THE EXPLAINTEES

DSAI 305 final project

Meet The Group

Salma

Sara

Dana

Laurence

Contents

Introduction

Alzheimer's disease (AD) is an accelerated neurological brain disorder, the most common type of dementia and is one of the significant challenges in the twenty-first century. It infects part of the brain called the **hippocampus**, causing **shrinking** in this part, which is responsible for thinking, reasoning and making new memories. Fortunately, AD can be detected 20 years or more before any symptoms appear, so it is essential to **detect**, **diagnose** and **categorize** AD early **to slow down or prevent it**

EDA techniques

Classes

1. Very mild Dementia
2. Mild Dementia
3. Non Demented
4. Moderate Dementia

Number of samples per each class

Very mild Dementia: 13725 images

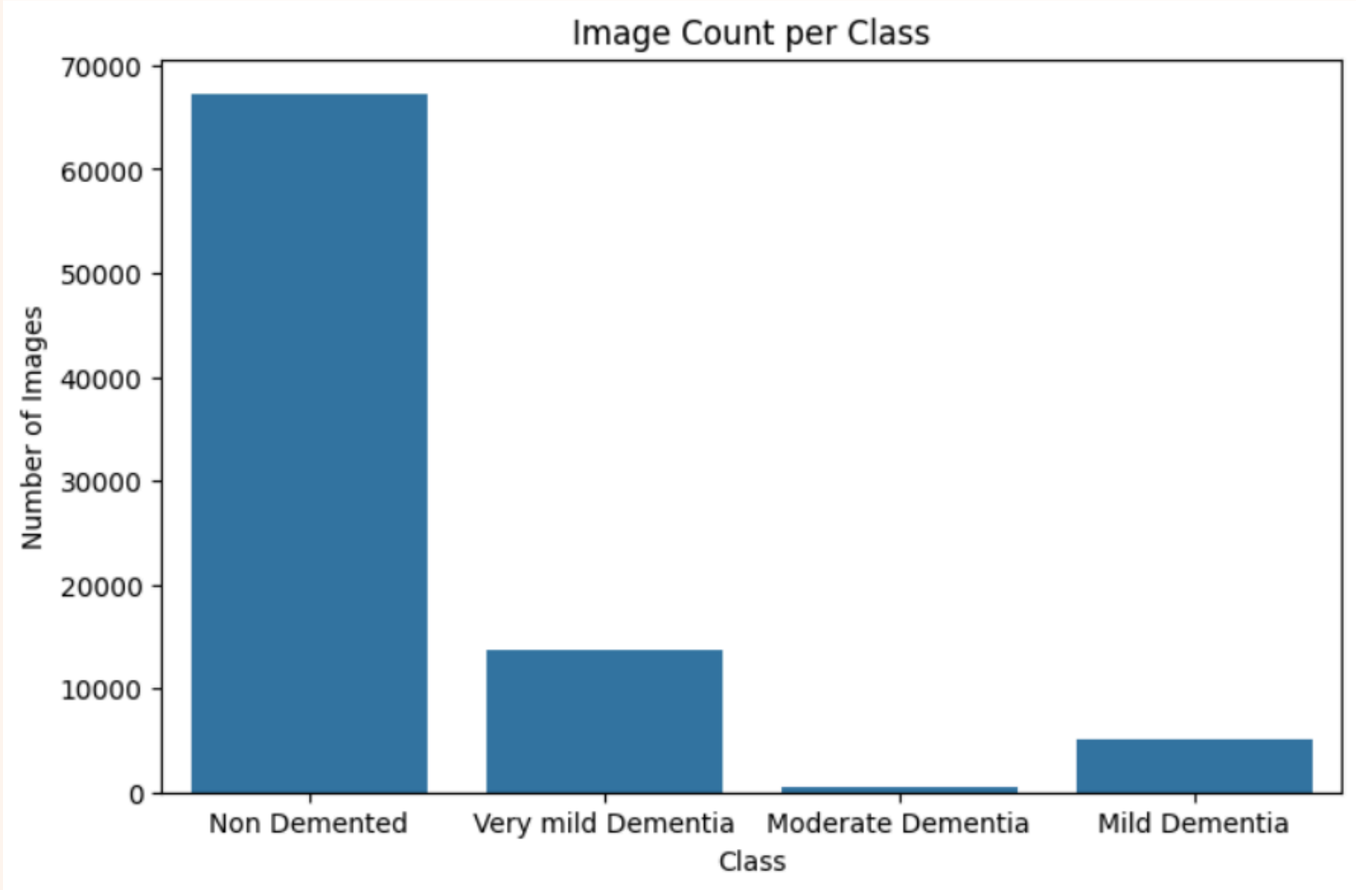
Mild Dementia: 5002 images

Non Demented: 67222 images

Moderate Dementia: 488 images

Total Images: 86437

Image count distribution



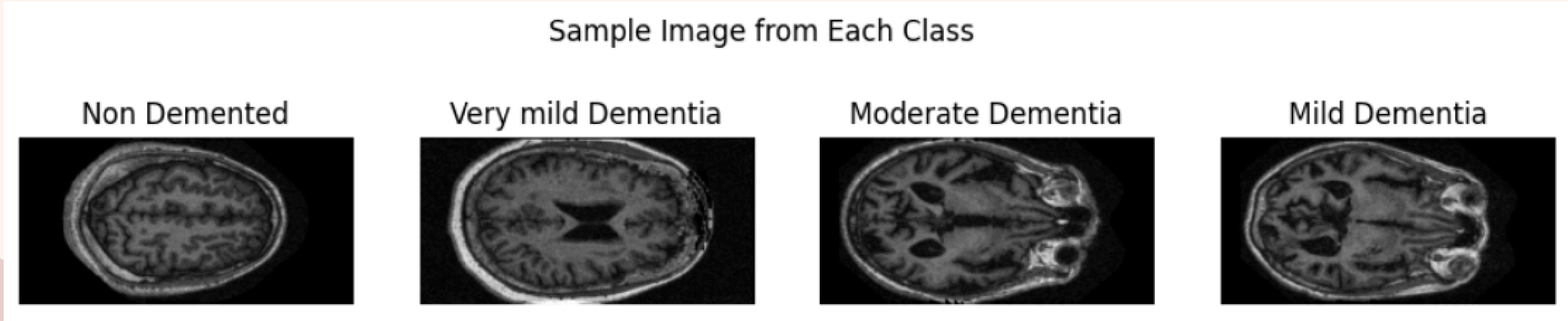
Images height and width statistics

Image Height - Mean: 248.0, Min: 248, Max: 248

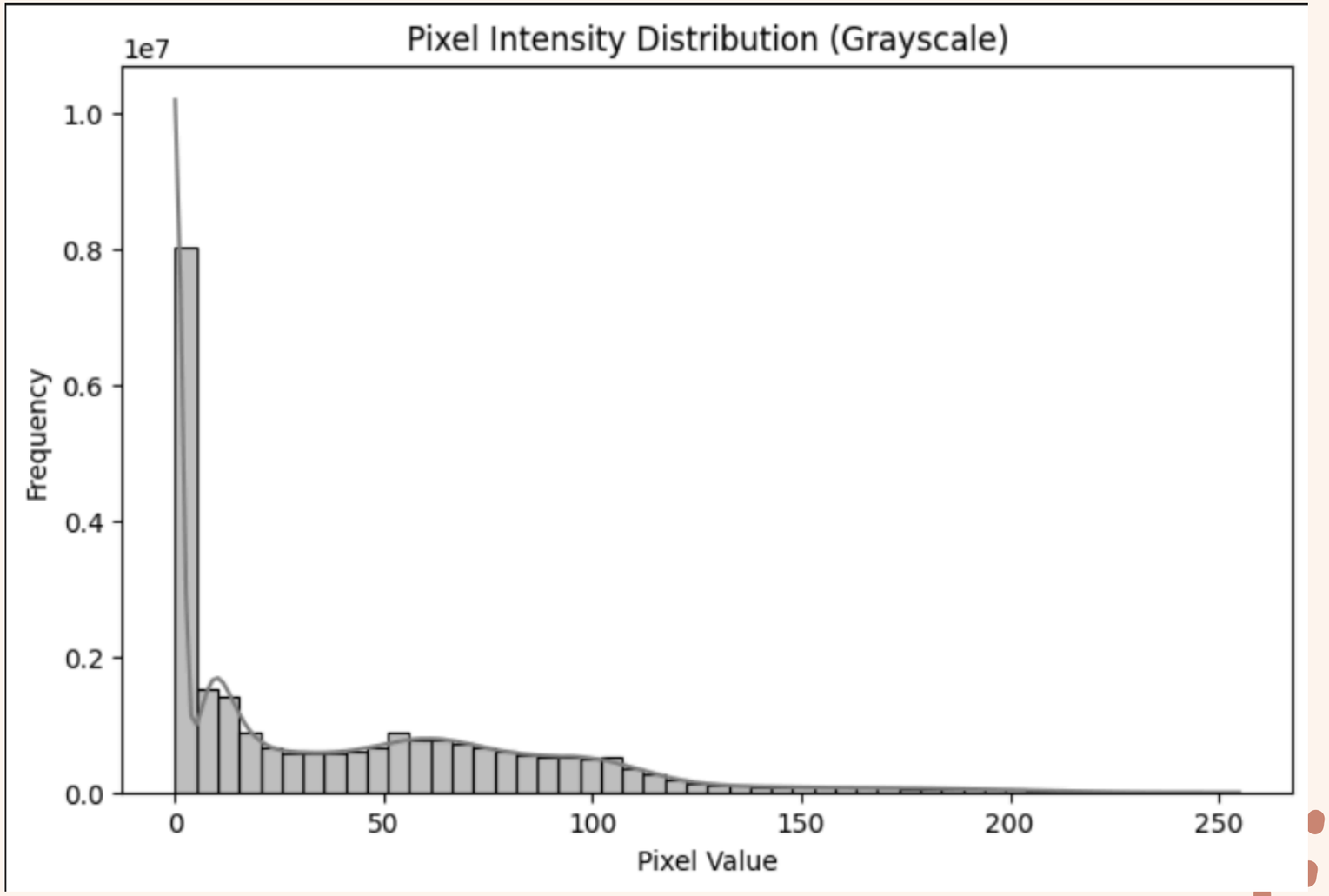
Image Width - Mean: 496.0, Min: 496, Max: 496

Channels: [3]

Sample image for each class



Pixels distribution



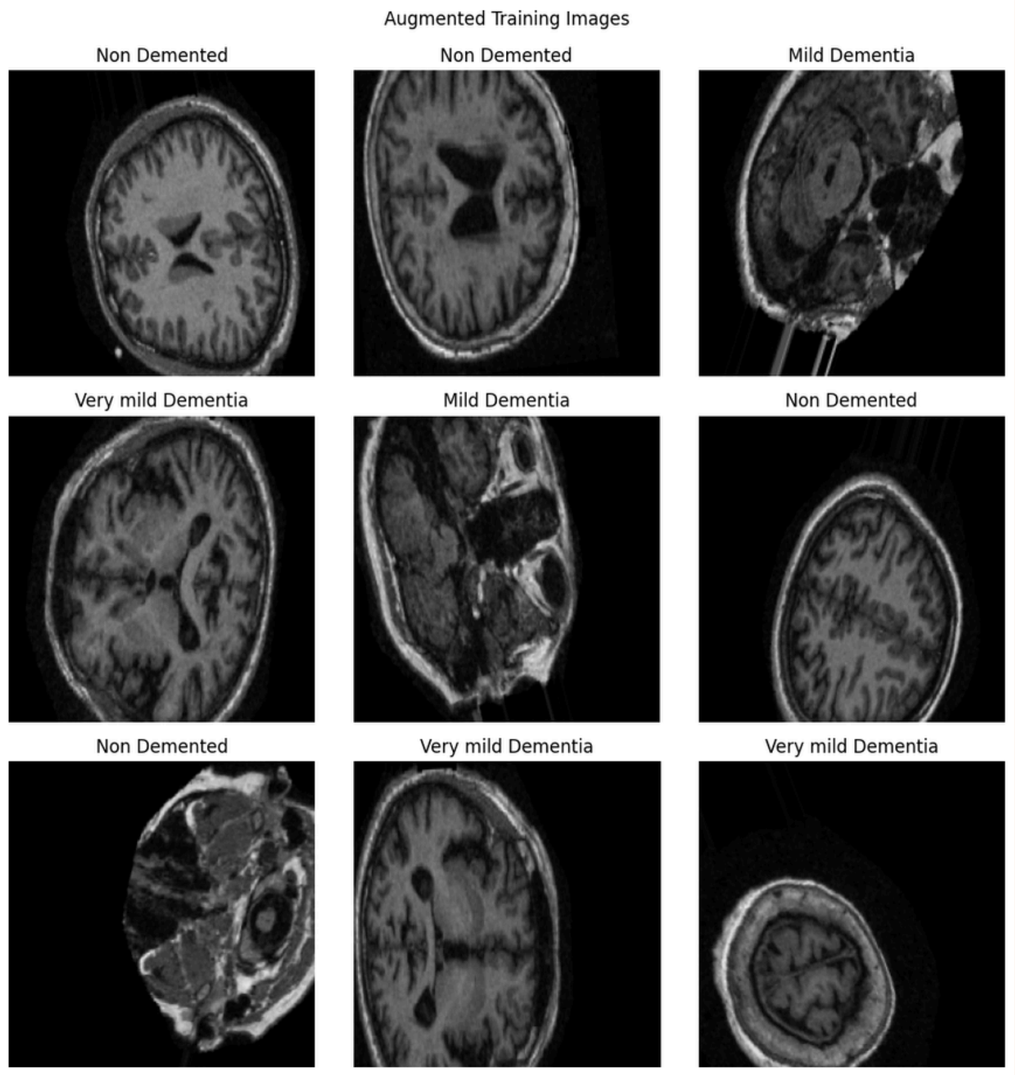
Checking for duplicates or corrupted images

Found 0 duplicate images

Found 0 corrupted images

Data preprocessing includes pixel normalization, image resizing (224x224) , and data augmentation through rotation, height shift, width shift, rotation, shear, zoom, horizontal flip,and vertical flip.

Visualizig images after data preprocessing (data augmentation)



Salma Wael 202201761

Model 1 :Dual-stream Convolutional Neural Network (CNN) with Residual Blocks

Purpose:

Improve feature extraction by processing the same input through two parallel CNN paths, then combining the outputs.

Structure :Two parallel CNN streams

- **Each stream includes :** 1 initial Conv layer
  - 3 Residual Blocks (ResBlocks) with skip connections
  - Global Average Pooling (GAP) after ResBlocks

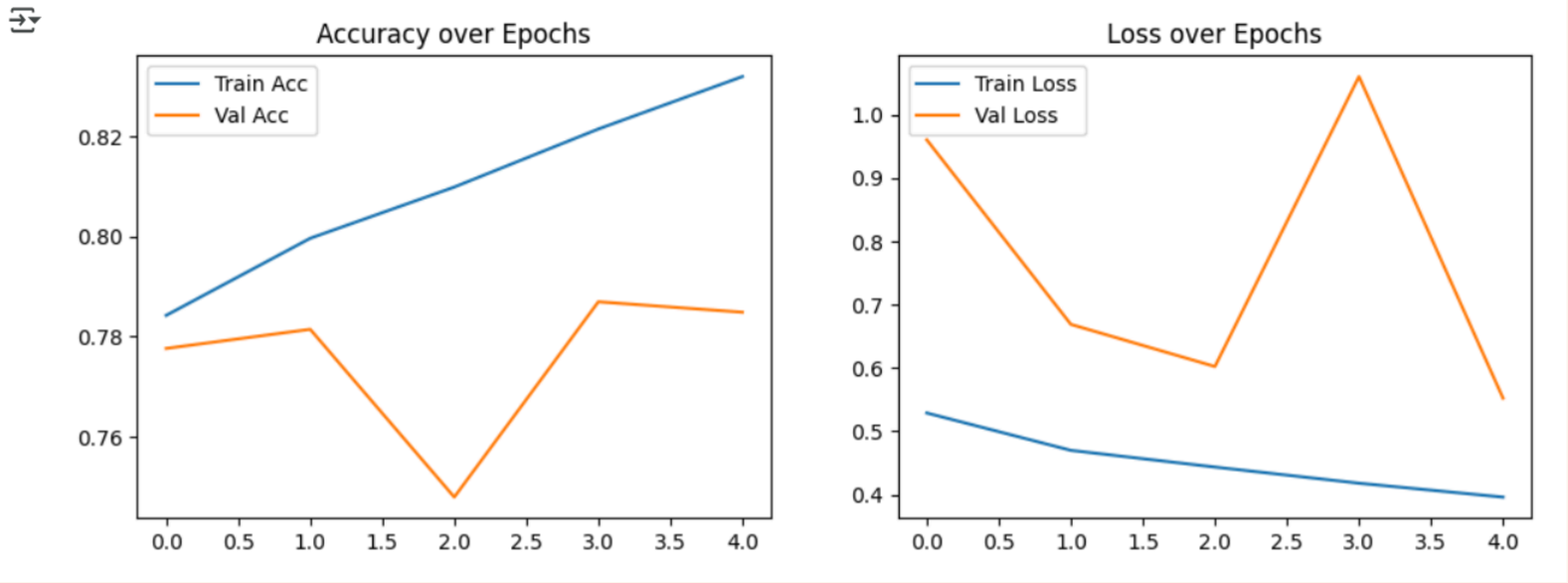
- **ResBlock structure:** 2 Conv layers with skip connection
- Kernel Sizes: 3×3
- Filters per ResBlock: ResBlock 1: 32 filters
  - ResBlock 2: 64 filters
  - ResBlock 3: 128 filters
  - Max Pooling: 2×2, stride 2×2
- Batch Normalization after Conv layers
- ReLU activation
- Concatenate both stream outputs
  - Dropout (0.5)
  - Fully Connected (FC) layer for final classification

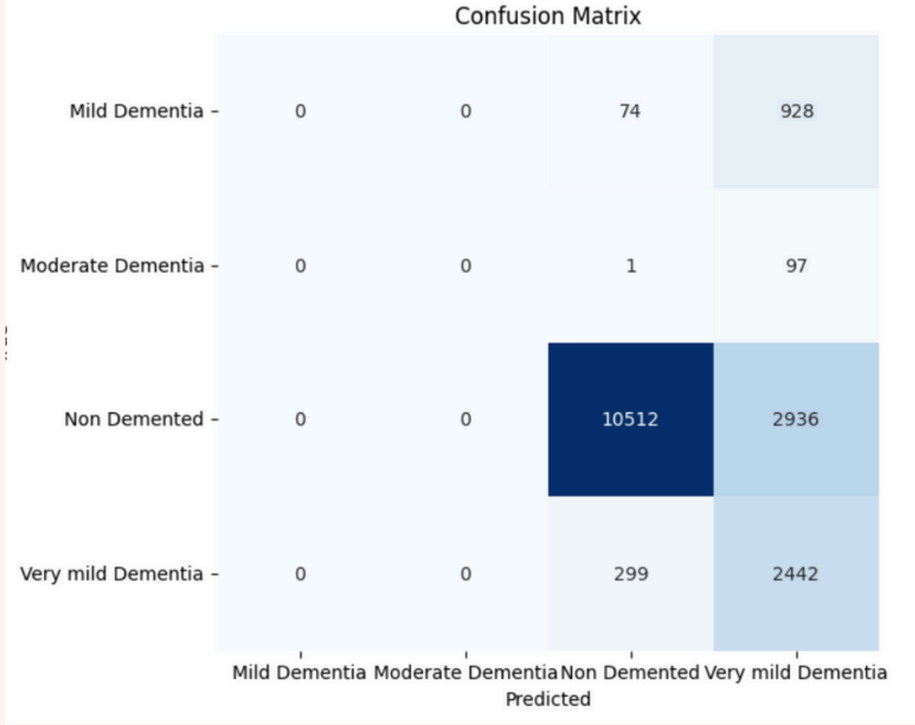
Salma Wael 202201761

Model 1 :Dual-stream Convolutional Neural Network (CNN) with Residual Blocks

Evaluation and Results :

```
ch 1/5
1/2161 ━━━━━━━━━━━━━━━━━━━━ 977s 442ms/step - accuracy: 0.7730 - loss: 0.5826 - val_accuracy: 0.7777 - val_loss: 0.9604
ch 2/5
1/2161 ━━━━━━━━━━━━━━━━━━━━ 981s 454ms/step - accuracy: 0.7972 - loss: 0.4783 - val_accuracy: 0.7815 - val_loss: 0.6688
ch 3/5
1/2161 ━━━━━━━━━━━━━━━━━━━━ 982s 455ms/step - accuracy: 0.8096 - loss: 0.4483 - val_accuracy: 0.7480 - val_loss: 0.6022
ch 4/5
1/2161 ━━━━━━━━━━━━━━━━━━━━ 954s 442ms/step - accuracy: 0.8212 - loss: 0.4217 - val_accuracy: 0.7870 - val_loss: 1.0603
ch 5/5
1/2161 ━━━━━━━━━━━━━━━━━━━━ 951s 440ms/step - accuracy: 0.8266 - loss: 0.4056 - val_accuracy: 0.7849 - val_loss: 0.5522
```



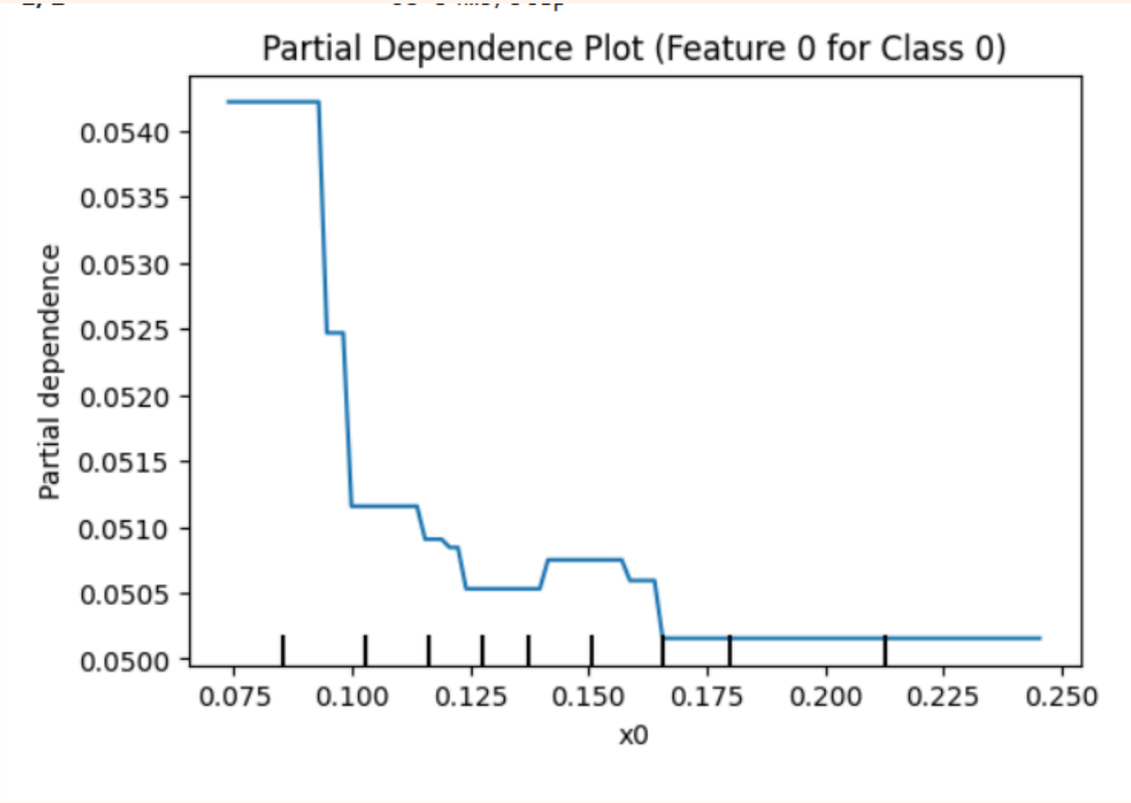


Salma Wael 202201761

Model 1 :Dual-stream Convolutional Neural Network (CNN) with Residual Blocks

Explainability Techniques :

1) PDP :



Interpretation:

Low values of Feature 0 contribute more to Class 0 predictions.

As Feature 0 increases, the model becomes less confident that the sample belongs to Class 0.

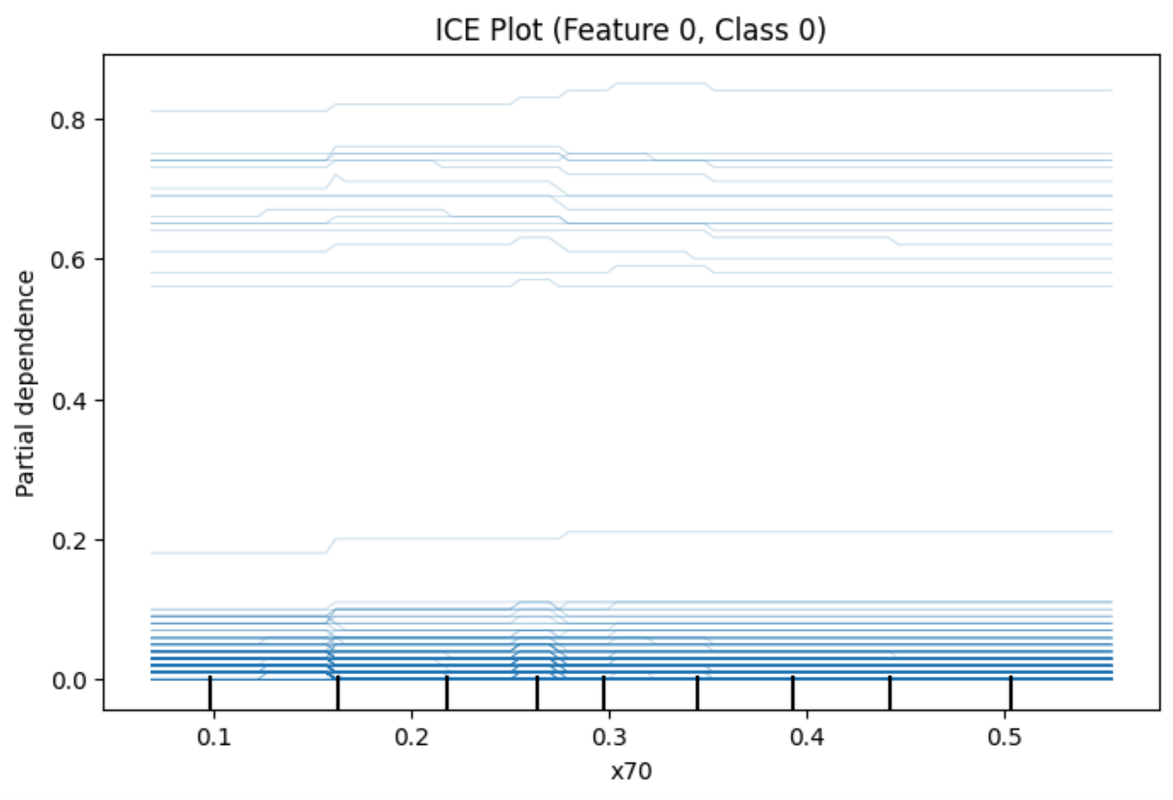
This is consistent with a threshold effect: small changes in this feature around a critical value (≈ 0.10–0.15) significantly affect the model's output.

Salma Wael 202201761

Model 1 :Dual-stream Convolutional Neural Network (CNN) with Residual Blocks

Explainability Techniques :

2) ICE :



Interpretation:

Many of the lines are relatively flat, meaning that changing Feature 0 does not significantly alter the model's output for those samples.
This suggests Feature 0 has limited importance in those cases.

Distinct Behavior in Some Cases:

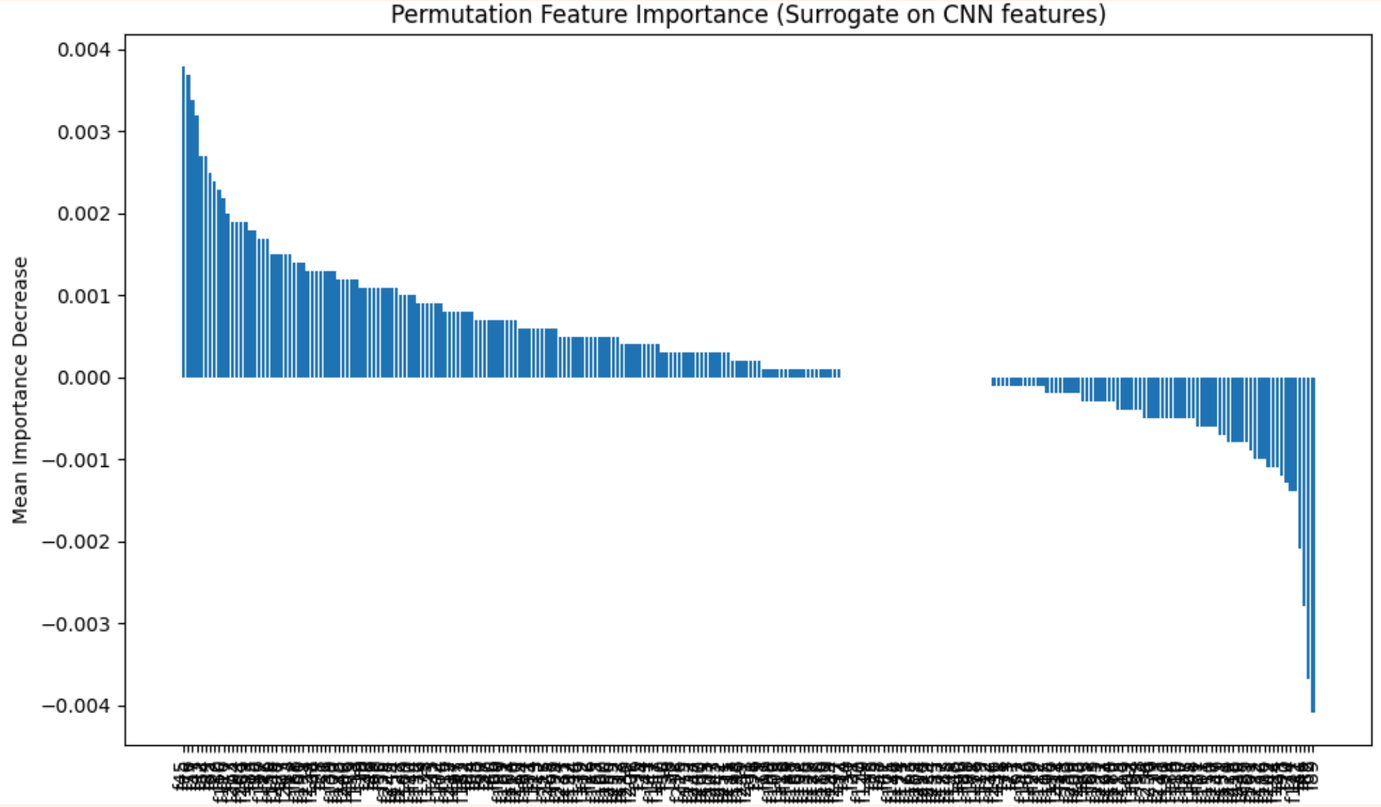
A few lines have slight upward or downward trends.
These may represent sensitive samples where Feature 0 does affect the prediction.
But the magnitude of change is relatively small.

Salma Wael 202201761

Model 1 :Dual-stream Convolutional Neural Network (CNN) with Residual Blocks

Explainability Techniques :

3) Permutation Feature Importance :



Interpretation:

Top Features (Left side of the plot):

The features on the far left cause the most significant performance drop when shuffled. These features are highly informative and most useful for the CNN-based model. Max importance is around 0.0037.

Middle Region:

Features near zero importance likely contribute little to the model. May represent redundant or less relevant extracted patterns.

Negative Importance (Right side):

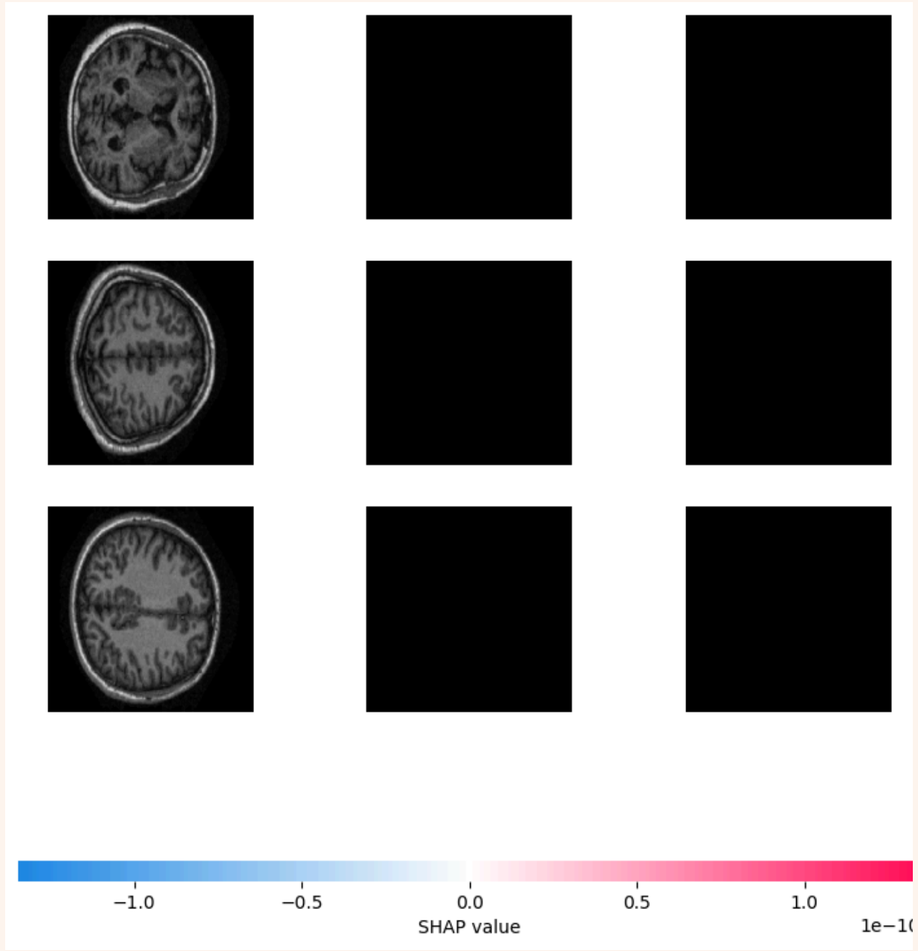
Features that caused a slight performance boost when shuffled. Indicates that the model might be overfitting to noise in these features or they are confusing.

Salma Wael 202201761

Model 1 :Dual-stream Convolutional Neural Network (CNN) with Residual Blocks

Explainability Techniques :

3) Shap :



Interpretation:

The SHAP overlays (middle and right columns) are completely black.
This means that the SHAP values are zero or extremely close to zero across all pixels.
Supporting evidence: The color bar shows an extremely narrow range: [-1.0e-10, 1.0e-10].

Salma Wael 202201761

Model 2 : XGBoost Classifier with VGG16 Features

Summary:

- Used VGG16 as a feature extractor combined with an XGBoost classifier, bridging deep learning with gradient boosting for classification.

Structure :

- Extract features from the images using a pre-trained CNN (like VGG16, ResNet50, etc.).
- Flatten those features into 1D vectors.
- Use XGBoost to classify based on those features.

Results:

The VGG16 + XGBoost model offered **strong results** by leveraging pre-trained feature extraction. The combination of deep learning for feature extraction (VGG16) and XGBoost for classification **balances predictive power and explainability.**

```
[ ]  # 5. Evaluation
     y_pred = clf.predict(X_test)

     class_indices = train_generator.class_indices
     idx_to_class = {v: k for k, v in class_indices.items()}

     print("Accuracy:", accuracy_score(y_test, y_pred))
```

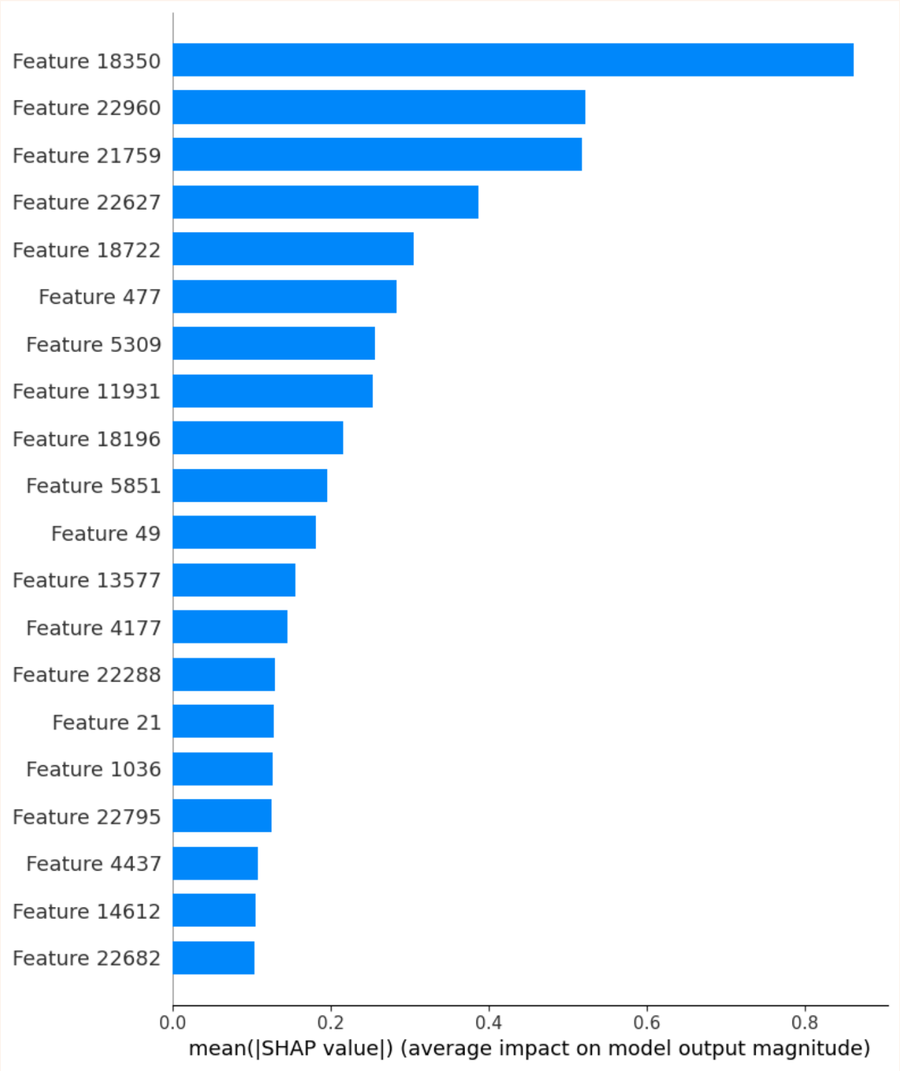
```
Accuracy: 0.95
```

Salma Wael 202201761

Model 2 : XGBoost Classifier with VGG16 Features

Explainability Techniques :

1) Shap :



Interpretation:

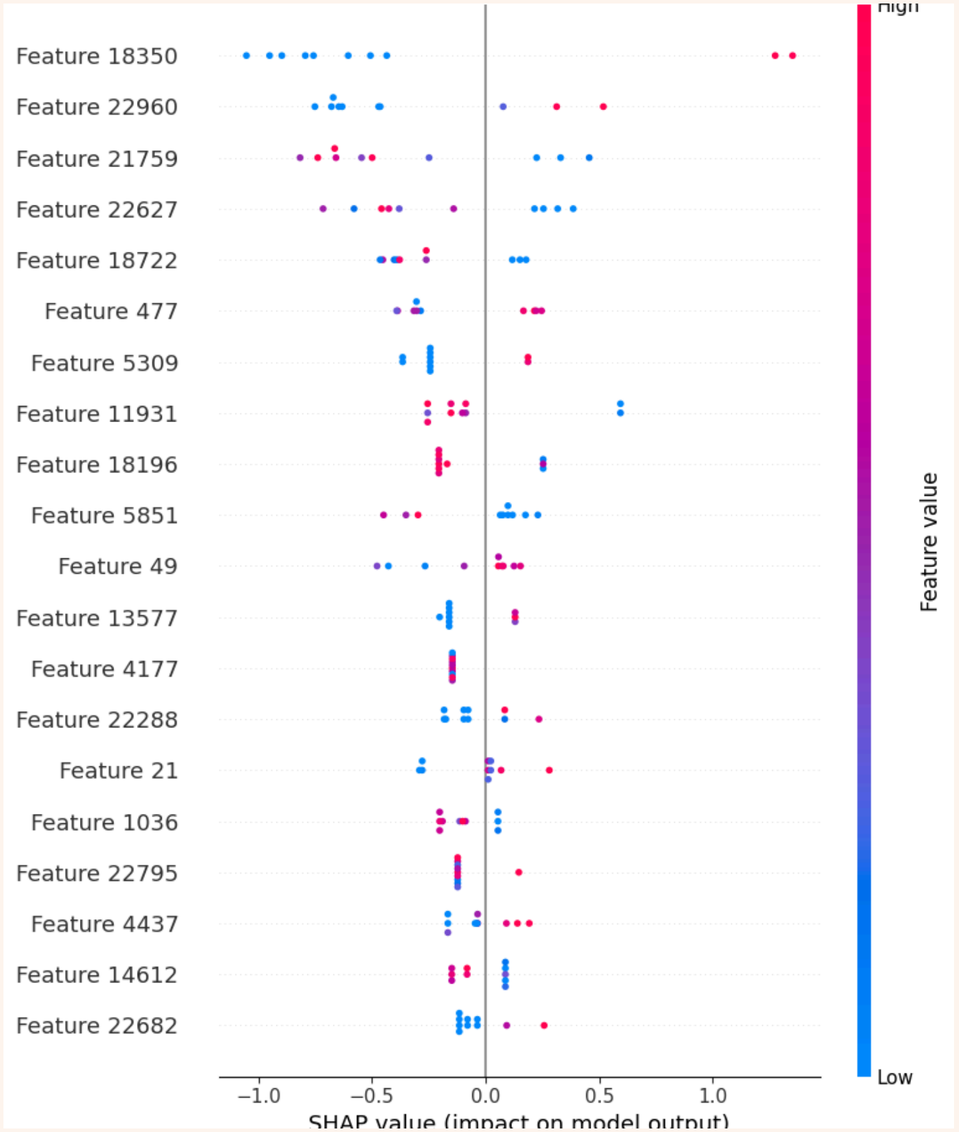
- Feature 18350 has the highest average impact on the model’s predictions. This means it is the most influential VGG16-derived feature in deciding the output class.
- Other important features include 22960, 21759, and 22627, though each has less impact than Feature 18350.
- The steep drop from the top 3 features suggests that a small subset of CNN-derived features dominate the model's decision-making process.
- The lower-ranked features still contribute but with significantly lower average impact.

Salma Wael 202201761

Model 2 : XGBoost Classifier with VGG16 Features

Explainability Techniques :

1) Shap



Interpretation:

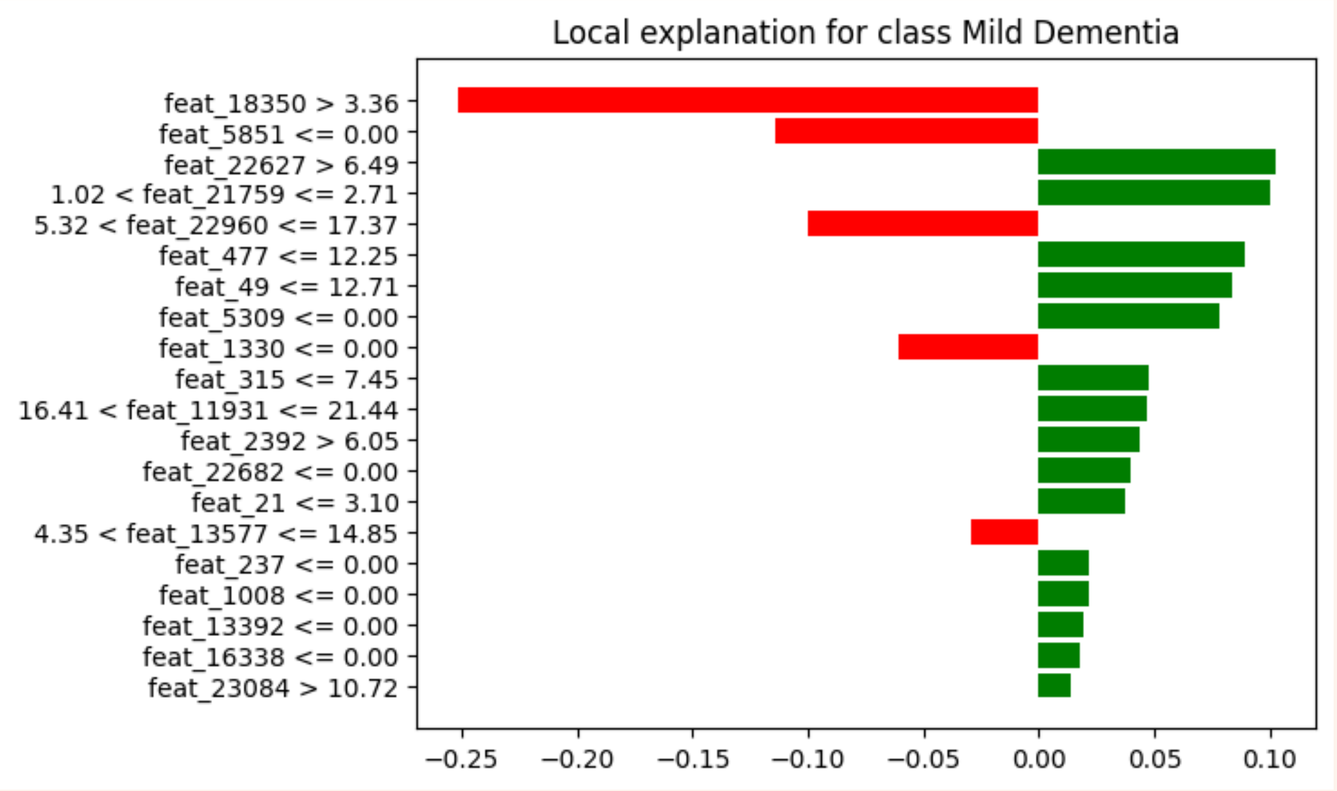
1. Feature 18350: Mostly low (blue) feature values → negative SHAP values → decrease model output.
A few high (red) values → positive SHAP values → increase output.
Strong directional influence, making it the most impactful feature overall.
2. Feature 21759: Shows a mixed pattern – both high and low values lead to positive/negative impacts.
This suggests a nonlinear or context-dependent effect on predictions.
3. Feature 22627: High values (red) tend to increase the output (positive SHAP values).
Low values (blue) tend to decrease output.
A more monotonic relationship is implied.

Salma Wael 202201761

Model 2 : XGBoost Classifier with VGG16 Features

Explainability Techniques :

2) Lime :



Top Influential Features in this Prediction:

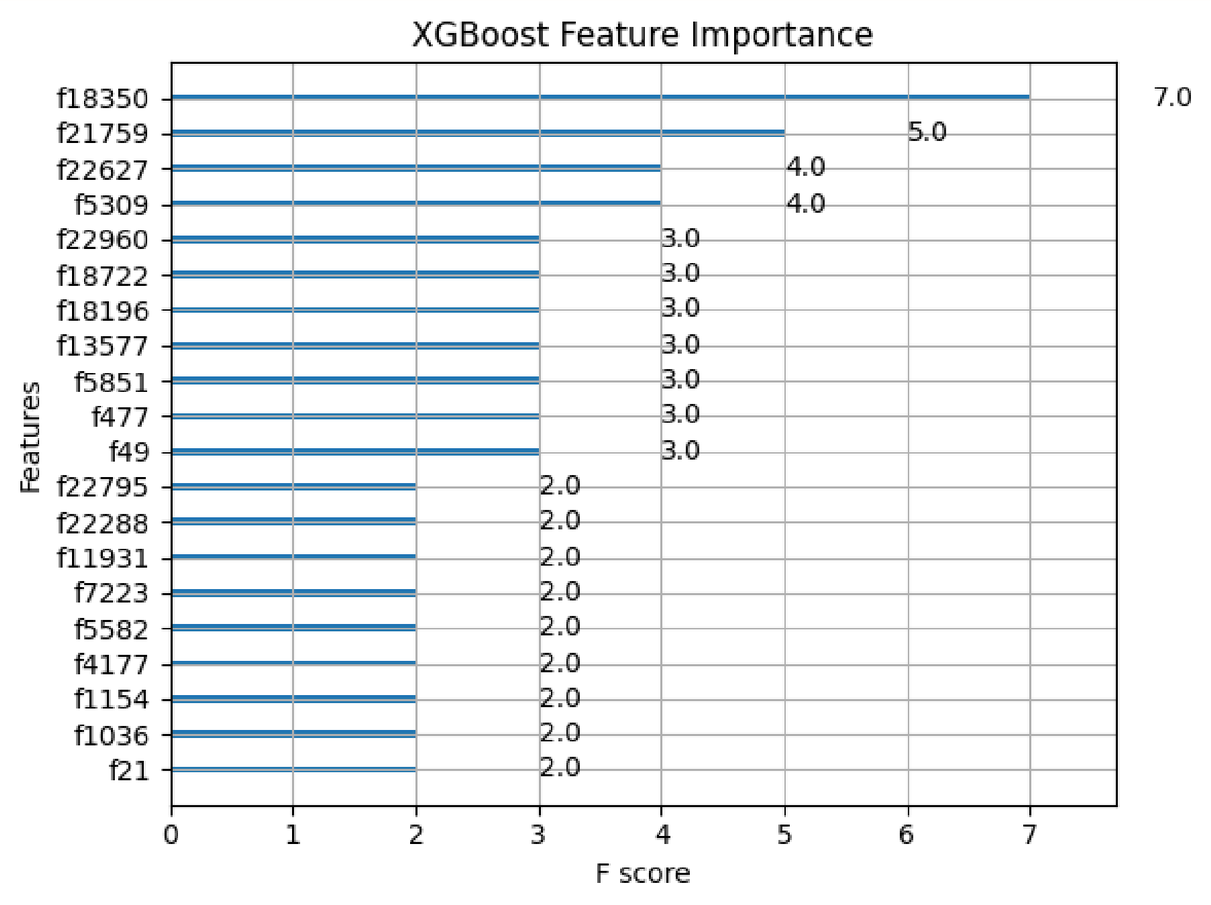
1. **feat_18350 > 3.36** : Strongest negative influence (~ -0.25) This pushed the model away from predicting "Mild Dementia".
2. **feat_22627 > 6.49 and 1.02 < feat_21759 ⩽ 2.71** Strongest positive influences, both contributed around +0.09 to the prediction.
3. **feat_22960 in range 5.32–17.37** : Negative influence – decreased the probability of "Mild Dementia".
4. **Other green bars (like feat_477, feat_5309, feat_11931 etc.)** : These had minor positive contributions to the "Mild Dementia" class.

Salma Wael 202201761

Model 2 : XGBoost Classifier with VGG16 Features

Explainability Techniques :

3) Feature Importance :



Interpretations :

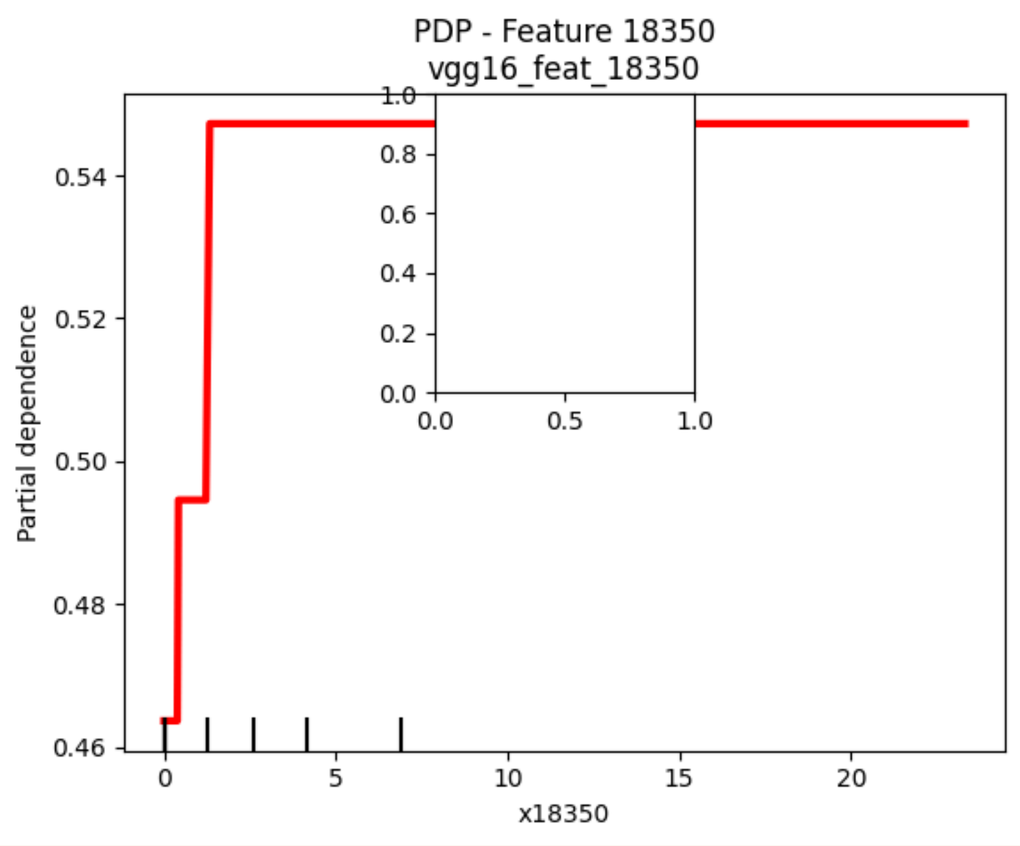
f18350 **is a key latent feature(Most important feature) in model split especially around detecting Mild Dementia.**

Salma Wael 202201761

Model 2 : XGBoost Classifier with VGG16 Features

Explainability Techniques :

4) PDP :



Interpretations :

- Feature 18350 is a highly influential CNN feature.
- Small increases in its value result in a nonlinear boost in the model's prediction for the target class.
- The model behaves almost like a threshold function around value 1–2 for this feature.
- Strong alignment with what the SHAP and tree visualizations showed: this feature is a key decision trigger.

Salma Wael 202201761

Model 3 : Basic CNN architecture

Summary:

A basic CNN architecture was used to classify the same image dataset.
This served as a baseline comparison to more complex or hybrid models.

Structure:

Layer	Type	Filters/Units	Kernel Size	Activation	Output Shape
1	Conv2D	32	(3×3)	ReLU	(224, 224, 32)
2	MaxPooling2D	-	(2×2)	-	(112, 112, 32)
3	Conv2D	64	(3×3)	ReLU	(110, 110, 64)
4	MaxPooling2D	-	(2×2)	-	(55, 55, 64)
5	Conv2D	128	(3×3)	ReLU	(53, 53, 128)
6	MaxPooling2D	-	(2×2)	-	(26, 26, 128)
7	Flatten	-	-	-	(86528,)
8	Dense	128	-	ReLU	(128,)
9	Dropout	5	-	-	(128,)
10	Dense	4	-	Softmax	(4,)

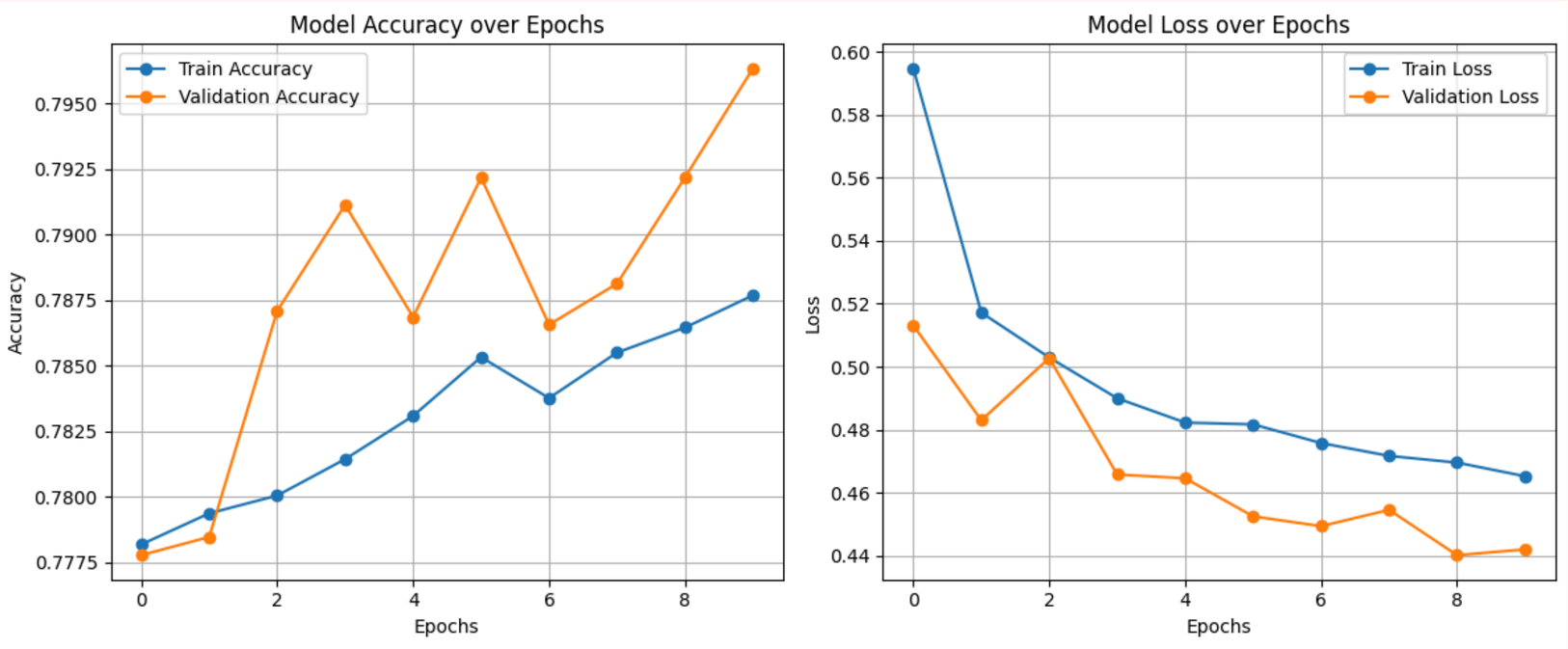
Salma Wael 202201761

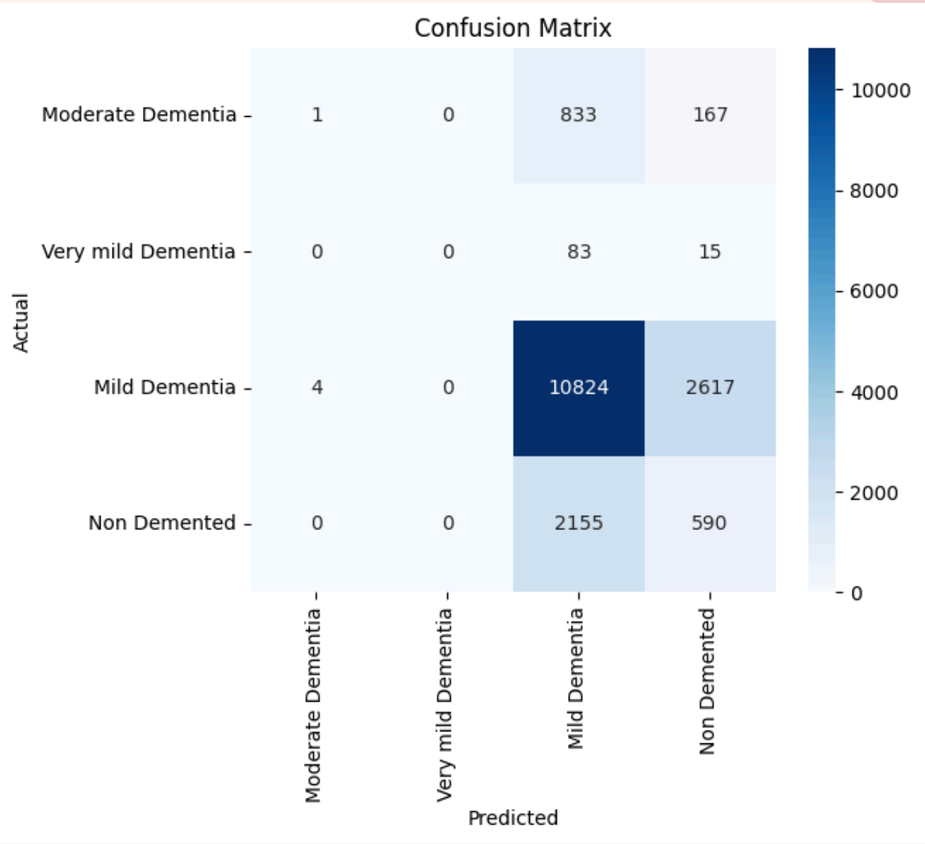
Model 3 : Basic CNN architecture

Evaluation and Results :

```
4323/4323 ━━━━━━━━━━━━━━━━━━━━ 34s 8ms/step - accuracy: 0.7936 - loss: 0.4446

Test Accuracy: 79.63%
Test Loss: 0.4420
```



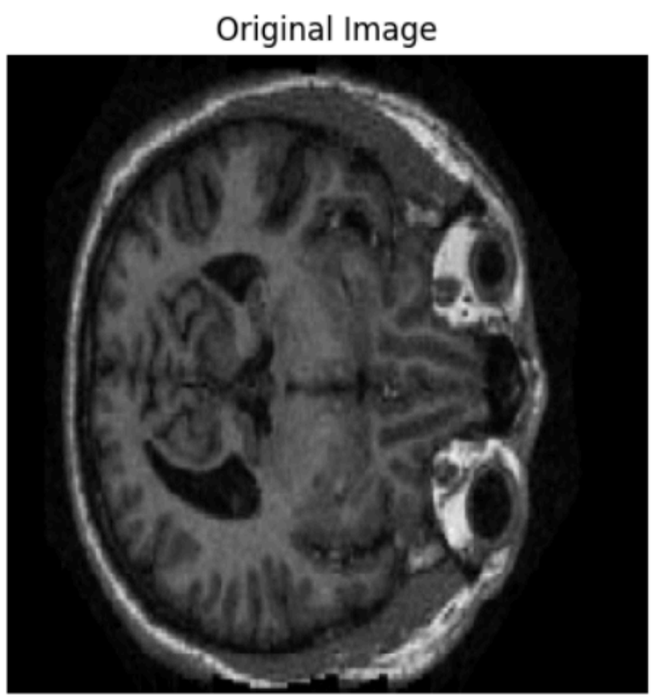


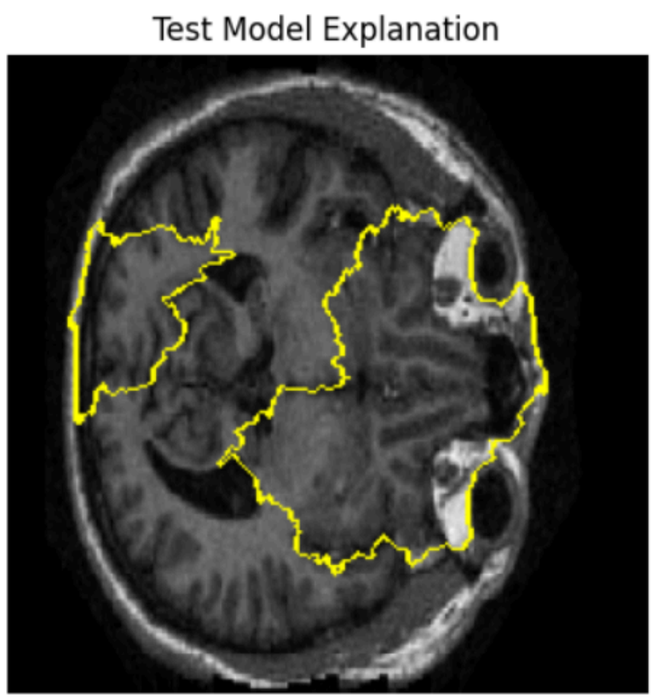
Salma Wael 202201761

Model 3 : Basic CNN architecture

Explainability Techniques :

1) Lime :





LIME worked with test model!

Interpretation:

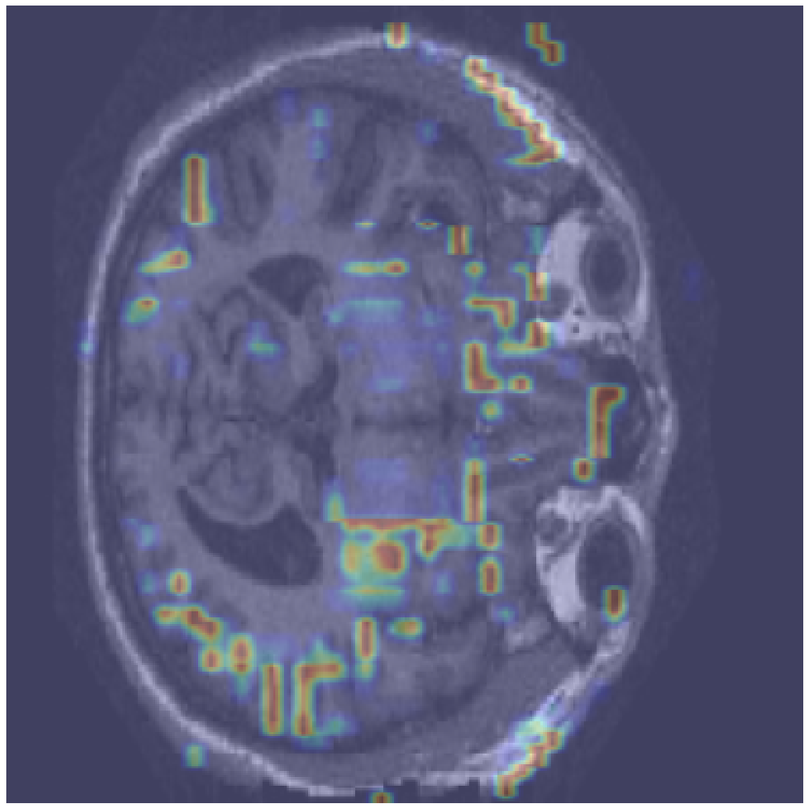
This is the same image with LIME
Yellow boundaries outline the superpixels (image patches) that contributed most to the prediction.
These highlighted regions are locally important—they indicate what the model "focused on" to make its classification for this particular image.

Salma Wael 202201761

Model 3 : Basic CNN architecture

Explainability Techniques :

2) GradCam :



Interpretation:

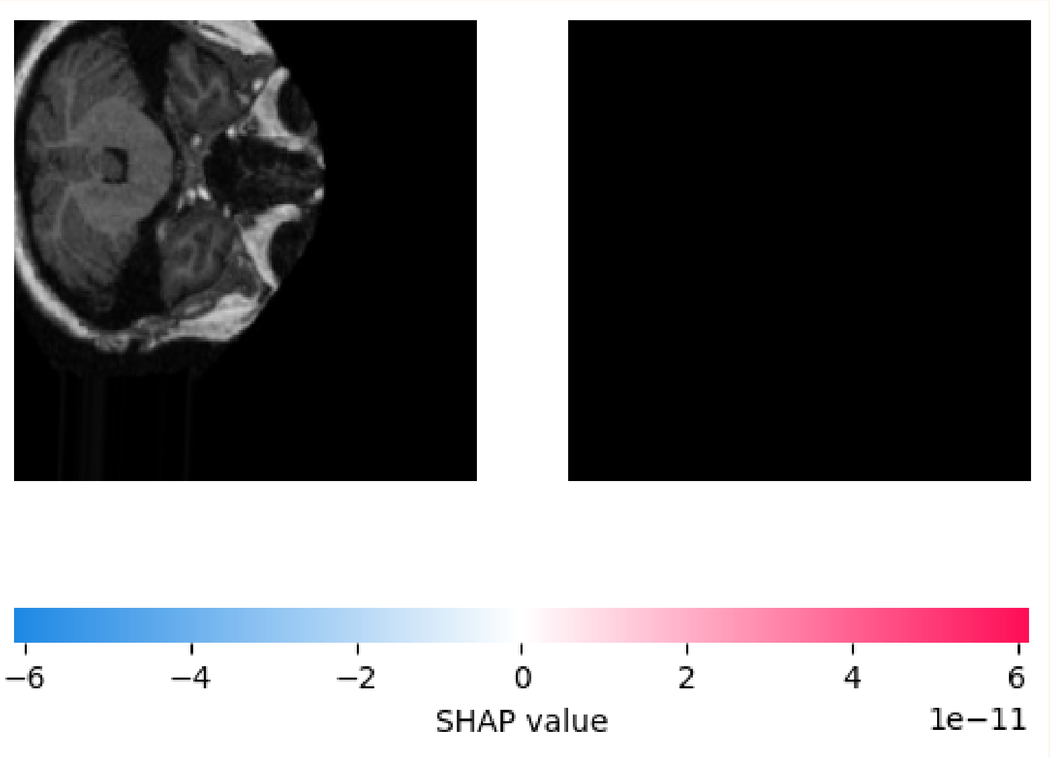
- The highlighted regions (reds/yellows) are the highly influenced in the model's decision that the CNN model relies on for classification .

Salma Wael 202201761

Model 3 : Basic CNN architecture

Explainability Techniques :

3) Shap :



Interpretation:

The SHAP overlays (middle and right columns) are completely black.
This means that the SHAP values are zero or extremely close to zero across all pixels.

model 1 overview

model type:

Custom CNN using Keras Functional API deep learning model capable of accurately classifying brain MRI images into tumor categories.

model architecture:

- Conv Layer 1:
- 32 filters, 3x3 kernel, ReLU activation
- Followed by 2x2 MaxPooling and Dropout (rate: 0.3)
- Conv Layer 2:
- 64 filters, 3x3 kernel, ReLU activation
- Followed by 2x2 MaxPooling and Dropout (rate: 0.3)
- Flatten Layer:
- Converts feature maps into a 1D vector
- Fully Connected Layer:
- Dense layer with 128 units, ReLU activation
- Dropout (rate: 0.5) to prevent overfitting
- Output Layer:
- Dense layer with softmax activation for multi-class classification

evaluation:

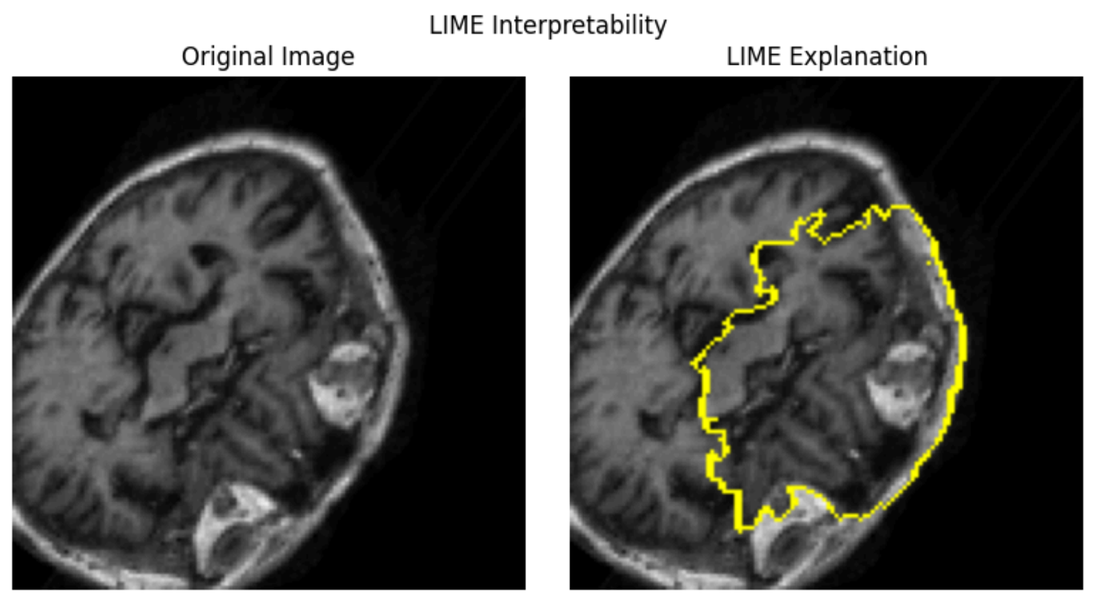
Final Validation Results:

- Validation Loss: 0.7745
- Validation Accuracy: 0.6403

model 1 overview

explainability techniques:

1)Lime :



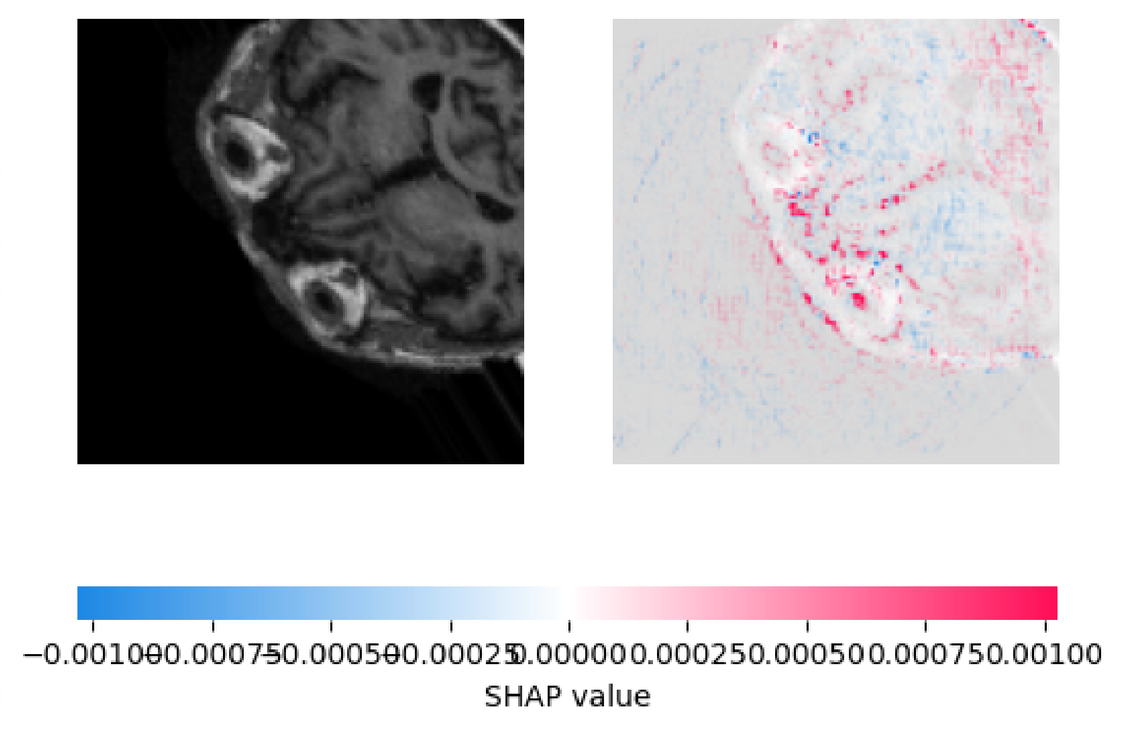
interpretation:

The LIME Explanation overlays a yellow boundary on the MRI, indicating the superpixels (regions) that most influenced the CNN’s decision.which could correspond to a tumor or abnormal region.here the right lower part is highlighted

model 1 overview

explainability techniques:

2)Shap :



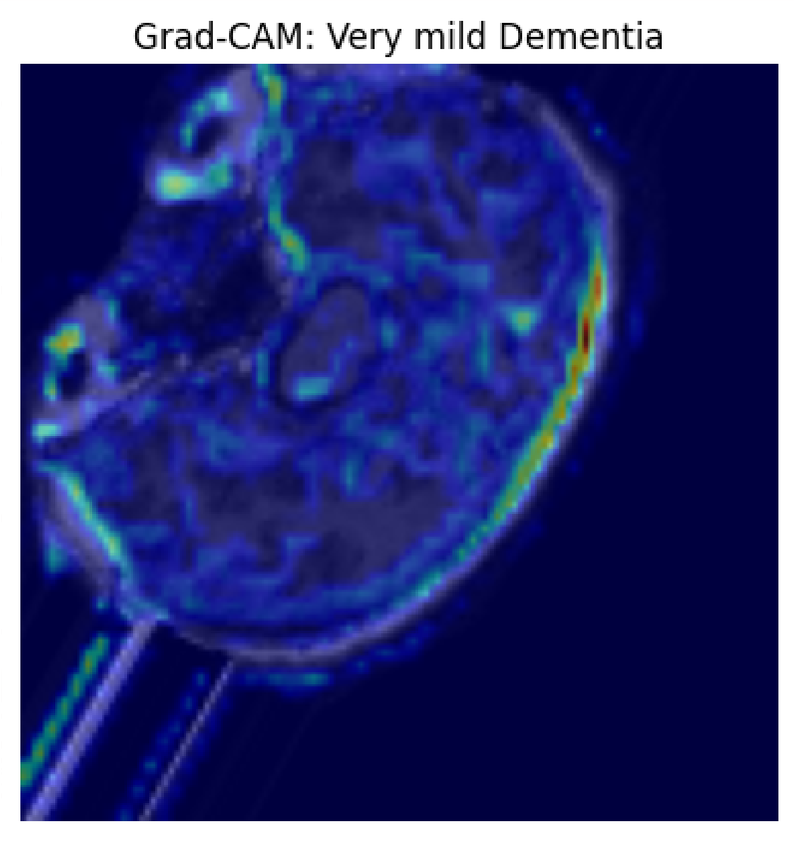
interpretation:

Red areas → positive contribution to predicting tumor
Blue areas → negative contribution (against tumor) tells us which pixels (or regions) in the image most influenced the model's prediction.

model 1 overview

explainability techniques:

3)Grad-Cam :



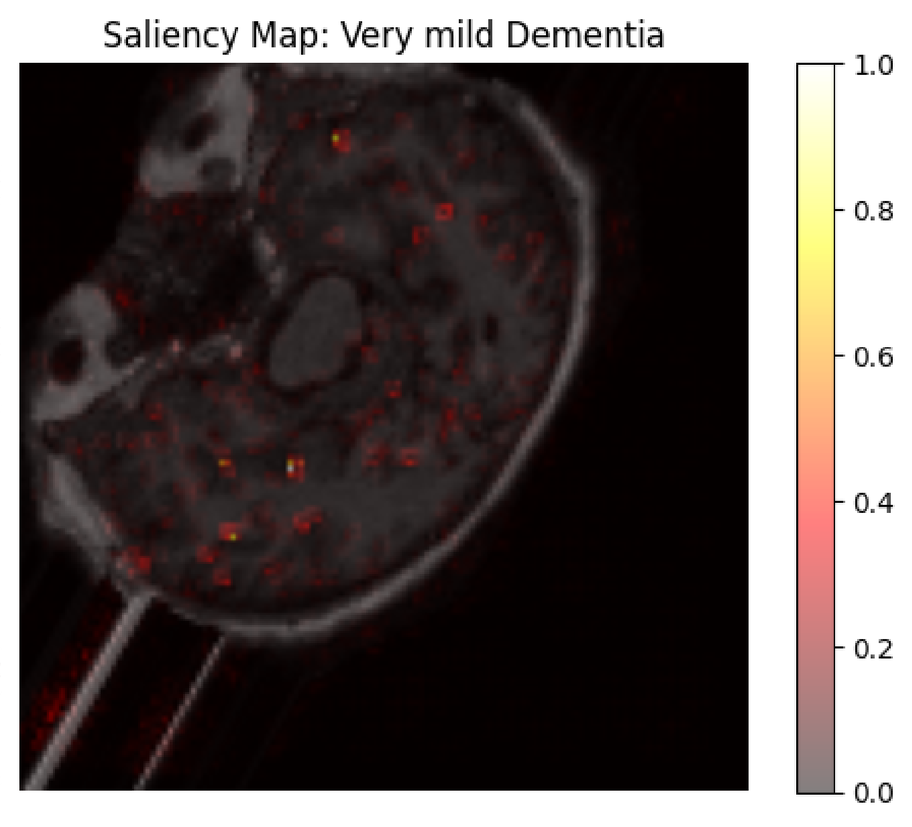
interpretation:

Warm colors (yellow/red) indicate high importance for the model’s prediction, while cooler colors (blue) show less important areas.In this case, the CNN focused on certain brain regions when predicting very mild dementia

model 1 overview

explainability techniques:

4)Saliency map :



interpretation:

Brighter (yellowish) areas indicate higher saliency—the model relies more heavily on those pixels.Darker regions mean those pixels had little influence.

model 2 overview

model type:

Deeper CNN with Batch Normalization a more stable and deeper model for tumor classification by integrating Batch Normalization and the SGD optimizer

model architecture:

Conv2D (32 filters) → ReLU → BatchNorm → MaxPooling
Conv2D (64 filters) → ReLU → BatchNorm → MaxPooling
Conv2D (128 filters) → ReLU → BatchNorm → MaxPooling
Flatten
Dense (128 units, ReLU)
Dropout (50%)
Output Layer (Softmax) → Multi-class prediction

evaluation:

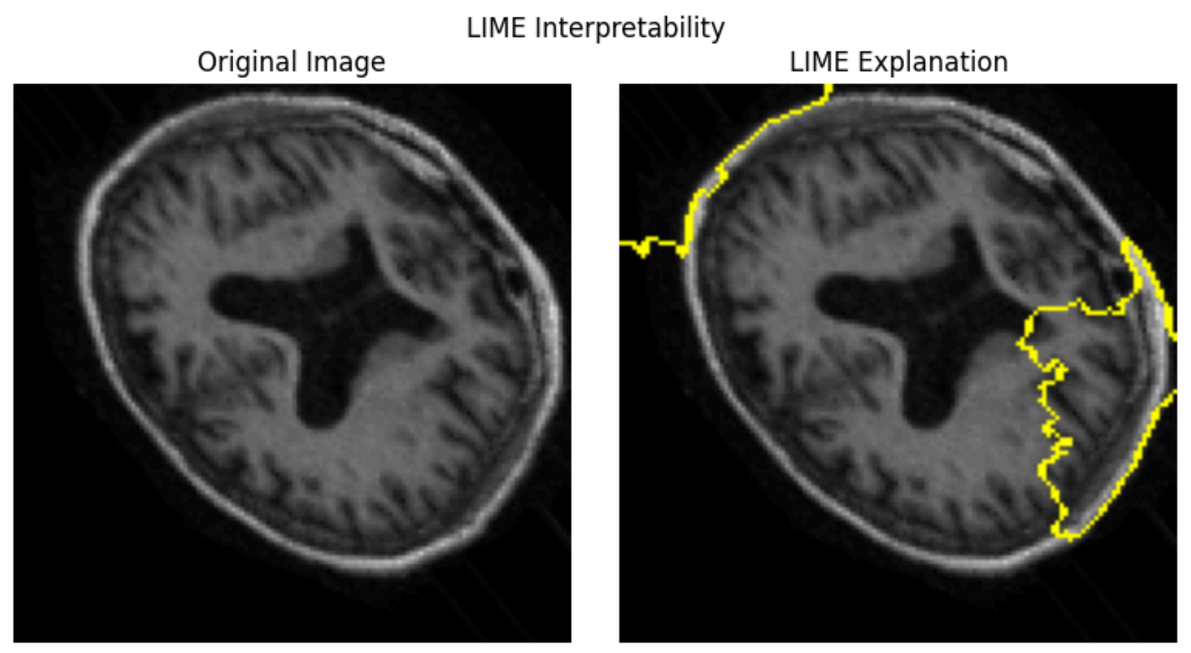
Final Validation Results:

- Validation Loss: 1.0275
- Validation Accuracy: 0.5105

model 2 overview

explainability techniques:

1)Lime :



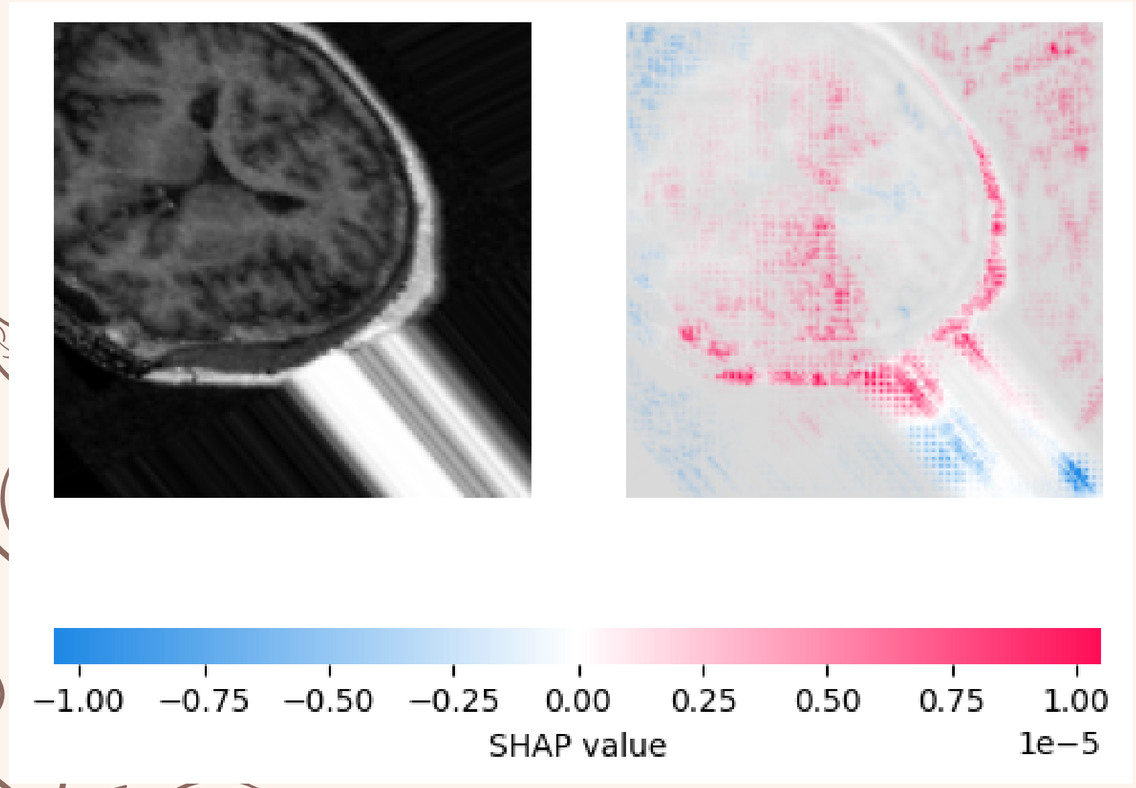
interpretation:

The LIME Explanation overlays a yellow boundary on the MRI, indicating the superpixels (regions) that most influenced the CNN’s decision.which could correspond to a tumor or abnormal region.here the right lower part is highlighted

model 2 overview

explainability techniques:

2)Shap :



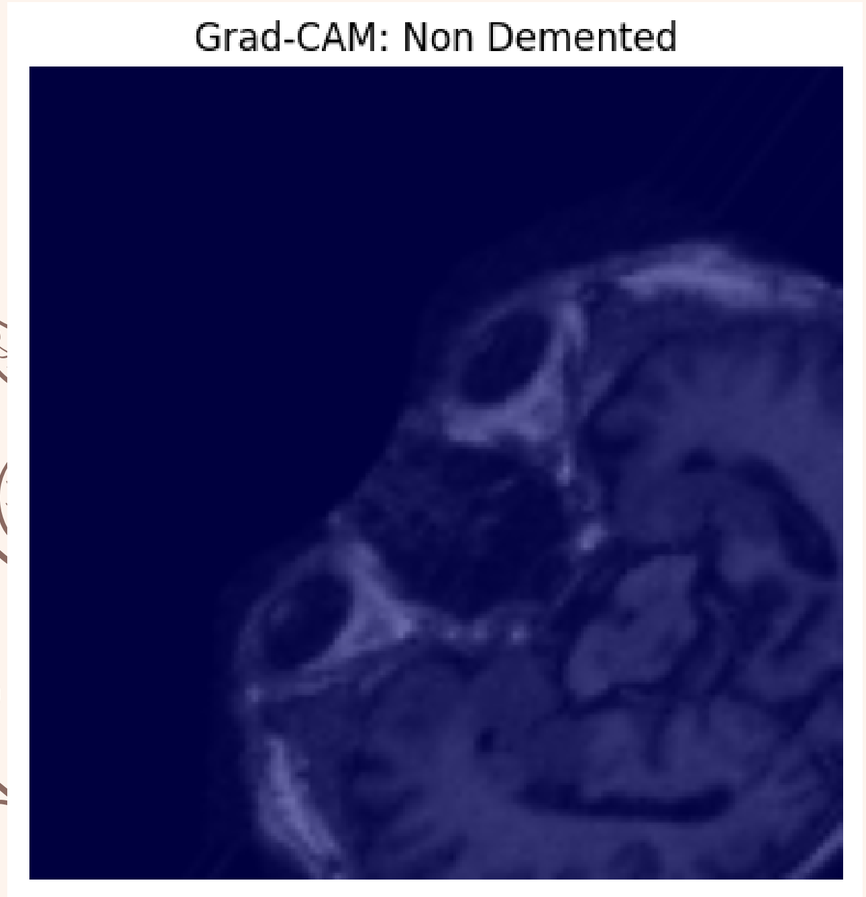
interpretation:

Red areas → positive contribution to predicting a tumor
Blue areas → negative contribution (against tumor)
Values range from -0.1 to +0.1 SHAP value (approx.)

model 2 overview

explainability techniques:

3)Grad-Cam :



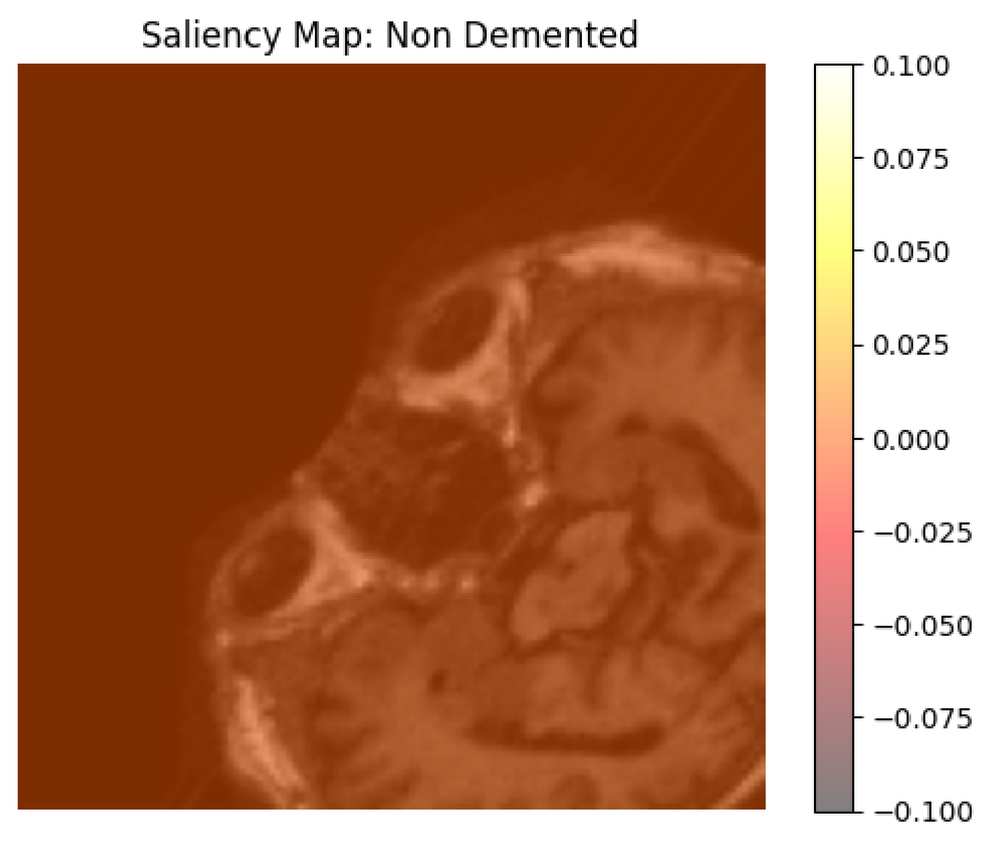
interpretation:

This heatmap shows minimal activation across the image.The almost entirely dark blue image suggests that no specific region was strongly activated by the model in deciding the label Non Demented.

model 2 overview

explainability techniques:

4)Saliency map :



interpretation:

The saliency map is low-contrast, with no significant red/yellow highlights. This indicates that no particular pixels had a strong influence on the prediction

model 3 overview

model type:

Deep CNN with Fully Connected Layers that emphasizes learning high-level features via stacked fully connected layers

model architecture:

Layers Summary:

- Conv2D (32 filters, ReLU) → MaxPooling
- Conv2D (64 filters, ReLU) → MaxPooling
- Conv2D (128 filters, ReLU) → MaxPooling
- Flatten Layer

Dense Layers:

- Dense (128 units, ReLU) → Dropout (50%)
- Dense (64 units, ReLU) → Dropout (50%)
- Dense (32 units, ReLU) → Dropout (50%)
- Output Layer: Softmax → Multiclass tumor prediction

evaluation:

Final Validation Results:

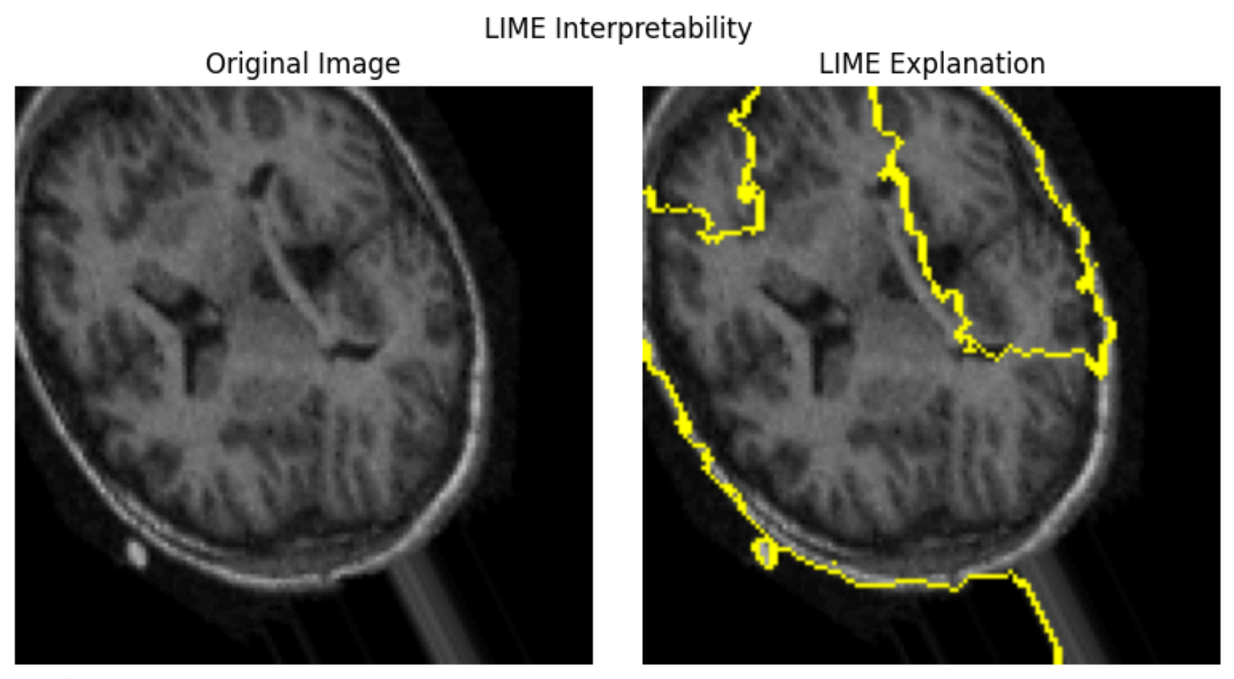
- Validation Loss: 0.7454
  Validation Accuracy: 0.6688

Confusion Matrix: [[0 0 390 610] [0 0 45 52] [0 0 1710 2290] [0 0 1196 1549]]

model 3 overview

explainability techniques:

1)Lime :



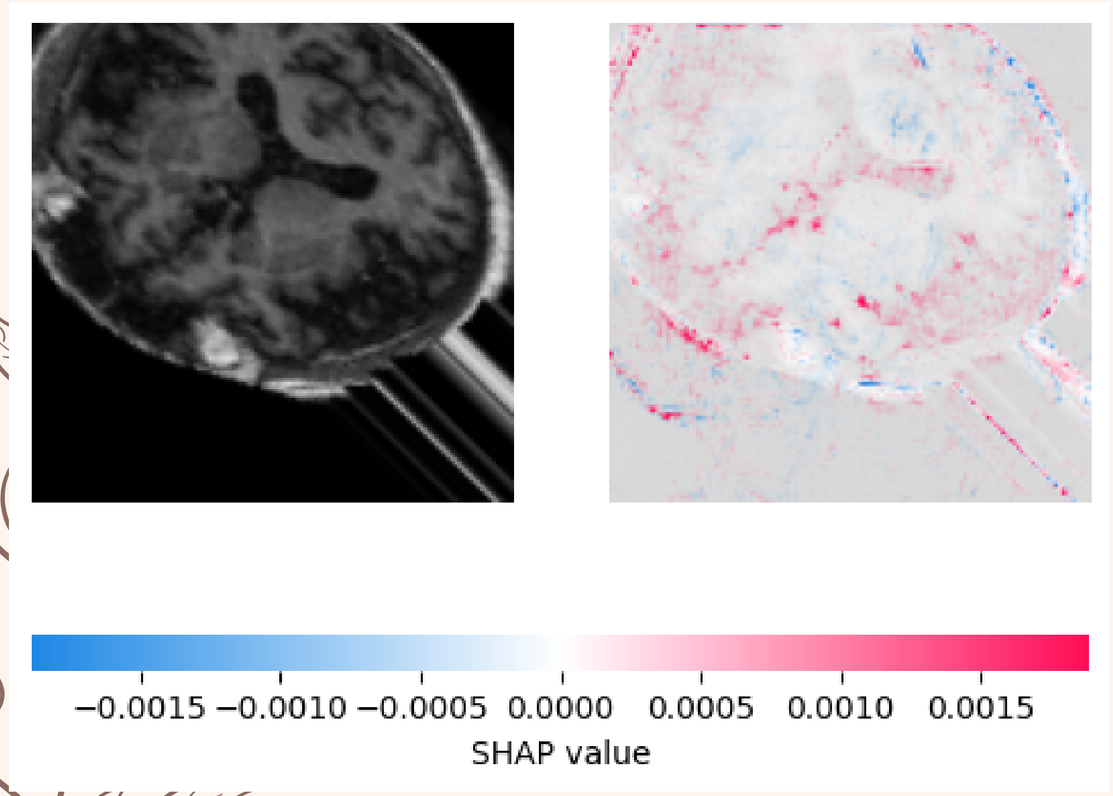
interpretation:

The LIME Explanation overlays a yellow boundary on the MRI, indicating the superpixels (regions) that most influenced the CNN’s decision.which could correspond to a tumor or abnormal region.here some upper and lower parts are highlighted

model 3 overview

explainability techniques:

2)Shap :



interpretation:

Red areas → positive contribution to predicting a tumor

Blue areas → negative contribution (against tumor)

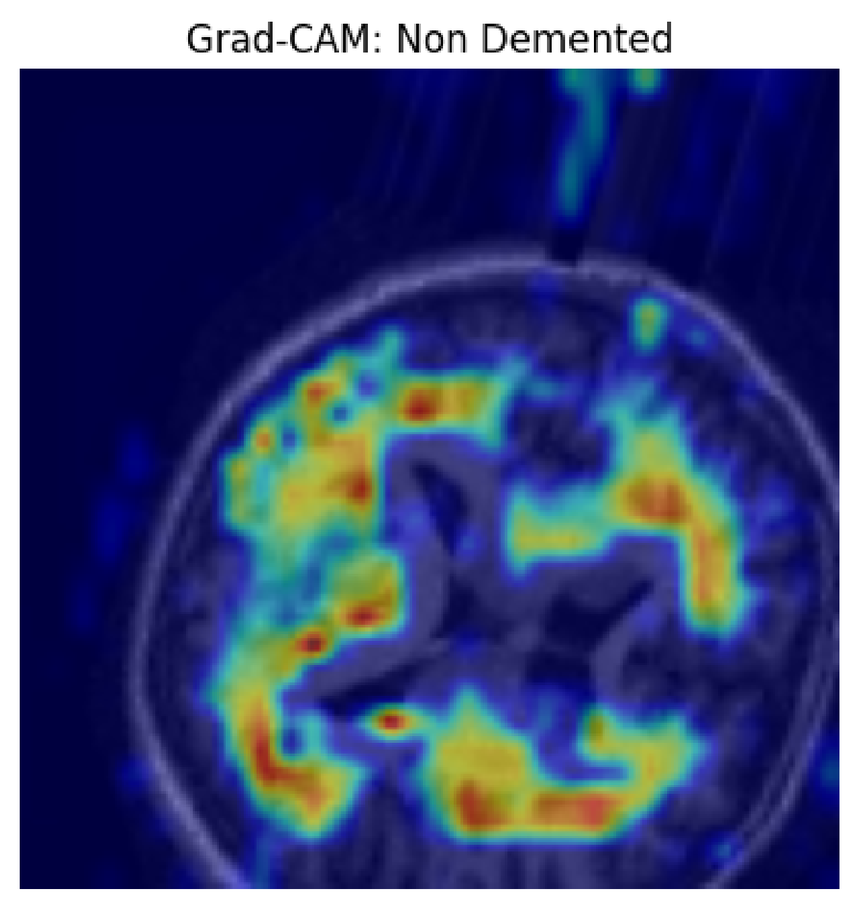
Values range from -0.1 to +0.1

SHAP value (approx.)

model 3 overview

explainability techniques:

3)Grad-Cam :



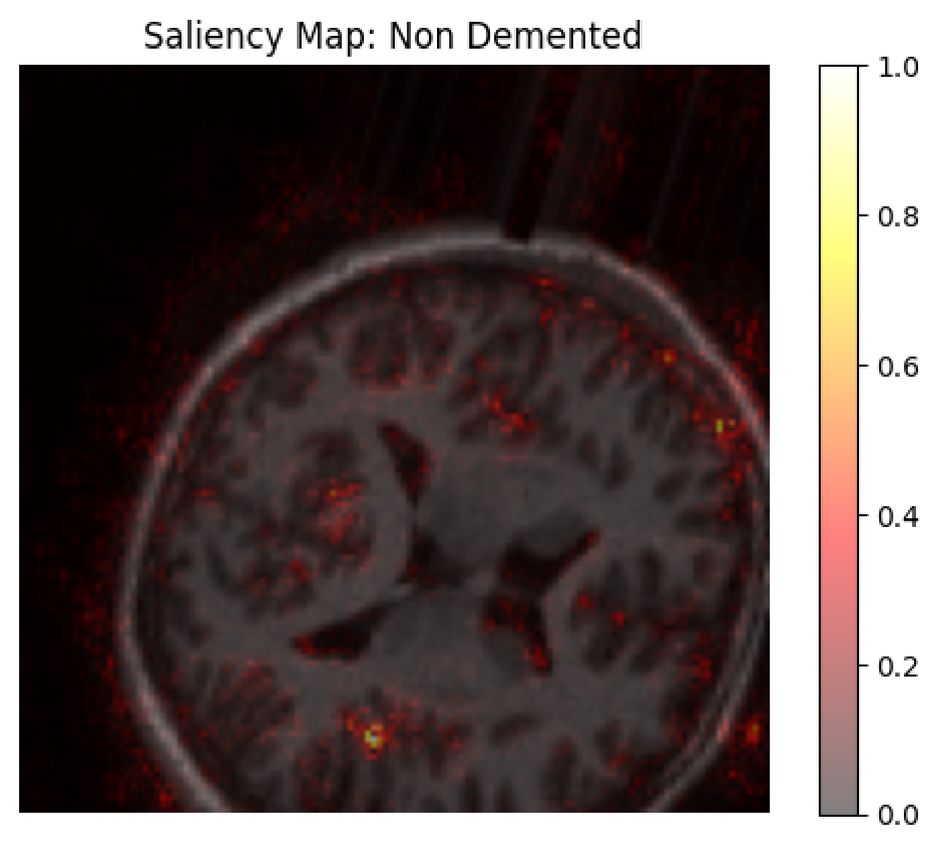
interpretation:

Red and Yellow areas: Highly activated—these are regions the model paid most attention to. Blue and Dark areas: Low or no activation—areas the model largely ignored

model 3 overview

explainability techniques:

4)Saliency map :



interpretation:

The red and yellow areas show where the model’s prediction of "non demented" is most sensitive to changes.
Brighter (yellowish) areas indicate higher saliency—the model relies more heavily on those pixels.Darker regions mean those pixels had little influence.

Paper 1 overview

“Alzheimer’s Disease Prediction and Classification Using CT Images Through Machine Learning“

This paper emphasizes the early detection of Alzheimer‘s Disease using CT images through machine learning techniques. The authors propose a system that combines the Visual Geometry Group (VGG)-16 and an Improved Faster Recurrent Convolutional Neural Network (IFRCNN) for feature extraction and classification. The study utilizes the Alzheimer’s Neuroimaging Initiative (ADNI) dataset and achieves an accuracy of 98.32%.

Paper 1

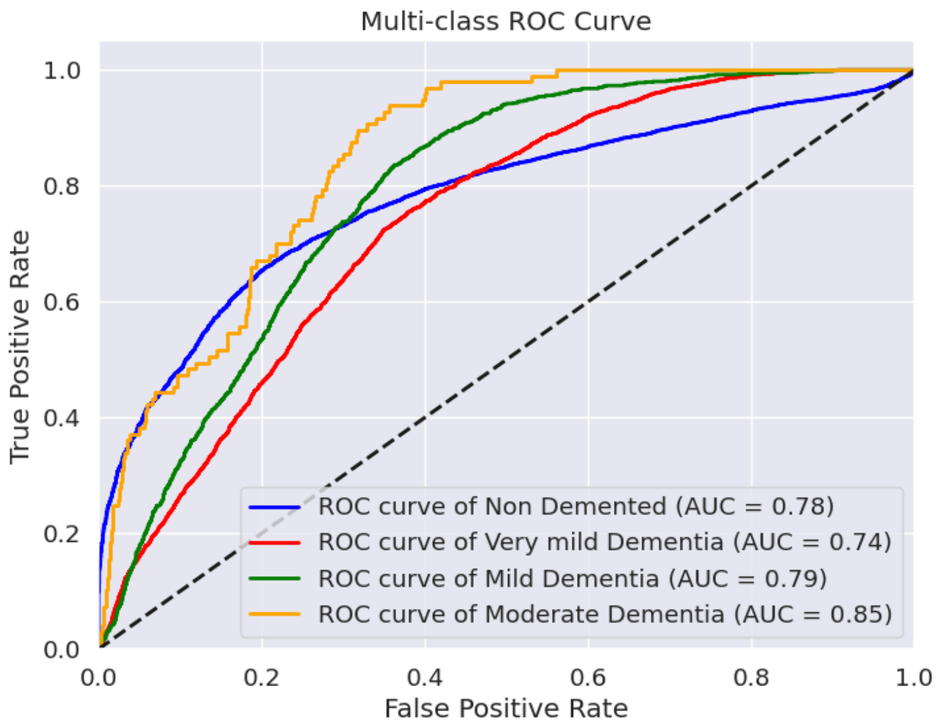
Classification Report:

	precision	recall	f1-score	support
Non Demented	0.81	0.91	0.86	13444
Very mild Dementia	0.31	0.21	0.25	2745
Mild Dementia	0.17	0.06	0.08	1000
Moderate Dementia	0.00	0.00	0.00	97
accuracy			0.74	17286
macro avg	0.32	0.29	0.30	17286
weighted avg	0.69	0.74	0.71	17286

Paper 1

Roc_curve:

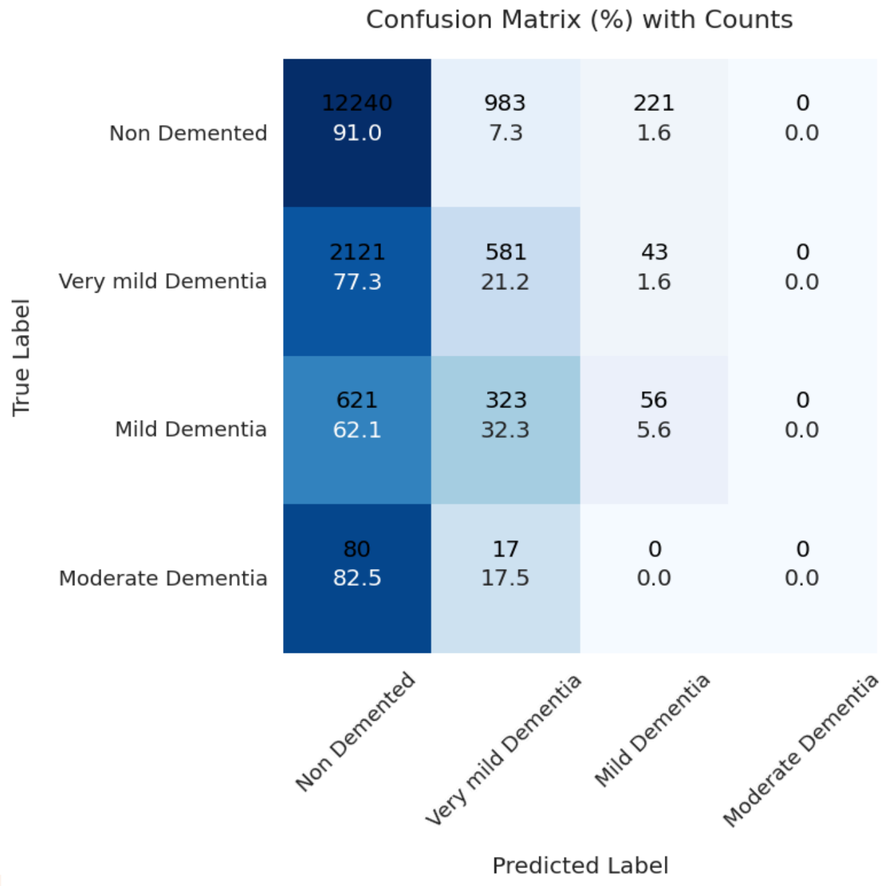
- All curves sit well above the diagonal reference line, confirming that our model performs well.
- This helps us understand which dementia stages require further refinement in our classification approach.



Paper 1

Confusion matrix:

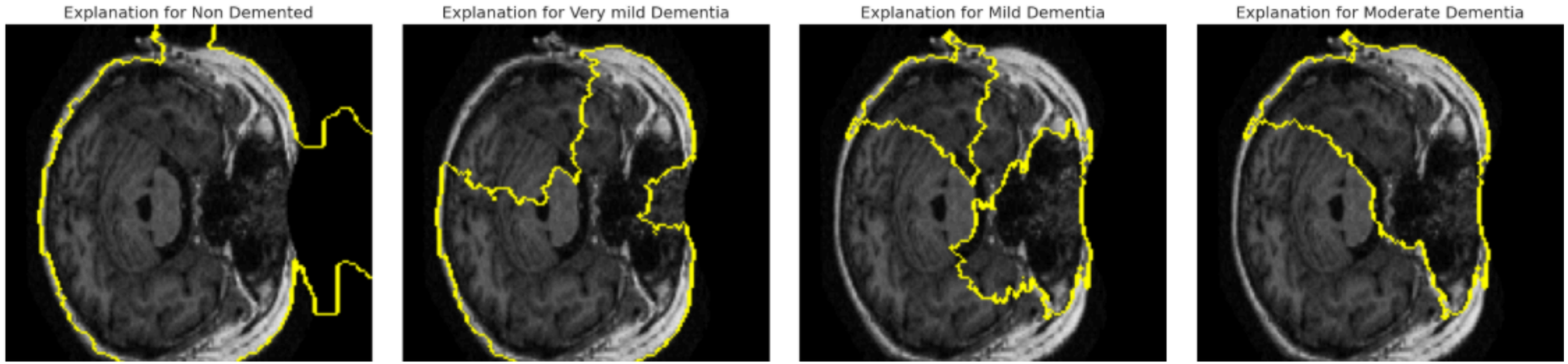
- Non Demented has been classified correctly by 91%.
- Very mild Dementia by 21.2%
- Mild Dementia by 5.6%
- Moderate Dementia by 0.0%



Paper 1

Lime:

- The yellow lines here show the most significant regions affect model training.



Paper 2 overview

“Deep Learning-Based Diagnosis of Alzheimer’s Disease Using Brain Magnetic Resonance Images: An Empirical Study“

This study evaluates the diagnostic performance of VUNO Med-DeepBrain AD (DBAD), a deep learning model, against medical experts. The model uses 2D coronal MRI slices of the medial temporal lobe and achieves comparable accuracy (87.1%) to medical experts (84.3%). The research highlights the potential of DBAD as a decision-support tool in clinical settings.

Paper 2

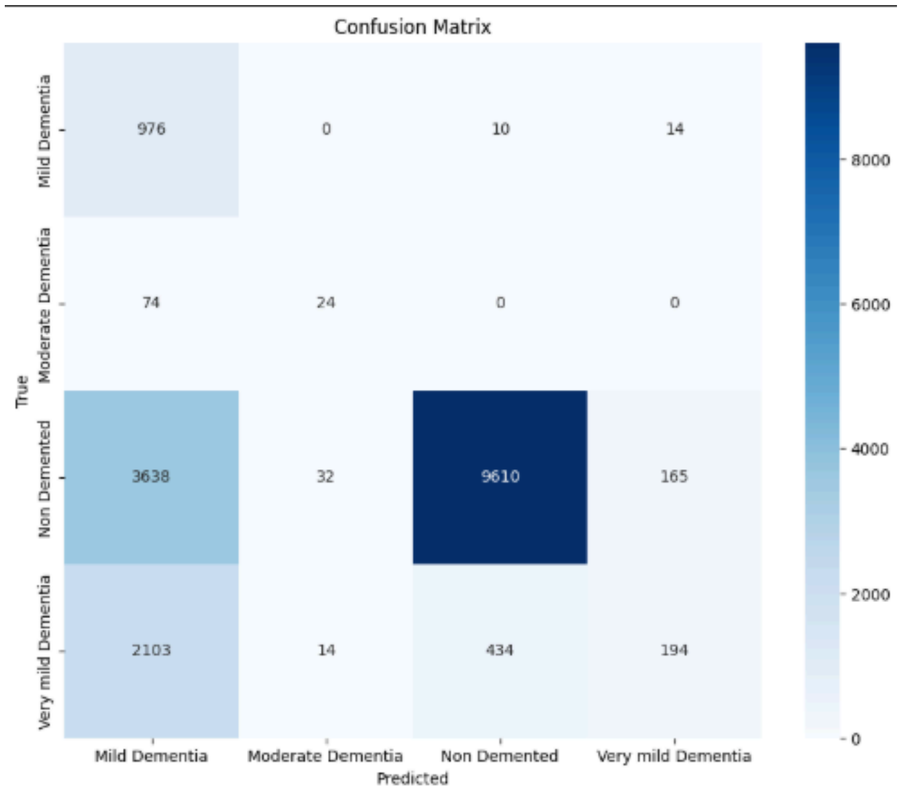
Classification Report:

	precision	recall	f1-score	support
Mild Dementia	0.14	0.98	0.25	1000
Moderate Dementia	0.34	0.24	0.29	98
Non Demented	0.96	0.71	0.82	13445
Very mild Dementia	0.52	0.07	0.12	2745
accuracy			0.62	17288
macro avg	0.49	0.50	0.37	17288
weighted avg	0.84	0.62	0.67	17288

Paper 2

Confusion matrix:

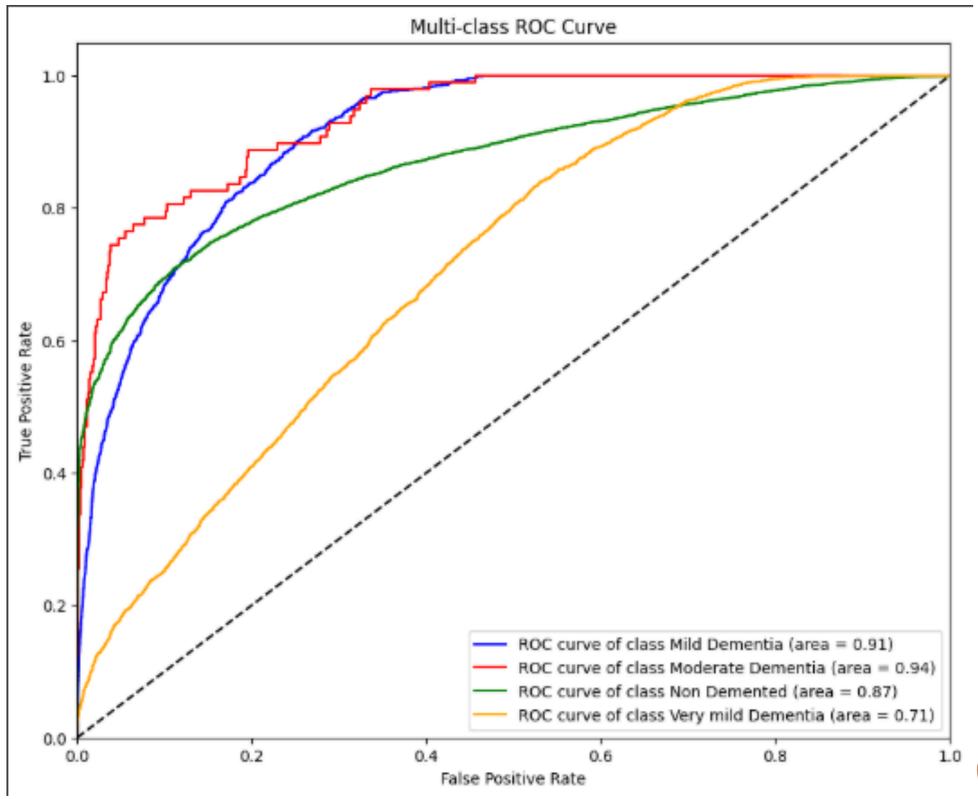
- For the non Demented, 9610 images have been classified correctly.
- Very mild Dementia, 194.
- Mild Dementia, 976.
- Moderate Dementia only 24.



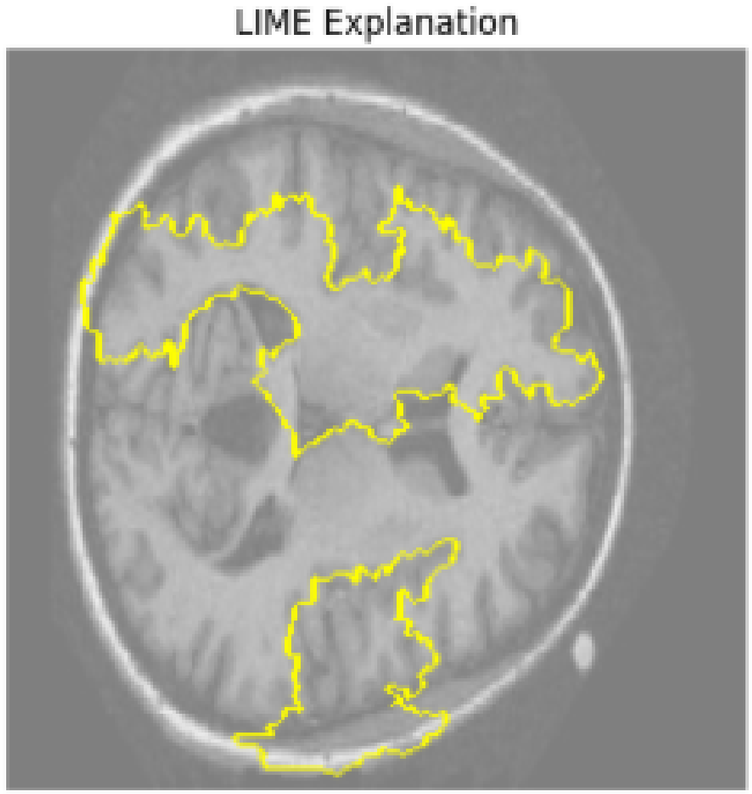
Paper 2

Roc_curve:

- All curves sit well above the diagonal reference line, confirming that our model performs well, except for the “very mild dementia” class.
- This helps us understand which dementia stages require further refinement in our classification approach.



Paper 2

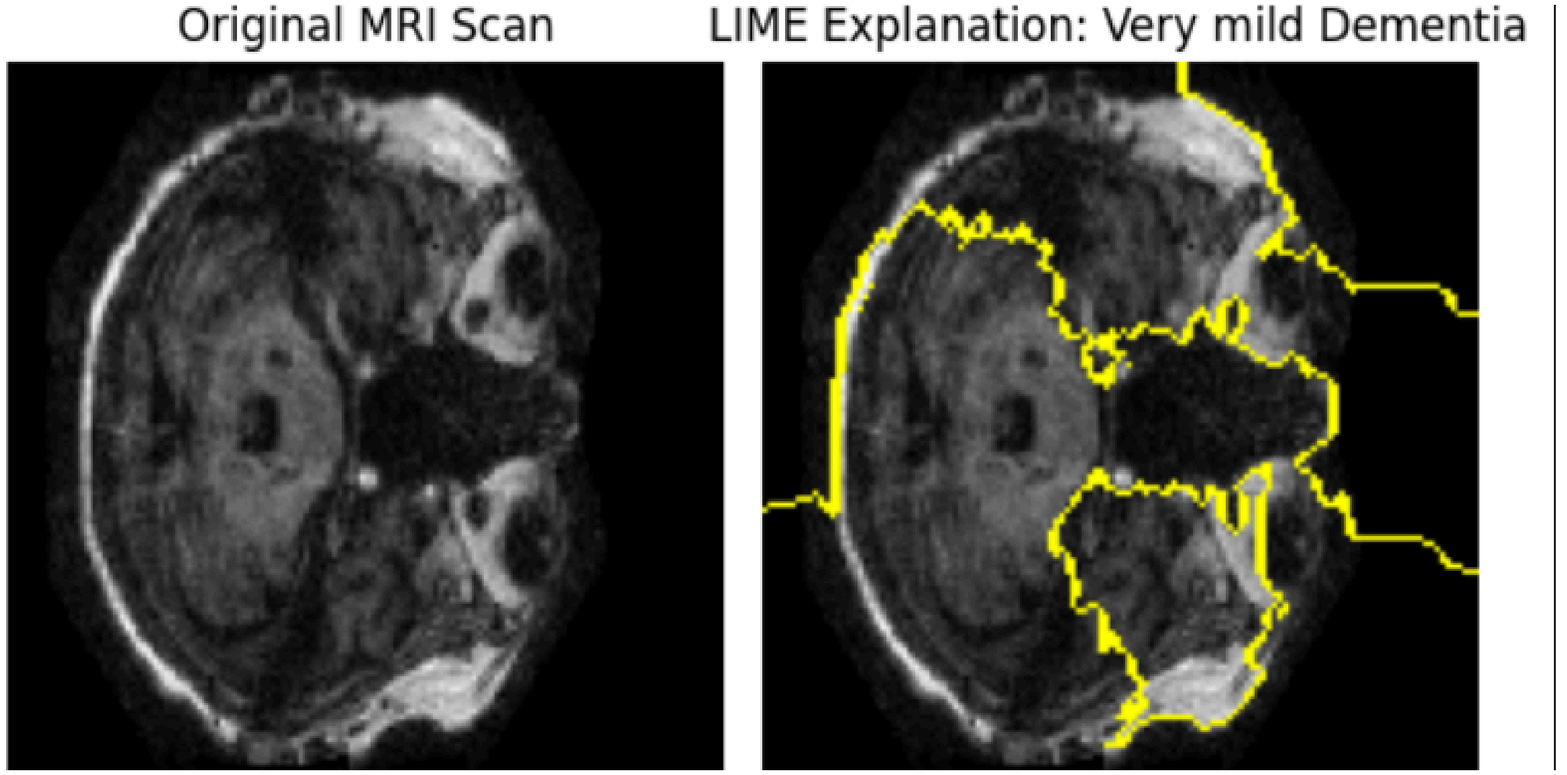


Paper 3 overview

“Early Detection of Alzheimer’s Disease Using Convolutional Neural Network Architecture“

This paper proposes a CNN-based approach for early AD detection using MRI images from the OASIS dataset. The model achieves over 95% accuracy by focusing on the hippocampus region, which is critical for memory and cognitive functions.

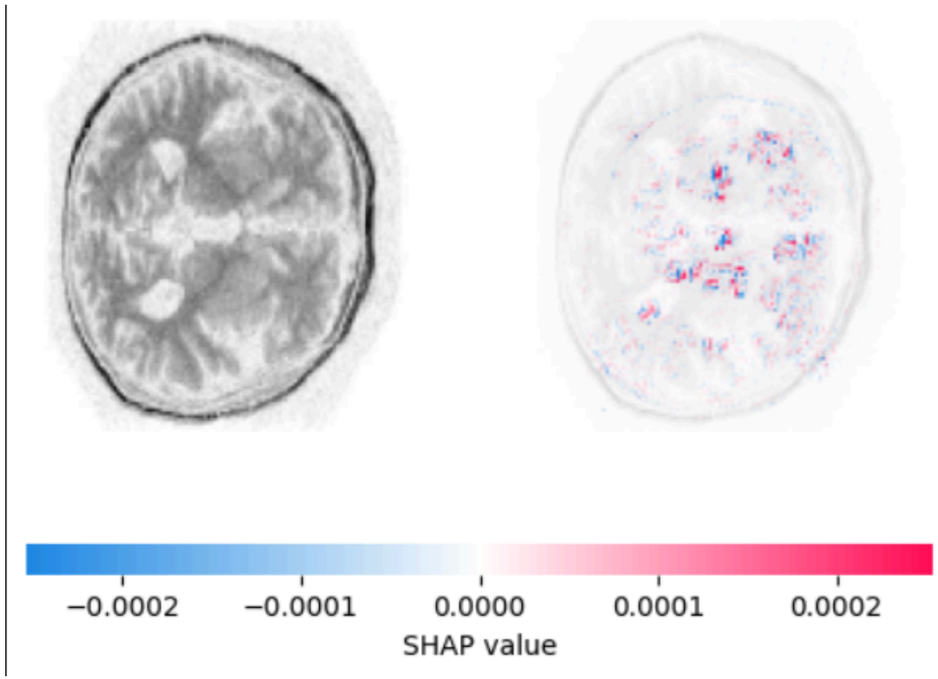
Paper 3



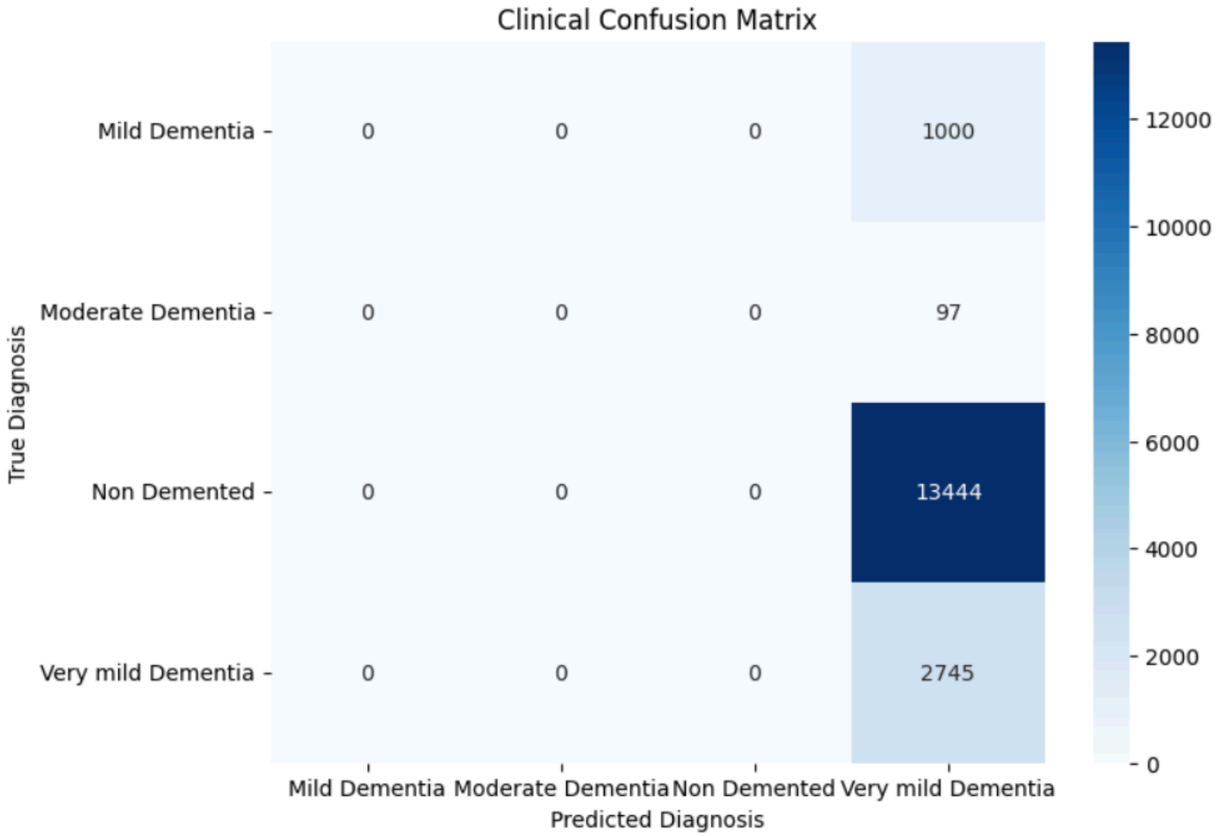
Paper 3

SHAP:

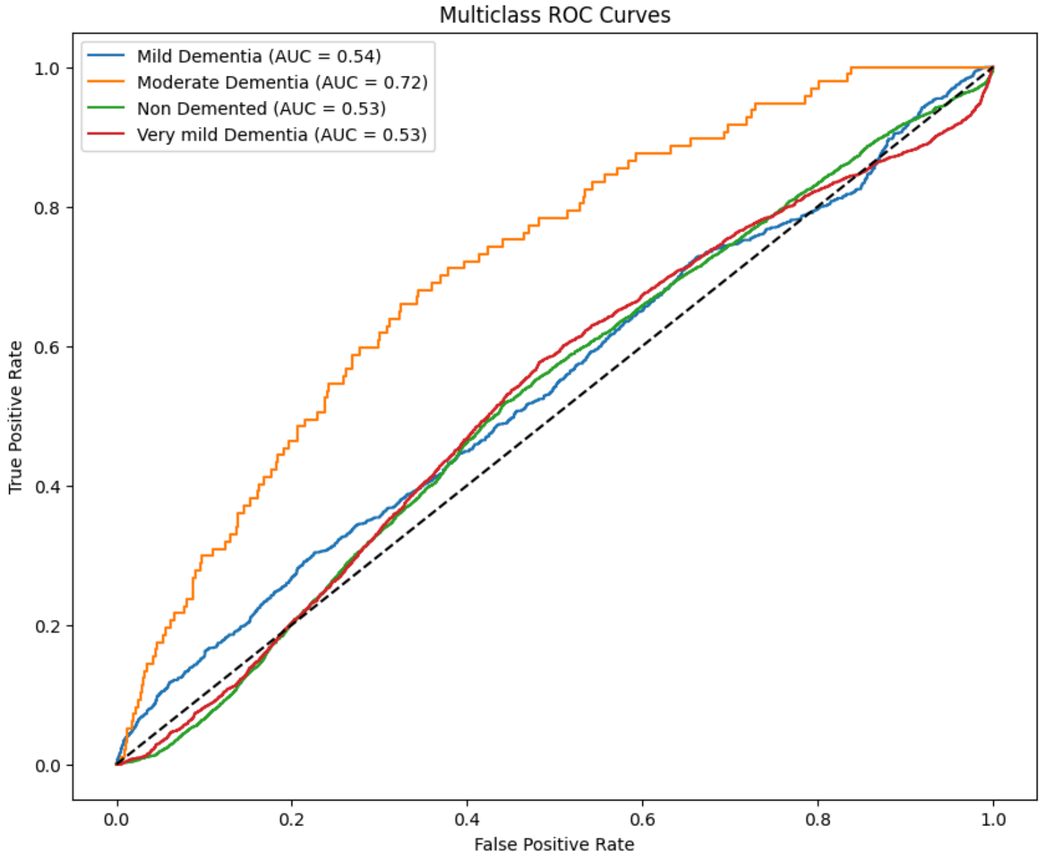
- The **RED** and **Blue** points show the most effective regions in the MRI, positively and negatively.



Paper 3


Clinical Confusion Matrix
Mild Dementia
Moderate Dementia
Non Demented
Very mild Dementia
True Diagnosis
0
0
0
1000
0
0
0
97
0
0
0
13444
0
0
0
2745
Mild Dementia
Moderate Dementia
Non Demented
Very mild Dementia
Predicted Diagnosis
12000
10000
8000
6000
4000
2000
0

Paper 3


Multiclass ROC Curves
Mild Dementia (AUC = 0.54)
Moderate Dementia (AUC = 0.72)
Non Demented (AUC = 0.53)
Very mild Dementia (AUC = 0.53)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Balancing classes

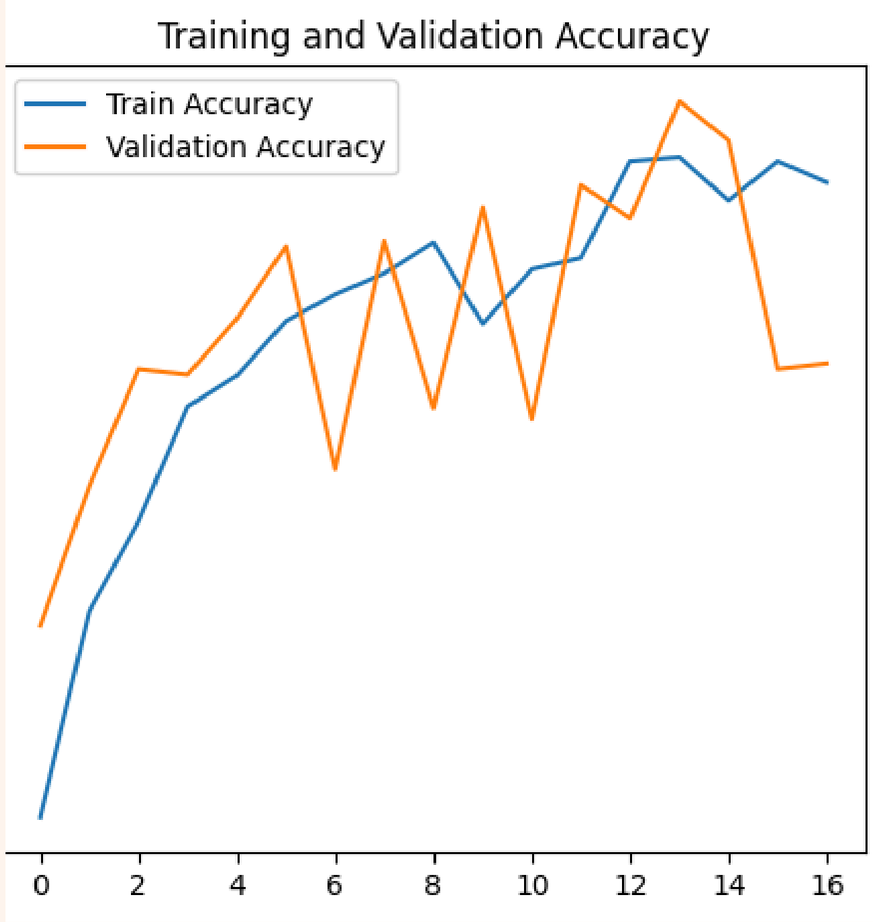
First model

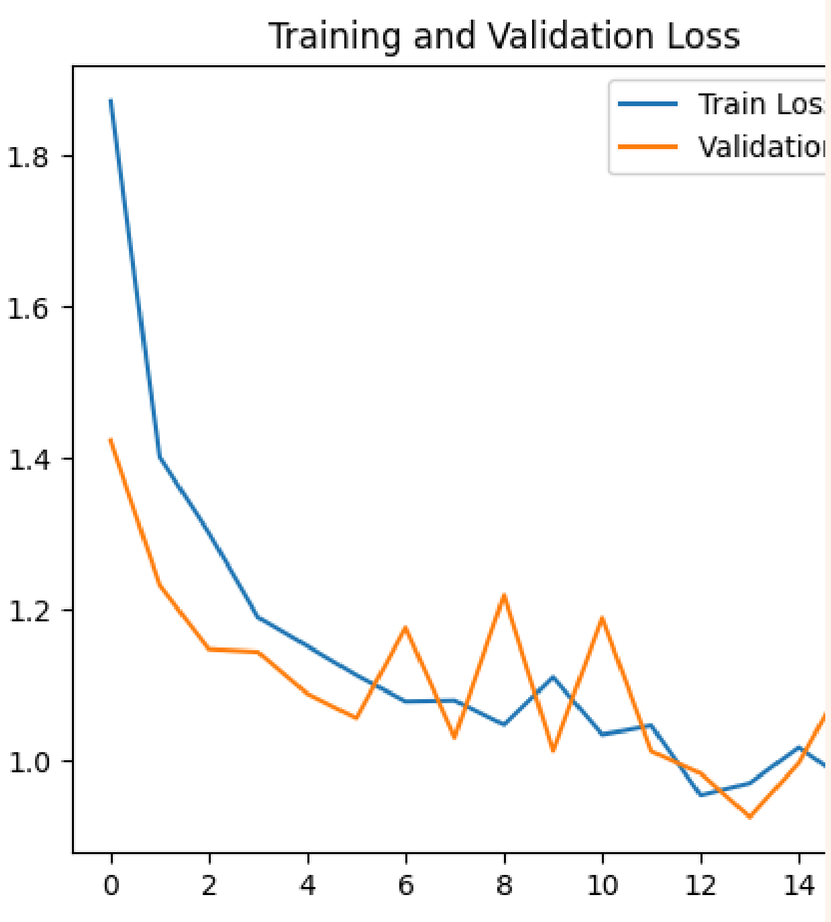
The first model includes VGG16 as a feature extractor and a classifer of 100 neurons and a softmax activation function as well as using RELU, ADAM optimizer, a regularization rate of 0.0006, and 20 epochs.

First model results using accuracy and loss

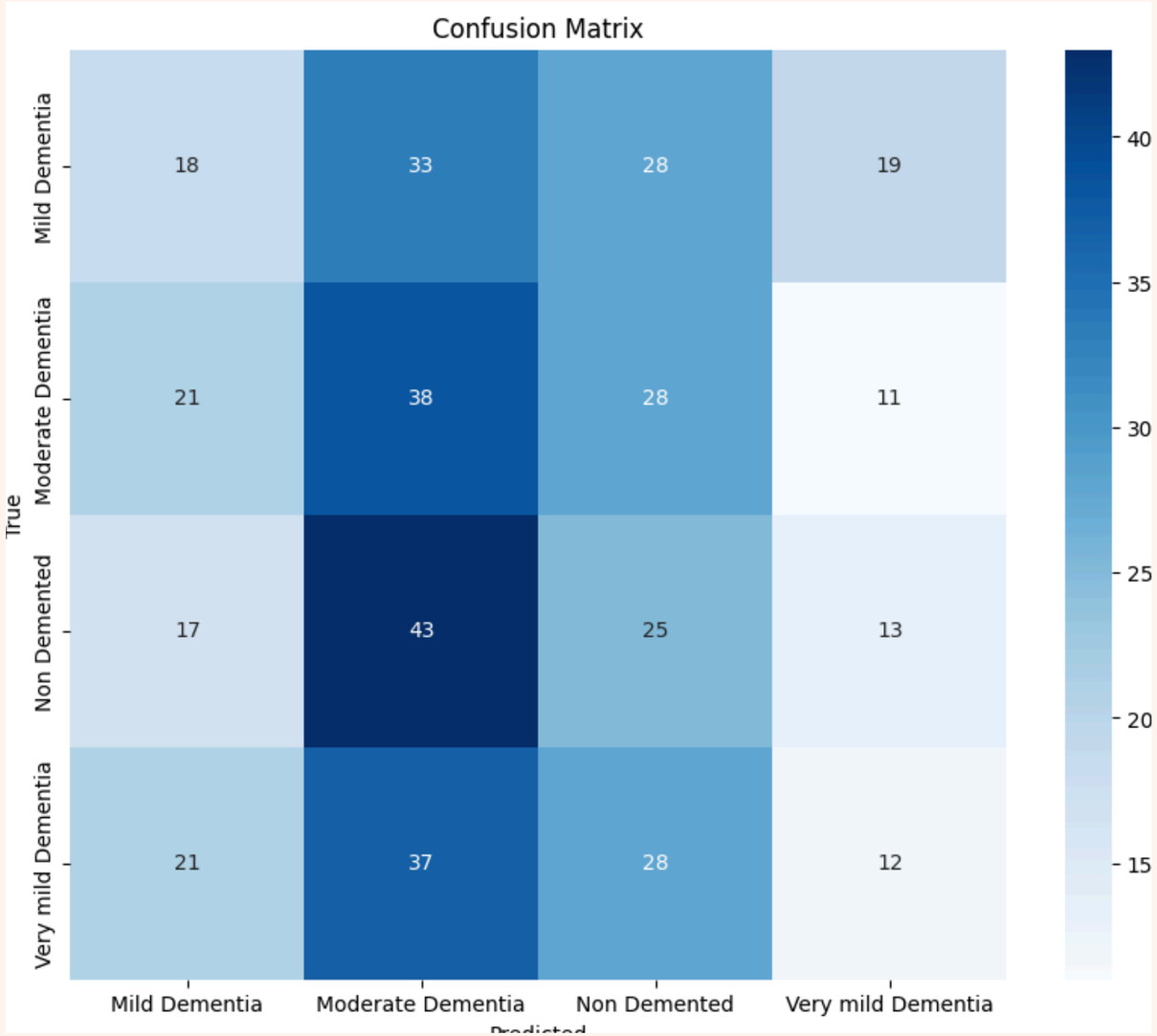
accuracy: 0.6777 - loss: 0.9208

Plotting training and validation accuracy and loss

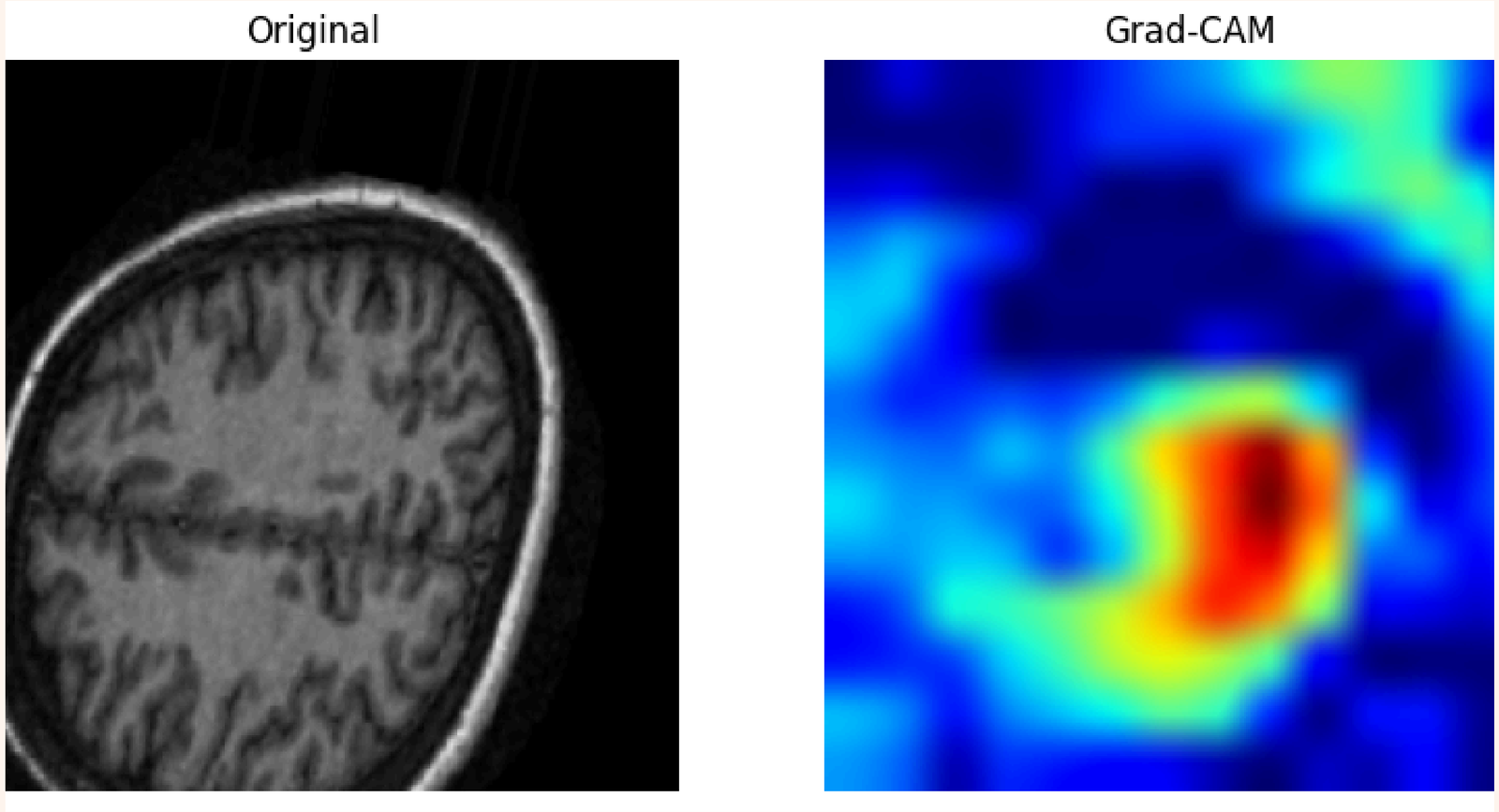




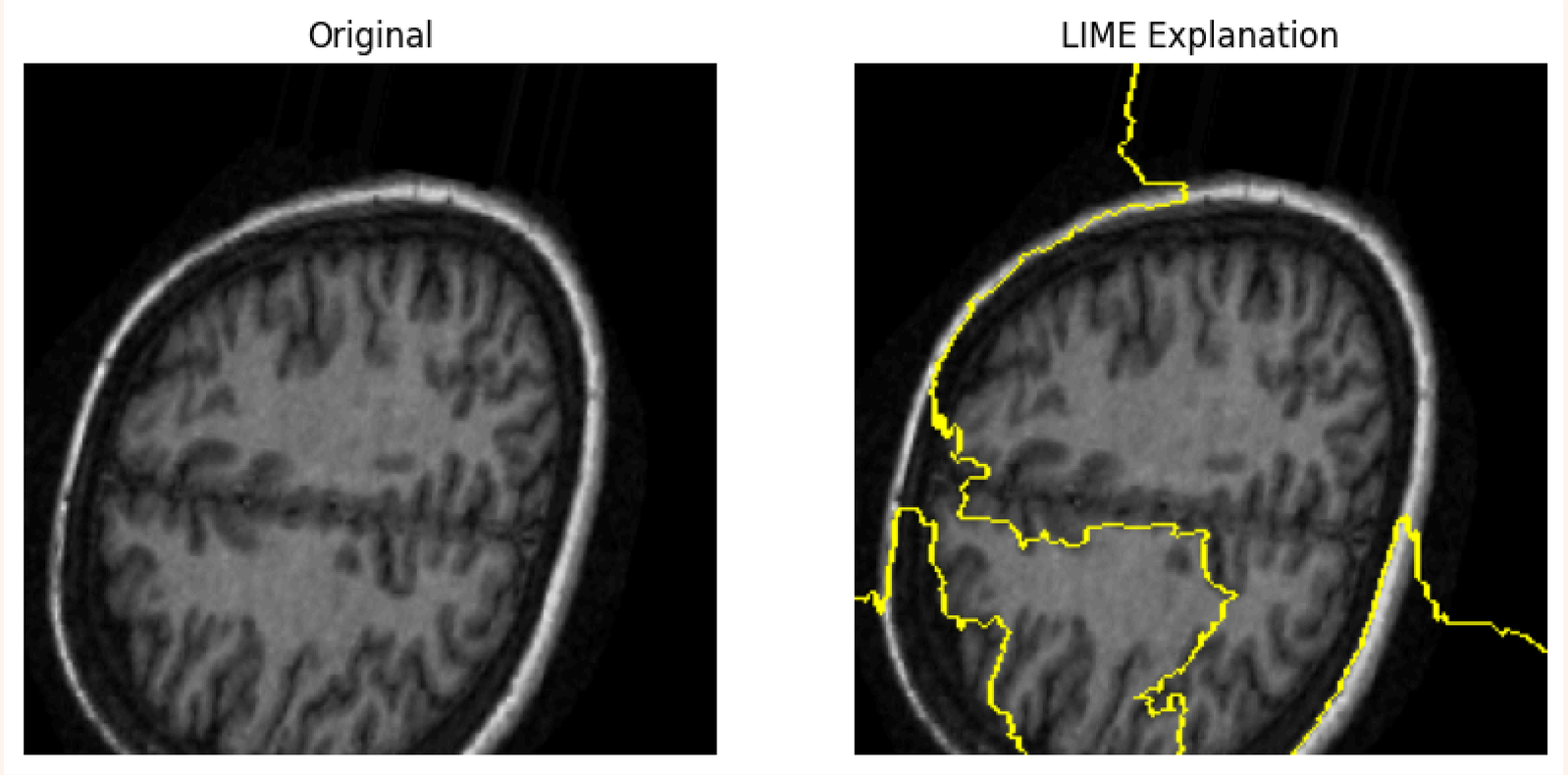
Confusion Matrix



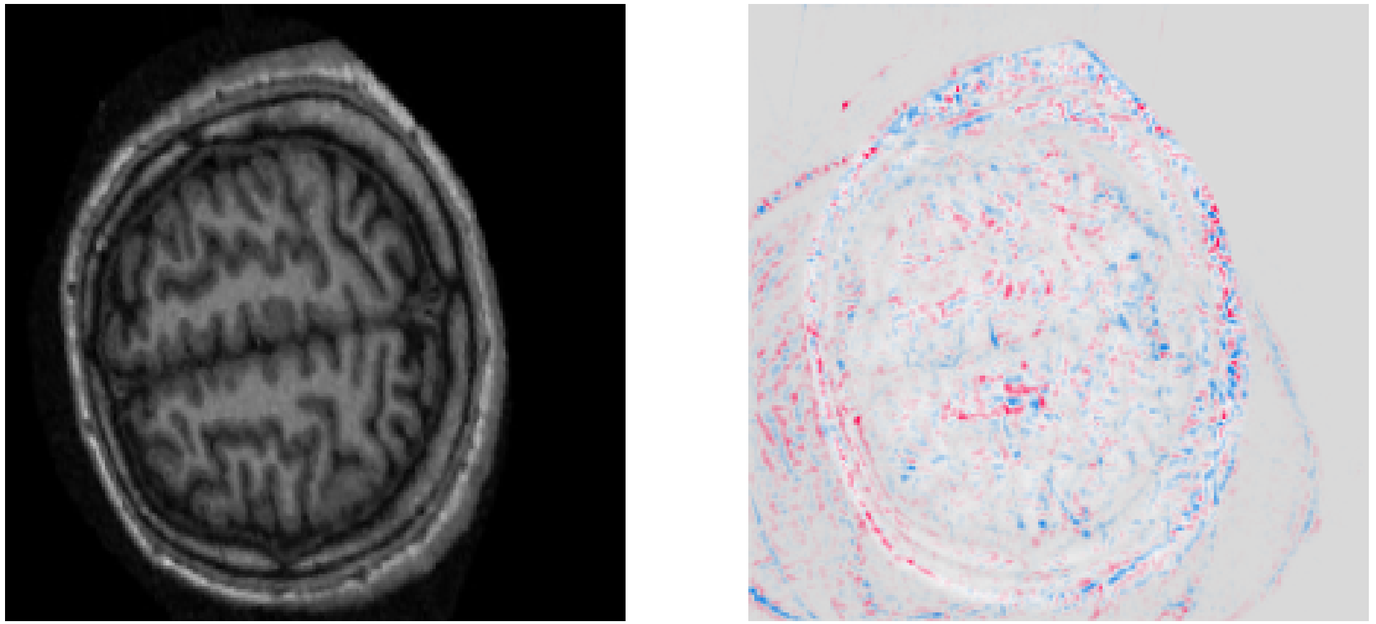
Grad-Cam

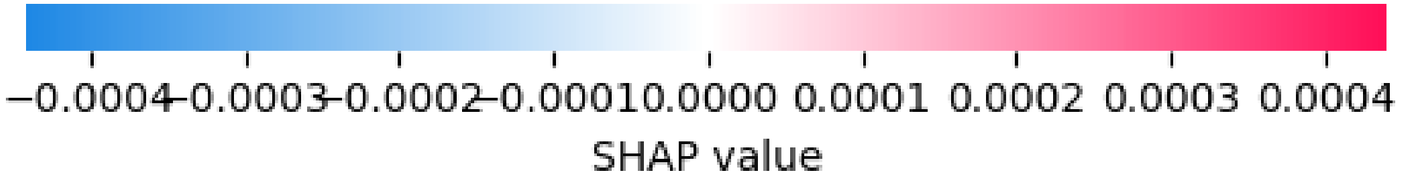


Lime



SHAP




−0.0004
−0.0003
−0.0002
−0.0001
0.0000
0.0001
0.0002
0.0003
0.0004
SHAP value

Model 2

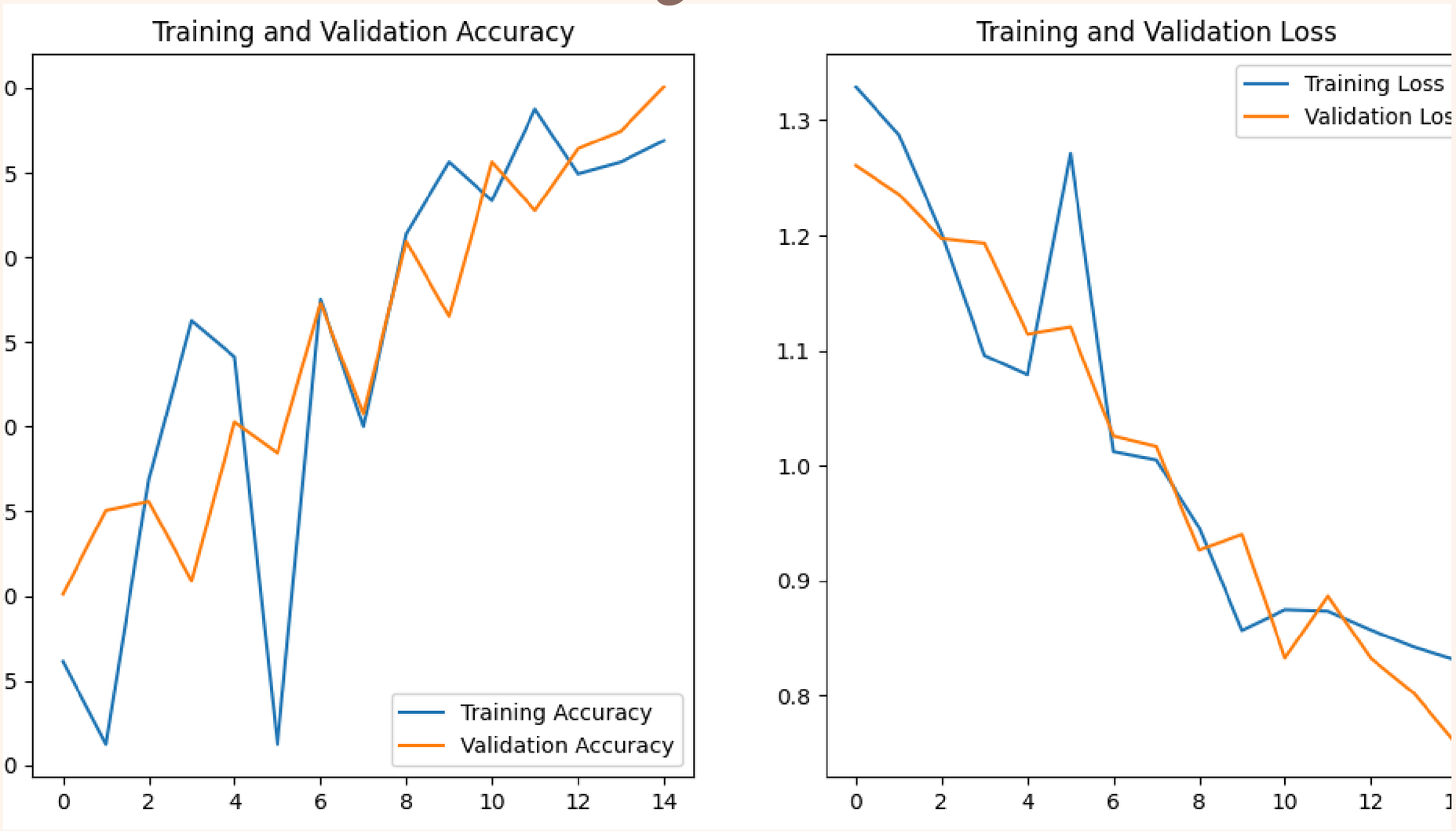
The second model includes Xception fine-tuned model. It includes two Dropout layers to reduce overfitting during training, with dropout rates of 0.3 and 0.25, respectively. Dense layer with 128 neurons and ReLU activation are added between the Dropout layers to enhance the model's capacity for non-linear representations. The last layer of the model is a Dense output layer with a softmax activation function to produce predictions for four classes. Adamax optimizer with a learning rate of 0.001 is used. The loss function used is categorical crossentropy.

Model 2 results using accuracy and loss

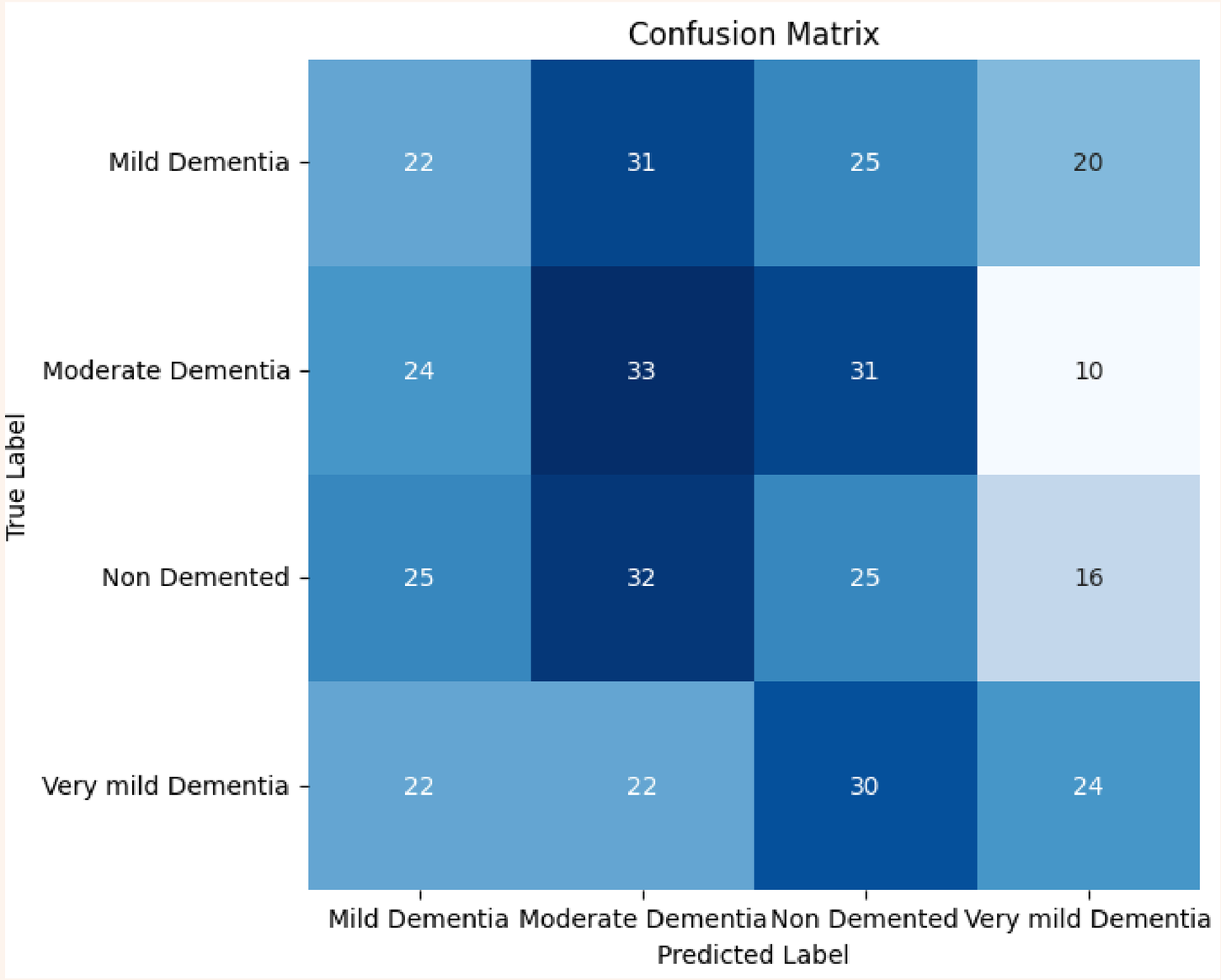
Test Loss: 0.7714812755584717

Test Accuracy: 0.6770833134651184

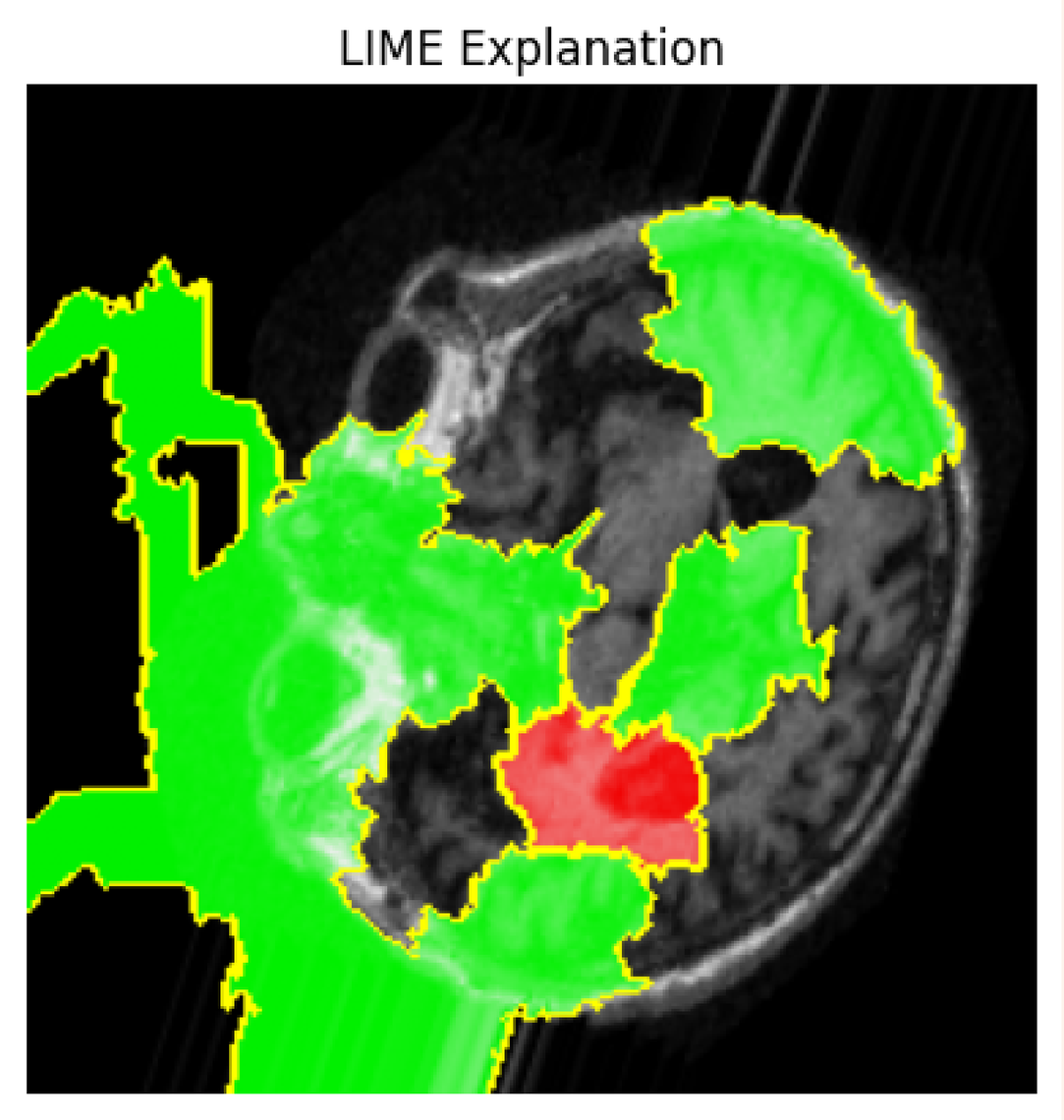
Plotting training and validation accuracy and loss



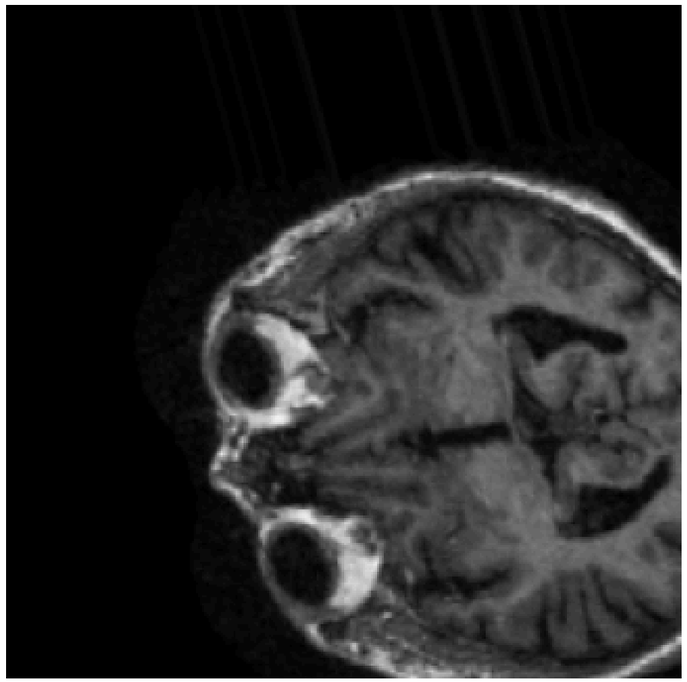
Confusion matrix

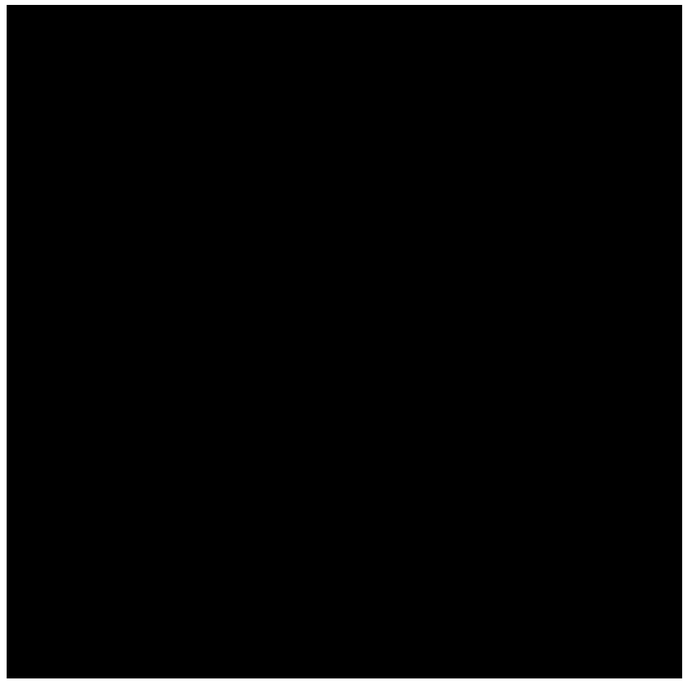


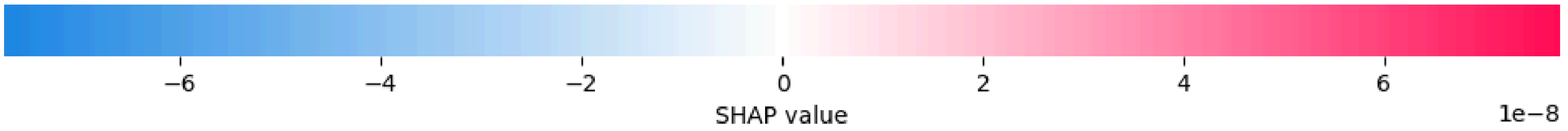
Lime



SHAP







Model 3

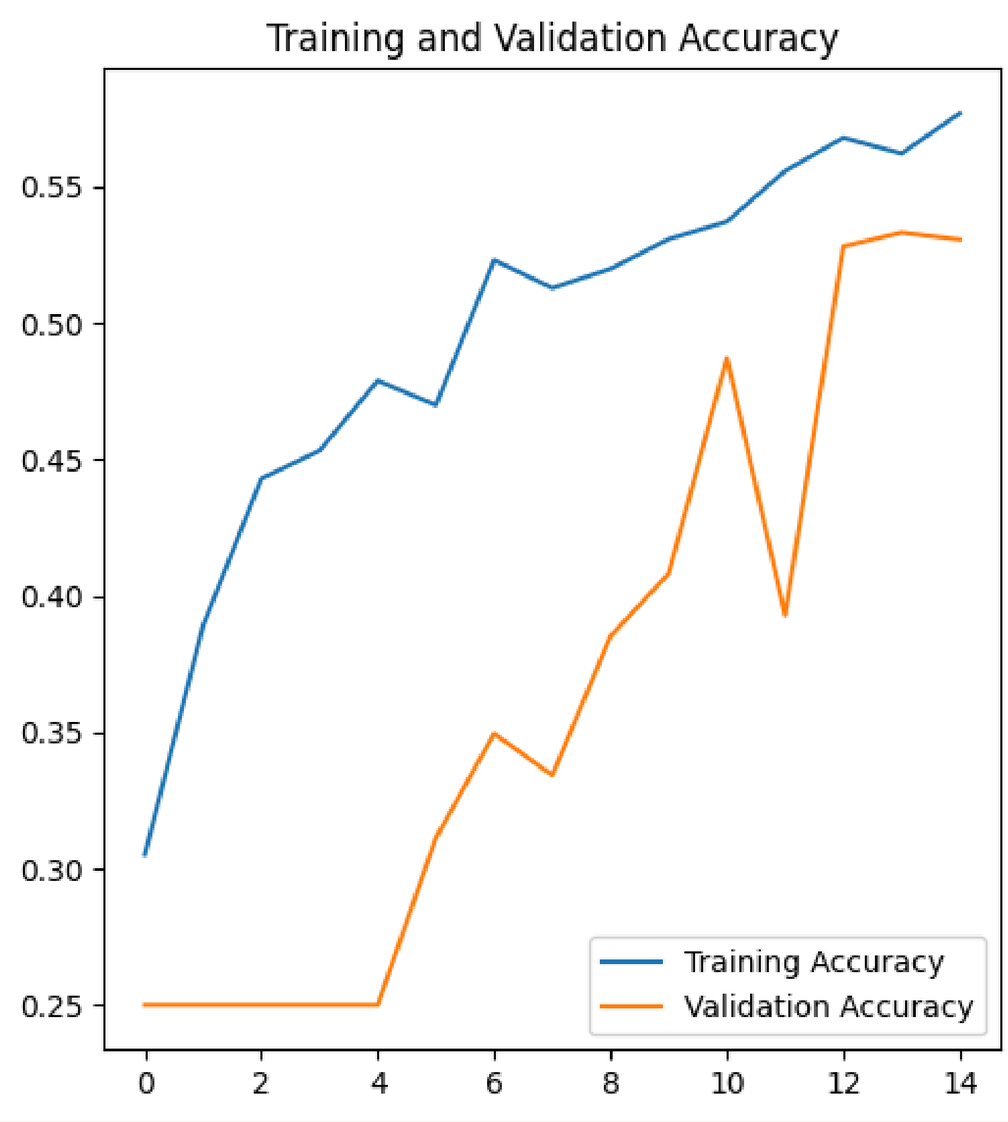
The proposed ADD-Net consists of four convolutional blocks, and each convolutional block has a Rectified Linear Unit (ReLU) activation function and a 2D average pooling layer, two dropout layers, two dense layers, and a SoftMax classification layer.

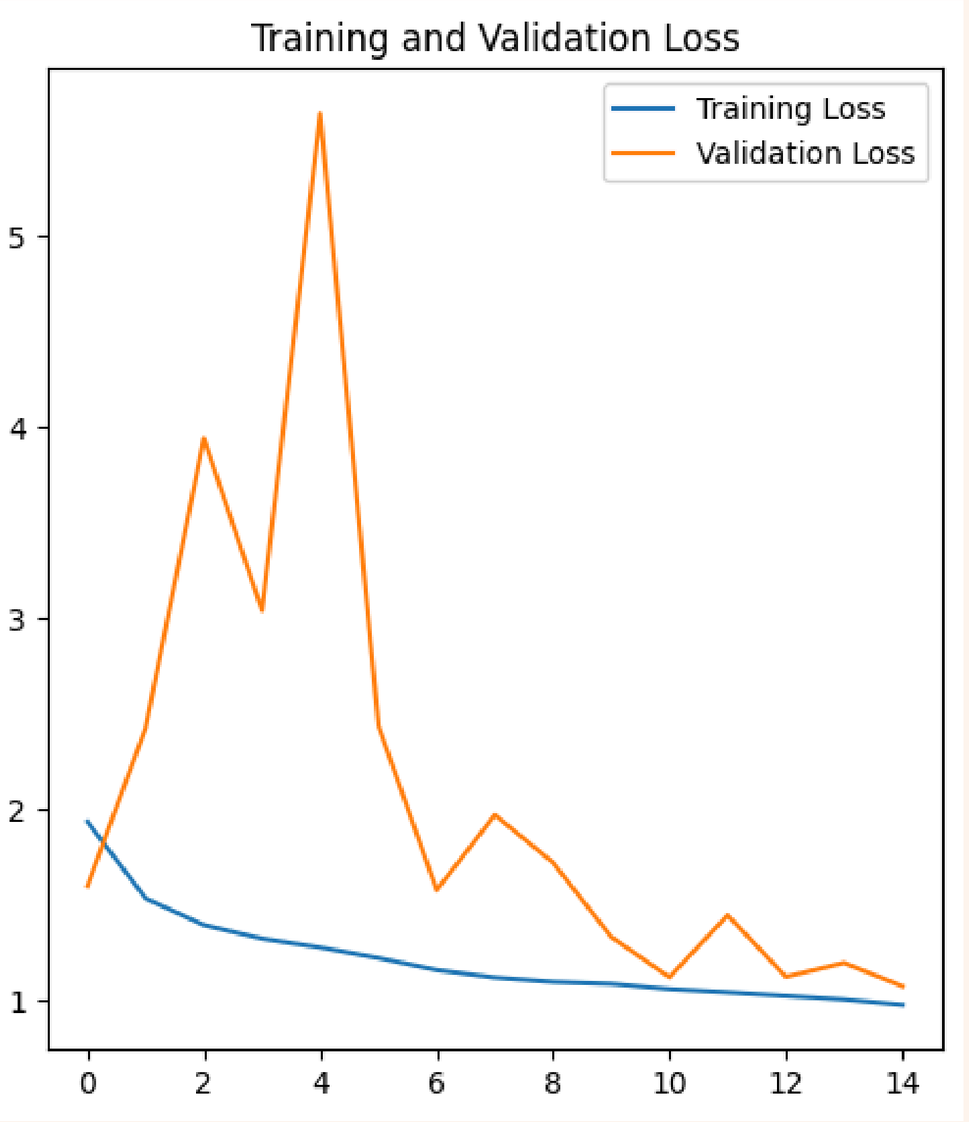
Model 3 results using accuracy and loss

Test Loss: 1.1229102611541748

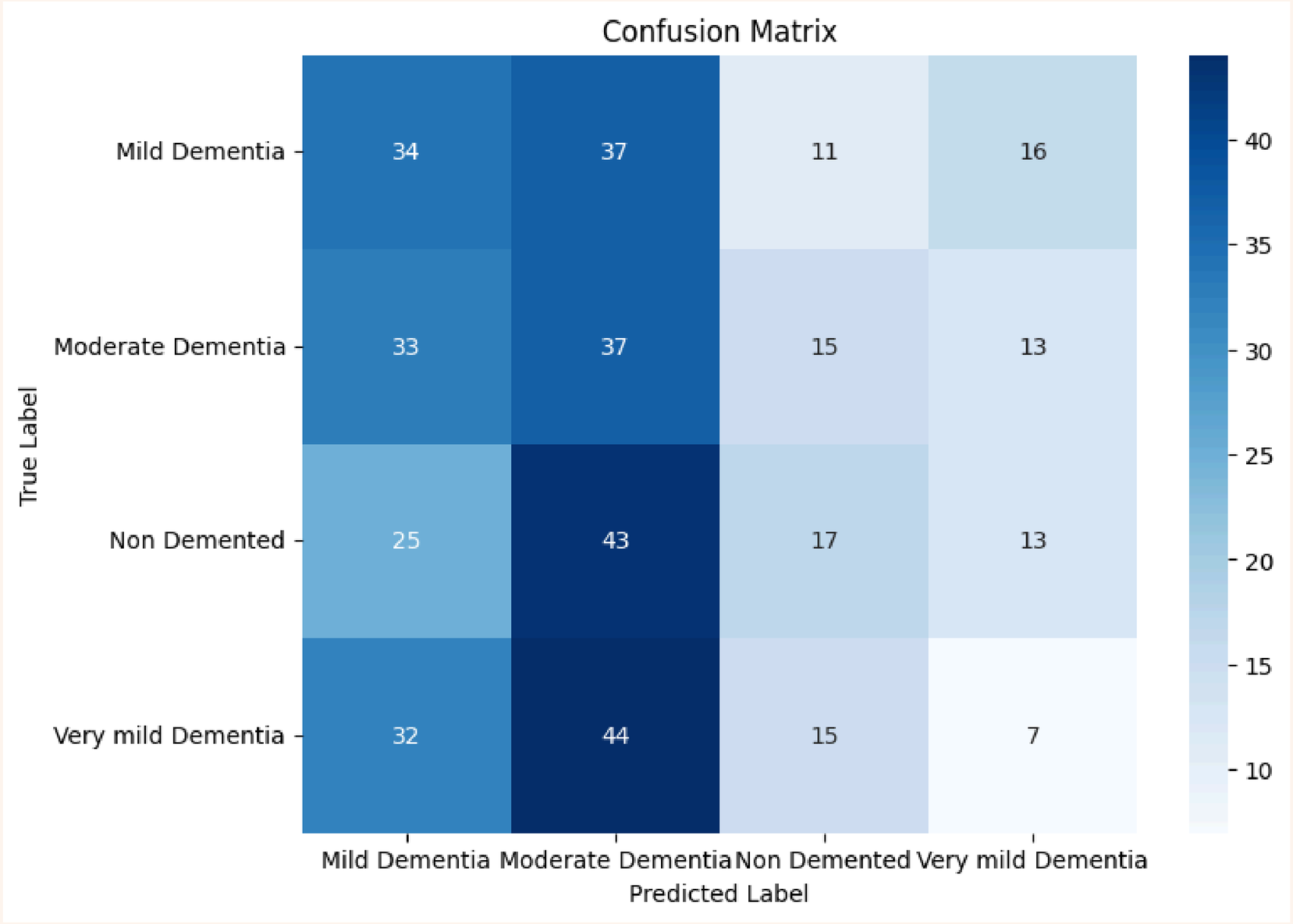
Test Accuracy: 0.48724490404129028

Plotting training and validation accuracy and loss

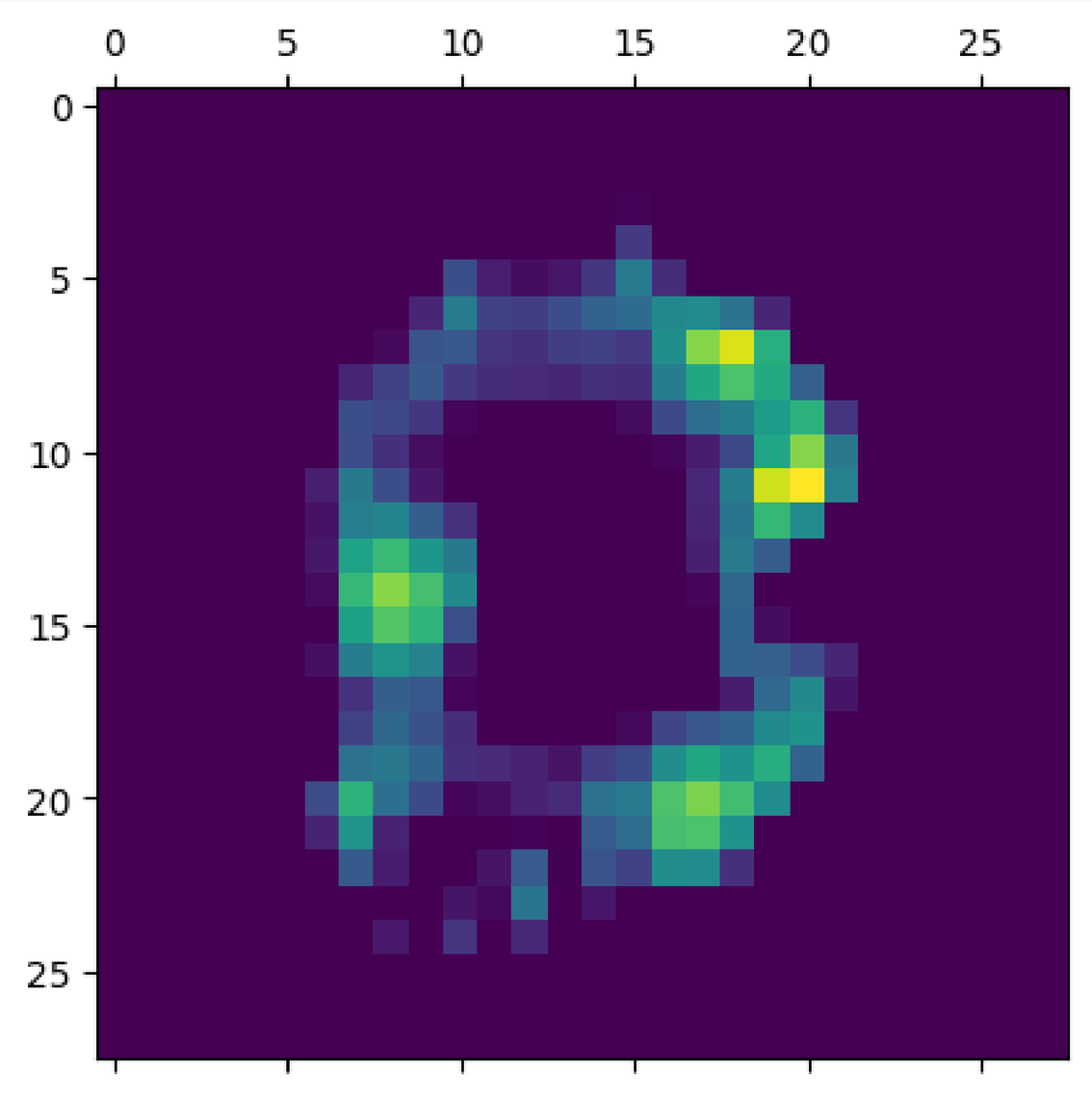




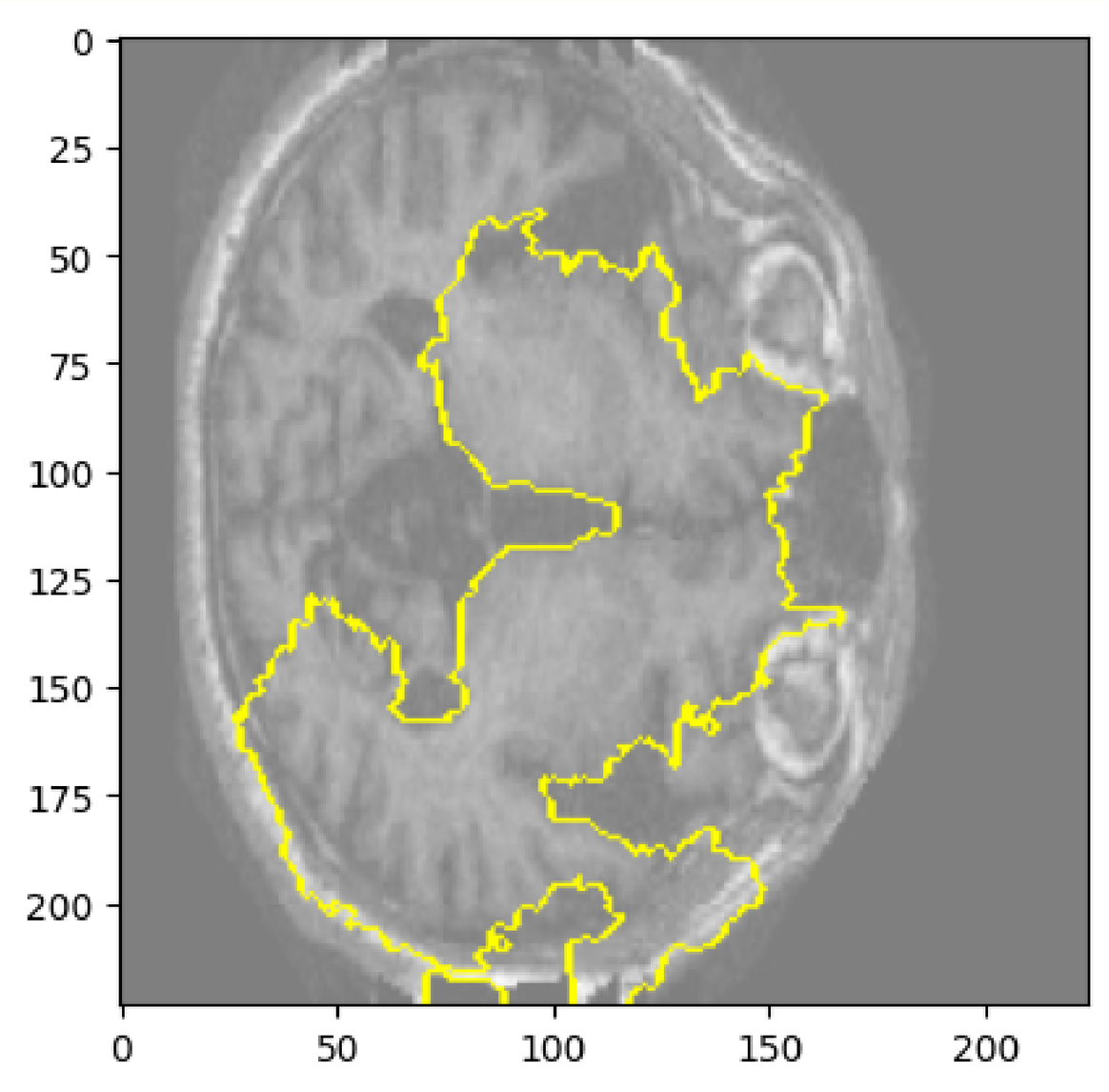
Confusion matrix



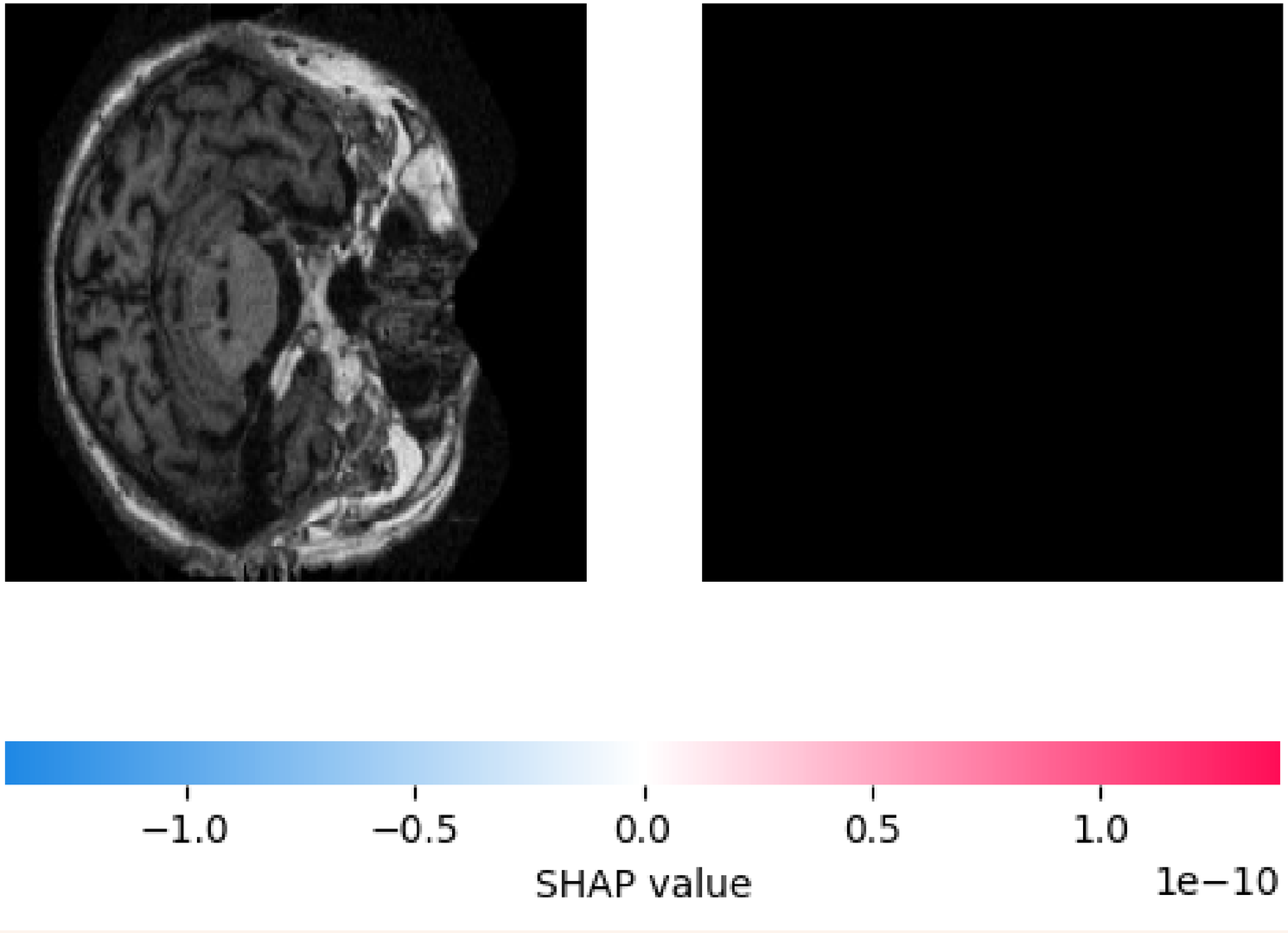
Grad-cam



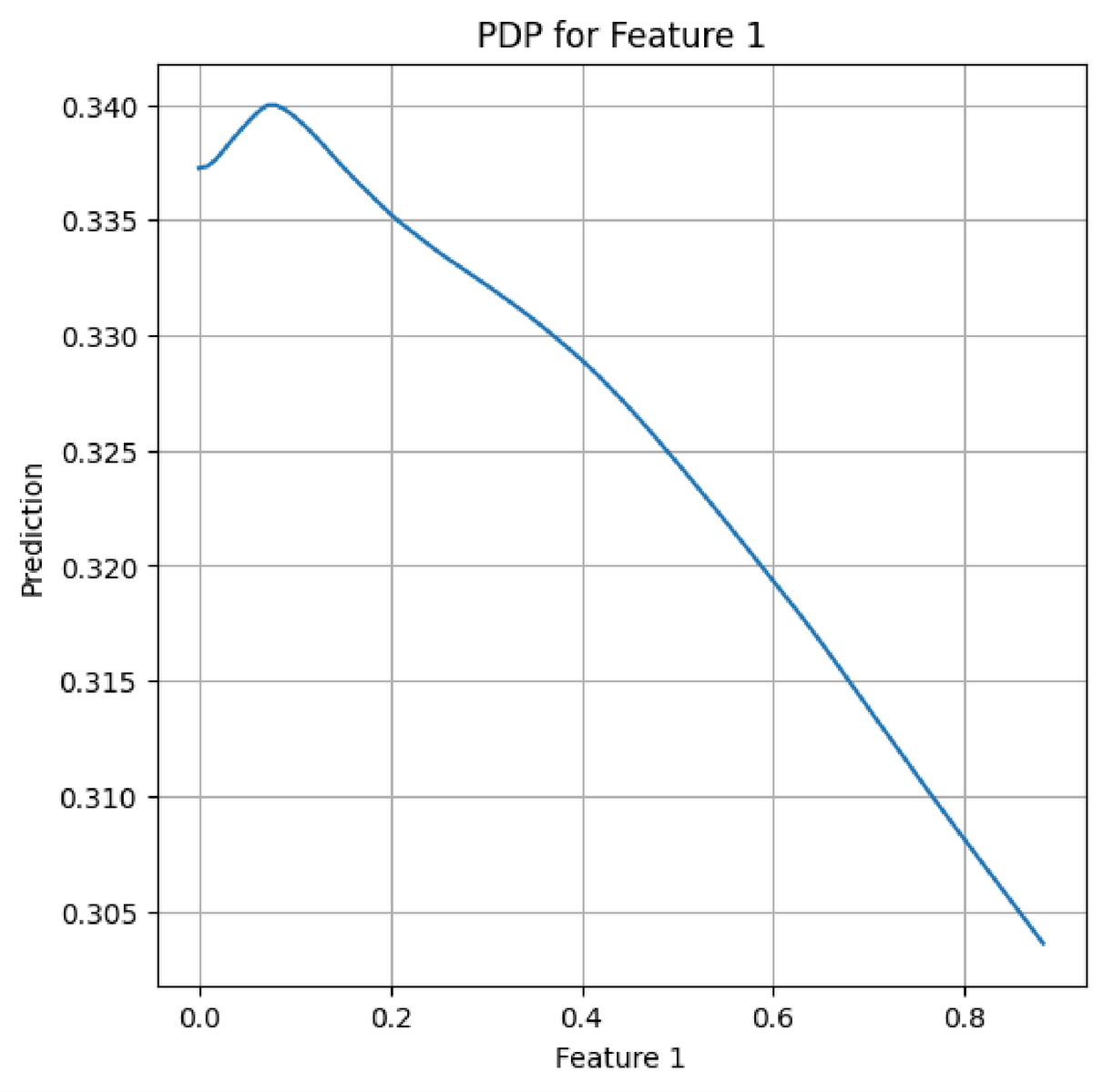
Lime



SHAP



PDP


PDP for Feature 1
Prediction
0.340
0.335
0.330
0.325
0.320
0.315
0.310
0.305
0.0
0.2
0.4
0.6
0.8
Feature 1

THANK YOU

Do you have any questions for us?