International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 14, Issue 6, June 2025)

Advancing Interpretability in Deep Learning Models Using Explainable AI Techniques

Shashank Mishra
Student
Department of Computer Science Engineering
CT University, Ludhiana, Punjab
shashankmishrakc@gmail.com

Dr. Mandeep Kaur
Assistant Professor
Department of Computer Science Engineering
CT University, Ludhiana, Punjab
mandeep17209@ctuniversity.in

Aditya SG Vyas
Chief Executive Officer,
IITI DRISHTI CPS Foundation,
IIT Indore
adivyas.88@gmail.com

Abstract: Deep learning models have revolutionized artificial intelligence (AI) with their exceptional predictive accuracy. However, their black-box nature has hindered adoption in sensitive applications where transparency is crucial, such as healthcare, finance, and autonomous systems. Explainable AI (XAI) addresses this challenge by providing mechanisms to interpret and trust model outputs. This paper investigates state-of-the-art XAI techniques applied to deep learning, offering a comprehensive review of methods such as LIME, SHAP, Grad-CAM, and integrated gradients. We define the core challenges in interpretability, present a structured methodology to evaluate XAI tools across image and tabular datasets, and provide experimental results comparing their effectiveness. The results affirm that combining multiple XAI methods can provide robust and reliable insights into model decisions. This research contributes to establishing guidelines for selecting appropriate XAI techniques, thereby advancing the interpretability and trustworthiness of deep learning models.

Keywords: Explainable AI (XAI), Deep Learning, Interpretability, SHAP, LIME, Grad-CAM,CNN, MLP

1. Introduction

Deep learning (DL), a subfield of machine learning, has transformed the AI landscape by enabling systems to learn complex patterns from vast datasets. From computer vision to natural language processing, deep neural networks have surpassed traditional algorithms in both accuracy and scalability. Yet, their opaque internal workings have sparked concerns regarding trust, bias, accountability, and regulatory compliance. Interpretability refers to the degree to which a human can understand the cause of a decision made by a model. With rising adoption of AI in mission-critical sectors, interpretability is no longer a luxury but a necessity. Explainable AI (XAI) has emerged as a suite of techniques and tools that make AI decisions understandable to humans without compromising model performance.

This paper focuses on advancing interpretability in deep learning models using XAI. We provide an extensive literature review, define the core problem of black-box decision-making, explore leading XAI techniques, and evaluate their performance across real-world datasets. Our goal is to assess which methods best balance accuracy and interpretability, providing actionable insights for practitioners.

2. Literature Survey

The research community has increasingly focused on interpretability in AI, leading to diverse strategies under the umbrella of XAI. The literature on XAI can be broadly categorized into model-specific and model-agnostic methods, and further into local and global explanations.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 14, Issue 6, June 2025)

2.1 Model-Agnostic Approaches

Ribeiro et al. (2016) introduced Local Interpretable Model-Agnostic Explanations (LIME), which approximates a complex model locally using an interpretable linear model. Lundberg and Lee (2017) developed SHAP (SHapley Additive exPlanations), based on game theory, to fairly attribute output predictions to input features.

2.2 Model-Specific Approaches

Model-specific methods rely on the internal architecture of deep models. For example, Grad-CAM (Selvaraju et al., 2017) visualizes salient regions of images by leveraging gradients of convolutional layers. Integrated gradients (Sundararajan et al., 2017) assign attribution scores by integrating gradients along the path from a baseline input to the actual input.

2.3 Global vs. Local Explanations

Local explanations clarify individual predictions, while global explanations aim to summarize overall model behavior. Techniques like decision trees and surrogate models serve as global interpreters. However, trade-offs often arise between model fidelity and comprehensibility.

2.4 Applications

XAI is widely applied in healthcare (e.g., explaining deep models for disease diagnosis), finance (e.g., credit scoring), and legal domains. For example, Caruana et al. (2015) used interpretable models for pneumonia risk prediction, where transparency was essential for clinical validation.

Despite progress, challenges remain in selecting the appropriate XAI tool, ensuring explanation fidelity, and quantifying interpretability.

3. Problem Definition

While deep learning models exhibit superior predictive power, their decision-making processes are typically opaque and difficult to interpret. This lack of transparency poses several problems:

- **Trust Deficit:** Users cannot verify or understand decisions, which is critical in healthcare and law.
- **Bias and Fairness:** Without interpretability, models may propagate or amplify biases.
- **Debugging Difficulties:** It becomes challenging to identify model errors or training issues.
- **Regulatory Challenges:** Legal frameworks like GDPR demand transparency in automated decision-making.

4. Proposed Methodology

Our methodology is designed to evaluate multiple XAI techniques across two types of datasets: image-based (CIFAR-10) and tabular (UCI Heart Disease). The process includes the following steps:

4.1 Model Selection

For this study, two types of deep learning models were selected based on the nature of the datasets: a Convolutional Neural Network (CNN) for image data and a Multi-Layer Perceptron (MLP) for tabular data. CNNs are particularly well-suited for visual tasks due to their ability to automatically extract spatial hierarchies and features from images, making them ideal for experiments on the CIFAR-10 dataset. The chosen CNN architecture includes multiple convolutional layers followed by pooling and fully connected layers to classify images into ten categories. On the other hand, MLPs are powerful feedforward neural networks capable of modeling non-linear relationships in structured data, making them appropriate for the UCI Heart Disease dataset. The MLP architecture employed consists of input, hidden, and output layers using ReLU activation and dropout for regularization.

Both models were trained to high accuracy levels to ensure that the application of explainable AI techniques could be meaningfully evaluated in the context of real-world performance. These models serve as black-box baselines for interpretability enhancement.

4.2 Dataset Preparation

To evaluate the effectiveness of various explainable AI techniques in interpreting deep learning models, two widely recognized datasets were selected: CIFAR-10 for image classification and the UCI Heart Disease dataset for tabular classification. The CIFAR-10 dataset comprises 60,000 color images of size 32×32 pixels, evenly divided into 10 classes such as airplanes, cars, birds, and cats. The dataset is pre-split into 50,000 training and 10,000 test images, and standard normalization techniques were applied to scale pixel values between 0 and 1. Data augmentation, including random horizontal flipping and cropping, was used to enhance generalization during training. For the UCI Heart Disease dataset, which contains 303 records with 14 attributes (such as age, cholesterol, blood pressure, and chest pain type), missing values were handled using mean imputation, and categorical variables were encoded using one-hot encoding. Continuous features were standardized to have zero mean and unit variance. The dataset was then split into training and testing sets in a 70:30 ratio to ensure balanced evaluation. This structured preparation ensured that both datasets were clean, well-formatted, and representative of real-world conditions for accurate model training and interpretability analysis.

4.3 XAI Techniques Applied

In this research, four prominent Explainable AI (XAI) techniques were applied to interpret the predictions of deep learning models: LIME, SHAP, Grad-CAM, and Integrated Gradients. LIME (Local Interpretable Model-Agnostic Explanations) works by perturbing the input data and training an interpretable surrogate model, such as a linear regression, to approximate the behavior of the original black-box model locally around a specific prediction. This method was used to explain both image and tabular data by highlighting influential features. SHAP (SHapley Additive exPlanations), based on cooperative game theory, assigns each feature an importance value for a particular prediction using Shapley values. It ensures consistency and local accuracy, making it suitable for complex tabular models like MLPs. Grad-CAM (Gradient-weighted Class Activation Mapping) is a model-specific technique used for CNNs that visualizes important regions in input images by computing the gradients of target classes flowing into the final convolutional layers. It produces heatmaps that intuitively indicate which areas of the image the model focused on. Integrated Gradients, another model-specific technique, calculates the average gradients along the path from a baseline input to the actual input, assigning attribution scores to each feature. This method was used for both tabular and image data to provide robust feature attribution. Together, these techniques offer a diverse and comprehensive view of model interpretability across different data types and architectures.

4.4 Evaluation Metrics

To systematically assess the performance and effectiveness of the selected XAI techniques, a set of evaluation metrics was employed, focusing on four key dimensions: fidelity, comprehensibility, computational efficiency, and visual coherence. Fidelity measures how accurately an explanation reflects the true decision-making process of the original model; higher fidelity indicates that the XAI method closely approximates the model's behavior. Comprehensibility assesses how easily a human—especially a non-expert—can understand and interpret the explanations; this is crucial for real-world usability, particularly in sensitive domains like healthcare. Computational Efficiency refers to the time and resources required to generate explanations for individual predictions. Methods with lower runtime and reduced computational overhead are preferred for real-time applications. Finally, for image-based models, visual coherence evaluates the clarity, relevance, and interpretability of the generated visual explanations, such as heatmaps or saliency maps.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 14, Issue 6, June 2025)

These metrics together provide a balanced framework for comparing the strengths and limitations of each XAI technique, ensuring both technical accuracy and practical usability are considered in the evaluation.

4.5 Implementation Tools

The implementation of the deep learning models and explainable AI techniques in this study was carried out using a combination of powerful open-source libraries and frameworks in Python. For model development, TensorFlow and Keras were used to build, train, and evaluate the Convolutional Neural Network (CNN) and Multi-Layer Perceptron (MLP) architectures due to their ease of use, scalability, and extensive community support. For explainability, specialized libraries were utilized: LIME was implemented using the lime package, which supports both tabular and image data; SHAP explanations were generated using the shap library, which provides efficient computation of Shapley values for a wide range of models; Grad-CAM was implemented using custom scripts alongside TensorFlow's Keras backend to extract intermediate feature maps and gradients; and Integrated Gradients were computed using the Captum library, developed by Facebook AI, which integrates seamlessly with PyTorch models and was adapted for compatibility with TensorFlow in this research. Data preprocessing and analysis were performed using standard Python libraries such as NumPy, Pandas, and Scikit-learn, while visualization tasks leveraged Matplotlib and Seaborn. The experiments were conducted on a GPU-enabled system to accelerate training and explanation generation, ensuring efficiency and reproducibility across the evaluation pipeline.

5. Experimental Results

5.1 Image Dataset (CIFAR-10)

The CNN achieved 84% accuracy. When applying Grad-CAM, regions of interest (e.g., wings for airplanes) were accurately highlighted. Integrated Gradients produced similar saliency but with less noise. LIME explanations were more scattered, and SHAP explanations, though accurate, were computationally intensive.

Technique	Fidelity	Visual Clarity	Time (sec/sample)
Grad-CAM	High	High	0.45
Integrated Gradients	Medium	Medium	0.60
SHAP	High	Medium	2.3
LIME	Medium	Low	1.5

5.2 Tabular Dataset (UCI Heart Disease)

The MLP achieved 89% accuracy. SHAP and LIME provided meaningful explanations. SHAP's feature importance (e.g., cholesterol, blood pressure) aligned well with domain knowledge. LIME was faster but less consistent.

Technique	Fidelity	Human Trust Score*	Time (sec/sample)
SHAP	High	8.7/10	1.2
LIME	Medium	7.5/10	0.9

6 Conclusion

This paper underscores the vital role of interpretability in advancing trustworthy AI through explainable deep learning. We reviewed and compared prominent XAI methods on image and tabular data, revealing that no single technique universally excels. Grad-CAM and SHAP provide high fidelity but differ in domain applicability. LIME offers speed at the cost of stability. Our findings suggest a hybrid strategy—using multiple XAI techniques—enhances reliability and depth of explanations.

References

1. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *KDD*.
2. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *NeurIPS*.
3. Selvaraju, R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *ICCV*.
4. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *ICML*.
5. Caruana, R., et al. (2015). Intelligible Models for Healthcare. *KDD*.
6. Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint* arXiv:1702.08608.
7. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITN*.