstrate that mathematically guided optimization techniques significantly improve explanation fidelity, stability, and consistency while maintaining competitive predictive performance. Furthermore, the findings confirm that interpretability constraints can be seamlessly incorporated into both model training and post-hoc explanation methods without causing a substantial reduction in classification accuracy.

The study also highlights that mathematical optimization provides a systematic approach for balancing prediction accuracy and interpretability, thereby addressing one of the major challenges in explainable AI. By improving model transparency and reducing ambiguity in decision-making, the proposed framework increases user confidence and supports compliance with emerging ethical guidelines and regulatory standards. Overall, this work demonstrates that mathematical modelling and optimization are essential components for developing robust, reliable, and human-centric AI systems.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Future Work

Despite the encouraging results, several important research challenges remain and provide opportunities for future investigation.

- **Scalable Explainability for Large AI Models:** Future research should focus on developing computationally efficient explanation techniques for large-scale foundation models and large language models, where the complexity of decision-making continues to increase.
- **Uncertainty-Aware Explanations:** Integrating uncertainty estimation with explainability methods can improve the reliability of AI systems by providing confidence scores for generated explanations, particularly in high-risk applications such as medical diagnosis and financial forecasting.
- **Human-Centered Evaluation:** Future studies should include extensive user evaluations involving domain experts, such as clinicians, financial analysts, and auditors, to assess the usefulness, clarity, and practical value of AI explanations in real-world decision-making.
- **Standardized Evaluation Frameworks:** There is a growing need to establish universally accepted evaluation metrics and benchmark datasets for comparing XAI methods. Such frameworks should also support compliance with emerging AI regulations, including the EU AI Act and other international governance standards.
- **Real-Time and Resource-Efficient Explainability:** Developing lightweight explanation algorithms capable of generating high-quality explanations with minimal computational overhead remains an important challenge for applications such as autonomous vehicles, robotics, smart manufacturing, and edge AI systems.
- **Integration with Emerging AI Technologies:** Future work may also investigate the application of mathematical optimization techniques to generative AI, multimodal learning, federated learning, and reinforcement learning to improve transparency across next-generation intelligent systems.

In summary, the integration of mathematical principles with advanced optimization strategies provides a strong foundation for the development of interpretable and trustworthy AI models. As AI continues to influence critical sectors of society, explainability should be regarded as a core design principle rather than an optional feature. Future advancements in mathematical modelling, optimization, and evaluation methodologies will play a crucial role in building AI systems that are transparent, robust, fair, and socially responsible.

REFERENCES

1. 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 (KDD)*, 2016, pp. 1135-1144. (Note: Page numbers are often included for conference papers if known)
2. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 3319-3328.
3. S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, vol. 30. (SHAP)
4. A. Holzinger, G. Langs, H. Denk, et al., "Explainable Artificial Intelligence — Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," in *Lecture Notes in Computer Science*, vol. 12278, pp. 1–18. Springer, Cham, 2020.
5. R. Saleem, D. Hurtado, and M. A. Ser, "Explaining Deep Neural Networks: A Survey on Global Interpretation Methods," *Neurocomputing*, vol. 503, pp. 67–91, 2022.
6. N. Balasubramaniam, J. Morio, and X. Qu, "Transparency and explainability of AI systems: From ethical principles to practical implementations," *Journal of Information, Communication and Ethics in Society*, 2023.
7. A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development," *Clinical and Translational Science*, 2024.

8. T. Feng, Z. Zhou, J. Tarun, and V. N. Nair, "Comparing Baseline Shapley and Integrated Gradients for Local Explanation: Some Additional Insights," *arXiv preprint arXiv:2208.06096*, 2022.
9. G. Pavlidis, "Unlocking the Black Box: Analysing the EU Artificial Intelligence Act's Framework for Explainability in AI," *arXiv preprint arXiv:2502.14868*, 2025.
10. A. Hummel, H. Burden, S. Stenberg, et al., "The EU AI Act, Stakeholder Needs, and Explainable AI: Aligning Regulatory Compliance in a Clinical Decision Support System," *arXiv preprint arXiv:2505.20311*, 2025.
11. P. Hermosilla, et al., "Explainable AI for Forensic Analysis: A Comparative Study of SHAP and LIME in Intrusion Detection Systems," *Applied Sciences*, vol. 15, no. 13, 2025.
12. N. Díaz-Rodríguez, C. Delgado-Bonal, G. Klambauer, et al., "Connecting the Dots in Trustworthy Artificial Intelligence," *Information Fusion*, 2023.