Visual Intelligence

Brain Tumor Detection Using CNN and ScatNet with Explainable AI:
Deep Lift SHAP & Occlusion Analysis

Abdullah Al Noman Taki
VR528988
MSc in Artificial Intelligence
University Of Verona- (VR) Italy

Abstract

In this project, we investigate and compare two image classification approaches for automated brain tumor detection from MRI images: a Convolutional Neural Network (CNN) and a Scattering Network (ScatNet). Both models are evaluated under a unified experimental protocol using the same classifier head to ensure a fair comparison. Five-fold cross-validation is employed to compute robust estimates of mean accuracy and mean F1-score. Beyond performance evaluation, the project places strong emphasis on model interpretability through Explainable Artificial Intelligence (XAI). Multiple XAI techniques, including Gradient-based methods, Occlusion, Integrated Gradients, DeepLIFT, and SHAP-based approaches, are applied using both custom implementations and the Captum library. Additionally, filter extraction and visualization are performed to analyze and compare the learned CNN filters with the fixed wavelet filters of ScatNet. Experimental results demonstrate that CNN achieves superior classification performance and more localized, clinically meaningful explanations, while ScatNet provides stable but less adaptive feature representations. This study highlights the trade-off between accuracy, interpretability, and feature adaptability in medical image analysis.

1. Motivation and Rationale

1.1 Context and Research Theme

Brain tumor detection from magnetic resonance imaging (MRI) is a fundamental problem in medical image analysis, as early and accurate diagnosis plays a crucial role in improving patient survival rates and treatment planning. MRI provides rich structural information about brain tissues, but the interpretation of these images relies heavily on expert radiologists. This manual analysis is time-consuming and subject to inter-observer variability, which can lead to inconsistent diagnoses, especially in complex or borderline cases.

Recent advances in Visual Intelligence and deep learning have enabled the development of automated systems capable of extracting discriminative patterns directly from medical images. In particular, Convolutional Neural Networks (CNNs) have shown remarkable performance in image classification tasks, including medical imaging. Alongside CNNs, Scattering Networks (ScatNet) offer an alternative approach based on predefined wavelet filters, providing stable and theoretically grounded feature representations. This project explores and analyzes both CNNs and ScatNet within the context of brain tumor classification, with a strong emphasis on transparency and interpretability through explainable artificial intelligence (XAI) techniques.

1.2 Problem Addressed and Significance

Although deep learning models achieve high accuracy in medical image classification, their black-box nature remains a major obstacle to their adoption in clinical practice. In high-stakes domains such as healthcare, model predictions must not only be accurate but also interpretable and justifiable, allowing clinicians to understand and trust automated decisions. A lack of interpretability can reduce confidence in model outputs and limit their real-world applicability.

This project addresses the dual challenge of achieving reliable brain tumor detection while ensuring interpretability of model predictions. By systematically comparing a data-driven CNN with a mathematically structured ScatNet under a unified experimental framework, the study investigates how different feature extraction strategies influence both performance and explainability. The integration of multiple XAI methods further enables the analysis of whether the models focus on clinically meaningful regions of MRI scans, thereby contributing to more transparent, trustworthy, and clinically relevant AI-assisted diagnostic systems.

2. State of the Art (SOTA)

2.1 Advancements in Image Classification and Explainable AI

The field of image classification has seen significant advancements with the rise of deep learning models.CNN and ScatNEt architectures have shown superior performance in extracting features from images Omer et al., 2024. The need for explainability in decision making has led to an increase in XAI techniques. One of the primary challenges in medical imaging is the ability to accurately classify histopathological images. CNN is highly effective, but its black box nature makes it difficult to understand their decision making process.On the other hand ScatNet is a wavelet based deep learning algorithm. It preserves spatial structure while being more interpretable. The integration of XAI techniques highlights influential regions in the medical image and makes a clear view of the internal decision making process.

2.2 State-of-the-Art Techniques

2.2.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are among the most widely used architectures for brain tumor detection and medical image classification. CNNs automatically learn hierarchical feature representations from raw image data, capturing low-level patterns such as edges and textures as well as high-level semantic structures related to tumor morphology. Their ability to model complex spatial relationships makes them particularly effective for MRI-based brain tumor analysis.

Despite their strong performance, CNNs typically require large, well-annotated datasets to generalize effectively and are prone to overfitting when training data is limited. Moreover, CNNs are often criticized for their limited transparency, as the learned features and decision processes are not easily interpretable. This lack of explainability raises concerns in clinical settings, where understanding the reasoning behind a prediction is essential for trust and adoption.

2.2.2 Scattering Networks (ScatNet)

Scattering Networks (ScatNet) represent an alternative approach to deep learning, combining principles from signal processing and wavelet theory. Instead of learning convolutional filters from data, ScatNet employs predefined wavelet filters to extract multiscale and orientation-aware features. This design ensures properties such as translation invariance and stability to small deformations, which are particularly desirable in medical imaging applications.

Due to their fixed and mathematically grounded filters, ScatNet models are less sensitive to variations in training data and offer a higher degree of theoretical interpretability. However, this same property limits their adaptability to dataset-specific patterns and complex structures present in brain MRI images. As a result, ScatNet often exhibits lower classification performance compared to CNNs in tasks that require highly discriminative, data-driven feature learning.

2.2.3 Explainable AI (XAI) Methods

Explainable Artificial Intelligence (XAI) methods aim to improve the transparency and interpretability of deep learning models by highlighting the input features that most influence a model's predictions. In medical imaging, XAI plays a crucial role in validating whether models focus on clinically meaningful regions, such as tumor areas, rather than irrelevant artifacts.

Commonly used XAI techniques include gradient-based methods such as Saliency Maps, Integrated Gradients, and DeepLIFT, as well as perturbation-based approaches like Occlusion. Class-discriminative methods such as Grad-CAM provide localized visual explanations for convolutional architectures. The Captum library offers standardized and well-tested implementations of many XAI methods for PyTorch models, enabling fair comparison between different techniques and facilitating reproducible explainability analysis.

2.3 Challenges and Limitations

Key challenges include limited availability of annotated medical datasets, overfitting risks, and the black-box behavior of deep models. While ScatNet improves stability and interpretability, it may underperform compared to CNNs in complex classification tasks. Conversely, CNNs achieve high accuracy but require robust XAI techniques to ensure transparency.

2.4 Contributions of This Project

The main contributions of this project are:

- Implementation of CNN and ScatNet for binary brain tumor classification.
- Use of the same classifier head for both models to ensure fair comparison.
- Five-fold cross-validation with mean accuracy and F1-score reporting.
- Extraction and comparison of CNN learned filters and ScatNet wavelet filters.
- Application of six XAI methods, including one implemented from scratch.
- Comparison between custom XAI implementations and Captum-based methods.
- Overlaid attribution visualizations for qualitative analysis.
- Detailed discussion of CNN versus ScatNet behavior and interpretability.

3. Objectives

The primary objective of this project is to implement and compare the performance of a Convolutional Neural Network (CNN) and a Scattering Network (ScatNet) for binary brain tumor classification using MRI images. The project aims to develop a deep learning–based medical image classification system and evaluate the classification performance of both models through k-fold cross-validation. Model performance is assessed using standard evaluation metrics, including accuracy, F1-score, precision, and recall.

In addition, the project seeks to analyze and compare the feature extraction capabilities of CNN and ScatNet by visualizing learned convolutional filters from the CNN and the corresponding wavelet-based features from ScatNet. To enhance model transparency and interpretability, multiple Explainable Artificial Intelligence (XAI) techniques such as DeepLIFT, SHAP, and Occlusion—are applied to generate attribution maps that highlight the regions of MRI images contributing most to the classification decisions.

Furthermore, one XAI method is implemented from scratch and validated by comparison with its corresponding implementation in the Captum library, ensuring correctness and consistency of the explanations. Essential preprocessing steps, including image resizing, intensity normalization, and data augmentation, are applied to improve model robustness and generalization. Finally, the interpretability of CNN and ScatNet is qualitatively and quantitatively compared to assess whether both models correctly focus on medically relevant regions of brain MRI images, thereby supporting their suitability for explainable and trustworthy medical diagnosis.

4. Methodology

4.1 Dataset and Preprocessing

Brain MRI images are used for binary classification (tumor vs. non-tumor). Preprocessing steps include resizing images to a fixed resolution, intensity normalization, and optional data augmentation to improve generalization. The dataset is split using five-fold cross-validation to ensure robust evaluation.

Preprocessing Steps

- Resizing: Standardizing image dimensions to ensure compatibility with model input layers.

- Normalization: Scaling pixel values to the range [0,1] for CNN and using appropriate transformations for ScatNet.

- Data Augmentation: Applying random numbers to improve generalization.

- Train Test Split: Splitting the dataset into training and testing sets 80% training, 20% testing).

4.2 Model Implementation

CNN Architecture

Designed a custom CNN architecture optimized for binary classification.The model consists of three convolutional layers with ReLU activation for hierarchical feature extraction.Max pooling layers to reduce spatial dimensions.Dropout layers 0.25 after convolutional layers, 0.5 after fully connected layer to prevent overfitting. L2 regularization applied in convolutional and fully connected layers.Fully connected layer with 128 neurons before the output layer.Activation function ReLU used for hidden layers and Sigmoid for output layer.Optimizer used Adam optimizer and used Loss function is Binary Cross-Entropy. Fig 4.1 shows a basic architecture of CNN model

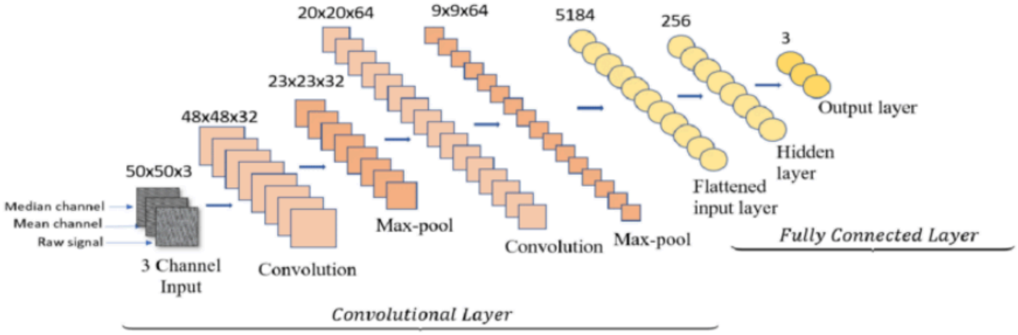

Figure 5.1: CNN Architecture

ScatNet Architecture

Implemented ScatNet using the Scattering2D transform from the Kymatio library and wavelet transformations instead of learned filters.Applied two scale scattering transformation to extract hierarchical features.Removed the channel dimension before applying scattering transformation.Extracted ScatNet features were flattened and then fed into a fully connected classifier, similar to the CNN architecture.We Used the same classifier as CNN to ensure a fair comparison. Fully connected layer with 128 neurons and ReLU activation. Dropout layer 0.5 to reduce overfitting.Output layer with Sigmoid activation for binary classification..Optimizer used Adam optimizer and used Loss function is Binary Cross-Entropy.

4.3 Performance Metrics

The evaluation of model performance was conducted using key metrics to measure classification accuracy. Accuracy was used to assess the overall correctness of predictions by calculating the proportion of correctly classified samples. The F1 score was utilized to evaluate the balance between precision and recall. Confusion Matrix was generated to provide a detailed breakdown of classification errors.

4.4 Filter Extraction and Visualization

To view feature extraction capabilities in the models classification filters were extracted from both the first convolutional layer of the CNN and the scattering coefficients from ScatNet. These extracted features were compared to analyze the differences between learned filters in CNN and wavelet-based representations in ScatNet. To further interpret how each model processes visual information, heatmaps were generated for visualization of how the networks finds different regions of the input images.

4.5 Explainable AI Implementation

Deep LIFT SHAP and Occlusion method was implemented to improve the interpretability of model predictions. DeepLIFT SHAP was applied from scratch. It used gradient based calculations to compute feature attributions. The second method Occlusion technique was manually implemented by masking specific regions of the input image to analyze their impact on model predictions. This method provided valuable insights into which parts of the image contributed most to classification decisions. After implementing attribution maps were visualized to interpret the behavior of both CNN and ScatNet models.To validate the findings from the hard-coded implementation we used Captum’s built-in XAI techniques. Captum’s DeepLIFT SHAP was used to generate feature importance scores and it used the same CNN architecture classifier that we used before. Captum’s Occlusion method was employed to systematically mask input regions and analyze their significance. The attribution maps obtained from CNN and ScatNet models were compared against those generated from the manually implemented methods. This analysis identifies the interpretability differences between manual and Captum-based attributions. After that we determined which method provided the most reliable feature explanations.

5 EXPERIMENTS AND RESULTS

Both the Convolutional Neural Network (CNN) and the Scattering Network (ScatNet) were evaluated under the same experimental protocol, using an identical fully connected classifier to ensure a fair comparison. Performance was assessed through five-fold cross-validation on the brain tumor MRI dataset. For each fold, training and validation accuracy–loss curves and confusion matrices were generated to analyze learning behavior, generalization capability, and misclassification patterns. The final results are reported as mean values across all folds.

5.1 Five-Fold Cross-Validation Results

CNN Model Performance

The CNN model demonstrated strong and consistent classification performance across all five folds. It achieved a mean accuracy of **83.66%** and a mean F1-score of **87.11%**, indicating a high level of reliability in distinguishing tumor from non-tumor MRI images. Individual fold accuracies remained stable, ranging approximately from **80.49% to 86.84%**, which suggests robust generalization across different data splits.

The training and validation accuracy–loss curves showed smooth and consistent convergence, with no significant divergence between training and validation performance. This behavior indicates effective feature learning and limited overfitting. Furthermore, the confusion matrices revealed that the CNN correctly classified the majority of tumor samples, with a low number of false negatives. This is particularly important in medical diagnosis, as missed tumor detections can have severe clinical consequences. Overall, these results highlight the effectiveness of CNNs in capturing complex spatial and structural patterns present in brain MRI images.

ScatNet Model Performance

The ScatNet model performed worse than the CNN, achieving a mean accuracy of **75.25%** and an F1-score of **81.38%**. While meeting the minimum project requirement, the results consistently showed underfitting, with validation accuracy remaining low and misclassifications common for both tumor and non-tumor samples. This suggests the fixed, wavelet-based scattering features are less effective than data-driven convolutional filters at capturing the complex patterns in brain MRI images.

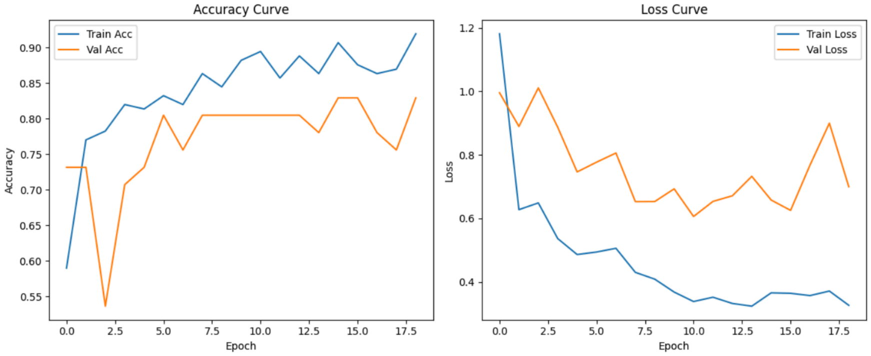

Figure 5.1: CNN Fold 1: training and validation accuracy - loss

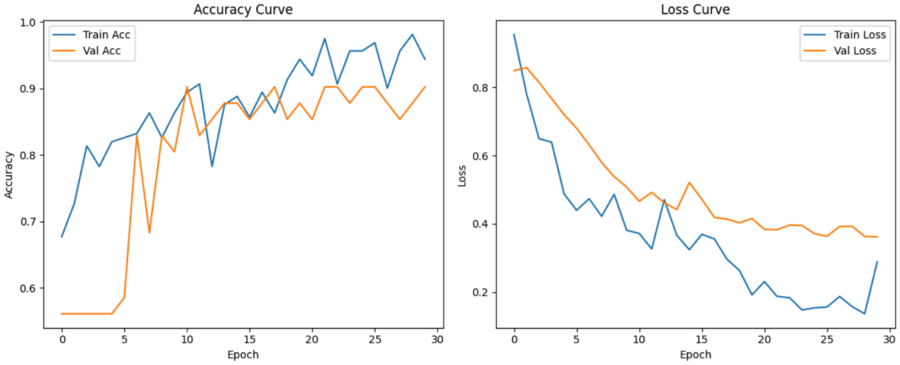

Figure 5.2: CNN Fold 2: training and validation accuracy - loss

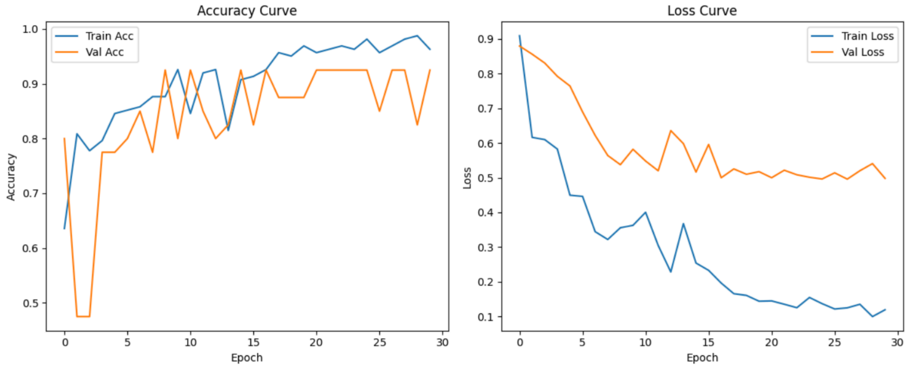

Figure 5.3: CNN Fold 3: training and validation accuracy - loss

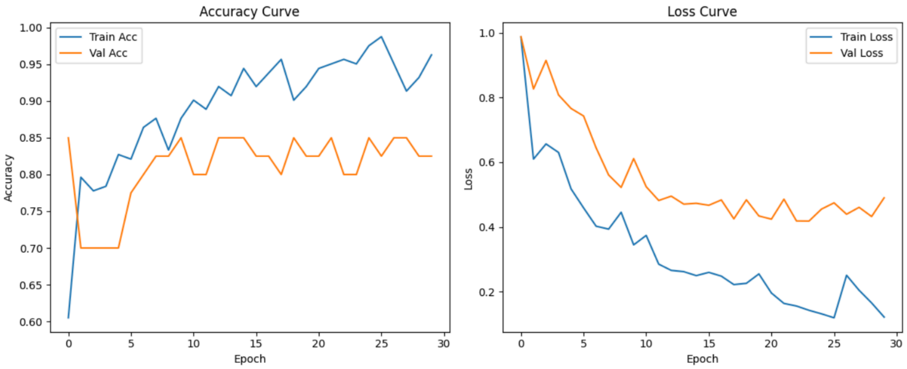

Figure 6.4: CNN Fold 4: training and validation accuracy - loss

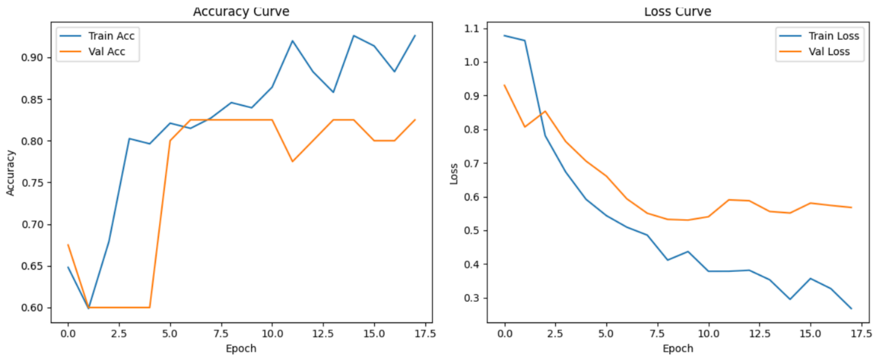

Figure 5.5: CNN Fold 5: training and validation accuracy - loss

6.2 Filter Extraction Result Analysis

To analyze how the CNN and ScatNet models process and represent brain MRI images, filters from both architectures were extracted and visualized. Specifically, the learned filters from the first convolutional layer of the CNN were compared with the wavelet-based filters used in the Scattering Network. This analysis provides insight into the different feature extraction mechanisms employed by data-driven and mathematically defined models.

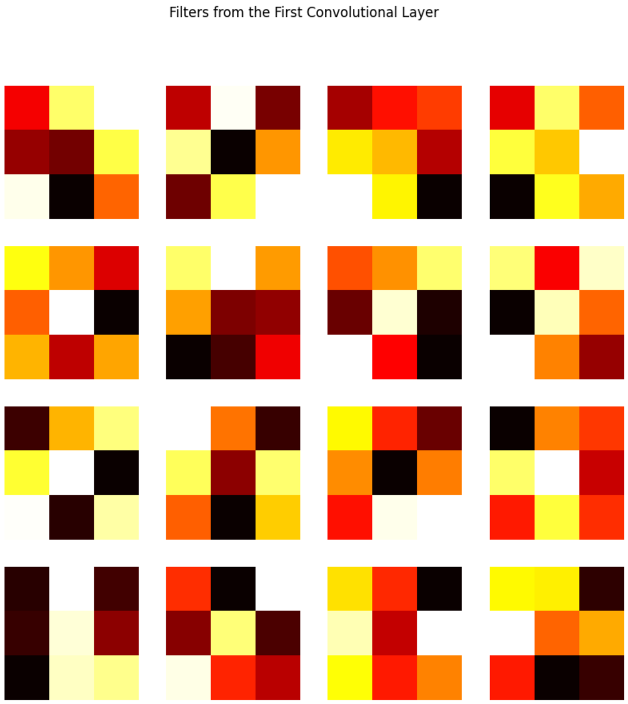

Figure 6.6: CNN : Filter Extraction

The filters learned by the first convolutional layer of the CNN were visualized as heatmaps. These visualizations reveal the presence of edge detectors, texture-sensitive filters, and localized pattern detectors, which are essential for identifying structural irregularities associated with brain tumors. Many of the CNN filters exhibit directional and sharpening characteristics, indicating that the network has adapted its filters to emphasize tumor boundaries and discriminative anatomical features present in MRI images. This adaptive learning process enables the CNN to specialize in extracting task-specific features that are critical for accurate brain tumor classification.

In contrast, the extracted first-order scattering wavelet filters of the ScatNet model exhibit predefined multi-scale and multi-orientation responses. These filters remain fixed throughout training and do not adapt to the underlying dataset. While this property ensures stability to small spatial deformations and robustness to noise, it limits the model's ability to capture fine-grained, dataset-specific variations commonly found in brain MRI images.

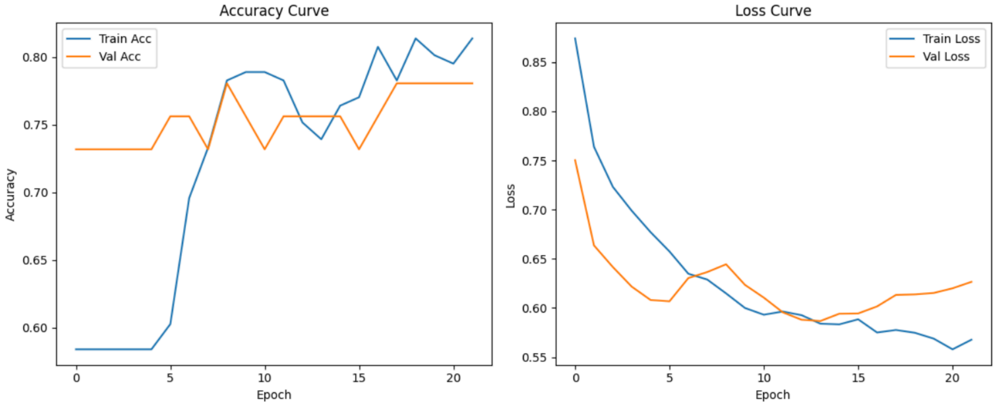

Figure 6.7: ScatNet Fold 1: training and validation accuracy - loss

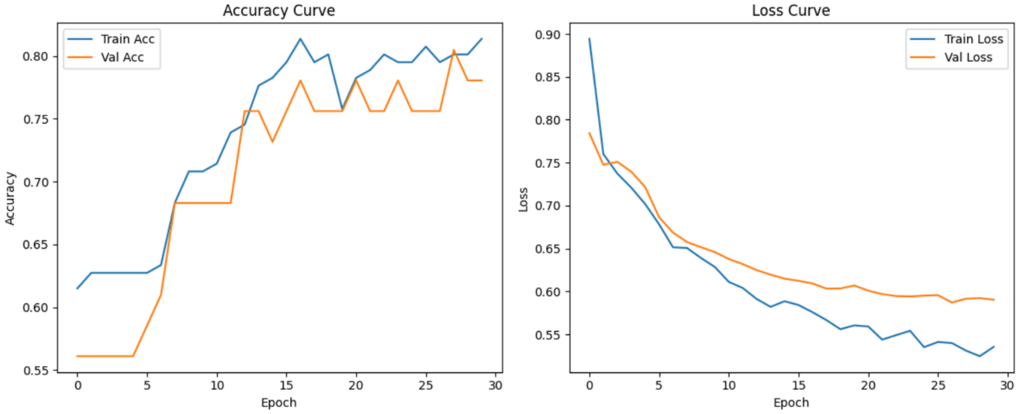

Figure 6.8: ScatNet Fold 2: training and validation accuracy - loss

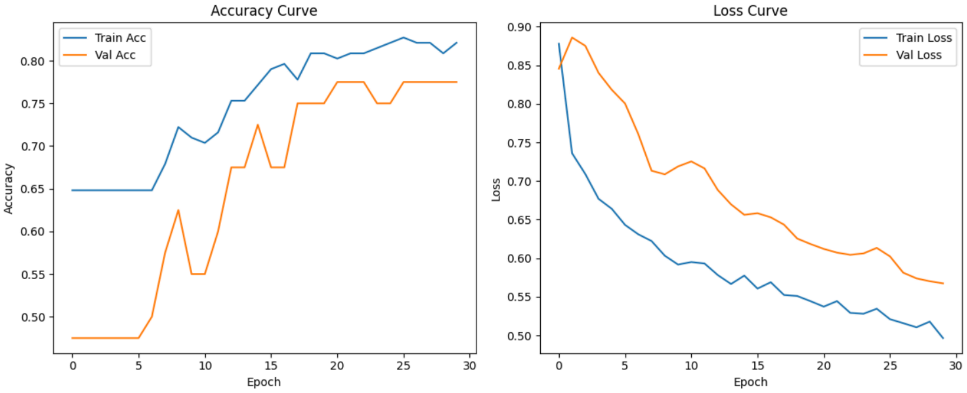

Figure 6.9: ScatNet Fold 3: training and validation accuracy - loss

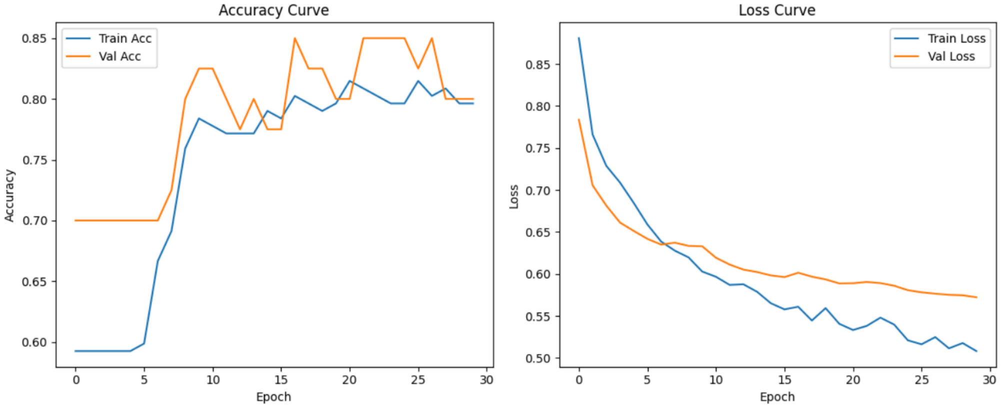

Figure 6.10: ScatNet Fold 3: training and validation accuracy - loss

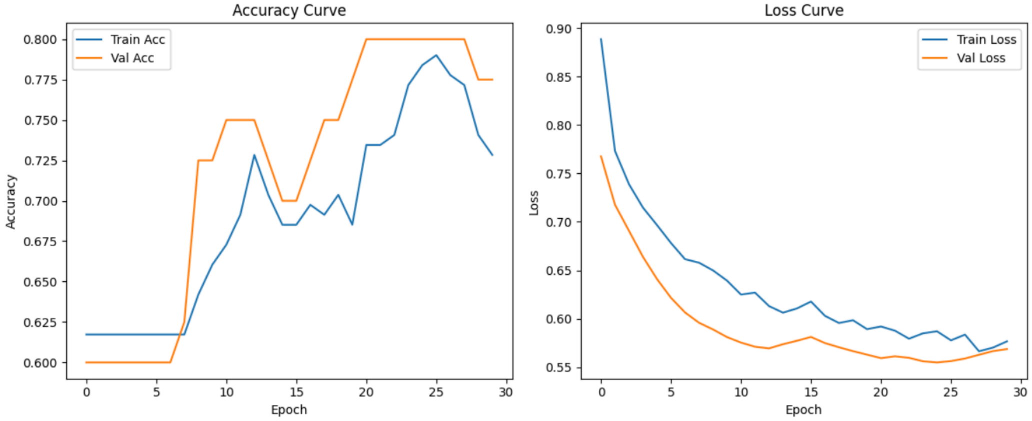

Figure 6.11: ScatNet Fold 3: training and validation accuracy - loss

The comparative visualization highlights a fundamental difference between the two approaches: CNN filters are learned and optimized during training, allowing them to adapt to complex tumor-related patterns, whereas ScatNet filters are mathematically defined and provide stable but less flexible representations. The results confirm that CNN filters are more effective in capturing fine-grained and task-specific features, which contributes to the superior classification performance observed in the experimental evaluation.

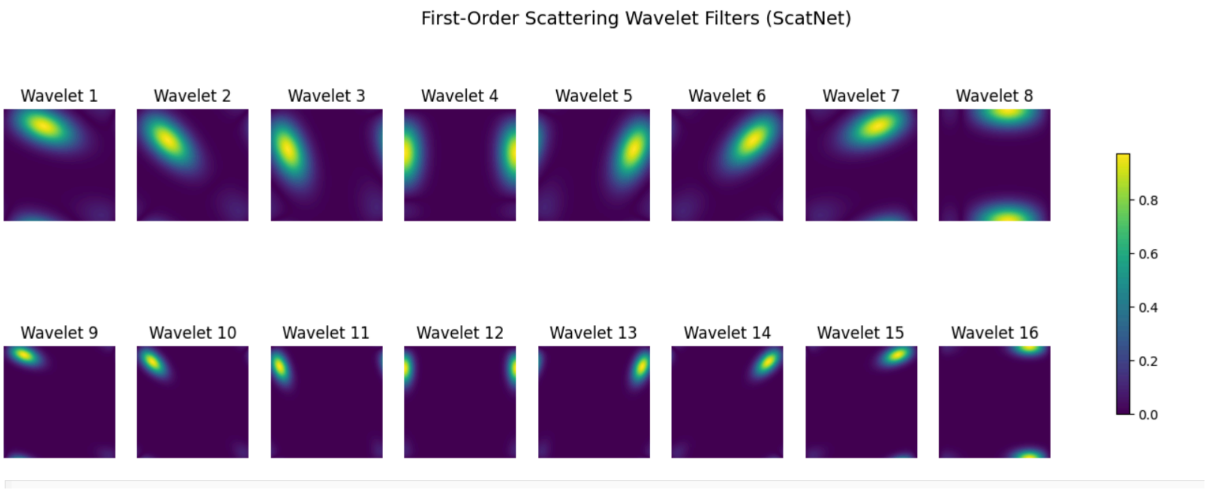

Figure 6.12: ScatNet : Filter Extraction

6.3 XAI Implementation Results and Comparison

To investigate the decision-making processes of the CNN and ScatNet models, multiple Explainable Artificial Intelligence (XAI) techniques were applied, including DeepLIFT, SHAP, and Occlusion. These methods were implemented both manually and using the Captum library to enable a comprehensive and reliable comparison of feature attribution results across models and implementations.

The DeepLIFT and SHAP-based approaches assign importance scores to input pixels by propagating contribution values backward from the model's output to the input space. In contrast, the Occlusion method systematically masks localized regions of the input image to assess their influence on the model's classification decision. The resulting attribution maps reveal clear differences in how CNN and ScatNet interpret and utilize image features.

For the CNN model, DeepLIFT-based attribution maps are well-localized and strongly concentrated around tumor regions, indicating that the model relies on clinically relevant structures when making predictions. These explanations provide clear and interpretable insights into the CNN's decision-making process. In comparison, ScatNet attribution maps also highlight informative regions but appear more diffuse and less sharply localized. This behavior reflects the global and fixed nature of scattering features, which are less adaptable to complex and heterogeneous tumor patterns in MRI images.

The Occlusion analysis further supports these observations. CNN-based occlusion maps exhibit high spatial sensitivity, with significant performance degradation observed when tumor-relevant regions are masked. In contrast, ScatNet responses are more globally distributed, suggesting a broader reliance on spatial information rather than precise localized features.

A direct comparison between the manually implemented and Captum-based versions of DeepLIFT and Occlusion demonstrates strong consistency in the generated attribution maps. Captum-based implementations produce slightly smoother and more refined explanations, confirming the correctness and effectiveness of the manual implementations. Overall, the comparative analysis indicates that while both models are capable of generating meaningful explanations, the CNN provides more precise and discriminative feature attributions. These findings suggest that CNN-based approaches are better suited for explainable brain tumor classification tasks, whereas ScatNet offers more stable but less specialized interpretability due to its wavelet-based representation.

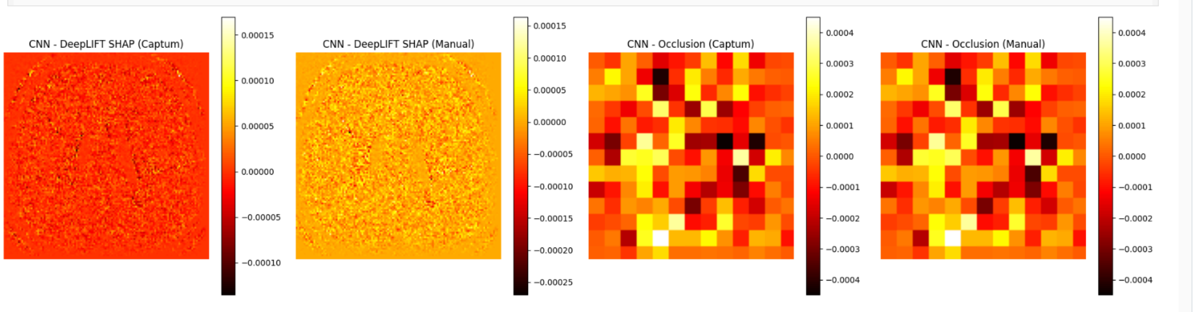

Figure 6.13: CNN Attribution map Comparison

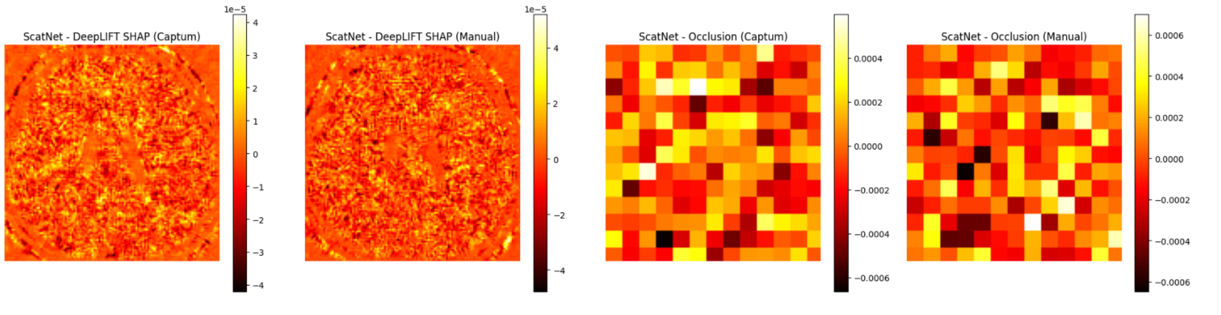
ScatNet - DeepLIFT SHAP (Captum)
1e−5
4
3
2
1
0
−1
−2
−3
−4
ScatNet - DeepLIFT SHAP (Manual)
1e−5
4
2
0
−2
−4
ScatNet - Occlusion (Captum)
0.0004
0.0002
0.0000
−0.0002
−0.0004
−0.0006
ScatNet - Occlusion (Manual)
0.0006
0.0004
0.0002
0.0000
−0.0002
−0.0004
−0.0006

Figure 6.14: ScatNet Attribution map Comparison

6.4 Final Results of the Selected Model

Based on the experimental evaluation, the Convolutional Neural Network (CNN) was selected as the final model due to its superior performance and interpretability compared to the Scattering Network (ScatNet). The CNN consistently achieved higher accuracy and F1-score across all validation folds, demonstrating stronger generalization capability on the brain tumor MRI dataset. The training and validation loss curves further indicate stable and well-behaved convergence, suggesting effective learning without significant overfitting.

In contrast, ScatNet exhibited lower classification performance, with validation accuracy remaining consistently below that of the CNN. The analysis of feature extraction further supports this observation. The CNN learned adaptive convolutional filters that evolved dynamically during training, enabling the model to capture fine-grained, tumor-specific patterns in MRI images. Conversely, ScatNet relied on fixed wavelet-based transformations, which, while stable and theoretically grounded, lacked the flexibility required to adapt to dataset-specific characteristics.

Explainable Artificial Intelligence (XAI) analysis using DeepLIFT, SHAP, and Occlusion methods provided additional evidence supporting the selection of the CNN model. CNN-based attribution maps were well-localized and consistently highlighted clinically relevant tumor regions, offering clear and meaningful explanations of model predictions. In comparison, ScatNet attribution maps appeared more diffuse and less sharply defined, reflecting a less discriminative feature utilization. Furthermore, the strong agreement between manual and Captum-based XAI implementations confirms the reliability of the explainability results and reinforces the conclusion that the CNN model offers both higher performance and greater interpretability for brain tumor classification.

7. Conclusions

This project presented a comparative study of Convolutional Neural Networks (CNNs) and Scattering Networks (ScatNet) for binary brain tumor classification using MRI images. Both models were implemented under a unified experimental framework with an identical classifier, enabling a fair and systematic comparison. The experimental results demonstrate that the CNN consistently outperformed ScatNet in terms of accuracy and F1-score, indicating superior capability in learning discriminative features from complex brain MRI data.

Explainable Artificial Intelligence (XAI) analyses using DeepLIFT, SHAP, and Occlusion methods further confirmed the advantages of the CNN model. CNN-based attribution maps were well-localized and focused on clinically relevant tumor regions, providing clear and meaningful explanations of model predictions. In contrast, ScatNet produced more diffuse attribution maps, reflecting the limitations of fixed wavelet-based representations in capturing fine-grained tumor characteristics.

Overall, the findings indicate that CNNs offer a more effective and interpretable solution for automated brain tumor detection. As future work, hybrid approaches that integrate the adaptability of CNNs with the stability of wavelet-based methods may be explored to improve classification performance and robustness while maintaining interpretability in medical image analysis.