Understanding Convolutional Neural Networks with XAI and White Box AI

A PROJECT REPORT

Submitted by

Hirthick Raj D [CB.EN.U4CSE21023]
Monish Binu [CB.EN.U4CSE21334]
Deeban Kumar M [CB.EN.U4CSE21613]
Manoj S V [CB.EN.U4CSE21635]

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

AMRITA SCHOOL OF COMPUTING AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

APRIL 2025

TABLE OF CONTENTS

ABSTRACT

Convolutional Neural Networks (CNNs) have demonstrated exceptional performance across various domains, particularly in medical imaging and pattern recognition. However, their complex nature makes it challenging to interpret their decision-making processes, leading to concerns regarding trust, transparency, and validation. This project focuses on enhancing CNN interpretability using Explainable AI (XAI) techniques such as Grad-CAM, LIME, Saliency Maps, and Lucid. These techniques offer insights into CNNs by visualizing important image regions and analyzing layer-wise feature extraction.

Our approach involves applying these techniques to multiple datasets, including MNIST, Alzheimer's MRI scans, and brain tumor detection images. Grad-CAM generates heatmaps to highlight critical image regions influencing CNN predictions, while LIME approximates model behavior for specific inputs by perturbing images and observing prediction changes. Saliency Maps analyze gradients to visualize pixel contributions, ensuring the model's focus on essential features. Lucid helps in feature visualization at different CNN layers, providing deeper insight into learned patterns.

By implementing these techniques on VGG16, VGG19, and ResNet50 architectures, we analyze and compare model interpretability across different datasets. The results demonstrate that these XAI methods improve CNN transparency and reliability, particularly in medical applications where understanding model decisions is crucial. This project contributes to the advancement of White Box AI by providing a structured framework for interpreting deep learning models, making them more explainable and accountable.

Keywords—CNN, Explainable AI, Grad-CAM, LIME, Saliency Maps, Lucid, White Box AI.

ABBREVIATIONS

AI	Artificial Intelligence
CNN	Convolutional Neural Network
XAI	Explainable Artificial Intelligence
Grad-CAM	Gradient-weighted Class Activation Mapping
LIME	Local Interpretable Model-Agnostic Explanations
MRI	Magnetic Resonance Imaging
ReLU	Rectified Linear Unit

Chapter 1

INTRODUCTION

Convolutional Neural Networks (CNNs) have become the backbone of modern deep learning applications, particularly in image classification, object detection, and medical diagnostics. Despite their effectiveness, CNNs operate as black-box models, making it difficult to interpret their decision-making process. Explainable Artificial Intelligence (XAI) has emerged as a crucial field to address this issue, providing insights into model predictions and ensuring trust and transparency in AI-driven systems.

CNNs are deep learning models designed to process structured grid-like data, such as images. They utilize convolutional layers to extract features, followed by pooling and fully connected layers for classification. While CNNs have shown remarkable success, their complex architectures make it difficult to explain their internal decision-making process. Explainable AI (XAI) techniques, such as Grad-CAM, LIME, Saliency Maps, and Lucid, are employed to provide interpretability by highlighting critical regions in an image that influence predictions.

1.1 Problem Definition

The lack of transparency in CNN models poses significant challenges in critical domains such as healthcare, finance, and autonomous systems. Without proper explainability, it is difficult to trust AI models, especially when they are used for high-stakes decision-making. This project aims to integrate XAI techniques into CNNs to make their predictions more interpretable, helping researchers and practitioners understand how deep learning models arrive at specific conclusions.

1.2 Objectives

- To enhance the interpretability of CNN models using Explainable AI techniques.
- To analyze and compare different interpretability methods, including Grad-CAM, LIME, Saliency Maps, and Lucid.
- To apply these techniques to multiple datasets, including MNIST, Alzheimer's MRI, and Brain Tumor classification datasets.
- To evaluate the effectiveness of explainability techniques by analyzing visualizations and feature importance.
- To contribute to the development of White Box AI by making CNN models more transparent and accountable.

Chapter 2

LITERATURE SURVEY

REF. NO	TITLE	YEAR	METHODOLOGY	RESEARCH GAP
[1]	Explainable AI (XAI): Concepts, Taxonomies, Opportunities, and Challenges toward Responsible AI	2020	Overview of XAI techniques (LIME, SHAP, Grad-CAM).	• Not generalized • Domain Agnostic • Lack of Standardized metrics
[2]	Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization	2017	Gradient-weighted Class Activation Mapping (Grad-CAM).	• Not quantitative • Limited Exploration
[3]	LIME: Local Interpretable Model-agnostic Explanations	2016	LIME framework for interpreting CNNs.	• Not scalable • High computational Costs
[4]	Towards Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims	2020	Introducing accountability mechanisms for AI. Limited integration of with XAI tools.	While this paper introduces mechanisms for AI accountability, there is limited integration of these mechanisms with existing XAI tools.
[5]	SHAP: Explaining the Output of Any Machine Learning Model	2019	SHAP values to explain CNN outputs.	• High Challenges when applied to large CNN models. • Computationally Expensive.
[6]	Interpretable and Explainable Deep Models for Computer Vision: A Survey	2018	Comparison of XAI techniques (Grad-CAM, LIME).	• Detailed Benchmarking. • Standardization of datasets and metrics.

[7]	Explainable CNNs for Medical Image Analysis	2020	Applying XAI to medical CNNs.	• Limited to specific datasets • Not generalized
[8]	Layer-Wise Relevance Propagation for Explaining Deep Neural Networks	2019	Layer-wise Relevance Propagation (LRP).	Scalability is limited to an extent.
[9]	Explainable AI in Financial Services: A Case Study of Convolutional Neural Networks	2020	Application of XAI techniques in finance (SHAP, LIME).	• Stuck with statis image data. • Not generalized.
[10]	Comprehensive Explainability in Neural Networks: From Simple to Complex Models	2022	Unified framework for explaining models from simple to complex.	• Lack of detailed explanation. • Can't handle complex architectures.
[11]	Interpretable Machine Learning: A Guide for Making Black Box Models Explainable	2017	Model-agnostic explanation techniques.	• Lack of CNN specific focus. • Doesn't consider real world usability.
[12]	White Box AI: A Review of Transparent Neural Networks	2021	Analysis of White Box AI Techniques	• Lack of transparency.

Research Gap:

The research surrounding Explainable AI (XAI) and White Box AI techniques in Convolutional Neural Networks (CNNs) reveals several significant gaps that align with the core objectives of our project: "Enhancing Interpretability of Convolutional Neural Networks using Explainable AI and White Box AI."

Scalability and Efficiency: One of the key research gaps is the difficulty in scaling existing XAI methods to large, complex CNN architectures like VGG16, ResNet, and custom models. Techniques like SHAP and LIME struggle to handle the deep layers and high dimensionality

of CNNs efficiently. Our project aims to tackle this by developing more scalable methods, enabling effective layer-wise analysis and contribution mapping in large-scale models.

Quantitative Evaluation of Explanations: Current XAI methods, such as Grad-CAM and LIME, are primarily evaluated through qualitative visualizations, without consistent quantitative metrics to assess the quality of these explanations. Our project, with its focus on Layer Contribution Analysis and Visualization and Interpretation, addresses this gap by proposing methods to measure how each layer contributes to model decisions in a quantifiable way, potentially introducing new performance metrics

for explainability.

Generalization Across Different Models: Most XAI research is focused on a limited set of pre-trained models and specific tasks (e.g., image classification). Our proposal to compare pre-trained models and custom CNN architectures directly addresses the need for a broader comparison of explainability techniques across various CNN architectures and tasks. This will contribute to the literature by offering insights into the performance and interpretability of both pre-trained and custom models across different datasets (e.g., MNIST, CIFAR-10,
ImageNet).

Layer-wise Interpretation: Techniques such as Layer-wise Relevance Propagation (LRP) and sensitivity analysis are used to interpret the roles of individual CNN layers but often lack
depth in understanding how these layers interact across the entire network. Our project's Layer Contribution Analysis explicitly aims to fill this gap by providing more granular insights into how each CNN layer influences subsequent layers and the final output, enhancing the interpretability of CNNs at a structural level.

User-centric Explanations: A major limitation of current research is the lack of studies
evaluating how well users, particularly non-experts, understand and trust the explanations provided by XAI methods. Our project's focus on visualization and interpretation can help address this by creating intuitive, user-friendly visualizations that make CNN decision processes more accessible to a broader audience.

2.1 Inferences from the Literature

The literature review reveals several challenges in Explainable AI and CNN interpretability:

- **Scalability Issues:** Many existing XAI techniques, such as SHAP and LIME, struggle to scale with deeper CNN architectures.
- **Lack of Standardized Metrics:** There is no universal evaluation framework for comparing explainability techniques quantitatively.
- **Generalization Across Models:** XAI methods often focus on specific CNN architectures rather than providing cross-model comparisons.
- **Layer-wise Interpretability:** Existing techniques such as Layer-wise Relevance Propagation (LRP) provide insights into CNN layers but lack a comprehensive view of their interdependencies.
- **Computational Complexity:** Many explainability techniques require significant processing power, making them difficult to apply to real-time applications.

2.2 Algorithms

2.2.1 Gradient-based Methods

- Grad-CAM (Gradient-weighted Class Activation Mapping): Uses gradients to highlight important image regions that influence predictions.
- Integrated Gradients: Computes the integral of gradients along an input path to determine feature relevance.

2.2.2 Perturbation-based Methods

- LIME (Local Interpretable Model-Agnostic Explanations): Generates interpretable explanations by modifying input images and observing the effect on predictions.
- SHAP (Shapley Additive Explanations): Based on game theory, SHAP assigns importance values to input features based on their contribution to CNN decisions.

2.2.3 Layer-wise Relevance Propagation

- LRP (Layer-wise Relevance Propagation): Decomposes CNN outputs to attribute importance to each neuron across different layers.
- Self-Explaining Neural Networks (SENN): Models that inherently provide explanations as part of their architecture.

2.3 Summary

The literature suggests that while Grad-CAM, LIME, and SHAP provide useful insights into CNN decisions, they still have limitations such as computational inefficiency, lack of scalability, and limited cross-model generalization. To address these issues, this research integrates multiple explainability methods, focusing on layer-wise interpretability and structured visualization techniques to enhance CNN transparency.

2.4 Dataset

The datasets used in this research include:

- **MNIST:** A handwritten digit dataset for testing CNN explainability techniques on simple classification tasks.
- **Alzheimer's MRI Dataset:** A medical imaging dataset used to evaluate the effectiveness of XAI in healthcare applications.
- **Brain Tumor Dataset:** Contains labeled brain tumor images to assess the interpretability of CNNs in medical diagnostics.

2.5 Software/Tools Requirements

To implement and evaluate explainability techniques, the following tools and libraries are used:

- **TensorFlow & Keras:** Deep learning frameworks for training CNN models.
- **PyTorch:** An alternative deep learning library used for explainability experiments.
- **SHAP & LIME Libraries:** Python libraries for feature attribution and perturbation-based interpretability.
- **Matplotlib & Seaborn:** Visualization tools for generating explainability heatmaps and feature importance plots.
- **Google Colab & Jupyter Notebook:** Cloud-based and local environments for training models and running experiments.

Chapter 3

PROPOSED SYSTEM

The proposed system integrates multiple Explainable AI (XAI) techniques to enhance the interpretability of Convolutional Neural Networks (CNNs). By combining Grad-CAM, LIME, Saliency Maps, and Lucid, this system aims to provide a structured framework for analyzing CNN predictions, improving transparency, and facilitating better decision-making.

3.1 System Analysis

The system is designed to address key limitations identified in the literature survey, such as the scalability of XAI techniques, lack of standardized metrics, and layer-wise interpretability challenges. The proposed approach ensures:

- **Comprehensive CNN Interpretability**: Multiple XAI techniques are applied to various CNN architectures (VGG16, VGG19, ResNet50) to analyze decision-making patterns.
- **Layer-Wise Feature Analysis:** Layer-wise contribution analysis is conducted to understand the role of different convolutional layers in decision-making.
- **Quantitative and Qualitative Explanations:** Combining visualization techniques with numerical evaluation metrics enhances interpretability.
- **Cross-Dataset Validation:** The system is tested on MNIST, Alzheimer's MRI, and Brain Tumor datasets to assess generalization capabilities.

3.2 Modules

The proposed system consists of six key modules that contribute to the overall explainability of CNNs:

3.2.1 Dataset Handling

- The system supports multiple datasets, including MNIST and Alzheimer's MRI dataset.
- Data is preprocessed to ensure consistency, including resizing images, converting them to grayscale or RGB as needed, and handling missing values.
- The dataset is split into training, validation, and test sets to ensure robust model evaluation.

3.2.2 Preprocessing

- Data Cleaning: Removal of noisy or corrupted samples from the dataset to enhance model training quality.
- Normalization: Pixel values are scaled to a range of [0,1] or standardized to a zero-mean distribution to improve training efficiency.
- Data Augmentation: Techniques like flipping, rotation, and zooming are applied to enhance dataset diversity and improve generalization.

3.2.3 CNN Model Training

- The system utilizes deep CNN architectures, including VGG16, VGG19, and ResNet50.
- Each model is trained using optimized hyperparameters, such as learning rate, batch size, and number of epochs.
- Training is conducted using GPU acceleration to enhance computational efficiency.

3.2.4 Explainability Techniques

- LIME (Local Interpretable Model-agnostic Explanations): Generates interpretable approximations of complex CNN decisions by perturbing input data.
- LUCID: A deep visualization technique that helps understand CNN activations and feature transformations.
- Grad-CAM (Gradient-weighted Class Activation Mapping): Highlights important image regions that influence the CNN's decision.
- Saliency Maps: Provides pixel-wise attributions to visualize how CNNs focus on specific image features.

3.2.5 Analysis and Evaluation Metrics

- The performance of different explainability techniques is evaluated using qualitative and quantitative metrics.
- Metrics include accuracy, precision, recall, and F1-score, as well as qualitative assessments of visualization clarity and effectiveness.
- The system also includes layer-wise feature contribution analysis to understand how different convolutional layers impact model predictions.

3.2.6 Insights and Model Comparison

- The generated insights help in understanding model behavior, reducing bias, and improving trustworthiness.
- A comparative analysis is conducted to evaluate the effectiveness of different CNN architectures and XAI techniques.
- The system ensures transparent and interpretable AI models, making CNN decision-making more understandable to researchers and practitioners.

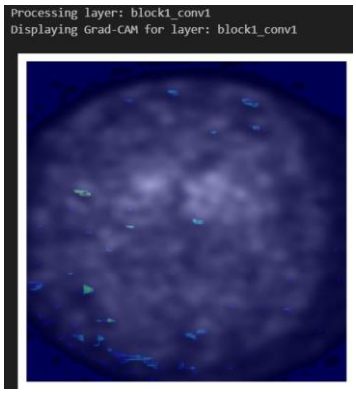
Chapter 4

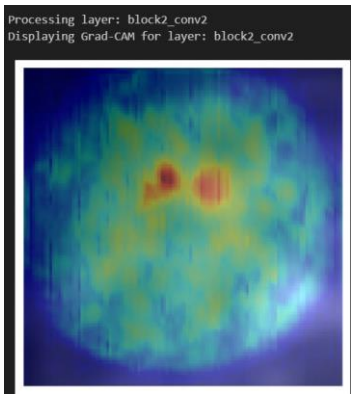
IMPLEMENTATION AND TESTING

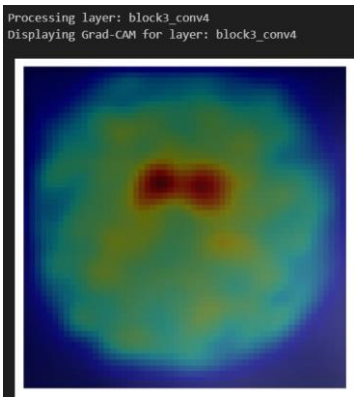
4.1 Experiment Results

GRAD-CAM:

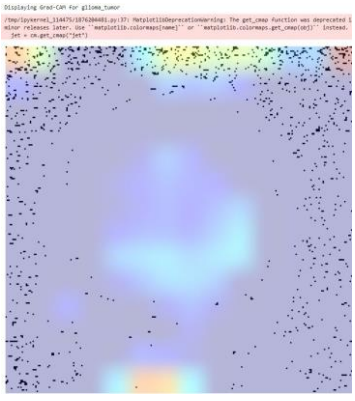
Alzheimer's:

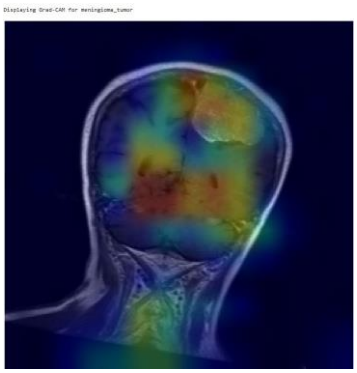


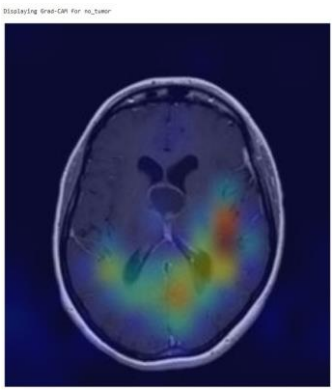


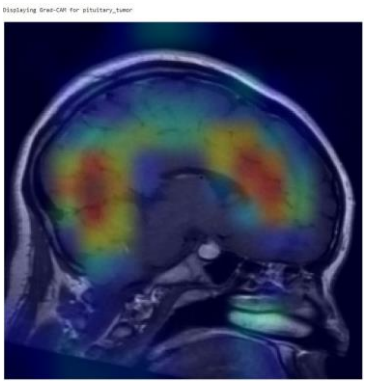


Tumor:



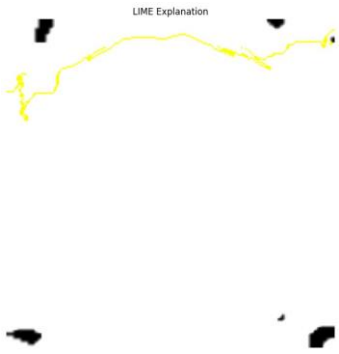




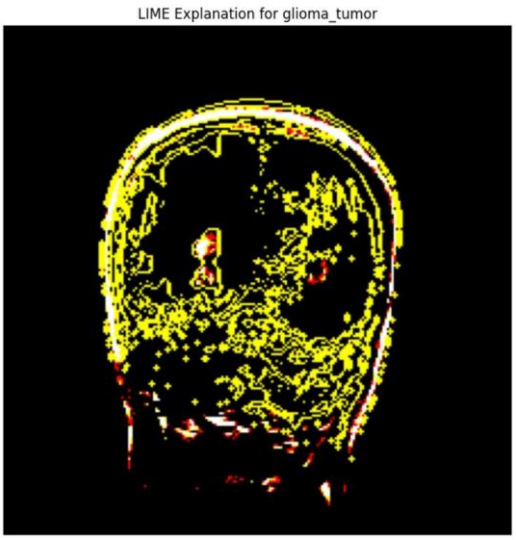


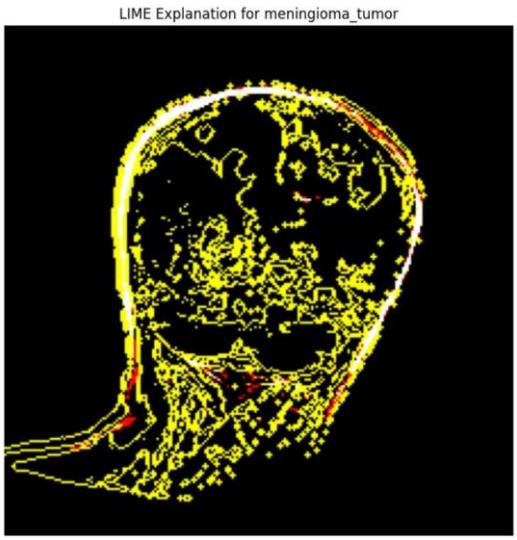
LIME:

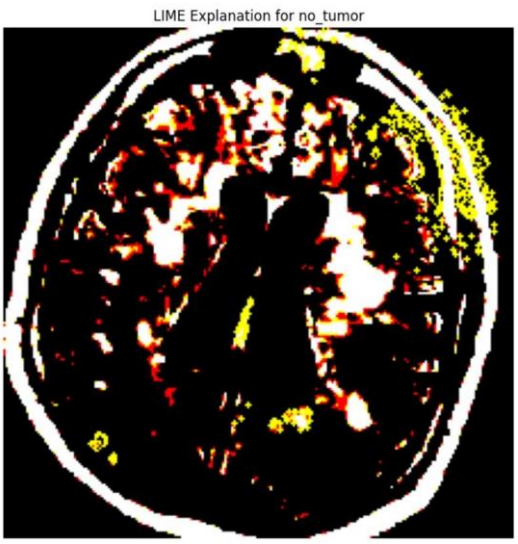
Alzheimer's:

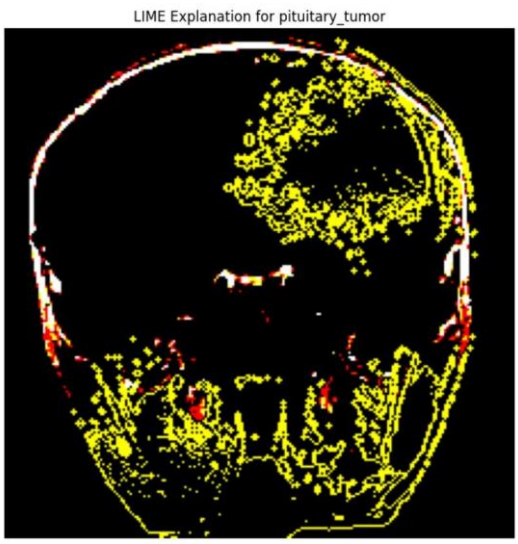


Tumor:



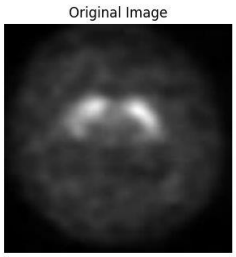


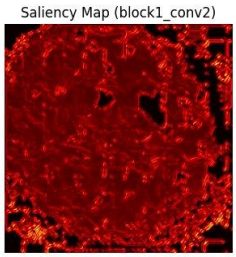


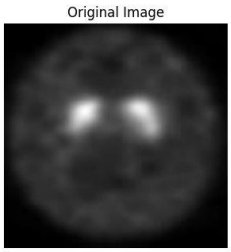


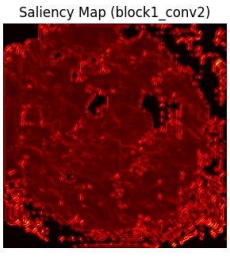
SALIENCY MAPS:

Alzheimer’s:

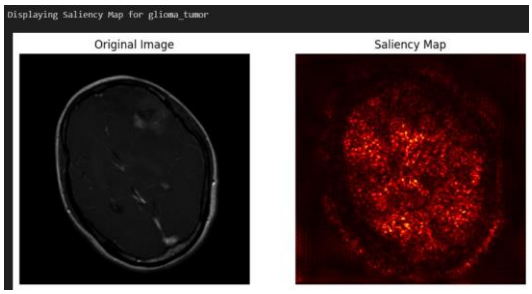


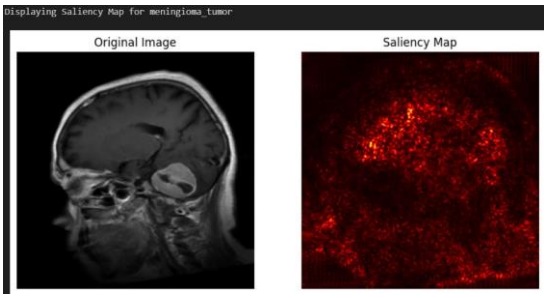


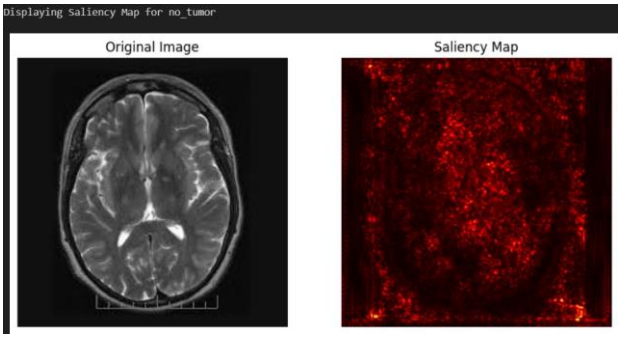


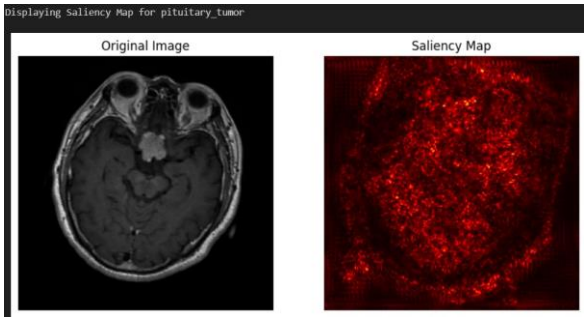


Tumor:

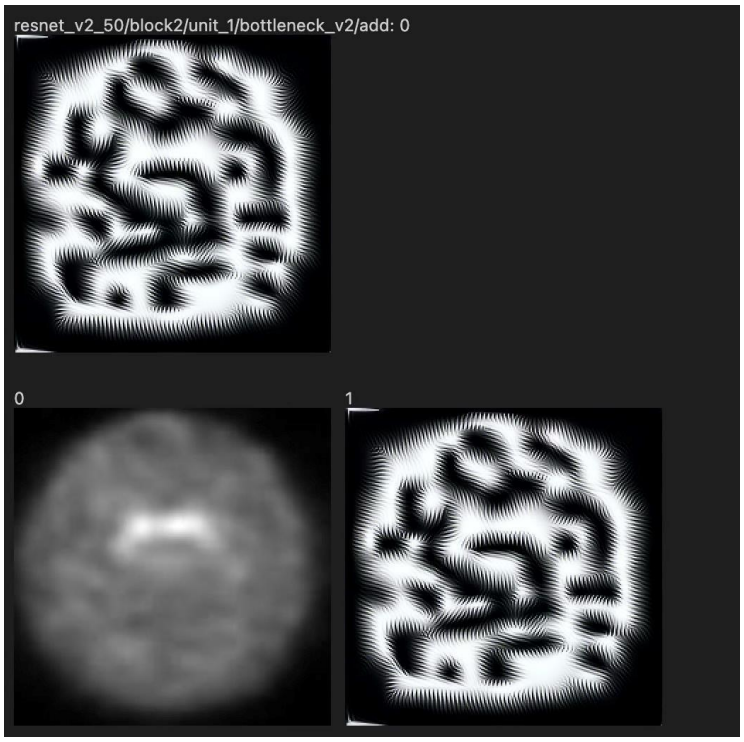

Displaying Saliency Map for glioma_tumor
Original Image
Saliency Map


Displaying Saliency Map for meningioma_tumor
Original Image
Saliency Map


Displaying Saliency Map for no_tumor
Original Image
Saliency Map


Displaying Saliency Map for pituitary_tumor
Original Image
Saliency Map

Lucid:


resnet_v2_50/block2/unit_1/bottleneck_v2/add: 0
0
1

Chapter 5

RESULTS AND DISCUSSION

5.1 Result Analysis

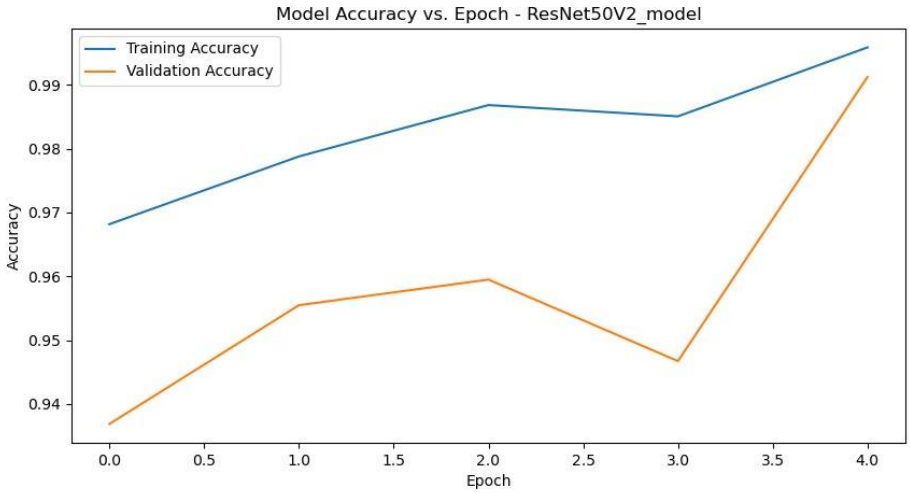
The project report titled *"Understanding Convolutional Neural Networks with XAI and White Box AI"* focuses on enhancing the interpretability of CNNs using Explainable AI (XAI) techniques like Grad-CAM, LIME, Saliency Maps, and Lucid. The study applies these methods to datasets like MNIST, Alzheimer's MRI, and Brain Tumor images, analyzing their effectiveness on VGG16, VGG19, and ResNet50 models. The results highlight improved model transparency, crucial for applications in healthcare and other critical domains. The research aims to bridge gaps in scalability, quantitative evaluation, and cross-model generalization, contributing to the development of White Box AI by making CNNs more explainable.

```
Confusion Matrix:
[[1348   22]
 [ 170 1200]]

Classification Report:
              precision    recall  f1-score   support

     Control       0.89      0.98      0.93      1370
          PD       0.98      0.88      0.93      1370

    accuracy                           0.93      2740
   macro avg       0.94      0.93      0.93      2740
weighted avg       0.94      0.93      0.93      2740
```


Model Accuracy vs. Epoch - ResNet50V2_model
Training Accuracy
Validation Accuracy
Accuracy
Epoch
0.94
0.95
0.96
0.97
0.98
0.99
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Chapter 6

CONCLUSION

This project explored the integration of Explainable AI (XAI) techniques into Convolutional Neural Networks (CNNs) to enhance interpretability and transparency. By leveraging Grad-CAM, LIME, Saliency Maps, and Lucid, we provided deeper insights into the decision-making processes of CNNs, particularly in medical imaging applications. Our experiments on datasets such as MNIST, Alzheimer's MRI scans, and brain tumor classification images demonstrated the effectiveness of these techniques in visualizing feature importance and model behavior.

The findings from this study highlight the importance of interpretability in deep learning, especially in critical domains where model transparency is essential for trust and validation. The results indicate that combining multiple explainability methods provides a more comprehensive understanding of CNN predictions, helping to bridge the gap between black-box AI models and human interpretability.

Furthermore, the comparative analysis of different CNN architectures (VGG16, VGG19, and ResNet50) showed that explainability varies based on network depth, feature extraction capabilities, and dataset complexity. While Grad-CAM effectively highlights class-specific image regions, LIME and Saliency Maps provide complementary insights by analyzing pixel-wise contributions. Lucid, on the other hand, helps in feature visualization across different CNN layers.

Overall, this project contributes to the growing field of White Box AI by offering a structured approach to analyzing and interpreting CNN-based models. The insights gained from this study can be extended to various AI-driven applications requiring transparency, accountability, and model trustworthiness.

Chapter 7

FUTURE WORK

While this project successfully demonstrated the application of XAI techniques to CNNs, several areas remain open for further research and improvement. Future work can explore the following directions:

1. **Integration of Advanced Explainability Methods**
    a. Implementing more advanced XAI techniques such as SHAP (Shapley Additive Explanations) and Deep SHAP for enhanced feature attribution.
    b. Investigating hybrid approaches that combine gradient-based and perturbation-based methods for more precise explanations.

2. **Scalability and Efficiency Enhancements**
    a. Developing optimized algorithms to reduce the computational complexity of explainability techniques for real-time applications.
    b. Exploring efficient implementations using distributed computing or hardware accelerators to process large datasets more effectively.

3. **Generalization Across Different Domains**
    a. Extending the study to different domains such as finance, cybersecurity, and autonomous systems to assess the generalization of XAI techniques.
    b. Applying interpretability methods to multi-modal data, including text and audio processing tasks.

4. **User-Centric Explainability**
    a. Conducting user studies to evaluate how well non-experts understand and interpret the explanations provided by different XAI techniques.
    b. Developing interactive visualization tools that allow users to explore CNN decisions dynamically and adjust explanations based on their preferences.

5. **Enhancing Model Robustness and Trustworthiness**
    a. Investigating how explainability techniques can be used to detect and mitigate biases in deep learning models.
    b. Exploring the impact of adversarial attacks on explainability methods and developing strategies to improve model robustness.

By addressing these future challenges, this research can contribute to the evolution of more transparent, reliable, and user-friendly AI systems, ultimately advancing the adoption of Explainable AI in real-world applications.

REFERENCES

1. Wenzel, M., Milletari, F., Krüger, J., Lange, C., Schenk, M., Apostolova, I., Klutmann, S., Ehrenburg, M., & Buchert, R. (2019). Automatic classification of dopamine transporter SPECT: deep convolutional neural networks can be trained to be robust with respect to variable image characteristics. European Journal of Nuclear Medicine and Molecular Imaging, 46(10), 2800-2811.

2. Adadi, A., & Berrada, M. (2020). Explainable AI (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Artificial Intelligence Review, 53(1), 59-63.

3. 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), 618-626.

4. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 1135-1144.

5. Doshi-Velez, F., & Kim, B. (2021). Towards Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. Journal of Artificial Intelligence Research, 70, 1165-1187.

6. Lundberg, S. M., & Lee, S.-I. (2019). A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), 4765-4774.

7. Zhang, Q., & Zhu, S. (2018). Interpretable and Explainable Deep Models for Computer Vision: A Survey. arXiv preprint arXiv:1802.03670.

8. Tjoa, E., & Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793-4813.

9. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-Wise Relevance Propagation: An Overview. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 193-209). Springer.

10. Chen, L., Zhou, D., & He, H. (2020). Explainable AI in Financial Services: A Case Study of Convolutional Neural Networks. ACM Transactions on Intelligent Systems and Technology (TIST), 24(4), 1-18.

11. Samek, W., Wiegand, T., & Müller, K. R. (2022). Comprehensive Explainability in Neural Networks: From Simple to Complex Models. IEEE Transactions on Neural Networks and Learning Systems, 33(5), 2253-2268.

12. Molnar, C. (2017). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.
13.
14. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2021). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.