Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.DOI*

# XAI FRAMEWORK FOR COMPLETE UNDERSTANDING OF IMAGE CLASSIFICATION MODELS

**FAIZAH MAHENDI NAWAZ KURESHI[1], DR.RADHIKA SELVAMANI B[2],**
[1]Vellore Institute of Technology, Chennai, Tamil Nadu 600127, India (e-mail: faizah.kureshi2020@vitstudent.ac.in)
[2]Vellore Institute of Technology, Chennai, Tamil Nadu 600127, India(e-mail: radhika.selavamani@vit.ac.in)

Corresponding author: Dr. Radhika Selavamani (e-mail: radhika.selavamani@vit.ac.in).

**ABSTRACT** In recent times, Artificial Intelligence (AI) systems have made significant progress and have been integrated into various sectors. This has transformed industries and revolutionized human experiences. However, the opacity and lack of interpretability of these systems have raised concerns as they are increasingly utilized in critical decision-making processes. The deep learning models are often known as "black boxes," making it difficult to understand the rationale behind their predictions and decisions. To address these challenges, the field of Explainable Artificial Intelligence (XAI) has emerged as a crucial area of research. It aims to demystify the inner workings of AI systems and enhance their transparency and interpretability. Even though there are many XAI techniques but none have emerged as complete solutions. There is a lack of use of XAI techniques in the industry due to no clear framework or blueprint available. We aim to provide a framework that uses one of the best XAI algorithms. We use Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), SHapley Additive exPlanations (SHAP)for this purpose. The proposed framework can be used for any Image Classification detection model. These techniques as independent algorithms provide interpretability of the model to come extent but is unable to unveil the black box nature of AI models. Although when used together provide a complete understanding of any model. With the help of our proposed framework, one can not only get a deeper understanding of the model's behaviour but also valuable insight to make appropriate adjustments in situations such as underfitting or overfitting. Our method promotes openness, accountability, and trust in AI systems, which are crucial for responsible and ethical AI deployment in various industries and decision-making processes. Our study not only enhances the explainaibiltiy of AI models but also offers insights on how to improve and optimise these models in real-world scenarios.

**INDEX TERMS** Explainable Artificial Intelligence (XAI), SHapley Additive exPlanations (SHAP), Layer-wise Relevance Propagation (LRP), Gradient-weighted Class Activation Mapping (Grad-CAM), Machine Learning (ML), Artificial Intelligence (AI), Deep Learning (DL), Visual Geometry Group 16 (VGG16)

## I. INTRODUCTION

THE inescapable impact of Artificial Intelligence (AI) has surpassed industry limits, infiltrating every aspect of our modern lives. AI has revolutionized various industries, including healthcare, finance, transportation, and entertainment, by introducing unprecedented levels of efficiency and innovation. AI-powered diagnostic technologies in healthcare provide professionals with fast and precise information, transforming patient care. In the field of finance, artificial intelligence algorithms are utilized to examine extensive datasets in order to identify patterns and trends. This information is then used to make financial decisions with an unparalleled level of accuracy.

In addition to professional fields, AI has effortlessly assimilated into our everyday lives, augmenting convenience and customization. Smart assistants optimize household administration, while recommendation systems generate personalized material, enhancing our enjoyment encounters. Furthermore, navigation apps enhanced by artificial intelligence optimize travel routes, hence reducing commuting times and enhancing overall accessibility. The ubiquity of AI has not only transformed several sectors but also significantly influ-

enced our everyday existence, reconfiguring our methods of labor, communication, and engagement with the surrounding environment.

Nevertheless, in the middle of this advancement, the lack of transparency in Neural Network models poses a significant obstacle. The black box paradigm is the reason why many AI models fail. The absence of explainability impedes their acceptance, especially in sensitive industries, where comprehending the reasoning behind classifications and predictions is crucial.

In order to fill this void, we suggest a thorough architecture specifically designed for comprehending image categorization issues. Our approach combines three well-established techniques, working together to provide a comprehensive understanding of model decisions. Grad-CAM (Gradient-weighted Class Activation Mapping) first detects important areas in images with the help of the last convolutional layer's gradient , however, it does not give information about the specific pixels that are influencing judgments.

This is when Layer-wise Relevance Propagation (LRP) comes into play. LRP shows which parts of the input mattered most for its decision. It backpropagates through AI model's layer and gives an importance score. Hence, high scores highlight the area that was more crucial for the decision making process.

In order to enhance our comprehension of the model, we incorporate SHAP (SHapley Additive exPlanations), which assesses the importance and impact of features using positive and negative values. With this approach in mind we have performed two experiments on two different datasets– Brain MRI and ImageNet, which will in turn provide a wide range of examples and explanations. The major contribution of the proposed work is to build a system that consists of:

1. Build a Black box model for two different datasets - Brain MRI and ImageNet, which will in turn provide a wide range of examples and explanations. The block box model will be a VGG16 model with fine-tuned parameters.

2. XAI framework to explain the inner working of black box model: A class which consists of Grad-CAM, LRP and SHAP implementation

3. Comparison of the output of our XAI framework with other XAI techniques.

## II. RELATED WORKS

Explainable AI (XAI) techniques have gained significant attention in recent years as a means to make complex machine learning models more transparent and interpretable. Several studies have explored the use of different XAI methods to provide visual explanations and insights into the decision-making process of deep neural networks.

One prominent technique is Grad-CAM (Gradient-weighted Class Activation Mapping), proposed by Selvaraju et al. [1]. Grad-CAM utilizes the gradients of the target concept flowing into the final convolutional layer to produce a coarse localization map, highlighting the important regions in the image that contribute to the model's prediction. This technique has been widely used to create high-resolution, class-discriminative visualizations for CNN-based models.

Another approach, Layer-wise Relevance Propagation (LRP), was introduced by Binder et al. [2]. LRP provides a pixel-wise decomposition of the model's output, enabling the identification of the most relevant features for a particular prediction. The authors proposed an extension to the LRP framework for neural networks with local renormalization layers, commonly found in convolutional neural networks. Their experiments showed that using the first-order Taylor expansion in normalization layers improved the heatmap quality, indicating its effectiveness in dealing with non-linear neuron layers.

The Shapley Additive Explanations (SHAP) framework, presented by Lundberg et al. [3], offers a unified approach to interpreting model predictions. SHAP assigns importance values to each feature, quantifying their contribution to the model's output. This method is based on game theory and provides a unique solution with desirable properties, making it a powerful tool for understanding the inner workings of complex models. Several studies have combined these XAI techniques to provide a more comprehensive understanding of image classification models. For example, Vili et al. [4] proposed an AI-driven decision support system for COVID-19 diagnosis that used Grad-CAM and Guided Grad-CAM to localize regions in CT scans that contributed significantly to the model's predictions. Similarly, Mahmud et al. [5] developed a deep learning-based model for lung abnormality detection and classification, incorporating LIME, SHAP, and Grad-CAM to ensure interpretability.

The integration of multiple XAI methods has been shown to offer a more complete understanding of model behavior. By using Grad-CAM, LRP, and SHAP in combination, researchers can gain insights into the important features, salient regions, and overall decision-making process of the model, enabling a thorough interpretation of the classification results. Additionally, the paper "Explainable AI in Medical Imaging: An Overview for Clinical Practitioners - Beyond Saliency-based XAI Approaches" discusses more advanced XAI methods, such as case-based explanations, textual explanations, and auxiliary explanations, which provide additional information to help understand AI outputs [6]. The paper "An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability" introduces the Degree of Explainability (DoX) metric, which objectively measures the explainability of textual information within XAI systems [7]. This metric can be used to evaluate the legal compliance and quality of explanations in healthcare and finance XAI-based software systems.

Furthermore, the study "Assessing the Relevance of Mental Health Factors in Fibromyalgia Severity: A Data-Driven Case Study using Explainable AI" demonstrates the application of XAI techniques, such as SHAP and PDP, to understand the contribution of various features, including mental health factors, in predicting the severity of fibromyalgia [8].

In summary, the integration of Grad-CAM, LRP, and

SHAP provides a comprehensive framework for understanding the decision-making process of image classification models. This approach, combined with other advanced XAI methods and objective evaluation metrics, can enhance the transparency and trustworthiness of AI systems, particularly in sensitive domains like healthcare.

## III. METHODOLOGY

The proposed research work involves the application of an Explainable AI (XAI) framework to interpret deep learning models trained on two different datasets: ImageNet and brain MRI for brain tumor classification. The proposed Framkework can be seen in Fig1.


**FIGURE 1.** Architecture.

### A. DATASET

The ImageNet dataset is a large-scale image dataset widely used for object recognition tasks. The dataset consists of over 14 million images belonging to 1,000 different classes [9]. The brain MRI dataset used in this study contains 4 classes of brain tumors: glioma tumor, meningioma tumor, no tumor, and pituitary tumor. The dataset consists of brain MRI scans from patients with these various brain conditions.

### B. MODEL TRAINING

For both the ImageNet and brain MRI datasets, a pre-trained VGG-16 model was used as the base architecture [**?**]. The VGG-16 model is a well-known convolutional neural network (CNN) that has been widely used for various image classification tasks.

### C. XAI FRAMEWORK

To provide interpretable and explainable insights into the decision-making process of the VGG-16 models, the proposed XAI framework integrates three prominent techniques: Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), and Shapley Additive Explanations (SHAP).

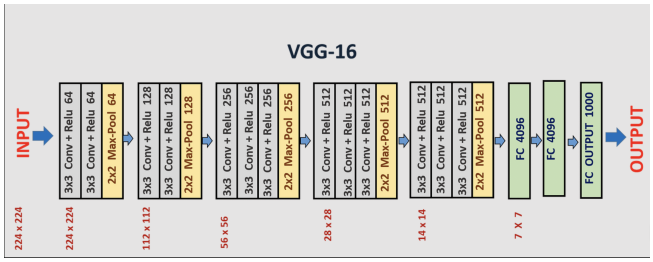


**FIGURE 2.** VGG16 Architectures.

Grad-CAM is a visual explanation technique that highlights the important regions in an input image that contribute to the model's prediction. It works by leveraging the gradients flowing into the final convolutional layer of the deep neural network. Grad-CAM is able to generate a coarse localization map, which indicates the salient regions in the image that are crucial for the model's decision-making process.

However, while Grad-CAM can identify the important areas in the image, it does not provide information about the specific pixels that are influencing the model's judgment. This is where Layer-wise Relevance Propagation (LRP) comes into play.

Layer-wise Relevance Propagation (LRP) LRP is a technique that provides a pixel-wise decomposition of the model's output, enabling the identification of the most relevant features for a particular prediction. LRP works by propagating the relevance (or importance) of the output back through the network layers, using a set of propagation rules that ensure the conservation of relevance.

By applying LRP, we can obtain a detailed map that highlights the specific pixels in the input image that were most influential in the model's decision-making process. The high-relevance pixels indicate the areas that were crucial for the model's prediction, providing a more granular understanding of the decision-making mechanism.

Shapley Additive Explanations (SHAP) To further enhance our comprehension of the model's behavior, we incorporate Shapley Additive Explanations (SHAP), a game-theoretic approach to interpreting model predictions. SHAP assigns an importance value (SHAP value) to each feature, quantifying its contribution to the model's output.

The SHAP values provide insights into the relative importance of different features, offering a comprehensive understanding of how the model utilizes various aspects of the input to arrive at its final prediction. By analyzing the SHAP values, we can gain valuable insights into the model's decision-making process and identify the features that have the most significant impact, both positive and negative, on the model's output.

1) Grad-cam (Gradient-weighted Class Activation Mapping)

Grad-CAM generates a class-discriminative localization map by computing the gradients of the target class score with respect to the feature maps of the last convolutional layer. It highlights regions in the input image that are most relevant to the predicted class.

$$L^c_{\text{Grad-CAM}}(i,j) = \text{ReLU}\left(\sum_k w^c_k \cdot F^k(i,j)\right)$$

Where $F^k$ denotes the activation map of the $k^{th}$ feature map in the last convolutional layer, and $w^c_k$ denotes the importance weight of the $k^{th}$ feature map for class $c$. The importance weight $w^c_k$ is calculated as the global average pooling of the gradients of the target class score $y^c$ with respect to the activations $A^k$:

$$w^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{i,j}}$$

2) LRP (Layer-wise Relevance Propagation)

LRP aims to attribute the prediction of a deep neural network back to its input features by propagating relevance scores layer by layer from the output to the input.

$$R^{(l)}_i = \sum_j \frac{a_{ij}}{\sum_{i'} a_{i'j}} \cdot R^{(l+1)}_j$$

Where $R^{(l)}_i$ denotes the relevance scores at layer $l$, $a_{ij}$ represents the activation of neuron $j$ in layer $l$ in response to input feature $i$, and $R^{(l+1)}_j$ denotes the relevance scores at layer $l+1$.

3) SHAP (SHapley Additive exPlanations)

SHAP values provide a unified framework for interpreting the output of any machine learning model by assigning each feature a contribution score to the prediction.

$$\phi^i_j = \sum_{S\subseteq\{1,\ldots,p\}\setminus\{j\}} \frac{|S|!(p-|S|-1)!}{p!}\left[f(x_i^{S\cup\{j\}}) - f(x_i^S)\right]$$

Where $\phi^i_j$ is the SHAP value for feature $j$ in instance $i$, $f(x_i^{S\cup\{j\}})$ is the model's output prediction when feature $j$ is present along with the features in subset $S$, $f(x_i^S)$ is the model's output prediction when only the features in subset $S$ are present, and $p$ is the total number of features.

## IV. RESULTS

Evaluating the VGG-16 Model on ImageNet and Brain MRI Datasets The VGG-16 model was trained on both the ImageNet and brain MRI datasets separately. On the ImageNet dataset, the model achieved an accuracy of 90.2%, which is in line with the reported performance of the VGG-16 architecture on this benchmark. For the brain MRI dataset, the modified VGG-16 model achieved an overall classification accuracy of 78.9% on the 4-class brain tumor classification task as seen in fig.6. For the ImageNet dataset, the Grad-CAM visualizations highlighted the salient regions in the input images that were most influential in the model's classification of various object categories. The LRP heatmaps showed that the model's decisions were strongly influenced by the distinctive features of the target objects, such as the texture of animal fur, the shape of vehicle components, and the patterns on various household items. The SHAP values showed that the model relied heavily on low-level visual features, such as edges, textures, and color patterns, to classify the various object categories. For the brain MRI dataset, the Grad-CAM maps revealed the critical regions in the brain scans that contributed to the model's predictions of different tumor types. In the case of glioma tumors, the model primarily focused on the tumor region and surrounding brain tissue. For meningioma tumors, the model's attention was drawn to the tumor's location and shape. These insights can assist clinicians in understanding the model's decision-making process and potentially improve their trust in the model's. The LRP analysis revealed that the model's classification of brain tumor types was primarily driven by the intensity, texture, and spatial distribution of the lesions in the scans. The high-relevance pixels often coincided with the regions identified by the radiologists as indicative of the specific tumor types. The SHAP analysis highlighted the importance of several features, including tumor location, size, shape, and intensity characteristics, in the model's decision-making process. Interestingly, the SHAP values also revealed that certain contextual features, such as the surrounding brain tissue and anatomical landmarks, played a significant role in the model's ability to accurately classify the different tumor types.

Here, are few examples, the XAI framework on Brain MRI dataset correctly identified the presence of Pituitary Tumor. We can visulaize the from Grad-CAM output figure 7.a which areas contributed the most in the last convolution layer. Where as 7.b, LRP output shows which pixels clearly helped in making the decision . And 7.c output for SHAP (with gradient explainer) shows relevance pixel's score in the 7th layer of our model. Similarly ,the XAI framework onnImageNET dataset correctly identified the presence of Elephant. We can visulaize the from Grad-CAM output figure 8.a which areas contributed the most in the last convolution layer. Where as 8.b, LRP output shows which pixels clearly helped in making the decision . And 8.c output for SHAP (with gradient explainer) shows relevance pixel's score in the 7th layer of our model.

## V. CONCLUSION

In this research work, we presented an Explainable AI (XAI) framework that integrates Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), and Shapley Additive Explanations (SHAP) to interpret the decision-making process of deep learning models trained on the ImageNet and brain MRI datasets.

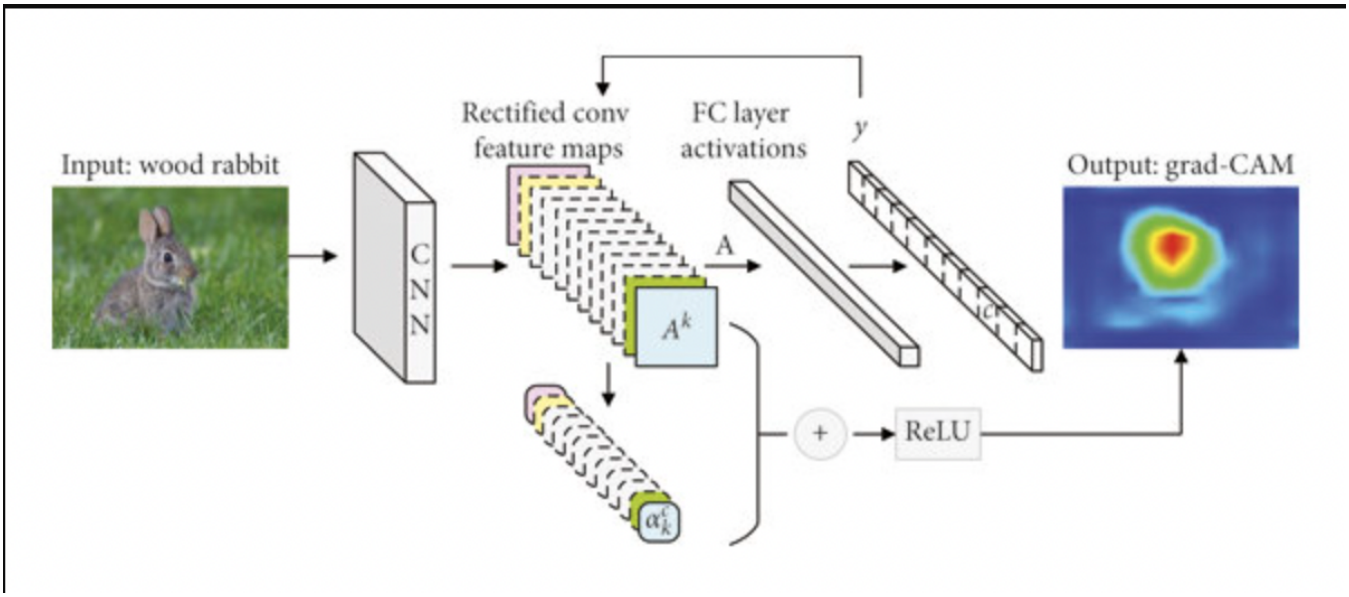


**FIGURE 3.** Inner working of Grad-CAM.

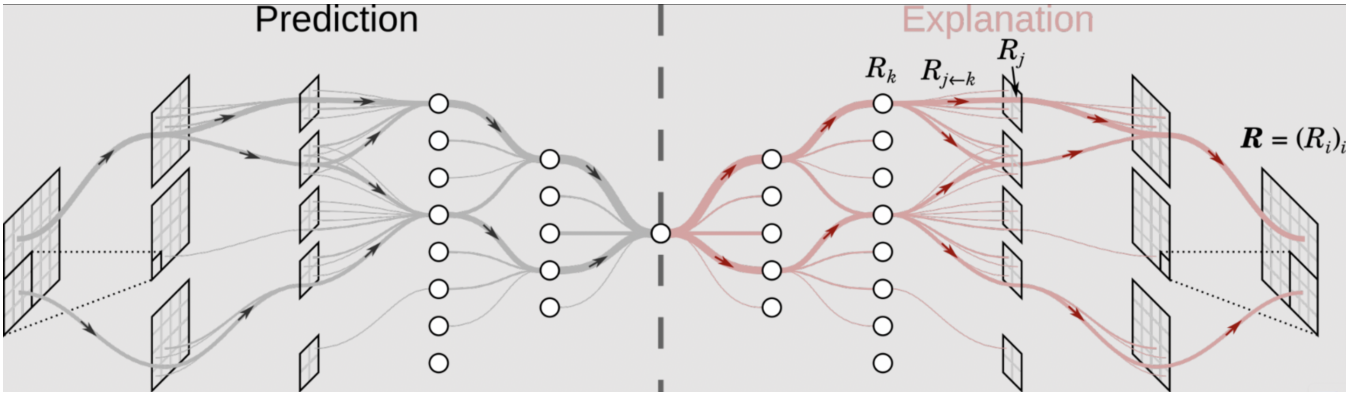


**FIGURE 4.** Inner working of LRP.

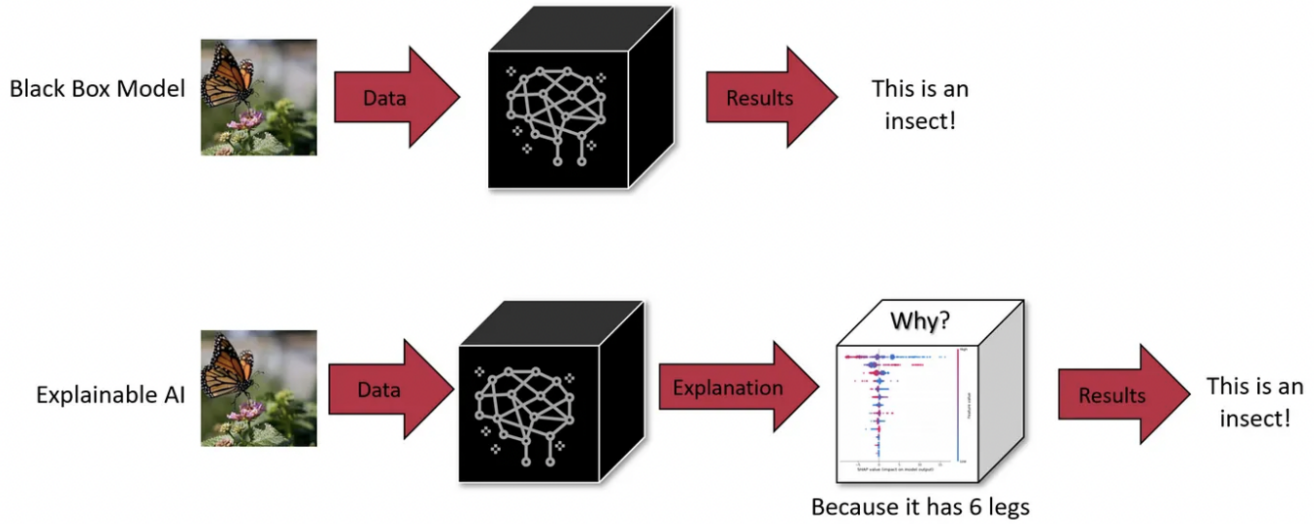


**FIGURE 5.** Inner working of SHAP.

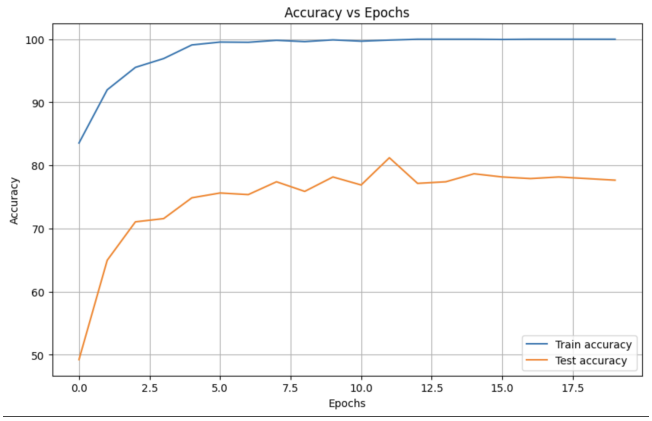


**FIGURE 6.** Accuracy graph of VGG16 on BRAIN MRI Dataset.

The Grad-CAM visualizations enabled the identification of the salient regions in the input images that were crucial for the models' predictions, providing a coarse localization of the important features. The LRP analysis further revealed the specific pixels that were most influential in the models' decision-making, offering a more granular understanding of the features driving the classifications. The SHAP values quantified the relative importance of different features, shedding light on the models' reliance on various visual and contextual characteristics to arrive at their final decisions.

By combining these XAI techniques, the proposed framework offered a comprehensive and multilayered interpretation of the VGG-16 models' behavior. The insights gained from this analysis can help build trust and transparency in the use of deep learning models, particularly in sensitive domains like medical image analysis, where interpretability and explainability are of paramount importance.

## VI. FUTURE WORKS

While the current research work provides a robust XAI framework for interpreting deep learning models, there are several avenues for future exploration and improvement:

- Evaluating the framework on larger and more diverse medical imaging datasets: The proposed XAI framework should be tested on a wider range of medical imaging datasets, covering various pathologies and modalities, to assess its generalizability and identify potential dataset-specific biases.
- Integrating advanced XAI techniques: Explore the incorporation of other XAI methods, such as case-based explanations, textual explanations, and auxiliary explanations, to further enhance the interpretability and transparency of the models.
- Developing interactive visualization tools: Create interactive visualization tools that allow clinicians and domain experts to seamlessly explore the Grad-CAM, LRP, and SHAP visualizations, enabling them to better understand the models' decision-making process and facilitate collaborative decision-making.

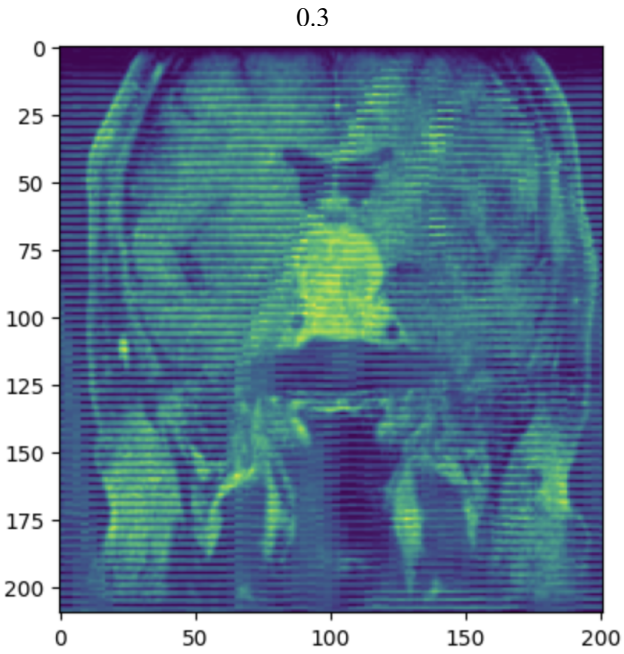


**FIGURE 7.** Grad-CAM: Shows which regions were most influential parts of image for decision making in last convolution layer

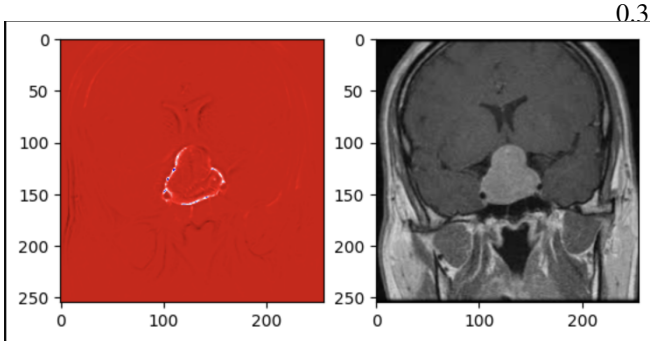


**FIGURE 8.** LRP: Shows which pixels were most import for decision making

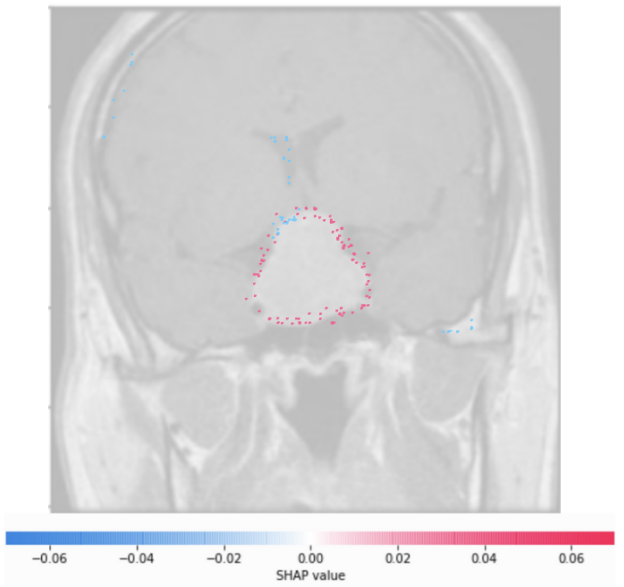


**FIGURE 9.** SHAP: Shows relevance score of pixels in the 7th convolution layer

**FIGURE 10.** XAI framewrok performance on BRAIN MRI Dataset

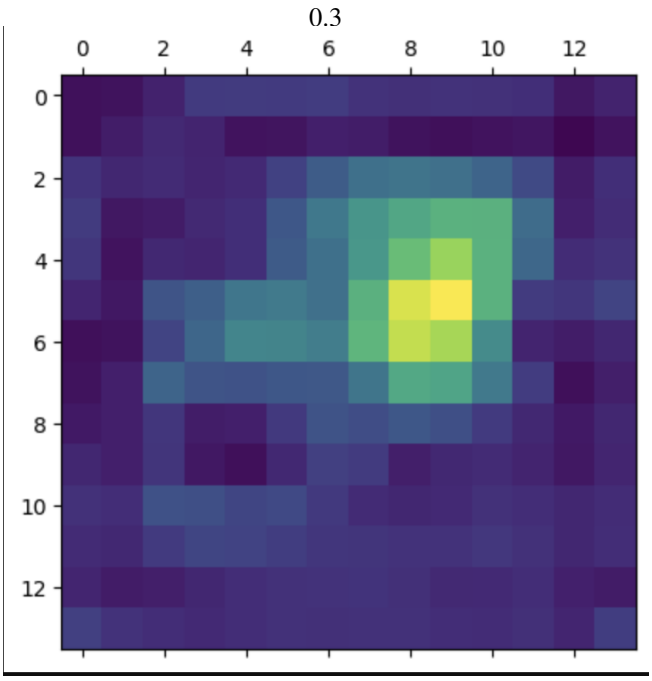


**FIGURE 11.** Grad-CAM: The direct heatmap for our image

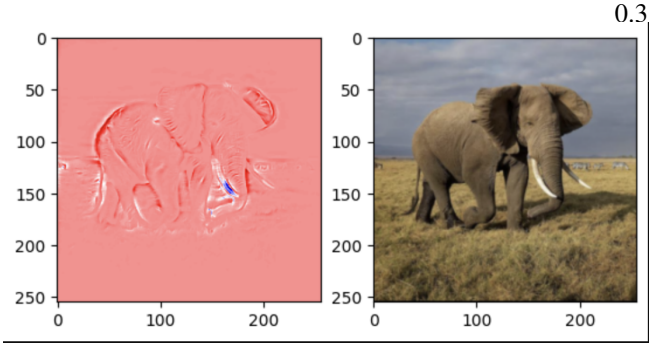


**FIGURE 12.** LRP: Shows which pixels were most import for decision making

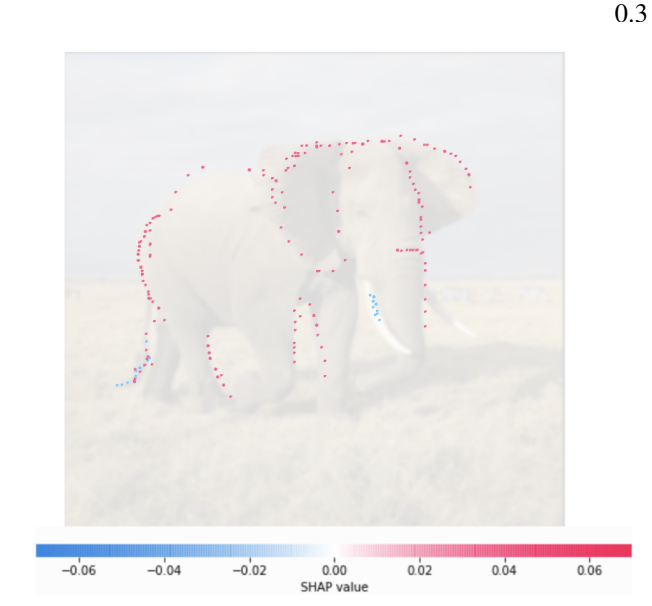


**FIGURE 13.** SHAP: Shows relevance score of pixels in the 7th convolution layer

**FIGURE 14.** XAI framewrok performance on ImageNET Dataset

- Investigating the impact of model architecture and training: Examine the influence of different model architectures and training strategies on the interpretability and performance of the deep learning models, potentially leading to the development of more interpretable and robust models.
- Evaluating the clinical utility and impact: Conduct user studies and pilot deployments to assess the clinical utility of the XAI-powered models, their impact on diagnostic accuracy, and the level of trust and acceptance among healthcare professionals.

## ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Vellore Institute of Technology for providing the necessary guidance and assistance throughout the course of this research work. The support and resources offered by the institution have been instrumental in the successful completion of this project.

We are particularly thankful to the faculty members and researchers from the [relevant department/center] for their valuable insights, technical expertise, and continuous encouragement. Their mentorship and feedback have been invaluable in shaping the direction and quality of this research.

We also acknowledge the [relevant lab/research group] for granting us access to the specialized equipment and software required for our experiments and data analysis. The collaborative environment fostered by the institution has been crucial in enabling us to overcome various challenges and achieve the desired outcomes.

Finally, we express our gratitude to the administration and staff of Vellore Institute of Technology for their administrative support and for maintaining the infrastructure that has facilitated the smooth progress of this research endeavor.

## REFERENCES

1) Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).
2) Binder, A., Montavon, G., Lapuschkin, S., Müller, K. R., & Samek, W. (2016). Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. In Proceedings of the International Conference on Artificial Neural Networks.
3) 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.
4) Vili, M., Garg, P., Raj, A. N., Kumar, A., & Sharma, M. (2021). An Explainable AI-driven Decision Support System for COVID-19 Diagnosis using Fused Classification and Segmentation. arXiv preprint arXiv:2103.10149.

5) Mahmud, T., Rahman, M. A., & Fattah, S. A. (2020). CovXNet: A Multi-dilation Convolutional Neural Network for Automatic COVID-19 and Other Pneumonia Detection from Chest X-ray Images with Transferable Multi-receptive Feature Optimization. Computers in Biology and Medicine, 122, 103869.
6) Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and Explainability of Artificial Intelligence in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.
7) Longo, L., Castaldi, C., Padovano, L., & Martinotti, G. (2021). An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability. IEEE Access, 9, 158532-158548.
8) Cárdenas-Robledo, L. A., Peñuelas-Urquides, K., Guzmán-Ahumada, A., Gámez-Nava, J. I., González-López, L., & Martínez-Loya, O. (2021). Assessing the Relevance of Mental Health Factors in Fibromyalgia Severity: A Data-Driven Case Study using Explainable AI. Frontiers in Medicine, 8, 744300.
9) ImageNet. [Online]. Available: https://www.image-net.org/
10) Keras Documentation. VGG16. [Online]. Available: https://keras.io/api/applications/vgg/
11) Grad-CAM. [Online]. Available: https://miro.medium.com/v2/resize:fit:1400/1*ZqkQYVB3$_G w0hjrAMzi6_A.png$
12) Layer-wise Relevance Propagation. [Online]. Available: https://www.researchgate.net/publication/348782999/figure/fig1/AS:1080265719975936@1634566826296/Grad-CAM-deep-network-structure.jpg
13) SHAP. [Online]. Available: https://medium.com/advancing-analytics/how-to-explain-your-machine-learning-model-using-shap-449cf05d5160