# Dynamic Explainability in AI for Neurological Disorders: An Adaptive Model for Transparent Decision-Making in Alzheimer's Disease Diagnosis

Anushka Shukla
anushkashukla11902@gmail.com
CSE GGITS

Shivanshu Upadhyay
shivanshuupadhyay798@gmail.co
m CSE GGITS

Priya Rachel Bachan
bachanpriya20@gmail.com
CSE GGITS

Udit Narayan Bera
uditnarayanbera@gmail.com
ECE GGITS

RV Kshirsagar
principal@ggits.org
ECE GGITS

Neeta Nathani
neetanathani@ggits.org
ECE GGITS

2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT) | 979-8-3503-0546-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/CSNT60213.2024.10546177

*Abstract*— **In this paper, we proposed a model that will solve the 'X' of the 'Xai' that is Explainable AI. The model is developed using deep learning and transfer learning algorithms using different methods to depict how the decisions and predictions of the Artificial Intelligence are made to be understandable for humans to interpret. The term deals with explaining how the models work and what all happens in each layer of neurons and the output is shown. The transparency in the process lets humans understand the way how the predictions are carried out, what are the parameters that the model is considering, what are the steps it takes to generate the final output. Here, we considered Alzheimer disease in the brain and brought out the results per layer of the model to comprehend the reason of the final result. This could made easy for humans to identify what are the errors, unknown biases and the number of possible paths the model can take in order to generate the more accurate output. This proposed model is able to identify and depict the processes going on while generating the result. The field tends to address the "black box" problem in the complex machine learning models. The analysis would be used by stakeholders to identify the real cause behind the interpretation of results. Considering these, there are many use cases for this new emerging field, some of them includes in the field of medical diagnosis, where the doctors would be able to identify the paradigm and produce the most accurate diagnosis that will cure the patient's disease, in the finance sector to identify the future trends, for the real time processing and many such fields.**

**Keywords—** ***Explainable AI, Alzheimer's Disease Diagnosis, Neural Network prediction, MobileNet V2.***

## I. Introduction

Alzheimer's disease is a progressive neurological disorder that impacts cerebral function, resulting in the deterioration of memory, cognitive capabilities, and overall bodily functioning [1]. It stands as the most prevalent cause of dementia. Individuals afflicted by Alzheimer's frequently encounter memory lapses that disrupt their daily activities, such as forgetting recently acquired information, important dates, and repetitive inquiries for the same details, or an increasing reliance on memory aids. Alzheimer's can induce significant alterations in behavior and personality. Figure 2 illustrates the specific region of the brain affected by Alzheimer's disease, which may manifest in symptoms such as confusion, agitation, mood fluctuations, social withdrawal, and heightened irritability. The examination is conducted through the application of neural network models. A neural network is a computational paradigm designed to emulate the structure and functionality of the neural networks in the human brain [2]. It serves as a framework in machine learning and artificial intelligence for processing information, executing pattern recognition, and making predictions or classifications. The architecture of the neural network is depicted in Figure 1.

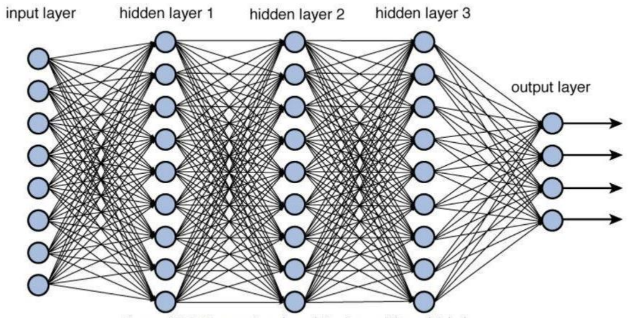


Fig 1. Deep Neural Network Architecture [2].

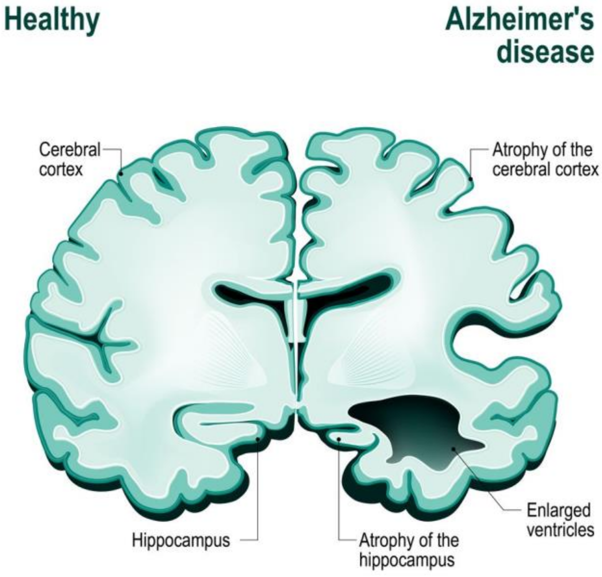


Fig 2. The brain affected with Alzheimer's Disease [3]

Authorized licensed use limited to: Univ of Calif Santa Cruz. Downloaded on June 15,2025 at 16:17:13 UTC from IEEE Xplore. Restrictions apply.

## II. METHODOLOGY

The dataset utilized in this study was sourced from Kaggle [4] and comprised images related to Alzheimer's disease, classified into four distinct categories. These categories were delineated based on the Clinical Dementia Rating (CDR) scale, which assesses observed symptoms and impairment levels. The classes encompass Non-demented (indicating no significant memory loss), Very mild-demented (reflecting slight memory loss), Mild-demented (indicating a mild level of cognitive impairment), and Moderate-demented (characterized by memory loss, compromised communication, and an inability to recognize individuals). Subsequently, we imported the dataset, as illustrated in Figure 3, employing an image data generator, and conducted the analysis through a series of preprocessing steps. These steps encompassed rescaling, rotation, width-shift, height-shift, shearing, flipping, among others.

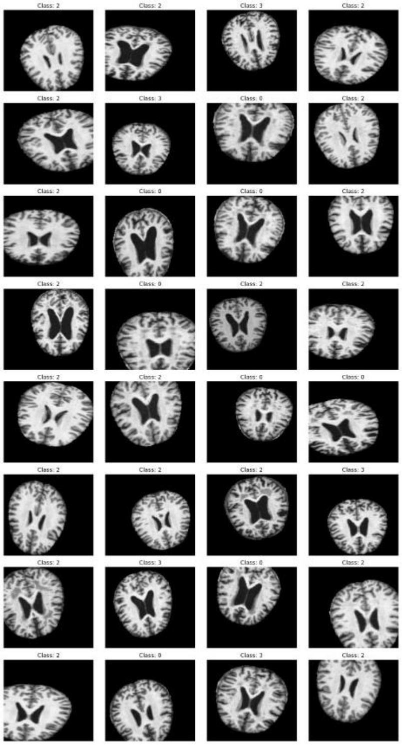


Fig 3. Input images for the study from dataset.

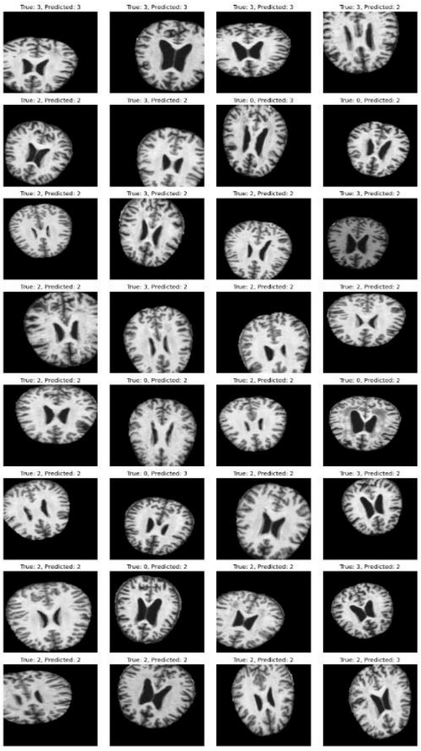


Fig 4. Output of MobileNet V2.

Subsequently, various transfer learning models were constructed, including VGG-19 (with training accuracy at 58% and testing accuracy at 53%), MobileNet V2, Inception V3 (with training accuracy at 55% and testing accuracy at 52%), ResNet-50 (with training accuracy at 45% and testing accuracy at 36%), and a custom model developed through deep learning algorithms (with training accuracy at 62% and testing accuracy at 53%). Moving forward, each model generated distinct outputs with varying accuracies. Notably, MobileNet V2 outperformed other models, achieving a training accuracy of 67% and testing accuracy of 60%, as depicted in Figure 4. The ensuing table presents the metrics associated with each model.

| S.no. | Neural Networks | Training Accuracy | Training Loss | Validation Accuracy | Validation Loss |
|---|---|---|---|---|---|
| 1 | Inception V3 | 0.6586 | 4.2752 | 0.5324 | 5.8107 |
| 2 | VGG 19 | 0.5053 | 1.0313 | 0.4812 | 1.0586 |
| 3 | Mobile Net V1 | 0.5728 | 1.2086 | 0.5286 | 1.2869 |
| 4 | Mobile Net V2 | 0.6741 | 3.9126 | 0.6000 | 6.0617 |
| 5 | ResNet 50 | 0.4439 | 1.3311 | 0.3623 | 1.4430 |
| 6 | Custom Model | 0.2600 | 6.0782 | 0.3400 | 6.1568 |

Table 1. Training and test validation and loss.

Authorized licensed use limited to: Univ of Calif Santa Cruz. Downloaded on June 15,2025 at 16:17:13 UTC from IEEE Xplore. Restrictions apply.

The process of the Alzheimer disease detection that proceeds with the loading of dataset, training, model selection, testing, evaluation and the final analysis is shown in the figure 5.

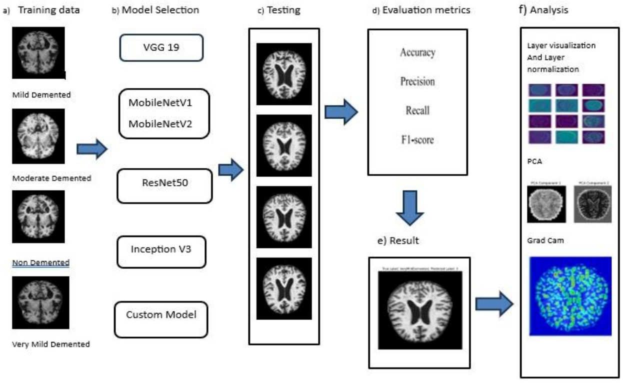
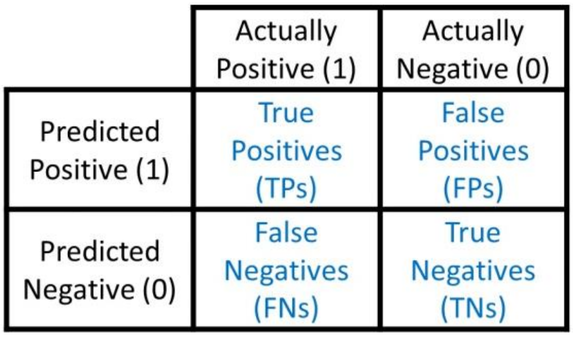


Fig 5. The model development process of Alzheimer's disease detection.

The confusion matrix is the evaluation metrics used for classification models [5]. The figure 6 shows the parameters involved in the confusion metrics.



| | Actually Positive (1) | Actually Negative (0) |
|---|---|---|
| Predicted Positive (1) | True Positives (TPs) | False Positives (FPs) |
| Predicted Negative (0) | False Negatives (FNs) | True Negatives (TNs) |

Fig 6. Confusion Matrix [5].

Here's what each term represents:

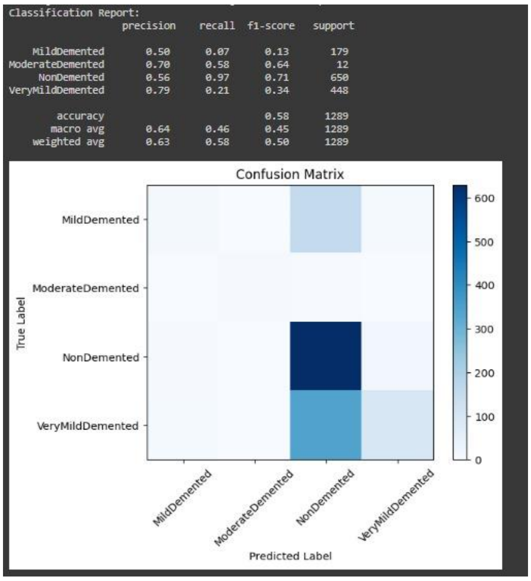
True Positives (TP): The cases where the model predicted the positive class correctly, and the actual value was also positive.

True Negatives (TN): The cases where the model predicted the negative class correctly, and the actual value was also negative.

False Positives (FP): T cases where the model predicted the positive class, but the actual value was negative.

False Negatives (FN): T cases where the model predicted the negative class, but the actual value was positive.

The confusion matrix of the Mobile Net V2 as shown in figure 7, depicts the evaluation of classification models. It includes certain parameters such as Precision, Recall, Accuracy, F1-score, and Support vectors. It provides the abstract of a prediction of a classification model compared to the actual true values.



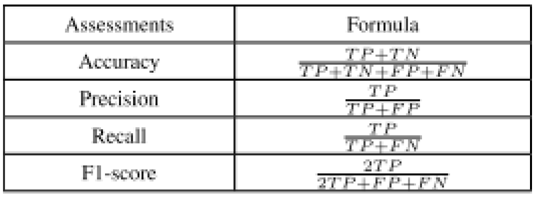
```
Classification Report:
                  precision    recall  f1-score   support

    MildDemented       0.50      0.07      0.13       179
ModerateDemented       0.70      0.58      0.64        12
     NonDemented       0.56      0.97      0.71       650
VeryMildDemented       0.79      0.21      0.34       448

        accuracy                           0.58      1289
       macro avg       0.64      0.46      0.45      1289
    weighted avg       0.63      0.58      0.50      1289
```

Fig 7. Evaluation metrics and confusion metrics of MobileNet V2.

Precision gauges the precision of the positive predictions made by the model, representing the ratio of accurately predicted positive observations to the total predicted positives. On the other hand, recall assesses the model's capability to accurately identify positive instances from the genuine positives in the dataset, denoting the ratio of correctly predicted positive observations to the total actual positives. Accuracy reflects the overall correctness of predictions, indicating the ratio of correctly predicted observations (both positive and negative) to the total observations. The F1 Score serves as the harmonic mean of precision and recall, offering a balanced assessment, particularly crucial in handling imbalanced datasets where one class prevails over the other. The formulas for the employed evaluation metrics are presented in Figure 8, while Table 2 provides additional details, such as input size, learning rate, batch size, epochs, and the optimizer used in the models.



| Assessments | Formula |
|---|---|
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Recall | $\frac{TP}{TP+FN}$ |
| F1-score | $\frac{2TP}{2TP+FP+FN}$ |

Fig 8. The formulas of Evaluation metrics [6].

| Parameters Networks | Initial Input Size | Initial Learning Rate | Batch Size | Epochs | Optimizer |
|---|---|---|---|---|---|
| Inception V3 | 224,224, 3 | 0.001 | 32 | 10 | Adam |
| VGG-19 | 224,224, 3 | 0.001 | 32 | 10 | Adam |
| MobileNet V1 | 224,224, 3 | 0.0001 | 32 | 10 | Adam |
| MobileNet V2 | 224,224, 3 | 0.0001 | 32 | 10 | Adam |
| ResNet 50 | 224,224, 3 | 0.001 | 32 | 10 | Adam |

Authorized licensed use limited to: Univ of Calif Santa Cruz. Downloaded on June 15,2025 at 16:17:13 UTC from IEEE Xplore. Restrictions apply.

| | | | | | |
|---|---|---|---|---|---|
| Custom Model | 224,224, 3 | 0.001 | 32 | 10 | Adam |

Table 2. The networks and their parameters.

MobileNetV2 is a convolutional neural network architecture that is designed for mobile and edge devices that can be used with limited computational resources. It is an evolution of the original MobileNet, developed by Google, aimed at achieving efficient and lightweight neural network models for tasks such as image classification, object detection, and more, particularly on mobile devices [7, 21]. The MobileNet V2 architecture is shown in figure 9.

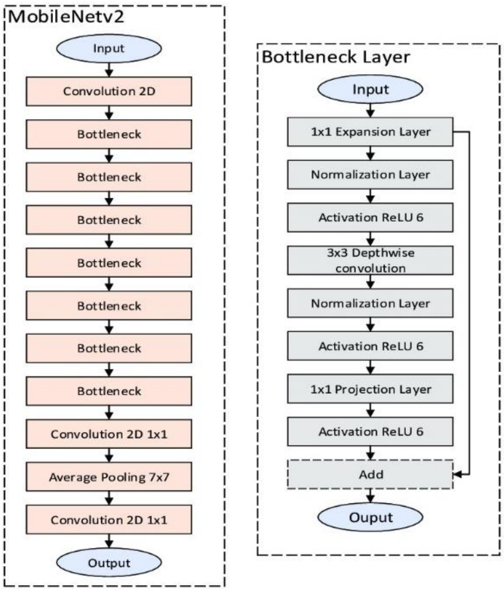


Fig 9. The architecture of MobileNet V2 [8].

The approach employed for interpreting the processes carried out within each layer involves visualizing the outputs generated by the neurons. MobileNet V2 comprises several key layers [9], including:

1. Inverted Residuals with Linear Bottlenecks: MobileNet V2 leverages depthwise separable convolutions, a pivotal component contributing to the model's computational efficiency. The process involves a depthwise convolution, which applies different filters to each input channel, and a 1x1 pointwise convolution that amalgamates information across channels. This dual operation maintains expressive power while demanding fewer computations and parameters than traditional convolutions.
2. Linear Bottlenecks: The linear bottleneck, inserted between depthwise separable convolution layers, incorporates a 1x1 convolution with linear activation, devoid of non-linearity. This linear bottleneck simplifies tuning and enhances the network's capability to capture nuanced features.
3. Inverted Residual Block Structure: The construction of the inverted residual block in MobileNet V2 involves linear activation, 1x1 pointwise convolution, depthwise separable convolution (with linear bottleneck), and a shortcut link connecting the input to the output. This structure promotes efficient learning and information transfer across the network.
4. Global Average Pooling and Final Dense Layer: Unlike conventional fully connected layers, MobileNet V2 employs global average pooling near the network's conclusion. The global average pooling reduces spatial dimensions to a single value per channel before a final dense layer with softmax activation is applied for classification. This design enhances efficiency, particularly for deployment on resource-constrained devices like mobile phones.

The outputs of certain layers are illustrated, such as the output of the 1st layer (Convolution Layer) capturing low-level features like edges and corners, depicted in Figure 10. Convolutional layers often integrate operations like max pooling to condense the spatial dimensions of extracted features while retaining essential information, thereby focusing on salient features and reducing computational complexity. The hierarchical representation generated by the convolutional layers is utilized by subsequent layers for classification or prediction tasks.

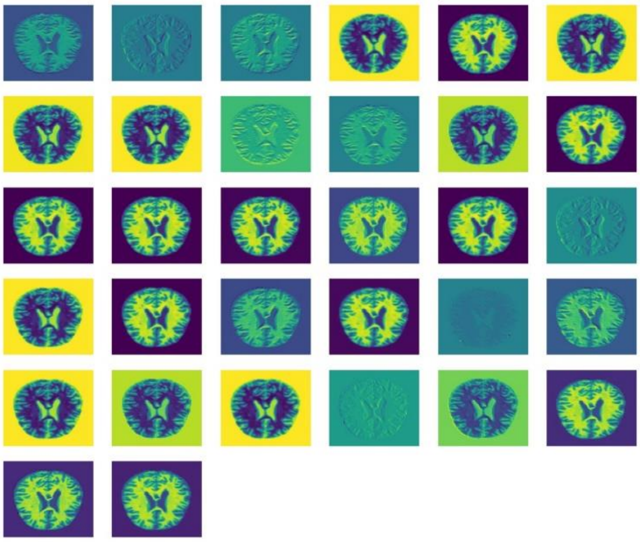


Fig 10. The feature of the output of $1^{st}$ layer of MobileNet V2

The $5^{st}$ layer (Expanded Convolution depthwise) performs the spatial filter independently for each input channel, instead of having a single convolution filter for each channel, it applies a separate filters to each input channel reducing the parameters. This is a technique used in neural network architectures to reduce the computational complexity of standard convolutions by decomposing them into two separate operations: depthwise convolution and pointwise convolution. The depthwise convolution applies a single filter per input channel, and the pointwise convolution performs a 1x1 convolution to combine the outputs as shown in figure 11.

Authorized licensed use limited to: Univ of Calif Santa Cruz. Downloaded on June 15,2025 at 16:17:13 UTC from IEEE Xplore. Restrictions apply.

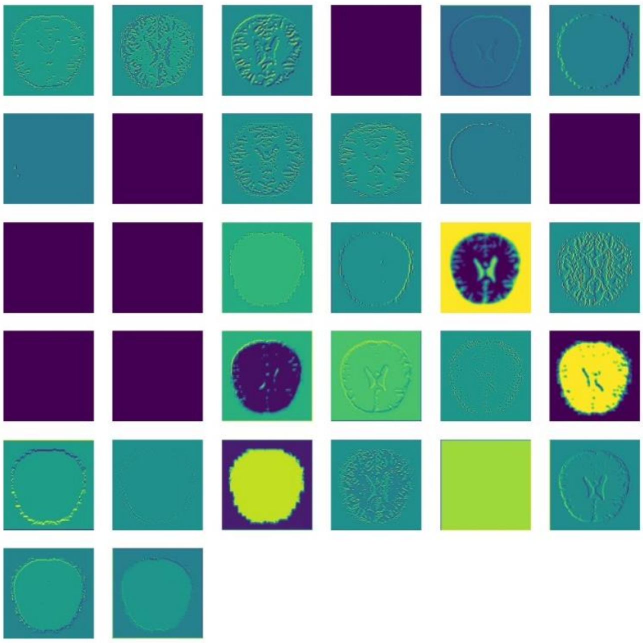


Fig 11. The features of the output of 5th layer of MobileNet V2

The 15th layer (Batch Normalization) normalizes the input of each layer, within a neural network by adjusting and scaling the activations. It does this by normalizing the inputs of each layer in a mini-batch (a subset of the training data) to have a mean close to zero and a variance close to one. shown in the figure 12.

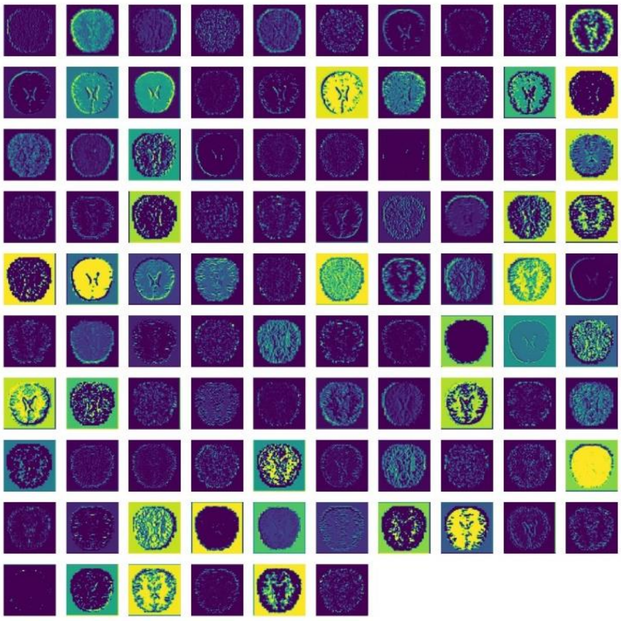


Fig 12. The output of 15th layer of MobileNet V2

The convolution layer 153th (Out Relu) is generating the final output that is flattened and is going to 155th (Dense layer) for final output [11]. It introduces non-linearity by allowing the network to learn complex patterns in the data. It replaces all negative pixel values in the feature map with zero while leaving positive values unchanged as shown in the figure 13.

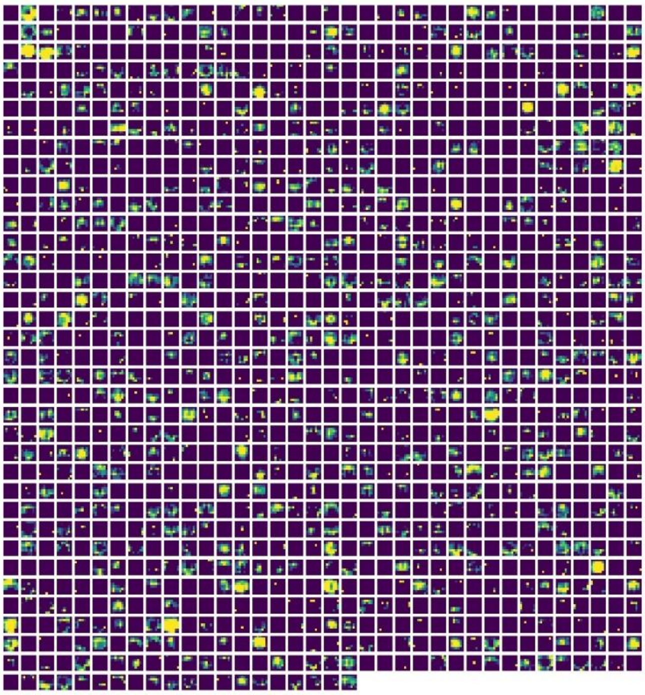


Fig 13. The features of the output of 153th layer of MobileNet V2

Now the normalization in each layer are performed. The layer normalization normalizes across the feature dimension for each individuals. It is different from batch normalization because the batch normalization normalizes the batch dimension while layer normalization normalizes the layer feature dimension.

Then the PCA (Principal Component Analysis) [12] was applied in the layer outputs. The PCA allows us to reduce the dimensionality while preserving the most of the variance in the data. It represents the layer output in the lower dimensional space [10, 17, 18]. It helps us to understand the distribution of features and patterns learned by the network. By examining we can gain insights about the relevant features for Alzheimer Classification. PC's can be more interpretable as shown in the figure 14.

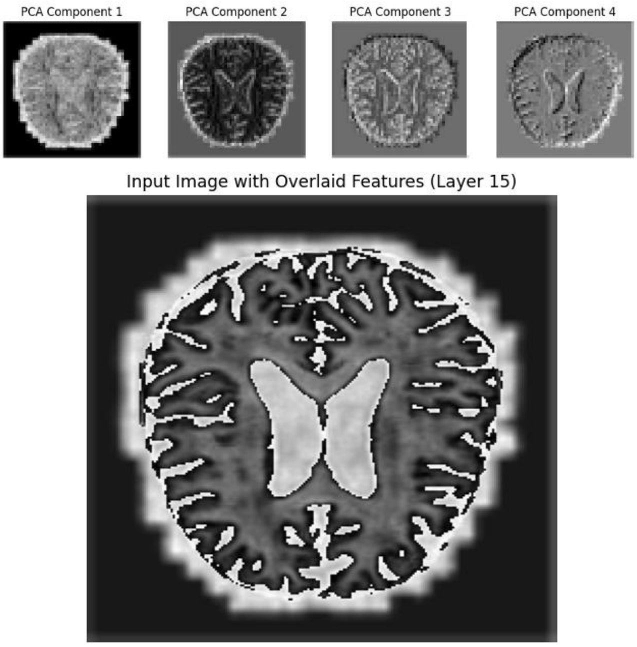


Fig 14. The PCA and its components [10].

Authorized licensed use limited to: Univ of Calif Santa Cruz. Downloaded on June 15,2025 at 16:17:13 UTC from IEEE Xplore. Restrictions apply.

Then the Grade CAM (Gradient weighted Class Activation Map) [14] was used which is a technique used for visualizing and understanding the areas of an input image that contribute the most to final classification decision of a neural network. Grad-CAM, or Gradient-weighted Class Activation Mapping, is used to identify and visualize the areas of an input picture that have the most influence on a neural network's final classification judgment. Any model having a convolutional architecture can use Grad-CAM [15, 16]; no changes to the model architecture are necessary. By examining the gradients of the projected class in relation to the feature maps, it draws attention to significant areas. Priorly we also used normal CAM method to interpret the most contributing part of an input image but it was unable to provide the relevant output which explains the most significant part of an image for final decision as given in the figure 15.

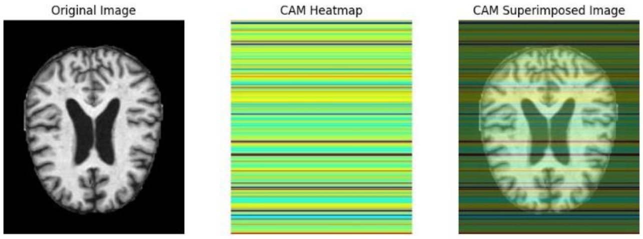


Fig 15. The most significant part of the brain.

So we proceeded with Grade CAM model interprets the target output layer [13] to identify the significant part which gave us the relevant output as shown in the figure 16.

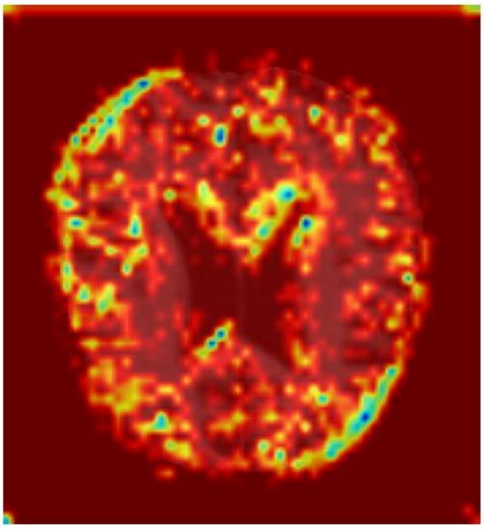


Fig 16. The output of Grade CAM [16].

The blue part in the image is depicting a least significant area and the red part depicts the high significant area and the rest yellow part is depicting the approximate significant values. The results of four different classes shown by Grade CAM technique is given in the figure 17.

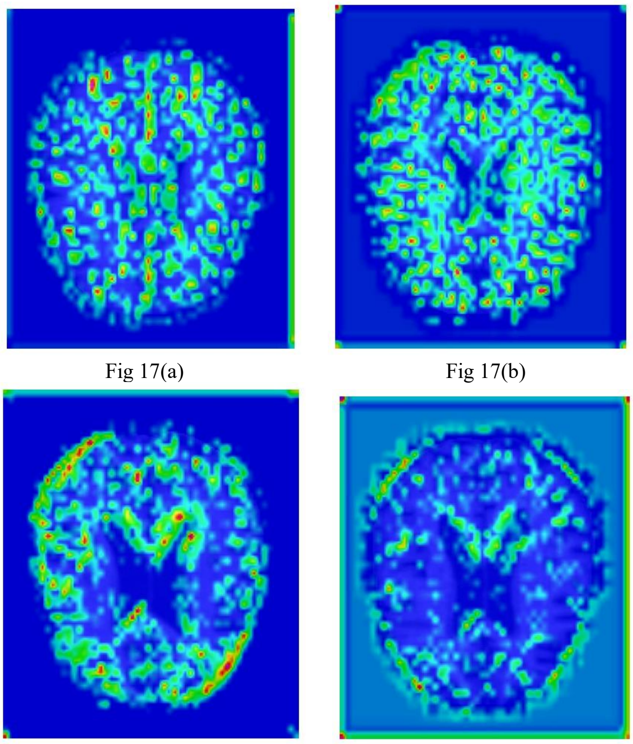


Fig 17(a)

Fig 17(b)

Fig 17(c)

Fig 17(d)

This comprehensive analysis through Grade CAM not only enhances our understanding of the neural network's decision-making process but also provides valuable insights into the intricate dynamics of feature importance during classification. By pinpointing the critical areas contributing to classification outcomes, the Grade CAM visualization unveils the neural network's inherent focus on specific regions within the input images. This nuanced understanding allows us to discern the model's interpretability and ascertain the significance of various anatomical features in the context of Alzheimer's disease classification [19-20].

The Grade CAM technique, by highlighting regions of high significance in vivid red and less critical areas in blue, offers a nuanced perspective on the neural network's attention allocation. This visual representation aids researchers and practitioners in identifying the key features that influence the model's decision, facilitating informed interpretations and potential refinement of the underlying neural network architecture.

Moreover, the utilization of Grade CAM transcends mere visualization; it serves as a powerful tool for not only validating the model's decision but also for potentially uncovering subtle patterns or anomalies that might be crucial for accurate diagnosis. This deeper level of insight contributes to the ongoing efforts to enhance the transparency and reliability of neural network-based diagnostic models, particularly in the context of complex medical conditions like Alzheimer's disease. The integration of interpretability techniques, such as Grade CAM, thus plays a pivotal role in bridging the gap between the model's complex computations and the practical application of its outputs in clinical settings.

Authorized licensed use limited to: Univ of Calif Santa Cruz. Downloaded on June 15,2025 at 16:17:13 UTC from IEEE Xplore.  Restrictions apply.

## III. CONCLUSION

In the pursuit of creating a bridge between artificial intelligence and human understanding, our proposed model emerges as a beacon of clarity in the intricate world of Explainable AI (XAI). By delving into the depths of deep learning and transfer learning, our model not only addresses the complexities of neural networks but also sheds light on the 'black box' problem that often shrouds machine learning models in mystery.

Focused on the realm of Alzheimer's disease, our model's after-effects of sleep-like transparency have the power to revolutionize the field of medical diagnosis. The ability to decode each layer of the neural network, revealing the decision-making processes, opens avenues for medical professionals to gain profound insights into the diagnosis of neurological disorders. It's not just about accuracy; it's about empowering doctors to comprehend the intricate dance of parameters, errors, and biases within the model, paving the way for more accurate and timely patient care.

As we presented the results, the MobileNet V2 model emerged as a frontrunner, showcasing impressive training and testing accuracies. Leveraging techniques such as PCA and Grade CAM, we went beyond traditional evaluation metrics, offering a deeper understanding of the model's inner workings. The Grade CAM's vivid visualization of significant brain areas during the Alzheimer's classification process exemplifies the power of transparency in machine learning.

Beyond the confines of medical diagnosis, the applications of our proposed model extend into various domains, from finance to real-time processing. The transparency it provides serves as a beacon for stakeholders, enabling them to make informed decisions based on a clear understanding of the model's predictions.

In conclusion, our journey through the intricacies of Explainable AI in the context of Alzheimer's diagnosis has not only unveiled the potential of our model but also highlighted the transformative impact it can have on the synergy between artificial intelligence and human interpretation. The after-effects of sleep, metaphorically embodied in our model's transparency, promise a new era where machine learning not only predicts but also enlightens, ensuring a future where AI and human intelligence work hand in hand for the greater good.

## REFERENCES

1. Diseases and Conditions, “Alzheimer’s disease”, Mayo Clinic, https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447, Aug. 2023.
2. Kinza Yasar, “What is Machine Learning and how does it work: The in-depth guide”, Tech Target, https://www.techtarget.com/searchenterpriseai/definition/neural-network, Mar. 2023.
3. Lakeside Manor, “How do you tell if a parent has Alzheimer’s disease?”, https://lakesidemanor.org/how-do-you-tell-if-a-parent-has-alzheimers-disease/, Jun. 2017.
4. S.Dubey, “Alzheimer’s Dataset (4 types of images)”, https://www.kaggle.com/datasets?search=alzheimers+image+dataset&fileType=csv, Sept. 2020.
5. Geeks for Geeks, “Confusion Matrix in Machine Learning”, https://www.geeksforgeeks.org/confusion-matrix-machine-learning/, Dec. 2023.
6. Xue, Dan, et al. "An application of transfer learning and ensemble learning techniques for cervical histopathology image classification." *IEEE Access* 8: 104603-104618, (2020).
7. Hariharan, Kulathumani. "Best Practices: Extending enterprise applications to Mobile devices." *The Architecture Journal, Microsoft Architecture Center* 14, 2008.
8. Tragoudaras, Antonios, et al. "Design space exploration of a sparse mobilenetv2 using high-level synthesis and sparse matrix techniques on FPGAs." *Sensors* 22.12, 2022: 4318.
9. Chandola, Yashvi, et al. "Deep Learning for Chest Radiographs." *Computer Aided Classification, Academic Press*, 2021.
10. “A Guide to Principal Component Analysis (PCA) for Machine learning.” https://www.keboola.com/blog/pca-machine-learning#:~:text=Principal%20Component%20Analysis%20(PCA)%20is,%2Dnoising%2C%20and%20plenty%20more,%20Apr.%202022.
11. Xu, Yuesheng, and Haizhang Zhang. "Convergence of deep convolutional neural networks." *Neural Networks* 153 (2022): 553-563.
12. Daffertshofer, Andreas, et al. "PCA in studying coordination and variability: a tutorial." *Clinical biomechanics* 19.4 (2004): 415-428.
13. Fu, Ruigang, et al. "Axiom-based grad-cam: Towards accurate visualization and explanation of cnns." *arXiv preprint arXiv:2008.02312* (2020).
14. “Grad-CAM reveals the why behind deep learning decisions - MATLAB & Simulink.” https://www.mathworks.com/help/deeplearning/ug/gradcam-explains-why.html
15. Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.
16. Selvaraju, Ramprasaath R., et al. "Grad-CAM: Why did you say that?." *arXiv preprint arXiv:1611.07450* (2016).
17. Maćkiewicz, Andrzej, and Waldemar Ratajczak. "Principal components analysis (PCA)." *Computers & Geosciences* 19.3 (1993): 303-342.
18. Granato, Daniel, et al. "Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective." *Trends in Food Science & Technology* 72 (2018): 83-90.
19. Chethana, Savarala, et al. "A Novel Approach for Alzheimer’s Disease Detection using XAI and Grad-CAM." *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*. IEEE, 2023.
20. Li, Qi, and Mary Qu Yang. "Comparison of machine learning approaches for enhancing Alzheimer’s disease classification." *PeerJ* 9 (2021): e10549.
21. Bukhres, Omran A., et al. "A proposed mobile architecture for a distributed database environment." *PDP*. 1997.

Authorized licensed use limited to: Univ of Calif Santa Cruz. Downloaded on June 15,2025 at 16:17:13 UTC from IEEE Xplore. Restrictions apply.