ISSN: 2182-2069 / E-ISSN: 2182-2077

Deep Feature Extraction and Classification of Alzheimer's Disease: A Novel Fusion of Vision Transformer-DenseNet Approach with Visualization

P. Archana Menon[1*], and Dr.R. Gunasundari[2]

[1*]Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India; Department of Cyber Security and Applied Computing, St. Teresa's College (Autonomous), Ernakulam, India. archananirmal1414@gmail.com, https://orcid.org/0000-0003-3103-2200

[2]Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore, India. gunasoundar04@gmail.com, https://orcid.org/0000-0003-4157-285X

Received: July 20, 2024; Revised: August 29, 2024; Accepted: October 1, 2024; Published: November 30, 2024

Abstract

Alzheimer's Disease (AD) classification from brain MRI images remains a strenuous mission due to the intricacy of the disease and limited dataset sizes. Eventhough Convolutional Neural Networks (CNNs) have excelled in the classification of brain diseases using MRI data, they are incompetent to apprehend global dependencies. Also, their results are not interpretable which is a major problem in medical domain. Transformer uses attention mechanisms to go with or even surpass CNNs on various vision tasks. This study proposes a novel fusion model integrating the complementary advantages of DenseNet-121 and Vision Transformer to address these challenges. By synergizing the strengths of both architectures, the proposed fusion model extracts comprehensive image features. To further optimize feature discrimination and computational efficiency, an ExtraTree classifier-based feature selection technique is incorporated. The performance of the proposed model is evaluated using standard metrics and compared with state-of-the-art techniques. Results demonstrate a superior classification accuracy of 99%, with the fusion model effectively differentiating between various AD stages. Furthermore, Class Activation Maps (CAMs) are utilized to visualize the model's decision-making process thereby enhancing trust in the model's predictions. We also provided a visual comparison of Grad-CAM, Grad-CAM++, Score-CAM, and Faster Score-CAM visualization techniques to assess the performance of these in highlighting discriminative regions for AD classification.

Keywords: Alzheimer's Disease, CAM, Deep Learning, DenseNet, Feature Extraction, Vision Transformer, Visualization.

1 Introduction

Alzheimer's disease (AD) is a deadly and traumatic brain illness that damages brain tissue and results in the death of neurons. Over 50 million people globally were estimated to have AD or associated dementias in 2020 (Zeisel et al., 2020). Since AD is neither totally curable nor irreversible, early diagnosis helps doctors treat patients at the appropriate time, which is essential to preventing significant

Journal of Internet Services and Information Security (JISIS), volume: 14, number: 4 (November), pp. 462-483.
DOI: *10.58346/JISIS.2024.I4.029*
*Corresponding author: Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India; Department of Cyber Security and Applied Computing, St. Teresa's College (Autonomous), Ernakulam, India.

brain damage. Giving patients therapy at the appropriate moment helps them control their behavioural problems, preserve their physical and mental well-being, and halt the progression of their illness (Gauthier, 2005).

Researchers are excited about using Deep Learning (DL), particularly Convolutional Neural Networks (CNNs), to analyze brain scans (neuroimaging data) and automatically diagnose AD and other neurological conditions (Razzak et al., 2018). CNNs have revolutionized the computer vision field by accurately classifying images. CNNs recognizes objects within images, and they've been used earlier to classify AD based on Magnetic Resonance Imaging (MRI) scans (Sarraf et al., 2016). Though they produce higher accuracy in image recognition, one of the major drawbacks of CNNs is the requirement of large and labelled training data. One effective approach to address the limitations posed by small medical image datasets is Transfer Learning (TL). This method makes use of the features learned from one problem on a new related problem (Menon & Gunasundari, 2022). In TL, a model that has already been trained on a bigger dataset will be fine-tuned for the particular task (Menon & Gunasundari, 2022). TL techniques have been effectively implemented previously for AD classification, showcasing its ability to enhance performance even with limited dataset.

Ramzan et al., (2020) used resting-state functional MRI (rs-fMRI), which is a neuroimaging mechanism for the classification of multiple progressive stages of AD. They investigated ResNet-18 architecture with and without finetuning in ADNI dataset and found that their network classified the subjects significantly after finetuning. The results were evaluated and is proved that if fMRI is used along with advanced DL techniques for the prediction of neurodegenerative diseases, it could bring out promising results.

The problem of overfitting due to smaller dataset in deep CNN models have been addressed (Mehmood et al., 2020) by introducing Siamese Convolutional Neural Network (SCNN). SCNN is stimulated by VGG-16 for dementia classification. They used OASIS dataset and performed data augmentation techniques for expanding the dataset. The test accuracy attained by the suggested model is remarkable.

Salehi et al., (2020) proposed a DL based solution for the early as well as accurate diagnosis of AD. They implemented DenseNet-169 and ResNet-50 CNN architectures for the classification of AD. DenseNet-169 have shown a superior performance on OASIS and ADNI datasets. It is stated that sophisticated DL techniques and the combination of data from various datasets could improve the model training which would lead to an enhanced performance by the model.

Helaly et al., (2022) suggests a DL framework that uses CNNs for the early diagnosis and classification of AD stages. To evaluate 2D and 3D brain scans, the framework implements TL technique with VGG19 architecture using the ADNI dataset. The suggested method successfully classifies different AD stages into many classes with a high degree of accuracy (above 93%). The authors also suggest a web application that uses these models for remote AD evaluation, considering the constraints imposed by the COVID-19 pandemic. This tool could help patients and physicians diagnose and determine the stage of the disease.

Suganthe et al., (2021) explores the use of CNNs in a DL pipeline for the diagnosis of AD and its phases based on MRI images. Because of minute variations in brain anatomy, it can be difficult to detect AD early and accurately (Faris et al., 2024; Sargunapathi et al., 2020). The suggested model divides people into four categories: mildly demented, moderately demented, very mildly demented and non-demented. It does this by combining the Inception and ResNet architectures. Even with a 79.12%

accuracy rate, the authors admit that their model might still be improved upon. This shows that improving deep learning techniques further could lead to more precise AD diagnosis.

In order to distinguish AD from age-related cognitive loss in senior populations (Rohini & Surendran, 2021) investigates the use of supervised ML. The article clearly highlights the importance of precise forecasting in healthcare for timely interventions and better patient outcomes. The researchers train Support Vector Machines (SVM) for categorisation using a diverse range of data sets such as demographics, neuroimaging, and cognitive evaluations. Their strategy successfully distinguishes between AD and ordinary age-related cognitive decline, as validated by the ADNI dataset. This approach demonstrates the potential of ML to improve diagnostic accuracy.

CNN architectures were dominant in computer vision because of their ability to capture local features in images. Recent experiments have shown that Transformers with self-attention mechanisms are a powerful replacement for CNNs (Gyamfi et al., 2022). As Transformers solves sequence to sequence task, they were first created for Natural Language Processing (NLP) (Vaswani et al., 2017). While CNNs extract features hierarchically, they emphasis on extracting local features. Whereas Transformers use a self-attention mechanism to analyse association among all elements in an input sequence, capturing long-range dependencies (Vaswani, 2017). Vision Transformers (ViT) are engaged in image classification, segmentation, object detection, generative modelling etc. ViT's ability to model complex interactions within data helps them to attain performance that is par with or greater than CNNs and holds immense promise in the domain of image recognition and classification.

Liu et al., (2023) observed the drawbacks of single modality models and came up with the Multi-Modal Mixing Transformer (3MT), a novel DL model for AD classification based on multi-modal data. They performed two experiments with 3MT using clinical and neuroimaging data- (i) classification of AD and Cognitive Normal (CN) persons, (ii) prediction of Mild Cognitive Impairment (MCI) conversion to progressive MCI (pMCI) or stable MCI (sMCI). The model employed Cascaded Modality Transformers architecture with cross attention and modality dropout mechanism to handle missing data. The model was evaluated using ADNI and AIBL datasets.

The 3D Hybrid Compact Convolutional Transformers (3D HCCT) model proposed (Majee et al., 2024) uses CNNs and Vision Transformers (ViTs) to record both internal characteristics and general relationships in 3D MRI scans for the detection of AD. The model was evaluated using ADNI dataset and could produce robust generalization capability and interpretability apart from higher accuracy.

Ilias & Askounis, (2022) introduced two multitask learning models- (i) to identify and classify dementia, (ii) identify the severity of dementia. The multi-task learning models achieved 86.25% accuracy by combining AD categorization with MMSE score prediction. In order to discriminate between AD and non-AD patients, the study's final step examines linguistic patterns in text data, and it finds notable linguistic differences between the two groups. These results highlight the need for interpretable models for improved clinical comprehension while also pointing to the promise of transformer networks and text analysis for AD diagnosis.

The aim of (Roshanzamir et al., 2021) is to explore the possibility of transformer-based language models for early risk assessment of AD. Complex models have historically required big datasets; however, this limitation can be addressed by transformer models that have been pre-trained on large volumes of text data. The model is trained using $BERT_{Large}$ and with an accuracy of 88.08%, this strategy outperforms earlier ones by 2.48%. According to these results, pre-trained language models present a viable way to evaluate AD risk using easily accessible text data, possibly eliminating the requirement for specialised feature engineering.

While TL has proven effective for CNNs in MRI classification, it's still a nascent field for Transformer-based approaches. This emphasizes the established role of transfer learning with CNNs and the lack of its widespread application with Transformers for MRI tasks. Although AD classification has had success with DenseNet architecture, the utilization of ViT in this domain remains relatively unexplored. As far as we are aware, this study is among the pioneering efforts to fuse DenseNet and ViT for a comprehensive analysis of brain MRI images in AD classification. This study presents an efficient and robust fusion approach which combines the strength of both DenseNet architecture and ViT to detect and classify AD patients.

DL models, particularly CNNs, while achieving impressive performance, often suffer from the "black box" problem, hindering interpretability (Menon & Gunasundari, 2024). Without understanding the model's decision-making process, it's hard to trust its outputs, especially in high-stakes medical applications. If a model is not interpretable, there are certain implications such as: the clinicians may be hesitant to adopt it; identifying and correcting model errors becomes more difficult and many regulatory bodies require explainable models for medical device approval. Overfitting is another critical challenge that can inflate accuracy metrics (Menon & Gunasundari, 2024). To address these issues, this paper proposes visualization techniques to comprehend how the model takes decision. Specifically, Grad-CAM, Grad-CAM++, Score-CAM, and Faster Score-CAM were utilized to uncover the underlying reasoning behind the model's classifications.

Significant contributions of the study are given below:

- Employing the DenseNet-121 architecture to extract discriminative features for AD classification.
- Exploring the application of Vision Transformers (ViTs) to extract features for classifying AD.
- Enhancing Feature Extraction Efficiency through ViT and DenseNet Fusion framework.
- Boosting Feature Selection for AD Classification through the ExtraTrees Classifier.
- Utilizing Class Activation Maps (CAMs) for Visual Explanation of model predictions. We also provide a visual comparison of Grad-CAM, Grad-CAM++, Score-CAM, and Faster Score-CAM approaches.
- Empowering clinicians in early and precise AD diagnosis and improved patient outcomes by the explainable framework.

The paper is structured as follows: Section II explains the resources used in this investigation. Section III introduces the proposed fusion model architecture, incorporating CAM visualization techniques for interpretability. The model's implementation, experimental findings, and analysis are detailed in Section IV. A summary of the contributions and possible directions for further study are outlined in Section V.

2 Preliminaries

A. Densenet 121

The vanishing gradient problem, which occurs when information about the input or gradient decreases as it propagates through several layers, has historically been caused by the deeper CNNs. Direct connections between layers are introduced by DenseNet to address this problem and enable more effective information flow in both forward and backward propagation. A DenseNet's feature maps are enhanced and more efficient. Feature reuse is made possible by the inputs that each layer receives from all layers that came before it. Consequently, DenseNet minimises the total number of attributes, improves feature propagation, and addresses the vanishing gradient issue (Kateb et al., 2023). In order

to produce a predicted label, a traditional CNN usually processes an input image through a number of layers. This forward pass, is seen in Figure 1 (Kumar, 2023).

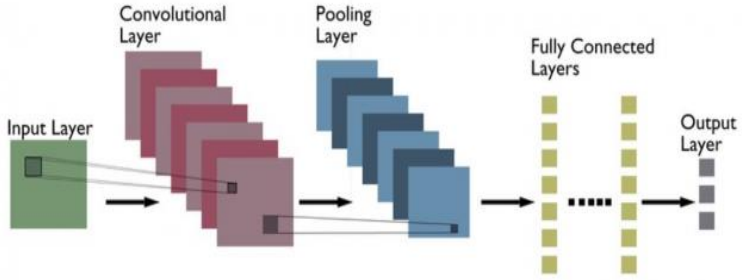


Figure 1: Simple CNN Architecture

With the exception of first layer, which processes the input image directly, every convolutional layer in a DenseNet architecture creates an output feature map by concatenating its own features with those from every previous layer (Kateb et al., 2023). DenseNet fundamentally differs from conventional CNN topologies, as seen in Figure 2 (Dense Block Explained). The term "Densely Connected Convolutional Network" comes from the way its layers are coupled to one another, all of the layers are directly connected to each other. Hence, there are N(N+1)/2 direct connections in an N-layer DenseNet (Kateb et al., 2023). This architecture makes it easier for data to move from earlier layers to later ones, allowing features to be reused across the network. There are several configurations of DenseNet designs and we utilise the DenseNet-121 model in this investigation where the number 121 indicates the approximate number of layers in the network.

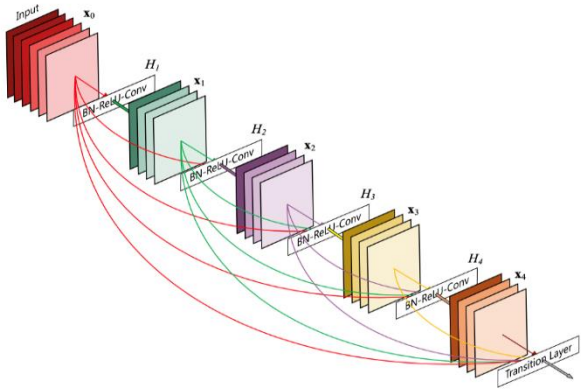


Figure 2: Densenet Architecture

B. Vision Transformer

ViT models process images as linear sequences of data points and predict class labels, enabling them to learn image patterns independently of their spatial structure as shown in Figure 3 (Dosovitskiy et al., 2020). Instead of processing the entire image, transformers divide it into smaller portions. Following channel concatenation to generate a single vector, each of these patches is linearly projected to the desired input dimension. The vector represents the complete patch and is utilized as a transformer input. Transformers can apprehend both local and global attributes by dividing the image into patches, allowing

them to represent long-term relationships and interactions between different portions of the image. The ViT architecture involves partitioning an image into smaller sections, flattening these sections into one-dimensional vectors, transforming the vectors into lower-dimensional embeddings, adding positional information into the embeddings to capture spatial relationships, feeding the resulting sequence into a Transformer encoder for feature extraction, and finally, training the model on a huge dataset for adapting the model to specific image recognition tasks (Azad et al., 2024).

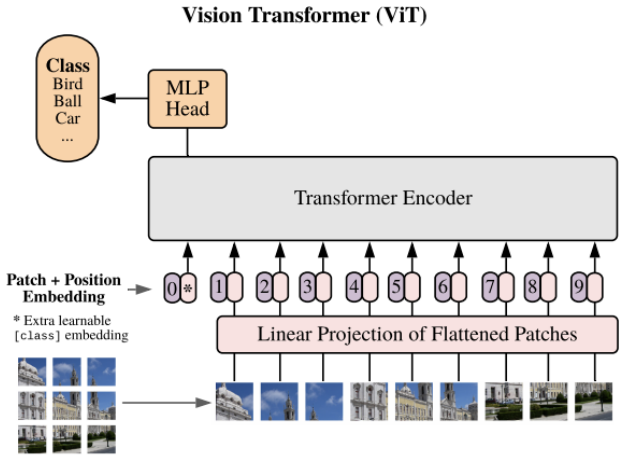


Figure 3: Overview of Vision Transformer

A ViT encoder is composed of multiple blocks, each containing (i) Layer Normalization, which adjusts to changes in images and stabilises training, (ii) Multi-Head Self-Attention (MSA), which creates attention maps to help focus on significant areas of the image and (iii) Multi-Layer Perceptron (MLP), which processes information and generates the final output (Azad et al., 2024).

C. Extratree Classifier

ExtraTrees, or Extremely Randomised Trees, is a robust ensemble learning technique which constructs multiple decision trees with random splits and averaging their predictions enhances the model's reliability and accuracy (Awadelkarim, 2024). Features are ranked using their scores based on their importance in the decision-making process and trees are built using these important features. Features are selected randomly from each tree which makes the trees diversified and in turn reduces the problem of overfitting and also improves generalisation. Top ranked features are considered as the most relevant features for the model (Awadelkarim, 2024). This strategy is beneficial when the dataset is huge and features are abundant. ExtraTree classifiers are fast in computation, efficient in removing irrelevant features and handles large and varied datasets.

D. Class Activation Maps (CAMs)

CAMs are a technique used by CNN models to determine the differential image areas for image classification. CAMs makes the model transparent by visualising the areas of an image that a CNN concentrates on during a prediction (Zhou et al., 2016). It explains the model’s decision-making process. CNN process images and extract features using various convolutional and Global Average Pooling (GAP) is applied on the last convolutional layer’s output to produce a feature vector. To generate class

probabilities, the feature vector is put into a layer which is fully connected. The feature maps employ the weights that link the GAP layer to the output neurons. The class activation map which is produced by calculating the weighted sum of the feature maps, shows the regions of the image that made the biggest contribution to the prediction. CAM sheds light on the logic of the model. It assists in locating possible problems with the model, such as misclassifications. CAM reveals which parts of the image are most pertinent for the prediction (Zhou et al., 2016). To enhance model interpretability, we employed four state-of-the-art visualization techniques: Grad-CAM, Grad-CAM++, Score CAM and Faster Score-CAM. These methods helped identify the specific image regions influencing the model's classification decisions, providing valuable insights into its reasoning process.

1) **Grad CAM:** It is a gradient-based CAM technique that identifies salient image regions contributing to a specific class prediction. Here, the gradients from the target class output node are propagated back to the final convolutional layer and generates a heatmap highlighting the most influential image areas (Selvaraju et al., 2020).
2) **Grad CAM++:** It offers an enhancement over Grad-CAM by providing more precise localization of relevant image regions and better handling of multiple objects within a scene. Unlike Grad-CAM, which can be influenced by the presence of multiple similar objects, Grad-CAM++ effectively highlights all relevant instances, providing more accurate and informative visual explanations (Chattopadhay et al., 2018).
3) **Score CAM:** This approach operates in two stages. First, activation maps are extracted from the final convolutional layer, where each map is treated as a mask over the input image. The masked image is then fed through the network to compute its class score. Second, these scores are used as weights for a linear fusion of the activation maps, generating the final class activation map. This gradient-free approach contrasts with methods like Grad-CAM and Grad-CAM++, which rely on backpropagated gradients for weight calculation (Wang et al., 2020).
4) **Faster ScoreCAM:** It is an optimized version of the Score-CAM technique designed to improve computational efficiency. It addresses the computational bottleneck of Score-CAM by focusing on a subset of channels within the activation maps (Li et al., 2023).

3 Methodology

This study's principal goal is to accurately categorize dementia levels in AD by constructing a robust classification model that efficiently merges ViT and DenseNet architecture utilising brain MRI images. Figure 4 depicts the workflow of the proposed model. Brain MRI images were sourced from Kaggle and underwent initial preprocessing. Subsequently, these preprocessed images are fed into both a DenseNet-121 model and a ViT for feature extraction. An ExtraTrees classifier is used to draw out features from the resulting high-dimensional feature vectors from both models in order to identify the most discriminative features. These reduced feature sets are integrated and input into a fully connected dense network for AD classification. AD is categorised into 4 classes- Moderately Demented, Mildly Demented, Very Mildly Demented and Non-Demented. Then the performances of the models are assessed and compared using evaluation metrices. Finally, CAMs such as Grad-CAM, Grad-CAM++, Score-CAM, and Faster Score-CAM were employed to enhance the model's interpretability by visualising the decision-making process.

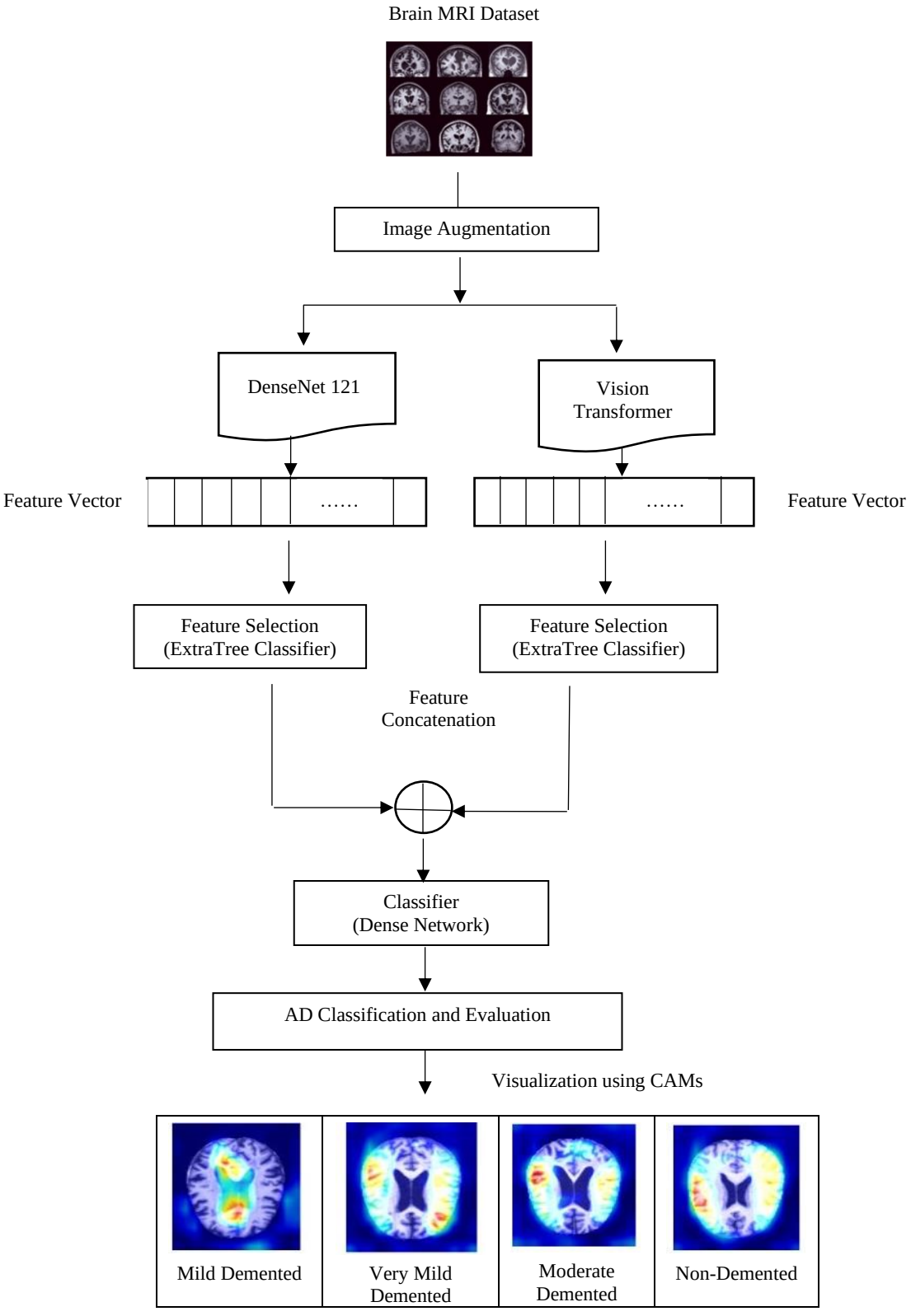


Figure 4: Flow Diagram of the Proposed Fusion Model

A. Data and Preprocessing

When it comes to diagnosing brain abnormalities, MRI is a useful diagnostic tool. An individual with AD exhibits noticeable alterations on their MRI, especially in the parietal and temporal lobes. The openly accessible Brain MRI dataset is collected via Kaggle (Kumar, 2022). There were 6400 Brain MRI pictures in the collection. Every image has been reduced in size to 128 by 128 pixels. Four classifications as shown in Table 1 comprises the entire dataset: Mildly Demented, Very Mildly Demented, Moderately Demented, and Non-Demented. To expose the model to different image variations and increase the amount of training data, methods for data augmentation like contrast, brightness, rotation, flipping, and saturation are applied on the original dataset. This also makes the model robust by reducing overfitting. The application of data augmentation techniques has resulted in a significantly expanded dataset of 10,737 images. Subsequently, the dataset is divided into training and testing sets and then given to the two pretrained models.

Table 1: Data Labelling

Label	**State**
Class 1	Mild Demented
Class 2	Very Mild Demented
Class 3	Moderate
Class 4	Non-Demented

B. The Fusion Model

Figure 5 demonstrates the architecture of the proposed fusion model for AD classification.

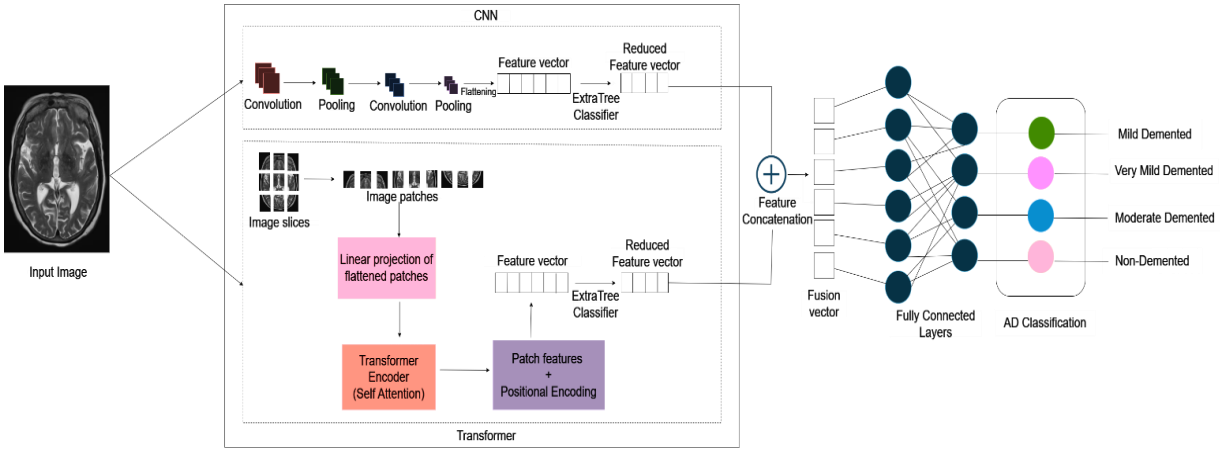


Figure 5: Architecture of the Fusion Model for AD Classification

Feature extraction from preprocessed images is the foundational step in image classification. This process involves transforming raw pixel data into numerical representations that capture essential image characteristics. Each image encapsulates wide range of data, from high-level semantic ideas like objects,

scenes, and relationships to low-level details like edges, textures, and colours. Extracting the features of Brain images effectively is crucial for AD classification task. In reality, CNNs have emerged as the standard for extracting features from the images. As the network gets deeper, its hierarchical architecture, which is composed of fully linked layers, pooling layers, and convolutional layers, allows them to learn progressively more complicated features. CNNs are very good at identifying spatial correlations and local patterns in images. Two well-known CNN architectures that have shown remarkable feature extraction performance are DenseNet and Vision Transformers. DenseNet enabled optimal information flow and effective feature reuse by introducing dense connections between layers. It has been demonstrated that this architecture produces cutting-edge outcomes on various image classification benchmarks. A more recent development is the use of NLP's attention mechanism in ViTs. ViTs can capture long-range relationships and global interdependence by treating images as sequences of patches, which improves the performance.

Deep neural networks, such as DenseNet and Vision Transformers, are capable of extracting a vast number of features from images. These surplus features can result in computational inefficiency, increased model complexity, and even overfitting. By selecting the most relevant features using feature selection techniques, these problems can be addressed. The ExtraTrees classifier is particularly useful for this purpose since it effectively ranks features according to their relevance. ExtraTree classifiers selects important features from the huge feature sets produced by both the DenseNet and ViT and diminish the size of the feature sets. The resultant feature vectors from DenseNet and ViTs are integrated and subsequently given to a fully connected network for the final classification task. While DenseNet and ViTs each have their strengths, they complement each other well. DenseNet excels in capturing local features and hierarchical representations, while ViTs thrive at modelling long-range interactions and global dependencies. By combining these models, we leverage the strengths of both architectures. Combining the features from these two models produce a feature space that is more expansive and discriminative, which could enhance classification performance. Because the combined model exposes the model to many feature views, it is less prone to overfit than a single architecture. Also, by combining the features from the two models, makes the fusion model more resilient to changes in the Brain image data. In order to classify AD into four final categories, the reduced, salient feature set obtained by ExtraTree classifier is fed into the four layered fully connected network. The performance of the model is thoroughly evaluated using common metrics including F1-score, accuracy, precision, and recall. Figure 6 depicts the different layers in the proposed architecture of the fusion model.

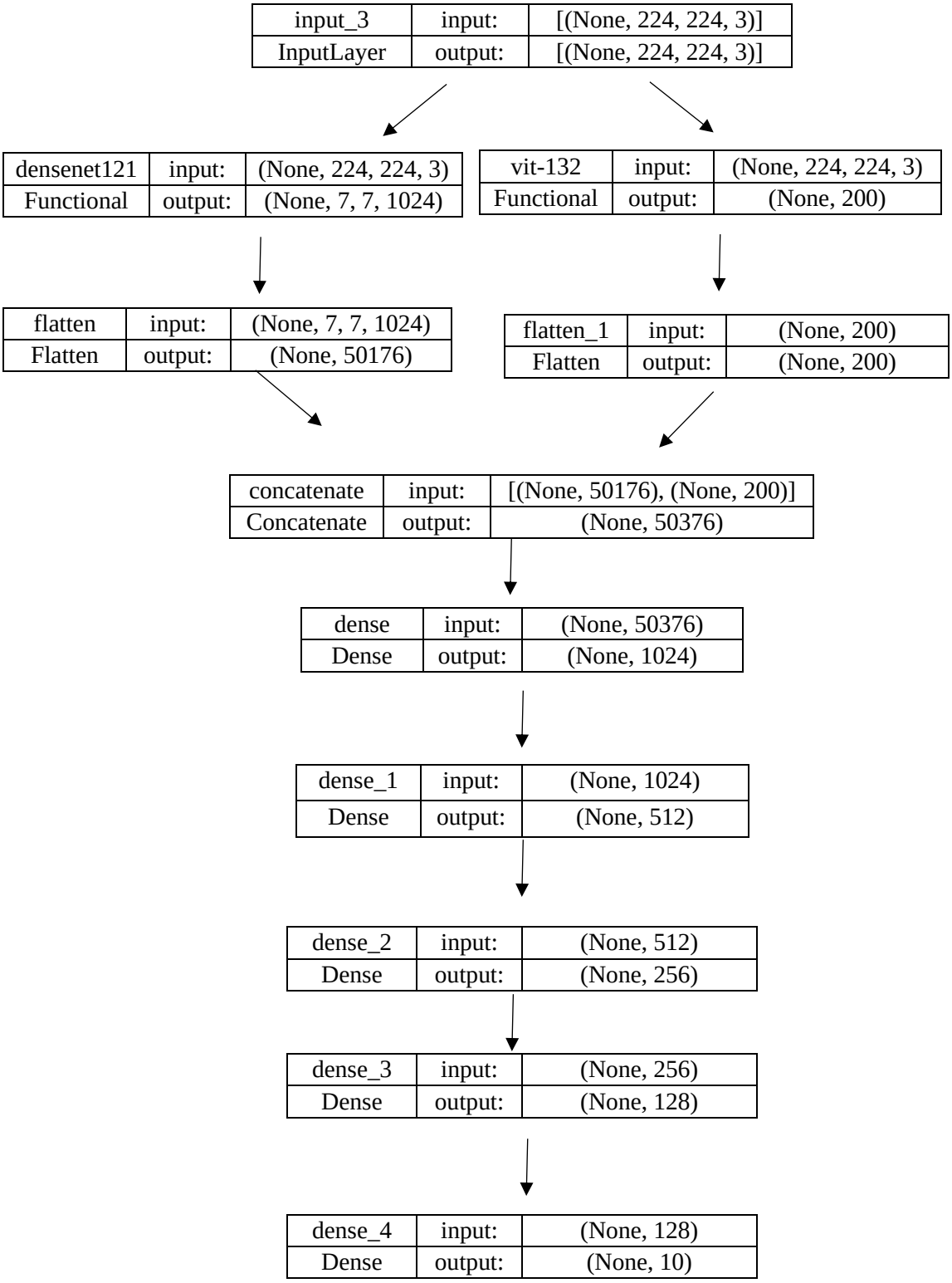


Figure 6: Layers in the Proposed Architecture

C. Visualization Using CAMs

Finally, heatmaps of the model's decisions are produced using four distinct CAM techniques: Grad-CAM, Grad-CAM++, Score-CAM, and Faster Score-CAM, in order to improve the interpretability of the model. CAM visualisation techniques not only improves model interpretability but spot the possible misclassifications too. We could confirm the model's predictions and identify the areas affecting the classification choice by verifying heatmaps on the original images. Images are classified into four phases of AD by the fusion model, and the CAM techniques uses these classified images as input to localise the demented regions.

4 Results And Discussions

A. Experimental Setup

Three distinct classification experiments were conducted: the first employing DenseNet-121, the second utilizing a Vision Transformer, and the third incorporating a fusion of both the models. To assess model performance, standard metrics such as accuracy, precision, recall, F1-score, confusion matrices, ROC curves, and precision-recall curves are employed.

The 6400 brain MRI images from the Kaggle dataset were divided into 4 classes. Among these, 896 images were of class Mild Demented, 64 were Moderate Demented, 3200 were Non-Demented and 2240 were Very Mild Demented. To address the class imbalance and improve model generalization, data augmentation strategies such as flipping, rotation, brightness, contrast, and saturation were applied. This resulted in an expanded dataset of 10,737 images.

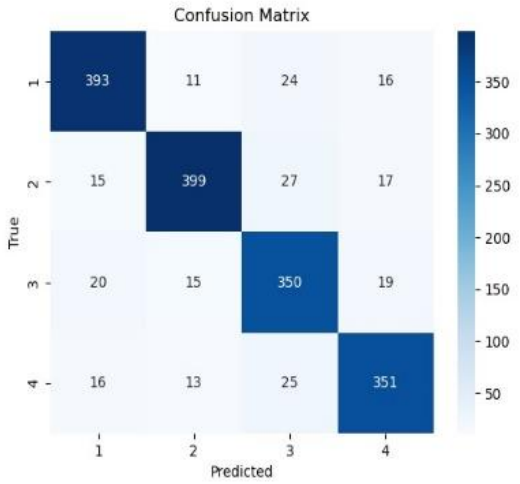
The dataset is split into training and testing sets. The dataset is trained separately in DenseNet 121 model and ViT for feature extraction. DenseNet supports feature reuse so that each layer just needs to learn the difference between its input and the outputs of preceding layers which reduces the number of parameters. DenseNet's efficient use of features through direct connections between layers, coupled with its ability to mitigate the vanishing gradient problem, makes it a powerful architecture for feature extraction in images. The self-attention mechanism allows a ViT to determine the significance of different regions of an image when processing any given part. This is accomplished by figuring the attention scores between various image patches. The attention scores are normalised by the softmax algorithm to make sure their sum equals 1. The resulting attention weights are multiplied by the value vectors to achieve the output. ViTs are able to comprehend the global context of an image by capturing long-range dependencies within it.

The output of DenseNet 121 and ViT are feature vectors. DenseNet-121 and ViT have extracted 6027 and 242 features respectively. Concatenating these two feature sets results in a 6269-dimensional feature vector. While combining features can be beneficial, it's essential to assess the importance of each feature. A 6269-dimensional feature vector might be computationally expensive. To solve these issues due to increased feature size, we have employed a feature selection technique called ExtraTree classifier. ExtraTrees greatly reduces the dimensionality of the feature space by choosing a subset of the most relevant features which contributes to achieve computational efficiency during model training and inference. The reduced feature set, resulting from the application of ExtraTrees classifier on the combined features from DenseNet-121 and ViT, is fed into a fully connected neural network comprising four layers. The final layer consisted of a softmax activation function for the classification of AD into four stages. By combining the complementary strengths of DenseNet and Transformers, we aimed to

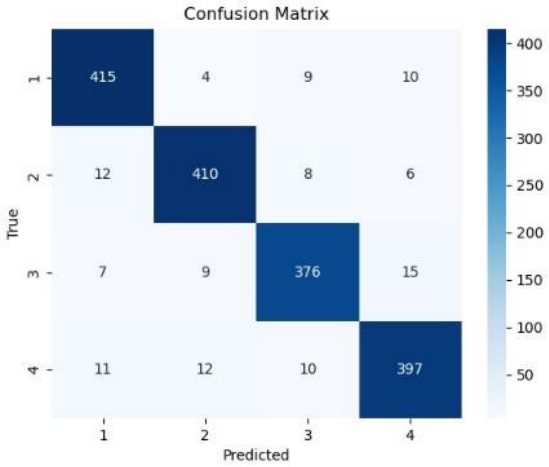
capture a comprehensive set of image features. DenseNet excels at extracting local patterns, while Transformer excels at capturing global dependencies. This image fusion approach improves classification performance by merging data from two powerful models, ensuring that no crucial image details are left unnoticed. This fusion approach ensures that no crucial image information is overlooked, potentially leading to improved classification performance.

B. Performance Evaluation

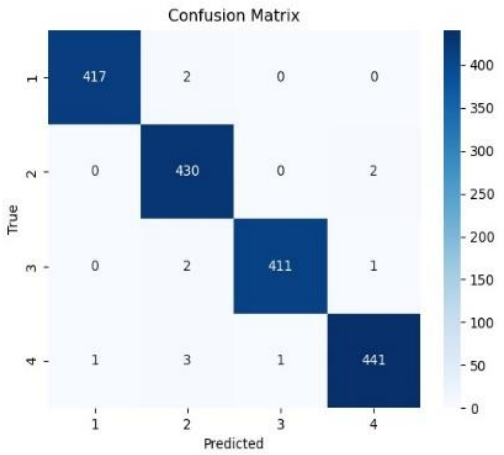
Confusion matrices are created to evaluate the model's performance. The diagonal members of each matrix reveal the accurate classifications, while the off-diagonal elements represent the misclassifications. The distribution of true and predicted labels determines the model's strengths and weakness in differentiating the four classes.



(a) DenseNet 121



(b) Vision Transformer

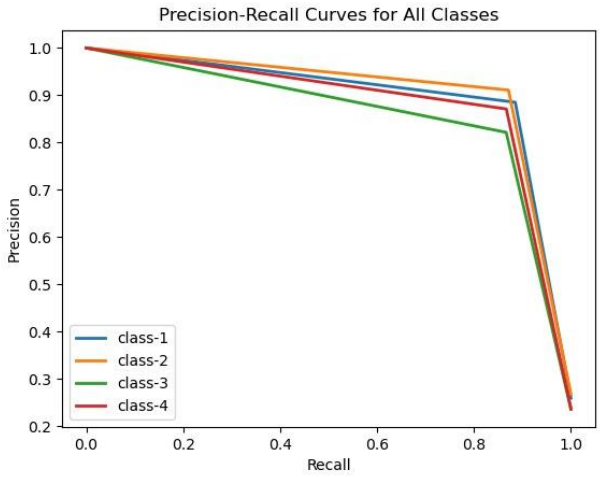


(c) Fusion Model

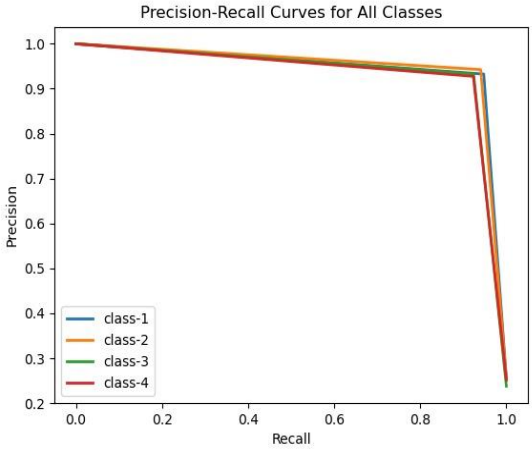
Figure 7: Confusion Matrices of Three Models in AD Classification

Figure 7 presents the confusion matrices for the three models demonstrating their performance in AD classification. It also gives a thorough explanation of classification accuracy for each of the four AD stages. The confusion matrix analysis highlighted strong performance across all the three models, with minimal misclassification of AD stages, indicating a low bias towards any one stage. Notably, the fusion model consistently outperformed the individual models, demonstrating higher classification accuracy.

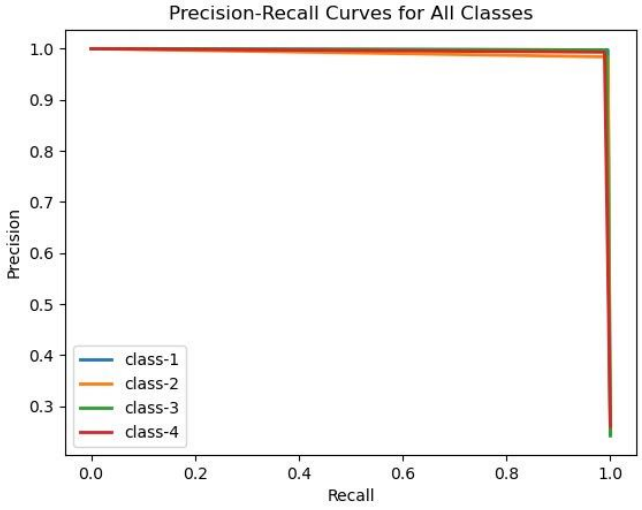
The Precision-Recall curve evaluates the performance of classification models, and are exceptionally good in dealing with imbalanced datasets. Figure 8 shows the Precision-Recall plots for the three models being evaluated. The largest area under the precision-recall curve of the fusion model demonstrates its efficiency compared to the rest of the models. This indicates the enhanced caliber of the fusion model in accurately identifying the positive instances while maintaining a low false positive rate.



(a) DenseNet 121



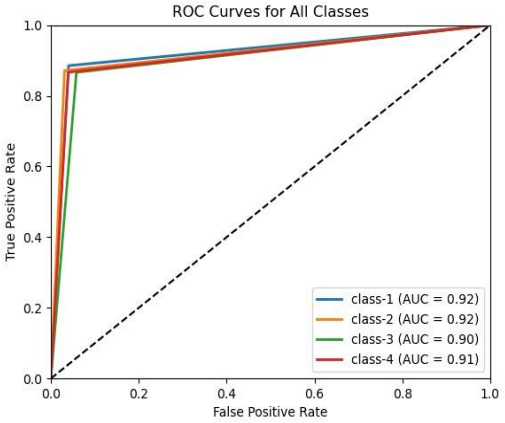
(b) Vision Transformer



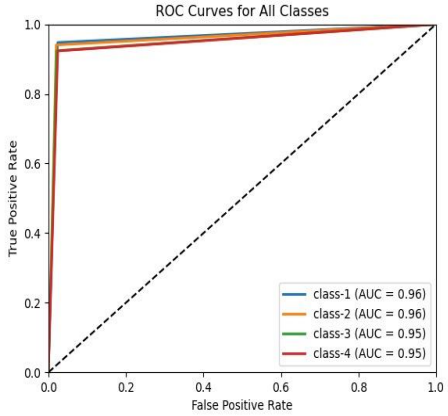
(c) Fusion Model

Figure 8: Precision-Recall Curves of Three Models in AD Classification

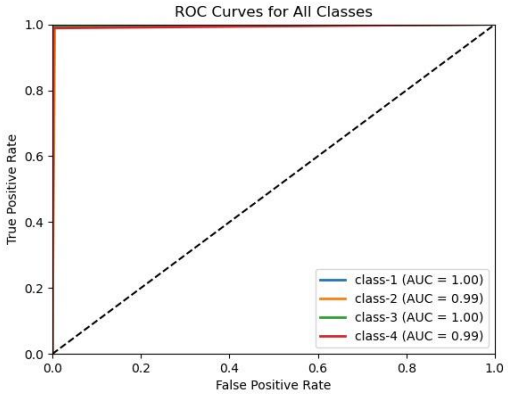
The ROC curves for the three models shown in Figure 9 demonstrates their capacity to distinguish between different stages of AD at various classification thresholds. These curves compare the true positive rate (sensitivity) to the false positive rate and provides a clear picture of the performance of each model. The ROC curve demonstrates that the fusion model achieved the highest overall accuracy in classification compared to the other two models as it has the highest area under the curve (AUC) and that the model could flawlessly categorise both positive and negative cases.



(a) DenseNet 121



(b) Vision Transformer



(c) Fusion Model

Figure 9: ROC Curves of Three Models in AD Classification

C. Comparison between the 3 Models

The suggested model's performance is contrasted with those of other individual networks using metrices such as accuracy, precision, recall and F1-score. As per the results of the study shown in Table 2, the proposed fusion model outperformed the other two models across different assessment criteria which further highlighted the effectiveness of the model. In addition, the recommended network receives a score of 99% for its accuracy. Higher precision value indicates that there is a higher probability of classifying an MRI image into the correct class. Similarly, a high recall denotes that the models are effective in identifying most of the Alzheimer's patients.

Table 2: Evaluation Metrics of the Models

Sl. No	Model	Performance Accuracy	Precision	Recall	F1- Score
1.	Densenet 121	87%	0.87	0.87	0.87
2.	Vision Transformer	93%	0.93	0.93	0.93
3.	**Fusion Model**	**99%**	**0.99**	**0.99**	**0.99**

D. Comparison of the Proposed Fusion Model with Related Models

A comparative analysis of the proposed method with ten existing approaches was conducted using precision recall, and F1-score as evaluation metrics. The findings which are compiled in Table 3, demonstrate how well the suggested fusion model performs in accurately classifying AD stages using brain MRI images. The suggested approach achieves an accuracy of 99%, precision of 99%, recall of 99%, and F1-score of 99%, outperforming the other strategies in these domains. This implies that the suggested approach has mastered the art of classifying AD cases with high accuracy and precision, identifying a significant percentage of true positive cases, and reducing the number of false negatives.

Table 3: Comparison of the Proposed Fusion Model with Existing Models

Reference	Year	Model	Performance Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed Study	**2024**	**Fusion Model**	**99%**	**99%**	**99%**	**99%**
(Menon & Gunasundari 2024)	2024	Ensembled Model (Inception V3+EfficientNet B7)	93	94	94	94
(Hajamohideen et al., 2023)	2023	Siamese CNN	93.85	-	-	-
(De Silva & Kunz, 2023)	2023	CNN	89	89	89	89
(Balaji et al., 2023)	2022	CNN+ LSTM	98	95	98	97
(Li et al., 2022)	2022	Trans-ResNet	93.9	-	-	-
(Helaly et al., 2022)	2022	VGG19	93.61	94	94	94
(Kabir et al., 2021)	2021	18-layer CNN	80.09	95.60	24.72	37.73
(Suganthe et al., 2021)	2021	Inception Resnet V2	79.12	70.64	28.22	39.91
(Qiu et al., 2020)	2020	Fusion- CNN	96.8	-	-	-
(Solano-Rojas et al., 2020)	2020	3D Densenet-121	87	-	-	-

E. Visualization of the Results

The literature review makes it clear that the majority of investigations have focused on the issue of AD classification (Mehmood et al., 2020; Salehi et al., 2020; Helaly et al., 2022; Rohini & Surendran, 2021; Liu et al., 2023; Ilias & Askounis, 2022; Roshanzamir et al., 2021; Ramzan et al., 2020; Majee et al., 2024). However, all those models were viewed as "black boxes," and we find it hard to believe in a model whose workings cannot be fully understood or explained. Furthermore, we cannot afford to make mistakes in disease detection and categorisation systems. The pixels in an image that most influence the model's classification can be shown using visualisation tools. Heatmaps are created in order to localize the area that needed to be classified. Visualization techniques highlights the exclusive pixels responsible for the decision making. This section presents a graphic comparison of Grad-CAM, Grad-CAM++, Score-CAM, and Faster Score-CAM. We anticipate that this will shed light on how to develop more effective models for AD applications. Remarkably, we discovered that every visualisation method

successfully localized the area on the MRI image accountable for that particular decision. Figure 10 displays a heatmap demonstrating how well Grad-CAM, Grad-CAM++, Score-CAM, and Faster Score-CAM identify the responsible areas. On the heatmap, the dark red line stipulates the area accountable for classifying the image into a particular class.

Our results demonstrate the fusion model's ability to capture discriminative features, suggesting its potential as a tool for identifying key biomarkers and improving the diagnosis of AD and potentially other neurodegenerative diseases. The heatmaps in Figure 9 visually demonstrate the fusion model's ability to accurately distinguish between the four AD stages during training. These maps show parts of the brain that are important for categorisation; red areas indicate if an image is classed as mild, extremely mild, moderate, or non-demented. CAM techniques successfully localises regions in these images that may be associated with dementia, which enhances the model's credibility and reliability for classifying AD. The generated heatmap shows the areas of the image that were most prominent in the model's decision. Brighter regions indicate higher importance.

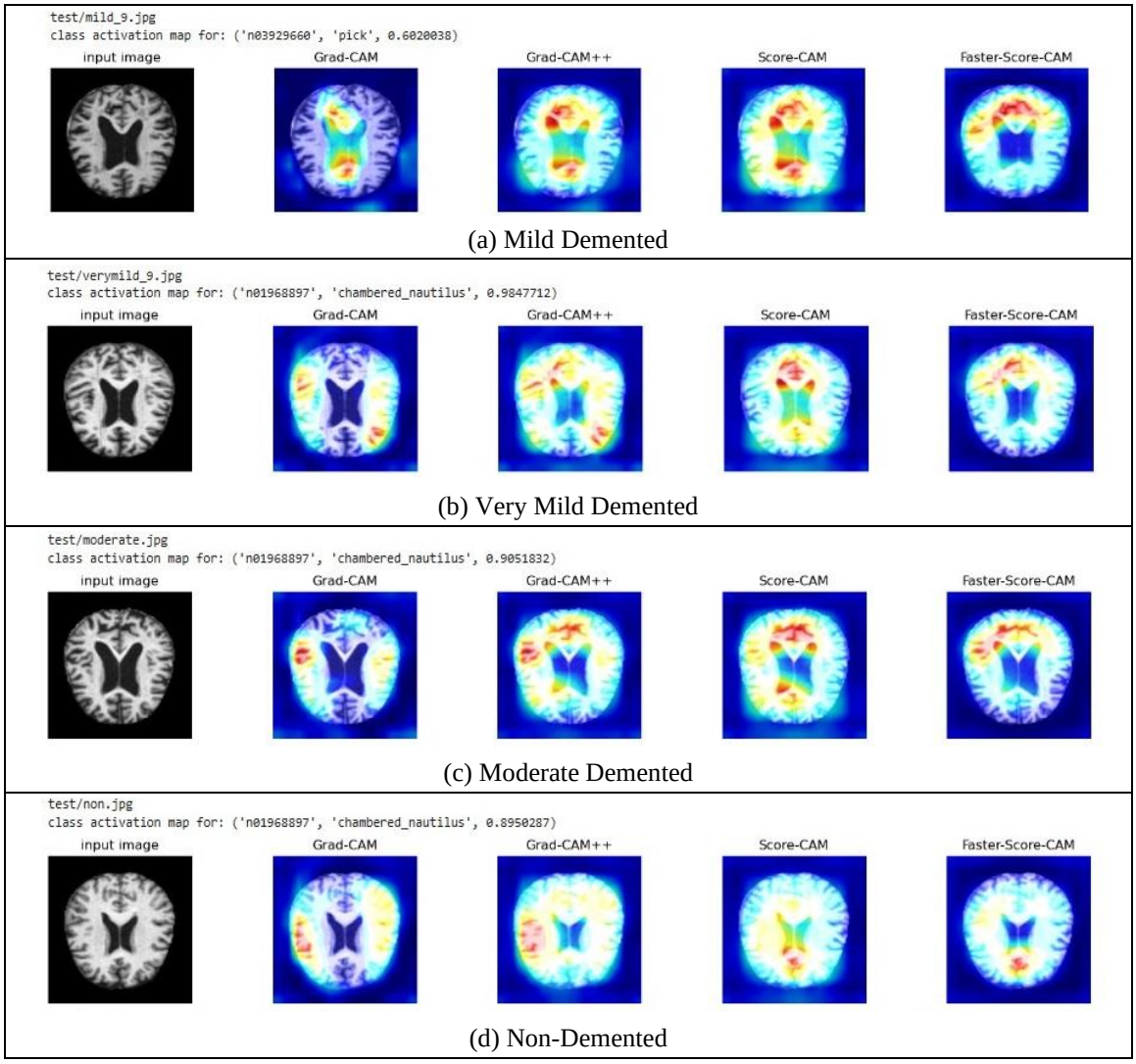


Figure 10: Heatmaps Generated by CAMs-A Visual Comparison of Activation Maps from GradCAM, GradCAM++, Score CAM and Faster SoreCAM for the four Classes - (a) Mild Demented, (b) Very Mild Demented, (c) Moderate Demented and (d) Non-Demented

Some of the key findings from this experiment are:

1. The proposed fusion model effectively addressed the difficulties presented by limited sample size and class imbalance in AD classification. By combining the complementary strengths of DenseNet and Vision Transformer, the model excelled in extracting both local and global features crucial for accurate diagnosis.
2. The incorporation of ExtraTrees classifier significantly improved feature discrimination, ultimately resulting in enhanced classification performance.
3. Leveraging pre-trained models**,** DenseNet-121 and ViT, as baseline classifiers for AD classification served as a benchmark for the proposed fusion model. While these established architectures demonstrated reasonable performance, the fusion approach consistently outperformed both models, highlighting the benefits of combining complementary feature extraction techniques.
4. Comparison of the suggested fusion model with the existing related models proved that the proposed model performs superior in the classification of AD. Also, the confusion matrix, and the curves- precision-recall and ROC- demonstrates the model's ability to distinguish between the four AD stages.
5. The confusion matrix shows the proposed model's performance in AD classification. The diagonal elements representing correct classifications were significantly higher compared to the off-diagonal elements. This reduces the confusion among the classes.
6. The model's ability to balance between precision and recall is represented using the precision-recall curve. The model with a 99% precision and recall indicates its accuracy in recognizing positive cases thereby reducing the false positives.
7. The large area under ROC represents the model's excellent performance. The ROC curve shows that the model could effectively differentiate positive and negative cases even as the classes are imbalanced in nature.
8. Finally, the heatmaps generated by a combination of CAM techniques identified specific regions in the brain MRI associated with AD. CAMs provide insights to the fusion model's decision-making process in AD classification.

5 Conclusion

This study offers an innovative and effective technique for the precise classification of AD stages using brain MRI images by integrating the strengths of DenseNet 121 and Vision Transformer architectures. The proposed fusion model outperformed state of the art techniques effectively by capturing both the global and local image features. The integration of feature selection techniques enhanced model's diagnostic capabilities and improves early detection and clinical management of AD. The versatility of model's ability in discriminating individuals with a range of brain diseases makes it effective for the treatment and management of various neurodegenerative disorders. We have employed four visualization techniques: Grad-CAM, Grad-CAM++, Score-CAM, and Faster Score-CAM, to identify the specific image regions influencing the model's classification of AD stages. These visualizations enhanced the interpretability of the proposed model and gained deeper insight of its reasoning and potential biases. The proposed fusion model combined with feature selection techniques and visualization approaches, offers a robust and interpretable technique for accurate classification of AD stages. These findings could also be implemented for the development of the advanced diagnostic tools for a wide range of brain diseases.

While this study showcases promising results in AD classification, several avenues for future research can enhance its clinical applicability and robustness. Evaluating the model on multiple independent datasets that encompass a variety of demographics and imaging protocols would ascertain the model's reliability and performance across different populations. Engaging with clinicians to ensure the model's effectiveness can offer insightful information into the practical challenges of implementing the model in diagnostic workflows. Collecting and analyzing ancestral data could provide valuable insights into the genetic and hierarchical underpinnings of AD. Conducting longitudinal studies to track the progression of AD in patients over time could offer critical insights into the model's predictive capabilities. By exploring these suggested avenues, researchers can significantly enhance the model's clinical applicability, ultimately contributing to better diagnostic and treatment strategies for AD.

6 Author Contributions

Archana Menon P conceived the original idea, implemented the frame work, performed analytical calculations and wrote the manuscript. R Gunasundari was involved in planning, verifying the results, reviewing the manuscript and supervising the entire work. Both the authors contributed to the final manuscript and discussed the findings.

Conflicts of Interest

The authors have no conflicts of interest to declare.

References

[1] Awadelkarim, S. (2024). Feature Selection using Extra Trees for Breast Cancer Prediction. *Indonesian Journal of Computer Science, 13*(2). https://doi.org/10.33022/ijcs.v13i2.3874

[2] Azad, R., Kazerouni, A., Heidari, M., Aghdam, E. K., Molaei, A., Jia, Y., Jose, A., Roy, R., & Merhof, D. (2024). Advances in medical image analysis with vision Transformers: A comprehensive review. *Medical Image Analysis*, *91*, 103000. https://doi.org/10.1016/j.media.2023.103000

[3] Balaji, P., Chaurasia, M. A., Bilfaqih, S. M., Muniasamy, A., & Alsid, L. E. G. (2023). Hybridized deep learning approach for detecting Alzheimer's disease. *Biomedicines, 11*(1), 149. https://doi.org/10.3390/biomedicines11010149

[4] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 839-847). IEEE. https://doi.org/10.1109/WACV.2018.00097

[5] De Silva, K., & Kunz, H. (2023). Prediction of Alzheimer's disease from magnetic resonance imaging using a convolutional neural network. *Intelligence-Based Medicine*, *7*, 100091. https://doi.org/10.1016/j.ibmed.2023.100091

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

[7] Faris, R. A., Mosa, Q., & Albdairi, M. (2024). Robust Classification for Sub Brain Tumors by Using an Ant Colony Algorithm with a Neural Network. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 15*(2), 270-285. https://doi.org/10.58346/JOWUA.2024.I2.018

[8] Gauthier, S. (2005). Alzheimer's disease: the benefits of early treatment. *European Journal of Neurology*, *12*(s3), 11–16. https://doi.org/10.1111/j.1468-1331.2005.01322.x

[9] Gyamfi, N. K., Goranin, N., Čeponis, D., & Čenys, H. A. (2022). Malware detection using convolutional neural network, a deep learning framework: comparative analysis. *Journal of internet services and information security*, *12*(4), 102-115. https://doi.org/10.58346/JISIS.2022.I4.007

[10] Hajamohideen, F., Shaffi, N., Mahmud, M., Subramanian, K., Al Sariri, A., Vimbi, V., ... & Alzheimer's Disease Neuroimaging Initiative. (2023). Four-way classification of Alzheimer's disease using deep Siamese convolutional neural network with triplet-loss function. *Brain Informatics*, *10*(1), 5. https://doi.org/10.1186/s40708-023-00184-w

[11] Helaly, H. A., Badawy, M., & Haikal, A. Y. (2022). Deep learning approach for early detection of Alzheimer's disease. *Cognitive computation*, *14*(5), 1711-1727. https://doi.org/10.1007/s12559-021-09946-2

[12] Ilias, L., & Askounis, D. (2022). Explainable identification of dementia from transcripts using transformer networks. *IEEE Journal of Biomedical and Health Informatics*, *26*(8), 4153-4164. https://doi.org/10.1109/JBHI.2022.3172479

[13] Kabir, A., Kabir, F., Mahmud, M. A. H., Sinthia, S. A., Azam, S. R., Hussain, E., & Parvez, M. Z. (2021, December). Multi-classification based Alzheimer's disease detection with comparative analysis from brain MRI scans using deep learning. In *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)* (pp. 905-910). IEEE. https://doi.org/10.1109/TENCON54134.2021.9707313

[14] Kateb, Y., Meglouli, H., & Khebli, A. (2023). Coronavirus diagnosis based on chest X-ray images and pre-trained DenseNet-121. *Revue d'Intelligence Artificielle*, *37*(1), 23-28. https://doi.org/10.18280/ria.370104

[15] Kumar, A. (2023). Different Types of CNN Architectures Explained: Examples. Analytics Yogi. https://vitalflux.com/different-types-of-cnn-architectures-explained-examples/

[16] Kumar, S. (2022) Alzheimer MRI preprocessed dataset, Kaggle. https://www.kaggle.com/datasets

[17] Li, C., Cui, Y., Luo, N., Liu, Y., Bourgeat, P., Fripp, J., & Jiang, T. (2022, March). Trans-resnet: Integrating transformers and cnns for alzheimer's disease classification. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE. https://doi.org/10.1109/ISBI52829.2022.9761549

[18] Li, J., Zhang, D., Meng, B., Li, Y., & Luo, L. (2023). FIMF score-CAM: fast score-CAM based on local multi-feature integration for visual interpretation of CNNS. *IET Image Processing*, *17*(3), 761-772. https://doi.org/10.1049/ipr2.12670

[19] Liu, L., Liu, S., Zhang, L., To, X. V., Nasrallah, F., & Chandra, S. S. (2023). Cascaded multi-modal mixing transformers for alzheimer's disease classification with incomplete data. *NeuroImage*, *277*, 120267. https://doi.org/10.1016/j.neuroimage.2023.120267

[20] Majee, A., Gupta, A., Raha, S., & Das, S. (2024). Enhancing MRI-Based Classification of Alzheimer's Disease with Explainable 3D Hybrid Compact Convolutional Transformers. https://doi.org/10.48550/arXiv.2403.16175

[21] Mehmood, A., Maqsood, M., Bashir, M., & Shuyuan, Y. (2020). A deep Siamese convolution neural network for multi-class classification of Alzheimer disease. *Brain sciences*, *10*(2), 84. https://doi.org/10.3390/brainsci10020084

[22] Menon, A. P., & Gunasundari, R. (2024). Intelligent Alzheimer's Disease Prediction Using Explainable Boosting Machine. *Journal of Theoretical and Applied Information Technology*, *102*(6), 2726- 2740.

[23] Menon, A., & Gunasundari, R. (2024, May). Visualizing Alzheimer's Prediction: An Ensembled Deep Learning and Grad-CAM Approach. In *2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS)* (pp. 1-7). IEEE.. https://doi.org/10.1109/RAICS61201.2024.10689918

[24] Menon, P. A., & Gunasundari, R. (2022). Transfer Learning in Medical Image Analysis: A Survey. *Karpagam JCS*, *17*(3), 97-102.

[25] Papers with Code - Dense Block Explained. Aug. 14, 2024. https://paperswithcode.com/method/dense-block

[26] Qiu, S., Joshi, P. S., Miller, M. I., Xue, C., Zhou, X., Karjadi, C., ... & Kolachalama, V. B. (2020). Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain*, *143*(6), 1920-1933. https://doi.org/10.1093/brain/awaa137

[27] Ramzan, F., Khan, M. U. G., Rehmat, A., Iqbal, S., Saba, T., Rehman, A., & Mehmood, Z. (2020). A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks. *Journal of medical systems*, *44*. https://doi.org/10.1007/s10916-019-1475-2

[28] Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of decision making*, 323-350. https://doi.org/10.1007/978-3-319-65981-7_12

[29] Rohini, M., & Surendran, D. (2021). Toward Alzheimer's disease classification through machine learning. *Soft Computing*, *25*(4), 2589-2597. https://doi.org/10.1007/s00500-020-05292-x

[30] Roshanzamir, A., Aghajan, H., & Soleymani Baghshah, M. (2021). Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, *21*, 1-14. https://doi.org/10.1186/s12911-021-01456-3

[31] Salehi, A. W., Baglat, P., & Gupta, G. (2020). Alzheimer's disease diagnosis using deep learning techniques. *International Journal of Engineering and Advanced Technology*, *9*(3), 874-880. https://doi.org/10.35940/ijeat.C5345.029320

[32] Sargunapathi, R., Vinayagamoorthy, P., Sumathi, P., & Sirajunissa Begum, S. (2020). Mapping of Scientific Articles on Brain Tumors: A Scientometric Study. *Indian Journal of Information Sources and Services*, *10*(2), 26–34. https://doi.org/10.51983/ijiss.2020.10.2.490

[33] Sarraf, S., DeSouza, D. D., Anderson, J., Tofighi, G., & Alzheimer's Disease Neuroimaging Initiativ. (2016). DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *BioRxiv*, 070441. https://doi.org/10.1101/070441

[34] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, *128*, 336-359. https://doi.org/10.1007/s11263-019-01228-7

[35] Solano-Rojas, B., Villalón-Fonseca, R., & Marín-Raventós, G. (2020). Alzheimer's disease early detection using a low cost three-dimensional densenet-121 architecture. In *The Impact of Digital Technologies on Public Health in Developed and Developing Countries: 18th International Conference, ICOST 2020, Hammamet, Tunisia, June 24–26, 2020, Proceedings 18* (pp. 3-15). Springer International Publishing. https://doi.org/10.1007/978-3-030-51517-1_1

[36] Suganthe, R. C., Geetha, M., Sreekanth, G. R., Gowtham, K., Deepakkumar, S., & Elango, R. (2021). Multiclass classification of Alzheimer's disease using hybrid deep convolutional neural network. *NVEO-Natural Volatiles & Essential Oils Journal| NVEO*, 145-153.

[37] Vaswani, A. (2017). Attention is All you Need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*, *30*, 5998–6008.

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. N. (2017). Attention is all you need [J]. *Advances in neural information processing systems*, *30*(1), 261-272.

[39] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., ... & Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 111-119. https://doi.org/10.1109/CVPRW50498.2020.00020

[40] Zeisel, J., Bennett, K., & Fleming, R. (2020). World Alzheimer Report 2020: Design, dignity, dementia: Dementia-related design and the built environment. https://www.alzint.org/resource/world-alzheimer-report-2020/.

[41] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929). https://doi.org/10.1109/CVPR.2016.319

Authors Biography

P. Archana Menon, is an Assistant Professor in the Department of Cyber Security and Applied Computing at St. Teresa's College, Ernakulam. She holds a Master's Degree in Computer Applications from Anna University. As a Research Scholar at Karpagam Academy of Higher Education, she has published many research papers. Her current research interests include Explainable AI, Machine Learning, Networking and Cyber Security.

Dr.R. Gunasundari, is a Professor in the Department of Computer Applications at Karpagam Academy of Higher Education. She holds a PhD in Computer Science. She has published articles in Scopus, SCI and peer-reviewed journals. Her areas of interest include datamining, Machine Learning, Artificial Intelligence etc.