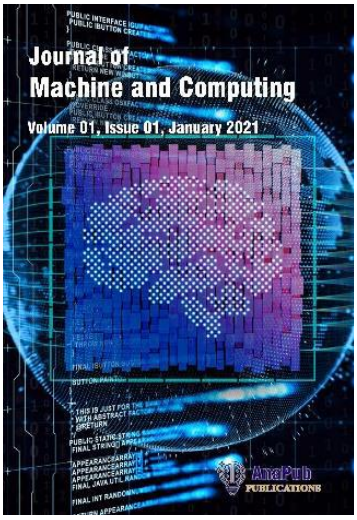
Journal Pre-proof



An Efficient Deep Learning Framework for Accurate Disease Classification

Aruna Kokkula and Chandra Sekhar P

DOI: 10.53759/7669/jmc202505121

Reference: JMC202505121

Journal: Journal of Machine and Computing.

Received 05 January 2025

Revised form 26 March 2025

Accepted 27 May 2025

Please cite this article as: Aruna Kokkula and Chandra Sekhar P, "An Efficient Deep Learning Framework for Accurate Disease Classification", Journal of Machine and Computing. (2025). Doi: https:// doi.org/10.53759/7669/jmc202505121.

An Efficient Deep Learning Framework for Accurate Disease Classification

Aruna Kokkula[1], P. Chandra Sekhar[2]

[1]Department of ECE, University College of Engineering, Osmania University, Hyderabad, India, 500007.

[1]Department of ECE, Maturi Venkata Subba Rao (MVSR) Engineering College, Nadargul, Hyderabad, Telangana 501510

[2]Department of ECE, University College of Engineering, Osmania University, Hyderabad, India, 500007.

[1]arunak_ece@mvsrec.edu.in, [2]sekharp@osmania.ac.in

Abstract

One of the leading causes of memory loss and thinking problems in older adults is a condition that affects human function over time. Detecting this condition early is important for better care and treatment. However, even with the latest technology in artificial intelligence (AI) and deep learning, the results are not convincing because the dynamic nature of the datasets. This study introduces a new deep learning approach that includes a tool called Grad-CAM, which helps explain how the AI makes decisions. Our goal is to build a reliable and understandable system that uses a special type of AI model called a convolutional neural network (CNN) to analyze online dataset images. The model includes techniques to reduce errors and handle different types of data, while Grad-CAM provides visual feedback showing what the model is focusing on. The system achieved 95% accuracy, performing better than other well-known models like Xception (94.40%) and InceptionV3 (93.20%). Overall, this work offers a highly accurate and transparent tool to support early detection of memory-related conditions, assist professionals in planning care, and open new possibilities for research in AI-supported health applications.

Keywords: Deep Learning, Grad-CAM, Convolutional Neural Networks, Classification, Explainable AI

1. Introduction

Alzheimer's disease (AD) is one of the most common and debilitating neurodegenerative disorders, imposing a major burden on life quality for the millions it afflicts globally [1]. It is one of the major causes of dementia in the elderly and is characterized by a progressive decline in cognitive function and memory loss. A timely and accurate diagnosis of Alzheimer's disease is critical to the management of the disease and can lead to improved patient outcomes. As a non-invasive imaging modality, Magnetic resonance imaging (MRI) has proved to be an essential strategy for studying the structural and functional changes in Alzheimer's [2]. On the other hand, the interpretation of manual diagnoses from MRI data leaves room for interpretive errors and necessitates considerable expertise, highlighting the necessity of automated and consistent methods.

Alzheimer's Disease (AD) is increasingly prevalent, bringing significant interest in possible diagnostic solutions utilizing artificial intelligence (AI) and machine learning (ML) [3]. The method has explored some different techniques, but deep learning specifically, has demonstrated great promise in the US for its ability to identify complex patterns and features from medical imaging data. Despite the above, the classification of Alzheimer's disease from MRI data remains a challenging task because, in the early stages of the disease, the subtle brain changes are often camouflaged by normal processes [4]. Moreover, the multi-dimensionality of MRI data demands paradigms capable of isolating disease-characteristic features and providing sufficient specificity.

Several reasons are challenging robust diagnostic model development for Alzheimer's disease [5]. Variations in MRI data due to variations in imaging protocols, scanner settings, and the demographics of the scanned patients make the task difficult. Moreover, the MCI stage may differ from early Alzheimer's disease only with a high level of precision and these features can overlap at this stage [6]. Existing advances themselves are hampered by the scarcity of large, properly annotated datasets, which further compound these issues by restricting the generalizability and robustness of available models. Overcoming these issues requires frameworks that can address data heterogeneity with high classification accuracy.

There is an increasing demand for an accurate, scalable, automated diagnostic framework for Alzheimer's disease [7]. Current methods usually fail to generalize across heterogeneous datasets and therefore can perform very differently in real-world clinical settings. This emphasizes the need for a solution that can extract relevant features from the complex MRI data and be able to adapt to different imaging conditions. In addition, this type of system would improve diagnostic capabilities and assist in early intervention strategies, which, in turn, could prolong disease progression and better the quality of life for patients.

This can be complemented or improved upon if a continuous stream of improvements on classification-based neural network architectures can be obtained [8]. Incorporating a variety of advanced techniques including convolutional neural networks (CNNs) and transfer learning, the framework is capable of handling and learning from MRI data, including extracting features inherent to the pathology by minimizing the effect of variability in the data. Utilizing this framework would yield a more solid and scalable solution, delivering clinicians an accurate and accurate tool for early detection of Alzheimer's disease.

2. Literature Survey

Shaymaa E. Sorour et al [9]. Proposed a deep learning technique-based early diagnosis of the Alzheimer's Disease-Deep Learning framework. Model development, which included pre-processing, training, and evaluation, was performed using brain magnetic resonance imaging scans. We explored five deep-learning models and grouped them according to whether they utilized data augmentation or not—the Convolutional Neural Network-Long Short-Term Memory model performed the best, producing an accuracy of 99.92 percent. The text-based features are designed specifically to optimize accuracy, recall, precision, F1score and computational efficiency. The findings underscore the promise of deep learning for Alzheimer's disease detection.

Doaa Ahmed Arafa et al. [10] provide a CNN-based deep-learning framework for Alzheimer's disease classification. The proposed paradigm encompasses four phases: preprocessing, data augmentation, cross-validation, and classification with feature extraction. We implemented two methods, simple CNN & Pre-trained VGG16 with transfer learning & fine-tuning. Results showed that the framework was effective with a limited number of labels and less domain-specific knowledge. Model: (acc: 99.95%, val_acc: 99.99%) and fine-tuned VGG16 model: (acc: 97.44%, val_acc: 97.40%) It focused on lowered computational complexity, limited over-fitting and reduced memory consumption, resulting in the suitability of the framework for AD diagnosis.

Ahmed A. Abd El-Latif et al. [11] developed a lightweight deep-learning model to detect Alzheimer's disease from MRI data. You are without deeper layers, which does it perform well. It is also less complex and consumes less time as compared to the other existing models with seven layers. On a 36 MB Kaggle dataset 99.22% accuracy on two classes and 95.93% accuracy on multi-class, higher than previous ly the model. Here, this study presents a novel combination of several methodologies of AD detection with the Kaggle dataset as providing new challenges to researchers. The results underline model efficiency, as well as accuracy, in AD classification tasks.

M. Khojaste-Sarakhsi et al. [12] gave a review of the recent progress on emerging architectures and techniques for Alzheimer's disease (AD) diagnosis, including explainable models, normalizing flows, graph-based deep architectures, self-supervised learning, and attention models. Three major categories of currently known challenges in the existing literature include data-related issues, methodology-related complexities, and clinical adoption challenges. The study ends with potential future directions and recommendations that may empower future studies in AD detection

Ahsan Bin Tufail et al. [13] devised a scheme based on multiple deep 2D convolutional neural networks (2D-CNNs), where different kinds of diversified features were extracted from the images of the local brain for Alzheimer's disease classification. Utilizing transfer learning architectures (Inception v3 and Xception) and custom CNN with separable convolutional layers to learn the generic imaging features, the model combined the features for final classification. T1-weighted MRI images from the OASIS database were used, ensuring consistent size and contrast across scans. Experimental results showed that transfer learning methods outperformed non-transfer learning approaches, highlighting their effectiveness in binary AD classification tasks.

Mian Muhammad Sadiq Fareed et al. [14] introduced Alzheimer's Disease Detection Network (ADD-Net), a CNN architecture designed for AD detection with fewer parameters, ideal for smaller datasets. ADD-Net distinguishes the early stages of Alzheimer's disease and generates class activation maps as brain heatmaps. It reduces computational costs while precisely classifying AD stages. To address the class imbalance in the Kaggle MRI dataset, synthetic oversampling was employed to balance the classes. Evaluation against DenseNet169, VGG19, and InceptionResNet V2 showed ADD-Net's superior performance across metrics, achieving 98.63% accuracy, 99.76% AUC, 98.61% F1-score, and a loss of 0.0549%. The results highlight ADD-Net's effectiveness over state-of-the-art models.

P. R. Buvaneswari et al. [15] proposed an approach for achieving high-performance automated classification of Alzheimer's disease. Seven morphological features, including grey matter, white matter, cortical surface, gyri and sulci contours, cortical thickness, hippocampus, and cerebrospinal fluid space, were extracted from 240 structural MRI (sMRI) scans using SegNet. These features were used to train a ResNet model for classification. The trained classifier demonstrated a sensitivity of 96% and an accuracy of 95% on 240 ADNI sMRI scans not included in the training set.

Ruhul Amin Hazarika et al. [16] Visualization of feature extraction was performed on deep learning models used for Alzheimer's disease classification on MR images from ADNI dataset 16. DenseNet-121 reached 88.78% average accuracy, though it was slower in terms of computational cost s it performs considerable convolution operations. To reduce its resource load, depth-wise convolution layers were replaced with regular convolution layers in the DenseNet-121 architecture. This change improved the computation, and resulted in an increase of the mean accuracy of the model to 90.22%, illustrating it has greater performance and easier usage.

3. Proposed Model

Alzheimer's disease is a progressive degenerative disease of the nervous system leading to loss of memory, impairment of cognitive functions, and changes in behavior. It is the most prevalent cause of dementia, causing a major burden on millions worldwide. Fortunately, early diagnosis is essential for managing symptoms and improving quality of life. Of the available modalities, MRI is essential in detecting structural and functional alterations in the function of the brain in the context of Alzheimer's. However manually analyzing the MRI data is error-prone, which requires an automated system built on advanced deep learning techniques. CNNs and transfer learning models have been working well for the accurate detection and classification of Alzheimer's disease even in its early stages.

The proposed CNN model which helps to classify the categories of Alzheimer's disease is depicted in Figure 1.

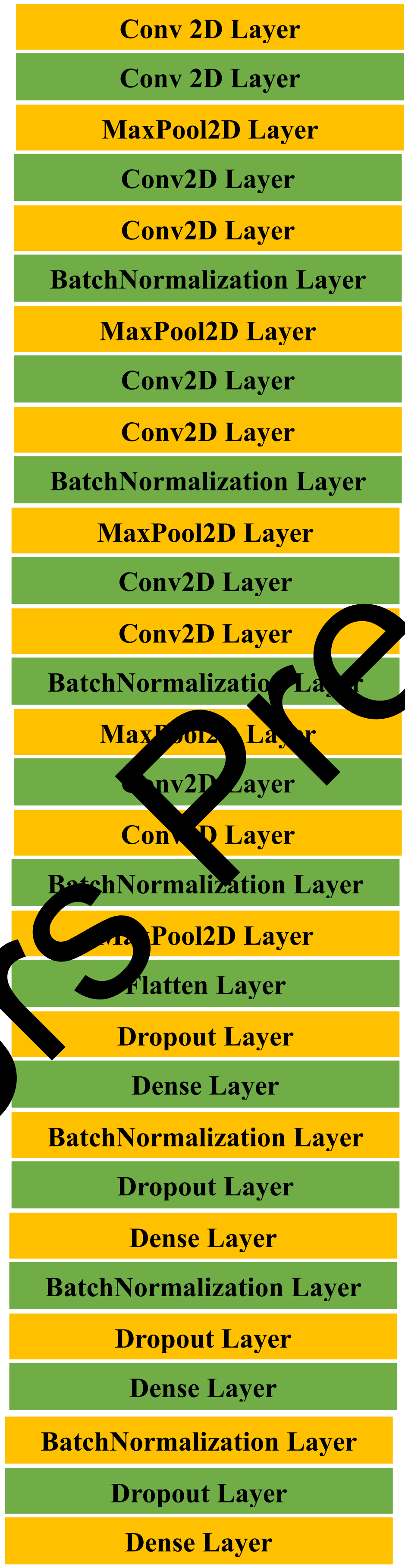

Conv 2D Layer
Conv 2D Layer
MaxPool2D Layer
Conv2D Layer
Conv2D Layer
BatchNormalization Layer
MaxPool2D Layer
Conv2D Layer
Conv2D Layer
BatchNormalization Layer
MaxPool2D Layer
Conv2D Layer
Conv2D Layer
Dropout Layer
Dense Layer
BatchNormalization Layer
Dropout Layer
Dense Layer
BatchNormalization Layer
Dropout Layer
Dense Layer
BatchNormalization Layer
Dropout Layer
Dense Layer

Figure 1: Proposed method Architecture

1. **Conv2D Layer (16, kernel_size=(3,3), activation='ReLU', padding='same'):** The effect of this block is that the first layer in a convolutional network is a convolutional layer, which takes the input data and applies 16 filters of size 3 x 3 High. This layer is responsible for extracting spatial features like edges and textures from the image. ReLU activation function adds non-linearity, allowing the network to learn complex behaviors. Using 'same' padding helps in keeping the aspect ratio of output feature maps equal to input feature maps so that whenever the model goes ahead with learning it can capture all the information from input as it can.
2. **Conv2D Layer (16, kernel_size=(3,3), activation='ReLU', padding='same'):** The second convolutional layer operates on these feature maps with the same parameters. Additional convolutional stacks allow for the addressing of finer details or more abstract features in the input image for downstream task representation.
3. **MaxPool2D Layer (pool_size=(2,2)):** The next layer is a pooling layer that halves the spatial dimensions of the feature maps. It downsamples by taking the maximum value in each 2×2 window of the input. This approach lowers the computational complexity, prevents overfitting, and keeps the strongest features that the previous convolutional layers have learned.
4. **Conv2D Layer (32, kernel_size=(3,3), activation='ReLU', padding='same'):** It increases the number of filters up to 32 for the network to recognize a higher number of more complex patterns in the input. The size of 3x3 for the filter allows for the capturing of local spatial relationships and the ReLU activation retains non-linearity.
5. **Conv2D Layer (32, kernel_size=(3,3), activation='ReLU', padding='same'):** It adds another 32 filters using convolutional layers. This allows the network to learn from higher-order statistics of the signal, providing a deeper and more abstract signal analysis.
6. **BatchNormalization Layer ():** It normalizes the outputs of the previous layer by scaling the activations and centering them. This technique, known as batch normalization, normalizes the inputs of every layer in a way that stabilizes the optimizers used preventing slow learning speed and being stuck in local minima.
7. **MaxPool2D Layer (pool_size=(2,2)):** The second pooling layer continues to reduce the spatial dimensions of the feature maps. This helps the network to only form high-level features as well as makes the architecture computationally efficient.
8. **Conv2D Layer (64, kernel_size=(3,3), activation='ReLU', padding='same'):** These are the filters from the convolution in the previous layer, this convolution layer has 64 filters that learn high-level concepts. More number of filters cause the layer to learn more diverse features.
9. **Conv2D Layer (64, kernel_size=(3,3), activation='ReLU', padding='same'):** This additional convolutional layer has 64 filters to further abstract the features. The stacking of several layers allows the model to create a hierarchical representation of the input.
10. **BatchNormalization Layer ():** It helps make the learning process more stable by reducing the sensitivity of the model to shifting input distribution and also normalizes the activations of the previous layer.

11. **MaxPool2D Layer (pool_size=(2,2)):** As such, the third pooling layer decreases the feature maps' spatial dimensions in a way that facilitates the network to focus on important features while omitting less important features.
12. **Conv2D Layer (128, kernel_size=(3,3), activation='ReLU', padding='same'):** This layer applies 128 filters to identify increasingly abstract and complex characteristics in the data. The high number of filters aids in learning minute details and intricate relationships.
13. **Conv2D Layer (128, kernel_size=(3,3), activation='ReLU', padding='same'):** A hundred and twenty-eight filters applied over the previous layer enhance this representation, enabling a more complex encoding of the class information in the data for the model.
14. **BatchNormalization Layer ():** It is used to normalize the convolutional output, thus making sure that the output of the convolutional layers gives consistent scaling for the training model and also stabilizes and enhances the training process.
15. **MaxPool2D Layer (pool_size=(2,2)):** The last pooling layer reduces the spatial dimensions dramatically and helps to prepare the feature maps before taking them to fully connected layers. This technique allows us to abstract the spatial information and capture the most relevant parts.
16. **Conv2D Layer (256, kernel_size=(3,3), activation='ReLU', padding='same'):** This is the first convolutional layer, with 256 filters, which is expected to detect high-level features and information from the input that captures complex patterns and relationships.
17. **Conv2D Layer (256, kernel_size=(3,3), activation='ReLU', padding='same', name='last_conv_layer'):** This layer fine-tunes the abstract features based on what the previous layer has produced. This specific layer is the 'last_conv_layer' as it is used in Grad-CAM to produce class activation maps based on gradients from this layer.
18. **Batch Normalization Layer ():** This layer normalizes the outputs of the last conv eyor to allow for stable gradients through backpropagation and higher generalization.
19. **MaxPool2D Layer (pool_size=(2,2)):** Reduces the feature map dimensions to prepare for the transition to the dense layers while retaining the most important high-level features.
20. **Flatten Layer ():** Flattens the multi-dimensional feature maps into a single 1D vector. This transformation is necessary for connecting the convolutional layers to the fully connected layers, which operate on vectors.
21. **Dropout Layer (rate=0.2):** Regularizes the model by randomly setting 20% of neurons to zero during training. This reduces the risk of overfitting by forcing the network to learn robust features.
22. **Dense Layer (512, activation='ReLU'):** Neurons in the fully connected layer are 512, which learns high-level features of input. The ReLU activation function enables the model to learn non-linear relationships.
23. **Batch Normalization Layer ():** This layer normalizes the outputs of the dense layer.
24. **Dropout Layer (rate=0.7):** Used 70% dropout rate to avoid overfitting by removing the dependence on specific neurons in the training.
25. **Dense Layer (128, activation='ReLU'):** The next layer is a dense layer of 128 neurons, allowing the model to better refine the feature representation and learn important patterns for classification.

26. **Batch Normalization Layer ():** It contains the dense layer output and normalizes the activations, which helps accelerate training.
27. **Dropout Layer (rate=0.5):** Implements 50% dropout for further regularization and reduces overfitting.
28. **Dense Layer (64, activation='ReLU'):** Further metas gave the dimensions 64 help convolve the process and identify the dimensions most discriminative.
29. **Batch Normalization Layer ():** It also normalizes the dense layer outputs for consistency.
30. **Dropout Layer (rate=0.3):** Implements 30% dropout to regularize the model before the final classification layer.
31. **Dense Layer (4, activation='SoftMax'):** The last dense layer consists of neurons that suit the output classes. It is a multi-class prediction model because SoftMax activates each class to give probabilities of each class.

3.1 Grad-CAM

Related work Gradient-weighted class activation mapping (Grad-CAM) is one of the techniques used to interpret the decision-making process of CNNs. Grad-CAM helps researchers determine and visualize salient features in input images by highlighting image regions most responsible for a model's predicted outcome. This technique calculates gradients of the predicted class score concerning the feature maps of the last convolutional layer and generates a heatmap indicating which regions of the input image give significant contributions to the predicted score.

1. **Feature Extraction:** The Grad-CAM algorithms pull gradients from the final convolutional layer of the CNN (e.g., called "last_conv_layer") and perform a backward pass to determine how relevant they were to the output prediction.

2. **Heatmap Generation:** It pools gradients to identify their significance and introduces a weighted map addition of feature maps. The produced heatmap identifies the important areas in the MRI image leading to the classification outcome.

3. **Superimposition:** It shows the heatmap placed on the original MRI image, which also gives an idea of where the model is concentrating its attention.

The base model used was a custom CNN consisting of Conv2D with multiple filters, MaxPooling, BatchNormalization, Dropout, and Dense layers. We applied Grad-CAM to this architecture to understand the classifier's decisions.

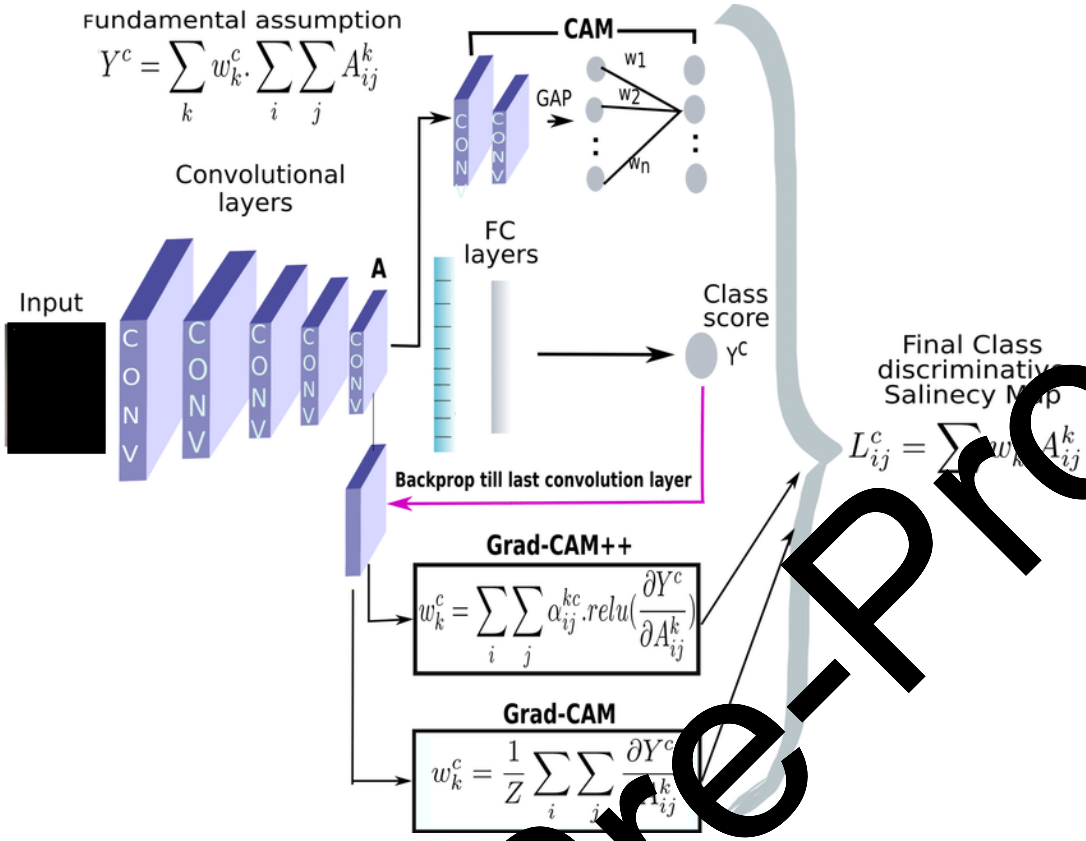


Figure 2: Grad-CAM Architecture

Figure 2: Grad-CAM Grad-CAM (Gradient-weighted Class Activation Mapping) [GS21] is once again a technique used to increase the interpretability of CNNs by highlighting the regions of input images most contributing to a model's prediction. We start with a standard CNN, where we feed images through multiple convolutional layers to derive features. The features are used for the fully connected layers where the class scores are calculated. Grad-CAM takes the gradients of the class concerning the final convolutional layer feature maps. This also generates a class-discriminative heatmap where only the prominent regions of the input image are preserved and the rest of the regions start to converge into the background.

The architecture uses the spatial information in convolutional layers to enhance interpretability. In this workflow, a heatmap that has been overlaid on the input image enables researchers to pinpoint the areas that contributed most to the model's classification. For example, in medical imaging applications, such as Alzheimer's disease classification, understanding the parts of the image focused by the model provides important insights into the diagnostic process. The proposed CNN model integrates perfectly with the Grad-CAM architecture for clinical practice to ensure that the prediction is not only accurate but also explainable.

Algorithm for Alzheimer's disease

Step 1: Input and Preprocessing

- The model shape is defined as $Input \in R^{H \times W \times C}$,

Where:

- H,W: Image height and width (e.g., 256 × 256 pixels).
- C: Number of channels (3 for RGB images).
- Input normalization ensures the pixel intensity values are scaled to the range [0,1]:

- $x_{ij} = \frac{I_{ij}}{255}$

  Here, I_{ij} represents the pixel intensity at position (i,j).

Step 2: Feature Extraction via Convolutional Layers

- Each convolutional layer applies a filter W_k(Kernel) over the input to compute feature maps A_k:

$$A_k^{(l)} = ReLU\left(W_k^{(l)} * A^{(l-1)} + b_k^{(l)}\right)$$

- $W_k^{(l)}$: Weights of the k-th filter in layer l.
- $A^{(l-1)}$: Input feature map to the layer.
- $b_k^{(l)}$: Bias for the k-th filter.
- $ReLU$: Activation function

Step 3: Downsampling with MaxPooling

- The MaxPooling operation reduces spatial dimensions by selecting the maximum value in non-overlapping windows:

$$A_{ij}^{pool} = \max_{m,n} \left(A_{(i+m)(j+n)}\right), m, n \in \ window\ size$$

  This step helps in reducing computations and focusing on dominant features.

  A_{ij}^{pool}: Output of the pooling operation.

  $Window\ size$: The size of the pooling window (commonly 2×2).

Step 4: Flattening

- After the final convolutional and pooling layers, the feature maps are flattened into a 1D vector z:

$$z = Flatten(A^L)$$

  Where A^L is the feature map from the last convolutional layer.

  $Flatten$: Converts multi-dimensional feature maps into a 1D vector for dense layer processing.

Step 5: Fully Connected Layers

- Fully connected (Dense) layers compute weighted sums of their inputs:

$$Z^{(l)} = ReLU\left(W^{(l)}Z^{(l-1)} + b^{(l)}\right)$$

  $W^{(l)}$: Weight matrix for layer l.

  $Z^{(l-1)}$: Input vector from the previous layer.

  $^{(Bl)}$: Bias term.

Step 6: Dropout for Regularization

- Dropout randomly sets a fraction p of activations to zero during training to prevent overfitting:

$$Z_i^{drop} = \begin{cases} Z_i, & with\ probability\ 1-p \\ 0, & with\ probability\ p \end{cases}$$

  P: Dropout rate (e.g., 0.2, 0.5, etc.).

Step 7: Output Layer with SoftMax

- The output layer computes class probabilities using the SoftMax activation:

$$\hat{y}_i = \frac{exp(Z_i)}{\sum_{j=1}^{C} exp(Z_i)}, i \in 1,2,\ldots,C$$

  C: Number of classes.

  $\hat{y}_i$: Predicted probability for class i.

Step 8: Loss Function (Categorical Crossentropy)

- The loss function measures the difference between predicted probabilities $\hat{y}$ and true labels y:

$$L = -\sum_{i=1}^{C} y_i log(\hat{y}_i)$$

y_i: True label for class i (one-hot encoded).

$\hat{y}_i$: Predicted probability for class i.

Step 9: Optimization (Adam)

- The Adam optimizer updates weights W using gradients ∇L:

$$W_{t+1} = W_t - \eta . \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}}$$

$\hat{m}_t$, $\hat{v}_t$: Corrected first and second moments of gradients.

η: Learning rate (e.g., 0.001).

$\in$: Small value to avoid division by zero.

Step 10: Early Stopping

- Early stopping halts training when validation loss does not improve for a specified number of epochs (p):

Stop training if min(L_{val},t) does not decrease for t>p

(L_{val},t) : Validation loss at epoch t.

The proposed model is a carefully designed convolutional neural network (CNN) optimized for multi-class classification of MRI data. The model incorporates multiple layers of convolutional operations with progressively increasing filter sizes (16, 32, 64, 128, 256) to extract hierarchical spatial features, enabling it to learn complex patterns in medical imaging data. Batch Normalization is strategically placed after convolutional and Dense layers to stabilize activations and improve training efficiency. Regularization is achieved through Dropout layers with varying rates (0.2, 0.[illegible] and 0.7) to prevent overfitting and enhance generalization. Additionally, the final convolutional layer is explicitly named `last_conv_layer` to support Grad-CAM, which provides interpretability by highlighting the critical regions in MRI scans that influenced the classification decision.

The proposed model brings along several key contributions making it appropriate for medical imaging classification tasks. Early Stopping allows for efficient training by stopping when validation loss is no longer improving, and Model Checkpoint allows the saving of the best weights based on validation. Adaptive updates to the model's parameters using the Adam optimizer with a learning rate of 0.001 ensure faster convergence. The overall evaluation metrics including categorical accuracy, AUC, and F1-Score describe the performance of the model. This model combines interpretability, robust regularization, and feature extraction, making your solution scalable and reliable in clinical applications and overcoming the obstacles posed by heterogeneous data, as well as ensuring explainable AI for healthcare applications.

4. Experimental Results

This subsection gives a thorough assessment of the results reported by the proposed method while the simulations are still in progress. This Simulations UNIX dataset was obtained from the Best Alzheimer MRI dataset [17]. The same data treatment detailed above was performed on this dataset for the current study.

The dataset consists of:

- Mild Impairment
- Moderate Impairment
- No Impairment
- Very Mild Impairment

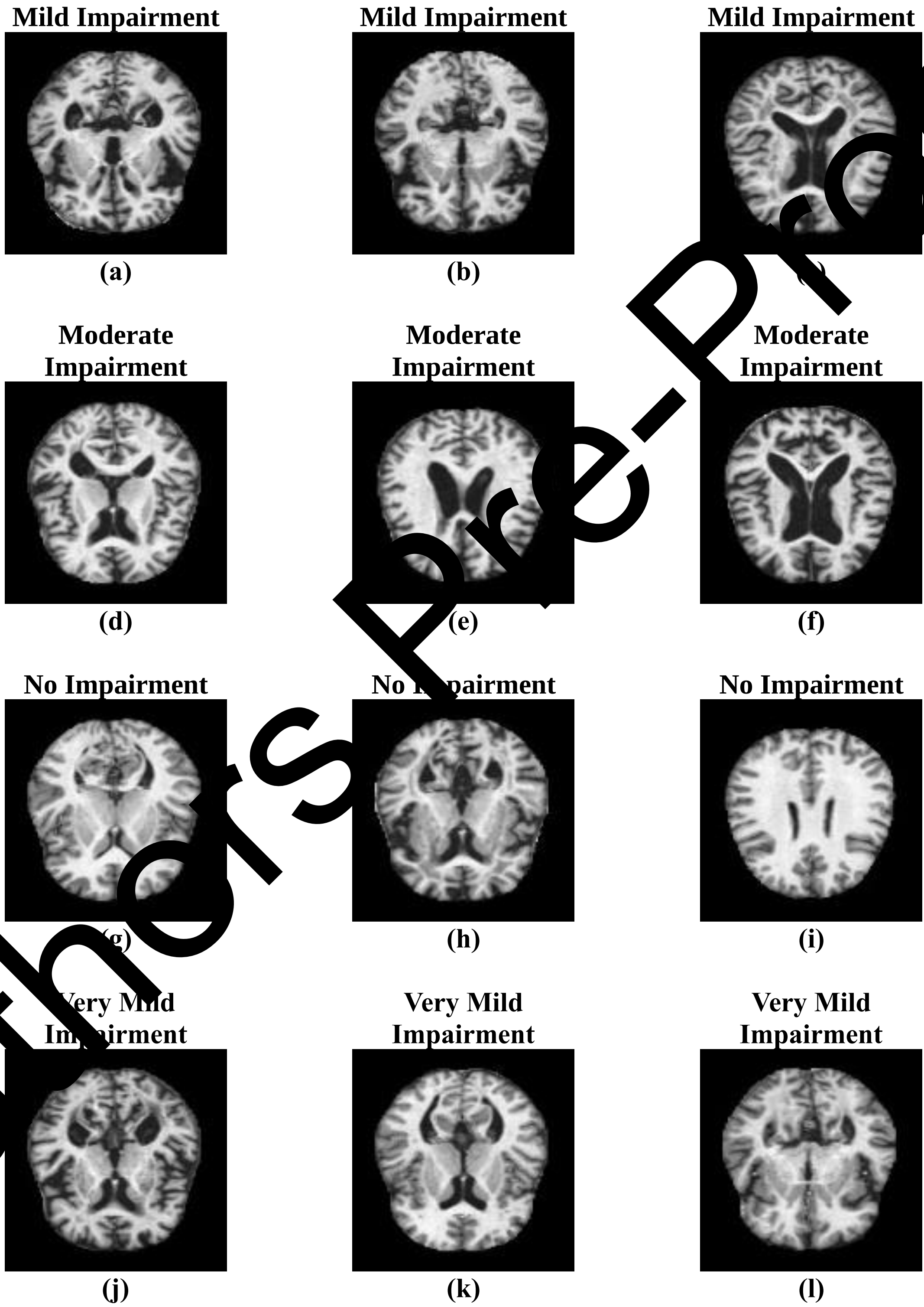
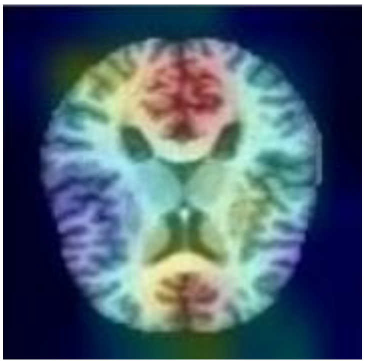


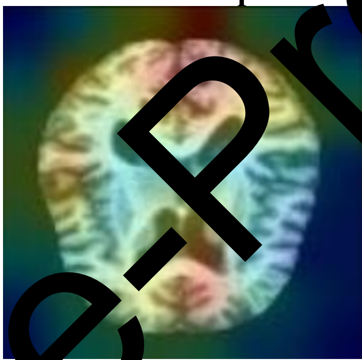
Figure 3: The sample images of the dataset



(a) Ground Truth: No Impairment
Predicted: No Impairment

(b) Ground Truth: No Impairment
Predicted: No Impairment

(c) Ground Truth: No Impairment
Predicted: No Impairment



(d) Ground Truth: Very Mild Impairment
Predicted: Very Mild Impairment

(e) Ground Truth: No Impairment
Predicted: No Impairment

(a) Ground Truth: Mild Impairment
Predicted: Mild Impairment

Figure 4: Grad-CAM Visualizations for Alzheimer's Disease Classification

Figure 4 depicts Grad-CAM visualizations showing the areas where the model was most focused on when classifying Alzheimer's disease stages using MRI scans. Each image comes with a heatmap superimposed with the origins of the MRI scan, where warmer colors (red, orange yellow) mean higher significance for the model's decisions, while cooler colors (green, and blue) mean lower significance for the decision-making process. The true label and predicted labels shown above each of the image's dependent on the classification output of the custom CNN model. The set of images provides evidence of the model learned to differentiate important areas of the human brain that are related to three degrees of impairment i.e., No Impairment, Mild Impairment, and Very Mild Impairment.

These numbers reinforce the need for something like Grad-CAM for interpretability in medical AI systems. The Explainable AI behavior can be validated based on visualizing the top pay between regions that lead to the classification made for a model by researchers and clinicians. The predicted regions are not matched with the image ground truth and in all cases the highlighted regions are consistent with clinicians' clinical expected regions, even providing validation that the focused areas are diagnostically reasonable. This ability to provide descriptive behavior improves trust in the Explainable AI itself and the model complements the structural changes observed in the brain with the different stages of Alzheimer's disease, thus making it useful for diagnostic aid but also for further studies in this research level.

Table 1: Classification Report

	Precision	**Recall**	**F1-Score**
Mild Impairment	0.93	0.96	0.94
Moderate Impairment	1.00	0.83	0.91
No Impairment	0.94	0.98	0.96
Very Mild Impairment	0.97	0.90	0.94
Accuracy	0.95		

Performance of the proposed model in terms of four categories, namely, Mild Impairment, Moderate Impairment, No Impairment, and Very Mild Impairment shown in Table 1 by using the classification report. The NO IMP category has the highest value of F1 score of 0.96 and recall of 0.98, while the model has very good precision, recall and F1 scores across the capabilities of the model. Both the Mild Impairment and Very Mild Impairment categories have high F1-scores of 0.94, indicating both high precision and recall. Where the Moderate Impairment class has a precision of 1.00, recall is lower at 0.83, leading to a n F1-score of 0.91. The metrics are yielding an overall accuracy of 95% indicating a pretty effective application of the model for the multi-classification of MRI data for patients with Alzheimer's disease.

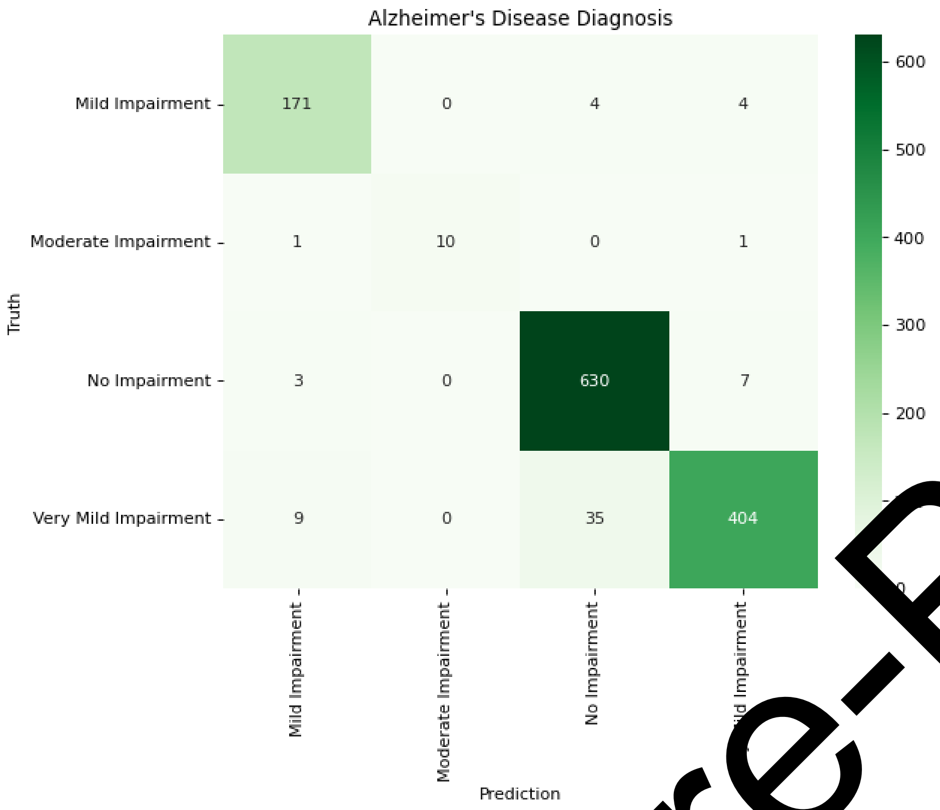


Figure 5: Confusion Matrix

The confusion matrix (Figure 5) is a detailed overview used in statistics to assess the performance of a multi-class classification model. Gives informative aspects of True positives (TP), False positives (FP), False negatives (FN) and True negatives (TN) in each category. It is a method that is especially useful in the context of Alzheimer's disease diagnosis, allowing us to see how well our model can distinguish between different impairment levels.

Within this specific matrix:

- **Mild Impairment**: The cell in the top-left corner (171) indicates the true positives (TP), which are cases accurately predicted as Mild Impairment. False negatives (FN) and false positives (FP) fill out the off-diagonal cells (4, 4), showing instances that were misclassified into the wrong group.
- **Moderate Impairment**: The TP count is 10, found in the second row, second column. The off-diagonal cells (1, 1) show misclassifications into Mild Impairment and Very Mild Impairment.
- **No Impairment**: The largest TP value is 630, found in the third row, and third column, reflecting the model's high accuracy for this class. Off-diagonal cells (3, 7) represent misclassifications into other classes.
- **Very Mild Impairment**: The TP count is 404 in the last row, last column, while the off-diagonal cells (9, 35) indicate misclassifications into Mild Impairment and No Impairment, respectively.

This confusion matrix demonstrates that the model performs exceptionally well for categories like No Impairment and Mild Impairment while showing some misclassification challenges for Very Mild Impairment and Moderate Impairment.

Table 2: Comparative Analysis

Methods	Accuracy
EfficientNetB0 [18]	39.48%
MobileNetV1 [19]	70.54%
VGG16[20]	72.87%
SqueezeNet [21]	88.60%
NASNETMobile [22]	92.80%
InceptionV3 [23]	93.20%
Xception [24]	94.40%
Proposed (GRAD-CAM's)	95.00%

Table 2 presents a comparative analysis of various deep learning methods used for Alzheimer's disease classification based on their accuracy. The proposed model utilizing Grad-CAM achieves the highest accuracy of 95.00%, outperforming several state-of-the-art architectures. Among the compared methods, Xception reaches an accuracy of 94.40%, closely followed by InceptionV3 at 93.20% and NASNETMobile at 92.80%. SqueezeNet and VGG16 achieve moderate accuracies of 88.60% and 72.87%, respectively, while MobileNetV1 and EfficientNetB0 yield lower accuracies of 70.54% and 39.48%. Such comparison indicates an excellent improvement for the presented model, hence attesting precision and strength for the successful classification of MRI images in the task of detecting Alzheimer's disease.

5. Conclusion

This research introduces and validates a new deep-learning system designed to identify different levels of memory-related conditions. The proposed model combines a custom-built AI structure with a tool called Grad-CAM, which helps explain how the system makes its decisions. Key features of the model include multiple layers for learning patterns, techniques to improve training stability, and methods to prevent the system from becoming too focused on specific training examples. Grad-CAM also provides visual feedback, showing which areas of the dataset images, the model used to make its decision, offering a clear explanation of its thinking process. The model reached an accuracy of 95%, outperforming other well-known systems like Xception (94.40%) and InceptionV3 (93.20%). This approach is unique because it not only delivers strong performance but also addresses the growing need for AI tools to be understandable and trustworthy. Our findings showed that the model can successfully recognize different stages of human function decline and could be useful in real-world decision-making processes. This system sets a new standard by bridging advanced technology with everyday use, supporting better outcomes for people, and increasing trust in AI-powered tools. The work shows that deep learning has great potential to improve how we detect and understand complex problems, while also making sure that the results are clear and reliable.

Results

[1] Memudu, Adejoke Elizabeth, Baliqis Adejoke Olukade, and Gideon S. Alex. "Neurodegenerative Diseases: Alzheimer's Disease." Integrating Neuroimaging, Computational Neuroscience, and Artificial Intelligence: 128-147.

[2] Du, Lixin, Shubham Roy, Pan Wang, Zhigang Li, Xiaoting Qiu, Yinghe Zhang, Jianpeng Yuan, and Bing Guo. "Unveiling the future: advancements in MRI imaging for neurodegenerative disorders." Ageing Research Reviews (2024): 102230.

[3] Kale, Mayur B., Nitu L. Wankhede, Rupali S. Pawar, Suhas Ballal, Rohit Kumawat, Manish Goswami, Mohammad Khalid et al. "AI-Driven Innovations in Alzheimer's Disease: Integrating Early Diagnosis, Personalized Treatment, and Prognostic Modelling." Ageing Research Reviews (2024): 102497.

[4] Mahmood, Tariq, Amjad Rehman, Tanzila Saba, Yu Wang, and Faten S. Alamri. "Alzheimer's disease unveiled: Cutting-edge multi-modal neuroimaging and computational methods for enhanced diagnosis." Biomedical Signal Processing and Control 97 (2024): 106721.

[5] Elazab, Ahmed, Changmiao Wang, Mohammed Abdelaziz, Jian Zhang, Jason Gu, Juan M. Gorriz, Yudong Zhang, and Chunqi Chang. "Alzheimer's disease diagnosis from single and multimodal data using machine and deep learning models: Achievements and future directions." Expert Systems with Applications (2024): 124[illegible]80.

[6] Singh, Soraisam Gobinkumar, Dulumani Das, Utpal Barman, and Manob Jyoti Saikia. "Early Alzheimer's disease detection: A review of machine learning techniques for forecasting transition from mild cognitive impairment." Diagnostics 14, no. 16 (2024): 1759.

[7] Nazir, Asifa, Assif Assad, Ahsan Hussain, and Mandeep Singh. "Alzheimer's disease diagnosis using deep learning techniques: datasets, challenges, research gaps and future directions." International Journal of System Assurance Engineering and Management (2024): 1-35.

[8] Upadhyay, Prashant, Pradeep Tomar, and Satya Prakash Yadav. "Comprehensive Systematic Computation on Alzheimer's Disease Classification." Archives of Computational Methods in Engineering (2024): 1-32.

[9] Sorour, Shaymaa E., Amr A. Abd El-Mageed, Khalied M. Albarrak, Abdulrahman K. Alnaim, Abeer A. Wafa, and Engy El-Shafeiy. "Classification of Alzheimer's disease using MRI data based on Deep Learning Techniques." Journal of King Saud University-Computer and Information Sciences 36, no. 2 (2024): 101940.

[10] Arafa, Doaa Ahmed, Hossam El-Din Moustafa, Hesham A. Ali, Amr MT Ali-Eldin, and Sabry F. Saraya. "A deep learning framework for early diagnosis of Alzheimer's disease on MRI images." Multimedia Tools and Applications 83, no. 2 (2024): 3767-3799.

[11] El-Latif, Ahmed A. Abd, Samia Allaoua Chelloug, Maali Alabdulhafith, and Mohamed Hammad. "Accurate detection of Alzheimer's disease using lightweight deep learning model on MRI data." Diagnostics 13, no. 7 (2023): 1216.

[12] Khojaste-Sarakhsi, M., Seyedhamidreza Shahabi Haghighi, SMT Fatemi Ghomi, and Elena Marchiori. "Deep learning for Alzheimer's disease diagnosis: A survey." Artificial intelligence in medicine 130 (2022): 102332.

[13] Tufail, Ahsan Bin, Yong-Kui Ma, and Qiu-Na Zhang. "Binary classification of Alzheimer's disease using sMRI imaging modality and deep learning." Journal of digital imaging 33, no. 5 (2020): 1073-1090.

[14] Fareed, Mian Muhammad Sadiq, Shahid Zikria, Gulnaz Ahmed, Saqib Mahmood, Muhammad Aslam, Syeda Fizzah Jillani, Ahmad Moustafa, and Muhammad Asad. "ADD-Net: an effective deep learning model for early detection of Alzheimer disease in MRI scans." IEEE Access 10 (2022): 96930-96951.

[15] Buvaneswari, P. R., and R. Gayathri. "Deep learning-based segmentation in classification of Alzheimer's disease." Arabian Journal for Science and Engineering 46, no. 6 (2021): 5373-5383.

[16] Hazarika, Ruhul Amin, Debdatta Kandar, and Arnab Kumar Maji. "An experimental analysis of different deep learning based models for Alzheimer's disease classification using brain magnetic resonance images." Journal of King Saud University-Computer and Information Sciences 34, no. 10 (2022): 8576-8598.

[17] L. Chugh, "Best Alzheimer MRI Dataset (99% Accuracy)," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/lukechugh/best-alzheimer-mri-dataset-99-accuracy. [Accessed: Jan. 15, 2025].

[18] Gasmi, Karim, Abdulrahman Alyami, Omer Hamid, Mohamed O. Altaieb, Osama Rezk Shahin, Lassaad Ben Ammar, Hassen Chouaib, and Abdulaziz Shehab. "Optimized Hybrid Deep Learning Framework for Early Detection of Alzheimer's Disease Using Adaptive Weight Selection." Diagnostics 14, no. 24 (2024): 2779.

[19] Heenaye-Mamode Khan, Maleika, Pushtika Reesaul, Muhammad Muzzammil Auzine, and Amelia Taylor. "Detection of Alzheimer's disease using pre-trained deep learning models through transfer learning: a review." Artificial Intelligence Review 57, no. 10 (2024): 275.

[20] Assmi, Ayoub, Khaoula Elhabyb, Achraf Benba, and Abdelilah Jilbab. "Alzheimer's disease classification: a comprehensive study." Multimedia Tools and Applications (2024): 1-24.

[21] Rani, K. Emily Esther, and S. Baulkani. "Alzheimer disease classification using optimal clustering based pre-trained SqueezeNet model." Biomedical Signal Processing and Control 100 (2025): 107032.

[22] Arslan, Naciye Nur, and Durmus Ozdemir. "Analysis of CNN models in classifying Alzheimer's stages: comparison and explainability examination of the proposed separable convolution-based neural network and transfer learning models." Signal, Image and Video Processing (2024): 1-15.

[23] Hatami, Mahla, Farzin Yaghmaee, and Reza Ebrahimpour. "Investigating the potential of reinforcement learning and deep learning in improving Alzheimer's disease classification." Neurocomputing 597 (2024): 128119.

[24] Singh, Yajuvendra Pratap, and Daya Krishan Lobiyal. "A comparative study of early stage Alzheimer's disease classification using various transfer learning CNN frameworks." Network: Computation in Neural Systems (2024): 1-29.