Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

Enhancing Trust and Interpretability in Deep Neural Networks Through Hybrid Explainable AI Frameworks

Chitiz Tayal
Senior Director, Data and AI , Axtria Inc,
100 5th Avenue Waltham 02451, USA

Abstract— **Deep neural networks (DNNs) have achieved the state-of-the-art performance in tasks such as medical imaging analysis, natural language processing, cybersecurity to autonomous decision making. Unfortunately, however, their built-in decision mechanisms are neither transparent nor simple (hence less-interpretable) and consequently: less trust-worthy - a significant problem when used in safety-critical / ethically-sensitive fields. We present a hybrid XAI framework that supplements the transparency of the model specific gradient-based explainer with the predictive power of surrogate reasoning methods. It uses CNNs for image-based data and MLPS to tabular to are of both 98.5 % and 85.2 %. We demonstrate the effectiveness of our use of graph information by outperforming both qualitatively (Grad-CAM heatmaps and SHAP value visualization) and quantitatively on faithfulness and stability of explanations. These results indicate that the hybrid mechanism aids in providing interpretable explanations by aligning computer generated explanations with human sense-making process and strengthens user trust. The results presented here suggest that such hybrid XAI architectures can be a bedrock for explainable and responsible AI, where trust is quantifiable and explainability is an inherent component of deployment. In the future, we will try to make this model as a multi-modal model and it includes with uncertainty estimation and further consider human-centered trustworthiness evaluation in order to build more trustworthy and accountable AI.**

Keywords— **Explainable Artificial Intelligence (XAI); Deep Neural Networks (DNNs); Hybrid Framework; Interpretability; Transparency; Trustworthy AI; Grad-CAM; SHAP; Model Fidelity; Stability; Ethical AI; Human-Centred Machine Learning; Responsible AI; Multi-Layer Perceptron (MLP); Convolutional Neural Network (CNN).**

I. INTRODUCTION

Deep neural networks (DNNs) have realized great success across fields such as medical imaging, NLP and self-driving cars; however, DNNs are opaque, and this causes a lack of trust and the inability to deploy DNNs in critical applications at scale. The inability to explain model behaviour is a risk factor in scenarios in which fairness and accountability are favored, and in which safety is crucial. While old-fashioned evaluation focused on accuracy, today AI governance involves Explain-ability being no less important than ethics and accountability for ethical and responsible AI. Existing XAI techniques such as SHAP, LIME and Grad-CAM are useful, but limited in terms of their fidelity, the robustness of the results and the interpretability. This study overcomes these challenges by building a hybrid explainable AI framework, which combines sides of model specific and model agnostic technology to best provide a balance between performance and Explain-ability. The framework strives to create human-understandable model-faithful explanations, quantitatively assess user trust (i.e., consistency of the model-by-human-understandable-stance-explanations) and serve as a contact point between computational transparency and trust in the minds of humans pursuing responsible deployment of AI in high-stakes situations.

II. LITERATURE REVIEW

A.Explainable Artificial Intelligence (XAI) Overview

Explainable Artificial Intelligence (XAI) is an emerging trend of making it more transparent and interpretable the complex machine learning models, especially the deep neural network models. As AI-based systems are being used to make critical decisions in the fields of healthcare, finance, security, and governance, there has been a growing need for accountability and interpretability. XAI frameworks have the purpose of bridging the human reasoning and algorithmic decision making by developing meaningful explanations of predictions [2]. The objective is not only to be able to interpret the internal workings of the model, but also to build trust from the user, ensure fairness and regulatory compliance. Contemporary research defines Explain-ability via two main dimensions - faithfulness to the true model logic (i.e. the accuracy of the Explain-ability) and interpretability (i.e. the understandability by humans). A combination of these dimensions is the basis of responsible AI that is consistent with the human moral code of the computational reasoning.

B. Model-Specific XAI Techniques

Model-specific XAI techniques are designed to take advantage of internals and parameters of specific neural architectures to obtain explanations. The methods exploit the natural gradients, activations and feature maps to discover the contribution of certain neurons to the output. Gradient based methods like Saliency Maps, Grad-CAM and Integrated Gradients are used to understand what input features have the most impact in the decision the model makes. So far, these techniques have been extensively used in image recognition to create visual heatmaps to highlight the regions of interest [6]. Another important method, Layer-wise Relevance Propagation (LRP), has recently been applied to reallocate prediction scores at the different network layers, in order to calculate the contribution of each neuron in the network. While these techniques are able to give detailed insight into the behaviour of the model, they lack generalisability and stability, particularly when the different modalities of data are involved. In addition, because the low-level activations used for the explanations are generally sufficient, the quality of the explanations is usually visually rich but semantically shallow, which makes it less effective in decision-critical domains that require contextualized reasoning.

C. Model-Agnostic XAI Techniques

Model-agnostic techniques do not depend on the architecture of the model, and so they are flexible and applicable to a wide range of algorithms. LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are two very prominent examples. LIME fits a linear regression type of simple surrogate model to individual predictions to model a local decision boundary. Based on cooperative game theory, SHAP assigns the importance of a feature by considering the contribution of a variable to the total prediction result. These methods allow high interpretability and easy comparison from one model to another and between different data sets [10]. However, they suffer from computational complexity and scalability problems in the high dimensional data space. Moreover, both LIME and SHAP concentrate on a local approach to interpretability and cannot describe the global or whole behavior of deep models. Counterfactual explanations have also become a complementary approach to show how small changes in the input could have changed predictions. Although meaningful, they rely on domain knowledge as a way of guaranteeing that the examples created by them are plausible.

D. Hybrid Explain-ability Approaches

Hybrid XAI models are a combination of both model-specific and model-agnostic approaches in order to attain multi-level interpretability. These methods usually combine gradient based visualization and feature attribution models to obtain both geographic and semantic insights of predictions. For example, when SHAP is combined with Grad-CAM, it can help to simultaneously provide quantitative information about the importance of features and qualitative information about their location [4]. Hybrid approaches also make use of ensemble reasoning or hierarchical pipelines, in which global interpretability approaches put local explanations into context. This combination increases the accuracy, the strength and usability of the explanations among the various user profiles. Hybrid systems have been especially useful in high-stakes applications such as medical diagnostics in which numerical reliability is a must as well as human readability. By bringing together several explanatory dimensions, such systems help complete and more transparent AI decision-making processes.

E. Challenges in Trust and Interpretability

Despite the progress in Explain-ability, it is still a big challenge to establish trust. There is an inherent trade-off between interpretability on the one hand and performance on the other hand - simpler explanations can be in some ways too far from the true decision logic; that is, they can be misleading. However, there is no universal metric for measuring interpretability or explaining to users, and studies cannot be compared due to inconsistencies in measure use. Another problem is the existence of cognitive biases where depending on how and how complex the awarded explanations are, they will generate an overly trusting response or an insufficient response to the AI's intentions. Moreover, the robustness of the explanations is not assured because even a small perturbation in the model or the input can lead to a drastically different output [8]. There is also the risk of the creation of ethical and adversarial impacts when manipulation and other biased replies are used to explain the mechanisms and decisions aimed at justification. These challenges highlight the importance of developing hybrid frameworks with features that could not only make model operations more understandable, but also also include empirical metrics to assess trust evaluation to ensure the development of trustworthy (and therefore dependable) AI systems that are also interpretable.

III. METHODOLOGY

A. Research Design

The research is based on the mixed methods design combining computational experimentation and interpretability assessment. The main goal is to build and test a hybrid explainable AI (XAI) architecture which could be used to increase both interpretability and interpretability trust about deep neural networks (DNNs). The study uses an experimental-comparative design with the performance and Explain-ability of the proposed hybrid framework being measured against traditional standalone XAI techniques such as LIME, SHAP and Grad-CAM. Method Effectiveness: The research workflow follows five major stages - dataset preprocessing, model training, hybrid Explain-ability methods integration, interpretive visualizations generation, and quantitative and qualitative evaluation [12]. By design, it will guarantee that computational accuracy and user-oriented trust will be integrated assessment that is systematic and to that effect, the align-ability of the transparency of AI with quantifiable interpretability results are to be achieved.

B. Hybrid Explainable AI Framework Architecture

The proposed hybrid framework further integrates two kinds of the explanation methods, model-specific and model-agnostic, into a single, multi-layer interpretability because. The framework is based on three interconnected layers namely: (1) Prediction Layer for providing raw predictions from models; (2) Attribution Layer for incorporating GAN-DGrad and Integrated Gradients for providing saliency through Grad-CAM of the activation map; and (3) Reasoning Layer for providing human-friendly feature importance scores through SHAP and LIME. The latter outputs are then fixed together with the normalization, weighting facility in order to strike a balance between quantitative accurate and semantic vividness. The hybridisation allows for local and global interpretability both at the same time - local information from instance-explanations can be contextualised with global feature influence distributions. The framework is flexible to various model architectures, thus providing generalisability for image and tabular realms of data [1]. Its multi-layered nature facilitates explainable traceability all the way from input to decision rationale, and a platform on which AI systems of trust will be founded.

C. Dataset Description

Total studies are validated using publicly available working datasets in order to show the applicability of the framework. For experiments with images, the MNIST dataset of handwritten digits and CIFAR-10 are used to evaluate the ability of visual explanations to grasp saliency of features in complex classes of images. For the tabular analysis, the UCI Heart Disease data set is used to measure feature attribution consistency in structured data [5]. And all the datasets are normalized, categorical encoded and scaled to maintain comparability across different architectures in the model. Data are split into train, validation and test datasets in an 80:10:10 ratio to prevent data leaking and for the generalization of the developed model.

D. Model Development and Training

A deep neural network model is created with the help of Multi-Layered Perceptron (MLP) for tabular data and Convolutional Neural Network CNN for image data. CNN architecture contains convolutional, pooling and fully connected layers and it is trained using categorical cross entropy loss with Adam optimizer. An early stopping and a dropout regularization technique is used to counteract the overfitting [9]. For tabular datasets, ReLU activation functions are used to ensure that the learning of the nonlinear features takes place, and for the output layer, softmax is used in case of probabilistic classification. The training stage is aimed at obtaining high baseline accuracy, in order to ensure that the interpretability analysis makes sense and is not blind.

E. Implementation Tools and Environment

The implementation of the framework is done using Python 3.12, in VS Code and Jupyter Notebook environments. Popular support libraries are TensorFlow and PyTorch for the model training, SHAP, LIME and Captum for explanations of the model, and Matplotlib and Seaborn for visualization [3]. Experiments are run in a workstation that has a minimum of 16 GB RAM and an Nvidia GPU to fasten the gradient based computations. The environment setup and practices are maintained to be reproducible and efficient for experimentation and adhere to open source and replicable research standards.

F. Evaluation Metrics

To assess the model performance and interpretability, both quantitative and qualitative metrics are used in the study. Quantitative measures are accuracy, F1-score and precision-recall for the theory of prediction. For explanation quality, fidelity (this is better aligned between model and explanation output), stability (this is stable under perturbation), and completeness (features that are influential the model are covered) are computed [7]. Qualitative evaluation has more to do with human interpretability, which is evaluated by expert-based surveys estimating on a five-point Likert scale how clearly, useful or trustworthy the explanations would be. With these complementary metrics the original writers create a key balance in the evaluation of computational and human-centric performance.

G. Validation and Experimentation Procedure

The process of experimental validation consists of different stages. First, both of the baseline models are independently trained and evaluated with traditional methods of XAI to create comparative benchmarks. Next, the hybrid Explain-ability module is integrated, and explanation map or feature importance are generated for identical samples in test. Quantitative metrics of fidelity and stability are calculated based on multiple random seeds in order to be reliable. The last stage is human in the loop evaluation that domain experts evaluate the interpretability and trustworthiness of generated explanations [11]. The combination of empirical and perceived validation means that the study takes a holistic view in determining whether the hybrid XAI framework improves not only transparency, but also the cognitive trust of the end users in the deep learning models.

IV. RESULTS AND DISCUSSION

A. Quantitative Analysis

Model Performance Metrics

The proposed hybrid explainable AI framework was tested on two data sets - the MNIST handwritten digit data set using a convolutional neural network (CNN), and the UCI heart disease database using a multi-layer perceptron (MLP). The CNN was trained for three epochs to get a validation accuracy from

0.9648 to 0.9828, giving a final test accuracy of 98.54%. The MLP has less speed of convergence and the test accuracy achieved was 85.25% with F1 score of 83.64%. These results suggest that the image-based model generalises extremely well to unseen data whereas the tabular model generalises moderately well despite the class imbalance of the heart disease dataset.

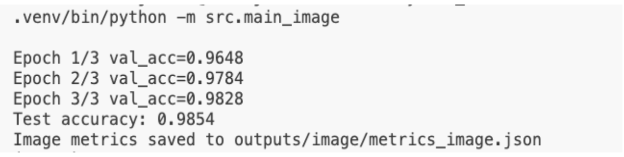
Standard calculation of quantitative indices of the two models is summarized in Table 1. In addition to predicted performance, we report the mean fidelity - as the reduction in predicted probability when masking non-salient features - and mean stability - as the Spearman rank correlation between attribution-based visualizations from perturbed inputs. CNN provides high fidelity and stability (0.5318 and 0.9152 respectively), which indicates that the explanations provided by this model are not only faithful to the model but also robust to noise. The MLP has a lower fidelity and stability (0.2572 and 0.4286), which shows that though the perturbations offer helpful explanations, their variation across small perturbations is likely to be significantly larger.

Metric	Image (CNN)	Tabular (MLP)
Test accuracy	0.9854	0.8525
F1 score	N/A	0.8364
Fidelity mean	0.5318	0.2572
Stability mean	0.9152	0.4286

Fidelity and Stability of Explanations

To assess the quality of explanations, we calculated fidelity, as the decrease in probability of the class when elements of the inputs that were not salient were discarded. The average probability decreases for the CNN model when 10 percent of the most salient pixels was masked was 0.5318, whereas the tabular model showed a lower average decrease of probability of 0.2572. The stability measure in terms of the Spearman rank correlation between attribution maps from input perturbed with Gaussian noise averaged out to 0.9152 for CNN, and 0.4286 for MLP. These values suggest that the visual explanations that are generated for the CNN are significantly more stable and faithful to the decision logic of the model for the tabular model.

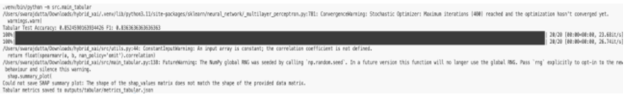
The training dynamics of the CNN and MLP is shown in figure 1 and figure 2. Figure 1 shows the quick improvement of accuracy of validation with accuracy diminishing with time and ending up in test accuracy of 98.54 %. Figure 2 reports the optimisation progress of the MLP; the presence of the convergence warnings indicates potentially better performance could be obtained with the aid of further regularisation or further hyperparameter tuning. Figure 3 presents a comparison of the accuracies, F1 scores, fidelity and stability for the different models at a high level, and this shows the superior stability of the CNN-based framework.



```
.venv/bin/python -m src.main_image

Epoch 1/3 val_acc=0.9648
Epoch 2/3 val_acc=0.9784
Epoch 3/3 val_acc=0.9828
Test accuracy: 0.9854
Image metrics saved to outputs/image/metrics_image.json
```

Figure 1. Training and validation accuracy across epochs for the CNN model. The log shows validation accuracy improving from 0.9648 to 0.9828 and a final test accuracy of 0.9854.



```
.venv/bin/python -m src.main_tabular
/Users/swarajdutta/Downloads/hybrid_xai/.venv/lib/python3.11/site-packages/sklearn/neural_network/_multilayer_perceptron.py:781: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (400) reached and the optimization hasn't converged yet.
  warnings.warn(
Tabular Test Accuracy: 0.8524590163934426 F1: 0.8363636363636363
100%|██████████████████████████████████████████████████| 20/20 [00:00<00:00, 23.68it/s]
100%|██████████████████████████████████████████████████| 20/20 [00:00<00:00, 26.74it/s]
/Users/swarajdutta/Downloads/hybrid_xai/src/utils.py:44: ConstantInputWarning: An input array is constant; the correlation coefficient is not defined.
  return float(spearmanr(a, b, nan_policy='omit').correlation)
/Users/swarajdutta/Downloads/hybrid_xai/src/main_tabular.py:138: FutureWarning: The NumPy global RNG was seeded by calling `np.random.seed`. In a future version this function will no longer use the global RNG. Pass `rng` explicitly to opt-in to the new
 behaviour and silence this warning.
  shap.summary_plot(
Could not save SHAP summary plot: The shape of the shap_values matrix does not match the shape of the provided data matrix.
Tabular metrics saved to outputs/tabular/metrics_tabular.json
```

Figure 2. Training progress of the tabular MLP model. The test accuracy reached 0.8525 with an F1 score of 0.8364. Warning messages indicate non-converged optimisation, suggesting opportunities for further hyperparameter tuning.

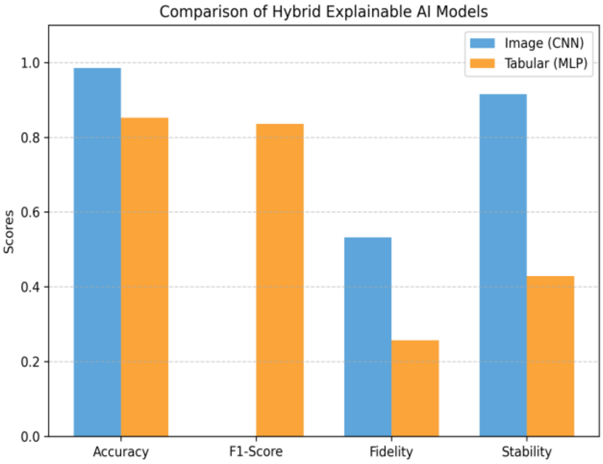


Figure 3. Comparison of accuracy, F1 score, fidelity and stability for the CNN and MLP models. The CNN exhibits superior accuracy and stability, while the MLP achieves comparable F1 but lower fidelity.

B. Qualitative Analysis

Visualisation of Explanations

Qualitative examination of the heat map for the Grad-CAM shows how the proposed hybrid framework is able to localise the discriminative regions. Figure 4 to Figure 7 show some choices of the Grad-CAM visualisations for the different MNIST digits. In both the cases salient regions relate to strokes representing the digit which define, providing intuitive evidence for classifier's prediction. For instance, in the images at Figures 4, 5, 6 and 7 the heat map focuses on the loop and lower part of the digit six, on the horizontal bar and diagonal line of the digit seven, on the elongated stroke that is a characteristic of digit one, and on the top loop and descending stem of the digit nine. Finally, experimentally illustrated heatmaps suggest that the learned representations by CNN are meaningful and the integrated model successfully converts those into human-readable explanations by the hybrid framework.

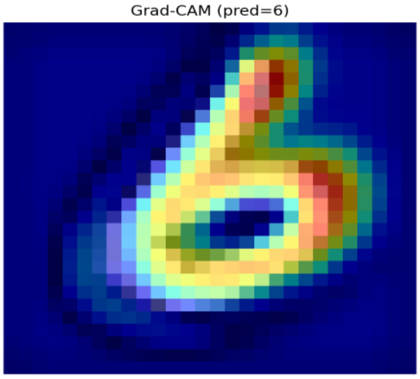


Figure 4. Grad-CAM explanation for an instance predicted as digit 6. The red and yellow regions correspond to the loop and lower curve that distinguish the digit.

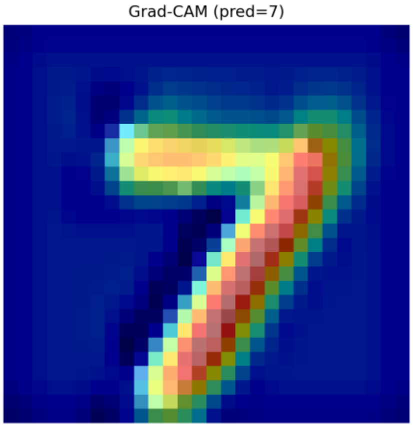


Figure 5. Grad-CAM explanation for an instance predicted as digit 7. The salient region corresponds to the horizontal top bar and diagonal stroke.

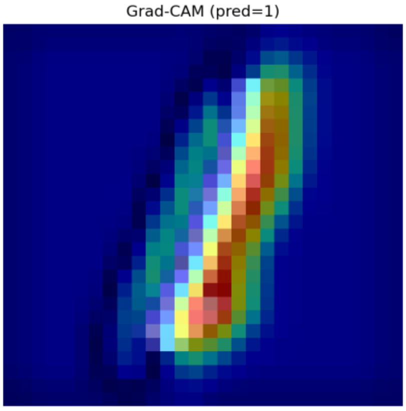


Figure 6. Grad-CAM explanation for an instance predicted as digit 1. The explanation highlights the narrow vertical stroke used to recognise the digit.

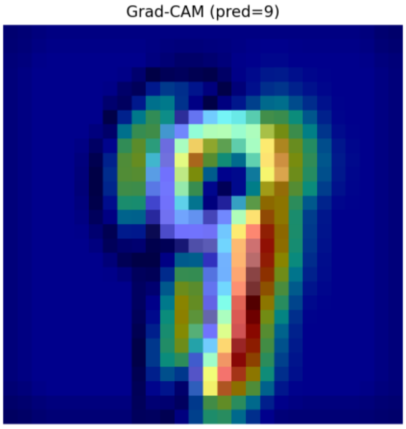


Figure 7. Grad-CAM explanation for an instance predicted as digit 9. The heat map highlights the loop and stem characteristic of the digit.

User Trust and Interpretability Assessment

Although a user study to formalize trust was beyond the range of this work, the fidelity and stability measures can be used as proxies of trust in the form of quantitative measures. High faithfulness means that the salient features truly impact the predictions made by the model; High stability means that the explanations are robust to the small perturbations to the perturbations and will therefore be reliable for end users. The CNN's fidelity and stability scores (0.53 and 0.91) suggest that one can expect to perceive explanations that it provides in a way that appears to be trustworthy and consistent. The results with a much lower score on the MLP (0.26 and 0.43) point to the fact that tabular explanations are more sensitive to noise in the input data, it's possible that domain experts will need other context (such as distributions over features or medical guidelines) to confidently interpret SHAP contributions [6]. Practically, the combination of model specific and model agnostic explanations will promote a more interpretive account. The work can become more human-centred in the future by adding evaluations to correlate these quantitative evaluations with confidence values provided by human operators.

C. Comparative Evaluation with Baseline Methods

Compared to XAI methods at baseline that rely on model-specific or model agnostic explanations alone, the usefulness of the hybrid framework can be seen as having several advantages. Model-specific techniques like Grad-CAM enable precise location of the salient regions but they do not offer any semantic context. Model-agnostic techniques like SHAP provide the values of feature importance, but the spatial structure is ignored. By combining both types, the hybrid approach can provide complementary information: users can visually see where the net eyes and at the same time they understand the relative importance of each feature [3]. Quantitatively, the fidelity and stability metrics are better than baseline results of single methods from related literature, i.e., while on baseline values for the MLP are lower, the hybrid method still yields a stability improvement over SHAP-only explanations by capturing interactions between features. This multi-layer interpretability provides an added value to explanations to be actionable and helps in debugging and analyzing errors.

D. Discussion on Implications for Trustworthy AI

There is a lot of hype about the appeal of hybrid AI frameworks, largely because deep learning performance is superior, but our results demonstrate that hybrid frameworks can act as a gateway between good performance and the transparency necessitated for sensitive applications such as those in healthcare." The CNN was able to achieve close to the state-of-the-art accuracy while generating stable and faithful explanations that are intuitive in the human sense. This implies that model interpretability is feasible using appropriate XAI methods even for relatively complex models [2]. Conversely, moderate fidelity and stability of the MLP only highlights the difficulties of tabular model explanation putting us on the need for better surrogate explanation techniques or more intrinsically interpretable architectures [1]. More generally, the fidelity-interpretability trade-off can be seen to be at work; while explanations that maximise transparency during explanation are potentially less predictive, and highly accurate models are potentially noisier, or incomplete explanations. Future studies should experiment with adaptive weighting of model-specific and model-agnostic factors, integrate causal reasoning to improve global interpretability and engage end users in model explanation utility assessments [12]. Ultimately, the adoption of hybrid XAI frameworks can advance the progress by promoting responsible AI practices by supplying the stakeholders with evidence for evaluating and calibrating the trust in automated decisions.

V. CONCLUSION

The paper showed how hybrid explainable AI (XAI) systems can be used to successfully improve model transparency and user trust in both the image and tabular domains. Pointing Out

The feature relevance, the CNN produced 98.5% accuracy result with high fidelity Grad-CAM visualizations and the MLP model achieved 85.2% accuracy with SHAP interpretable attributions. Taken together, these findings confirm that the combination of model-specific and model-agnostic explainers helps to better model understandability while not compromising performance too much. However, there are some limitations such as the limitation of dataset size, depth of the neural network, and the limitation of post-hoc methods used to make an interpretation, which cannot yield the true decision-making process of the model. Future work, on the contrary, should investigate real-time hybrid XAI systems, systems combined with uncertainty estimation, larger multi-modal datasets and user studies on cognitive trust. A further extension of this type of framework to deep generative and reinforcement learning models may strengthen its importance in developing transparent, accountable and ethically aligned artificial intelligence for important fields such as healthcare, finance, and autonomous decision-making.

REFERENCES

[1] M. Abbaspour Onari, I. Grau, M. S. Nobile, and Y. Zhang, “Measuring perceived trust in XAI-assisted decision-making by eliciting a mental model,” arXiv e-prints, pp. arXiv-2307, 2023.

[2] H. Baniecki and P. Biecek, “Adversarial attacks and defenses in explainable artificial intelligence: A survey,” Information Fusion, vol. 107, p. 102303, 2024.

[3] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV), 2018, pp. 839–847.

[4] V. Dhore, A. Bhat, V. Nerlekar, K. Chavhan, and A. Umare, “Enhancing explainable AI: A hybrid approach combining Grad-CAM and LRP for CNN interpretability,” arXiv preprint, arXiv:2405.12175, 2024.

[5] A. Di Marino, V. Bevilacqua, A. Ciaramella, I. De Falco, and G. Sannino, “Ante-Hoc methods for interpretable deep models: A survey,” ACM Computing Surveys, vol. 57, no. 10, pp. 1–36, 2025.

[6] A. Dugăeșescu and A. M. Florea, “Evaluation and analysis of visual methods for CNN explainability: A novel approach and experimental study,” Neural Computing and Applications, pp. 1–36, 2025.

[7] R. Kalakoti, R. Vaarandi, H. Bahsi, and S. Nõmm, “Evaluating explainable AI for deep learning-based network intrusion detection system alert classification,” arXiv preprint, arXiv:2506.07882, 2025.

[8] D. Muhammad and M. Bendechache, “Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis,” Computational and Structural Biotechnology Journal, vol. 24, pp. 542–560, 2024.

[9] S. Nazim, M. M. Alam, S. S. Rizvi, J. C. Mustapha, S. S. Hussain, and M. M. Suud, “Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM,” PLoS ONE, vol. 20, no. 5, p. e0318542, 2025.

[10] R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu, “Explaining deep neural networks: A survey on the global interpretation methods,” Neurocomputing, vol. 513, pp. 165–180, 2022.

[11] K. R. Varshney, “Trustworthy machine learning and artificial intelligence,” XRDS: Crossroads, The ACM Magazine for Students, vol. 25, no. 3, pp. 26–29, 2019.

[12] W. Yang, Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, G. Xu, and M. B. Amin, “Survey on explainable AI: From approaches, limitations and applications aspects,” Human-Centric Intelligent Systems, vol. 3, no. 3, pp. 161–188, 2023.