`

Deep Learning Approaches for Alzheimer’s MRI Classification

Abstract

This project report presents the development and evaluation of two deep learning models for Alzheimer’s MRI classification. The primary objective is to create robust diagnostic tools that can assist in the automated detection and assessment of Alzheimer’s disease using MRI scans. Two classification tasks were addressed: a binary classification task that distinguishes between “Impairment” and “No Impairment,” and a multiclass classification task that stratifies subjects into four categories—Mild Impairment, Moderate Impairment, No Impairment, and Very Mild Impairment.

For both tasks, the project leverages transfer learning by fine-tuning state-of-the-art DenseNet architectures. Specifically, DenseNet121 serves as the backbone for the binary model, while DenseNet169 is employed for the multiclass model. The input MRI scans undergo rigorous preprocessing that includes resizing (224×224 pixels for the binary model and 256×256 pixels for the multiclass model), min–max normalization, and data augmentation (incorporating rotations, translations, zooming, and brightness adjustments) to enhance model robustness. In cases where the images are originally in grayscale, a 1×1 convolution layer is used to convert them into three-channel inputs to match the expected input format of the pre-trained networks.

Experimental results demonstrate that the binary classification model achieved near-perfect performance, with training accuracies exceeding 98% and validation accuracies nearing 99%. This success is reflected in extremely low loss values and exceptional precision, recall, and F1-scores, making it a reliable tool for preliminary screening. In contrast, while the multiclass model showed promising overall accuracy (approximately 73.3% on validation data), it encountered challenges in differentiating subtle distinctions, particularly between the “No Impairment” and “Very Mild Impairment” classes. Detailed analysis via confusion matrices and per-class metrics highlighted these difficulties, suggesting the need for further refinement in feature extraction and optimization strategies.

`

Overall, the report outlines a comprehensive approach that combines advanced preprocessing, transfer learning, and systematic evaluation to address the complexities inherent in Alzheimer's MRI classification. The insights gained from this project pave the way for future enhancements, including the exploration of more sophisticated architectures, multimodal data integration, and extensive clinical validation, ultimately contributing to the development of reliable, automated diagnostic systems for Alzheimer's disease.

Introduction

Alzheimer's disease is a progressive neurodegenerative disorder that affects millions of people worldwide, leading to memory loss, cognitive decline, and severe impairment in daily functioning. Early and accurate diagnosis is critical to managing the disease, planning treatments, and improving patient outcomes. Traditionally, Alzheimer's diagnosis has relied on clinical evaluations and manual interpretations of imaging data, such as MRI scans. However, these conventional approaches are time-consuming, subjective, and often limited by inter-observer variability.

In recent years, advances in deep learning have opened new avenues for medical image analysis, offering automated and objective tools for diagnostic support. The ability of deep neural networks to learn complex features directly from raw images has made them particularly attractive for applications in medical diagnostics. This project leverages these advances to develop a comprehensive deep learning framework for Alzheimer's MRI classification. Specifically, two distinct classification models were developed: one for binary classification to distinguish between "Impairment" and "No Impairment," and another for multiclass classification to further stratify cases into four levels of impairment—Mild Impairment, Moderate Impairment, No Impairment, and Very Mild Impairment.

The binary classification model serves as an effective screening tool by providing a clear-cut decision boundary. Utilizing a DenseNet121 architecture pre-trained on large-scale image datasets, the model is fine-tuned to capture features specific to Alzheimer's-related changes in MRI scans. The choice of DenseNet121 is driven by its ability to reuse features through dense connections, which helps in extracting intricate details from the scans. The images undergo rigorous preprocessing—including resizing to 224×224 pixels, min–max normalization, and data augmentation techniques such as rotations, translations, and brightness adjustments—to enhance model robustness and improve generalization.

In contrast, the multiclass classification model is designed to provide a more nuanced assessment of Alzheimer's severity. By leveraging a DenseNet169 architecture, which offers increased capacity compared to its counterpart used in the binary model, this approach aims to capture subtle variations among different levels of impairment. Images for this task are resized to 256×256 pixels to preserve finer anatomical details, and similar preprocessing steps ensure consistent input quality. The model outputs a probability distribution over the four classes through a softmax activation, and its performance is evaluated using metrics such as per-class accuracy, precision, recall, and F1-score.

Both models employ transfer learning as a core component of their design. By utilizing pre-trained networks, the project benefits from features learned from extensive datasets, which are then fine-tuned on the Alzheimer's MRI data. This approach not only accelerates the training process but also mitigates the challenges associated with limited labeled data in medical imaging. Additionally, the use of class weighting helps to address any residual imbalance in the datasets, ensuring that the models are not biased toward the majority class.

The significance of this project lies in its dual approach. The binary model, with its high accuracy and near-perfect classification metrics, demonstrates the feasibility of using deep learning for rapid Alzheimer's screening. Meanwhile, the multiclass model, despite facing challenges in differentiating between subtle classes, provides valuable insights into the complexities of Alzheimer's progression and highlights areas for future improvement. The detailed experimental results, including training curves, loss metrics, and confusion matrices, offer a comprehensive evaluation of both approaches.

This report outlines the motivation, methodology, and experimental findings of the project, providing a clear overview of how deep learning techniques can be applied to Alzheimer's MRI classification. The subsequent chapters will delve into related work, detailed methodology, experimental results, discussion, and conclusions, offering a holistic view of the project and its potential implications for automated diagnostic systems in neurodegenerative diseases.

`

Related work

The application of deep learning to medical image analysis has experienced rapid growth in recent years, particularly in the field of Alzheimer's disease diagnosis using MRI scans. Early approaches in Alzheimer's classification relied on traditional machine learning methods, which typically involved manual feature extraction techniques such as voxel-based morphometry (VBM) and regional volumetric measurements. These features were subsequently fed into classifiers like support vector machines (SVMs) or logistic regression models. Although these traditional methods provided initial insights into disease progression, they were limited by their dependency on handcrafted features and often suffered from variability in performance due to differences in feature extraction protocols.

The emergence of convolutional neural networks (CNNs) has significantly transformed the landscape of medical imaging, as CNNs automatically learn hierarchical feature representations directly from raw imaging data. In recent studies, CNN-based models have been successfully employed for Alzheimer's classification, often achieving higher accuracy and robustness compared to classical methods. Architectures such as VGG, ResNet, and DenseNet have been explored extensively, with DenseNet in particular gaining attention due to its efficient feature reuse via dense connections. This architecture has been shown to capture subtle variations in MRI scans, which are crucial for identifying early signs of Alzheimer's disease.

Transfer learning has become a widely adopted strategy in medical imaging applications due to the limited availability of labeled medical data. By leveraging models pre-trained on large-scale datasets like ImageNet, researchers have been able to fine-tune these networks on Alzheimer's MRI datasets, thereby reducing the training time and improving generalization. For instance, DenseNet121 and DenseNet169 have been adapted for Alzheimer's classification tasks, with the former typically used for binary classification (impaired vs. non-impaired) and the latter employed for multiclass classification tasks that require more nuanced differentiation between stages of impairment.

Data augmentation and class weighting are additional strategies that have been adopted to address common challenges such as overfitting and class imbalance. Data augmentation techniques—such as rotations, translations, zooming, and brightness adjustments—

`

simulate a broader range of imaging conditions, thereby enhancing model robustness. Similarly, class weighting is employed to balance the influence of classes that are underrepresented, ensuring that the model does not become biased toward the majority class.

Several studies have demonstrated the potential of deep learning for Alzheimer's MRI classification. Prior work has shown that CNNs can achieve high accuracy in differentiating between Alzheimer's patients and healthy controls, while also providing insights into the regions of the brain that contribute most significantly to the classification decisions. These studies lay the groundwork for the current project, which builds on these established techniques by implementing two complementary classification models—one for binary classification and one for multiclass classification—using DenseNet architectures.

In summary, the body of related work underscores the transition from traditional machine learning approaches toward more sophisticated deep learning frameworks for Alzheimer's MRI classification. The success of CNN-based models, combined with strategies such as transfer learning, data augmentation, and class weighting, forms the foundation upon which this project is built. This project aims to extend these approaches by developing two models that not only achieve high accuracy but also provide a detailed stratification of Alzheimer's severity, thereby offering a more nuanced tool for clinical diagnosis and research.

`

Methodology

A. Binary Classification Methodology

Data Preprocessing

The binary classification dataset consists of 5,120 MRI images equally divided between two classes: “Impairment” and “No Impairment.” Each image is resized to 224×224 pixels. To ensure consistent input, min–max normalization is applied using the formula:

x_norm = (x – x_min) / (x_max – x_min)

This normalization scales pixel values to the [0, 1] range. Since the original images are in grayscale, a 1×1 convolution layer is employed to convert the single channel into a three-channel format, thereby aligning the inputs with the DenseNet121 model’s requirements.

Model Architecture

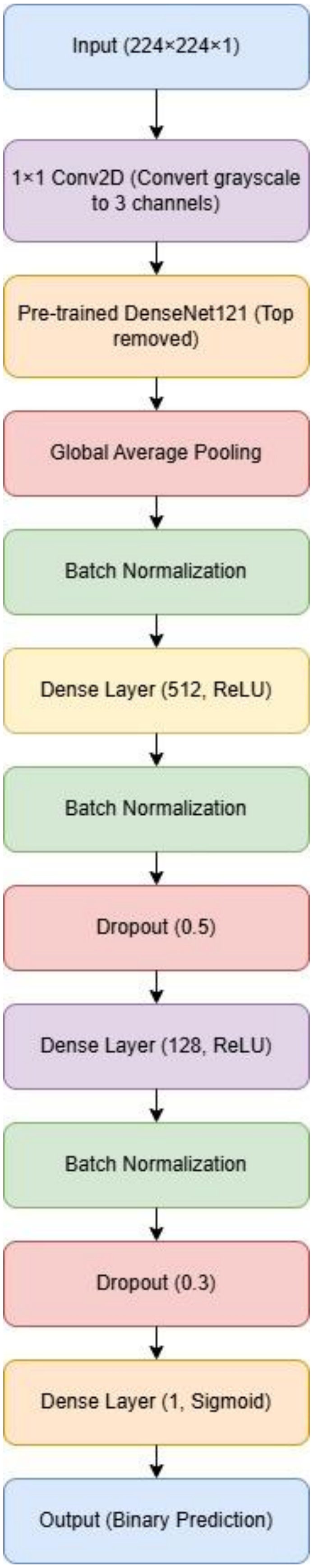
The binary classification model utilizes a DenseNet121 architecture, pre-trained on a large-scale dataset, with the top (classification) layers removed. The network architecture comprises the following stages:

a. Feature Extraction:

The pre-trained DenseNet121 serves as a fixed feature extractor. Its dense connectivity allows the network to reuse features effectively, capturing intricate patterns present in the MRI scans.

b. Global Average Pooling:

After feature extraction, global average pooling is applied to convert the spatial feature maps into a fixed-length vector, reducing the number of parameters and mitigating overfitting.



c. Fully Connected Layers:

This vector is then passed through one or more dense layers. Batch normalization is applied after each dense layer to stabilize learning, while dropout layers (with dropout rates of 0.5 and 0.3, respectively) help reduce overfitting.

d. Output Layer:

The final layer consists of a single neuron with a sigmoid activation function to generate a probability value, p, indicating the likelihood of "Impairment."

Loss Function and Optimization

The binary model is optimized using the binary cross-entropy loss function, defined as:

$$L = -(1/N)\, \Sigma\, [y \cdot \log(p) + (1 - y) \cdot \log(1 - p)]$$

where N is the number of samples, y is the true binary label, and p is the predicted probability. The model is trained using the Adam optimizer with an initial learning rate of 0.0005, incorporating adaptive learning rate scheduling and early stopping based on validation loss. Data augmentation techniques—including rotations, translations, zooming, and brightness adjustments—are applied during training to improve model robustness. Class weights are computed using:

$$w_i = N / (C \cdot n_i)$$

where C is the number of classes (2 in this case) and n_i is the number of samples in class i.

Fine-Tuning

After the initial training phase, selective unfreezing of the DenseNet121 layers is performed to fine-tune the model on Alzheimer's MRI data. This fine-tuning process further refines the model's feature extraction capabilities and leads to incremental performance improvements.

`

B. Multiclass Classification

Data Preprocessing

For the multiclass classification task, the dataset comprises 10,240 MRI images equally divided into four classes: “Mild Impairment,” “Moderate Impairment,” “No Impairment,” and “Very Mild Impairment.” Each image is resized to 256×256 pixels to preserve more detailed anatomical features. The same min–max normalization technique is applied:

$$x\_norm = (x - x\_min) / (x\_max - x\_min)$$

Additionally, a 1×1 convolution layer converts the single-channel grayscale images into a three-channel format, ensuring compatibility with the DenseNet169 model.

Model Architecture

The multiclass model is based on a DenseNet169 architecture with partial layer freezing to retain robust pre-trained features while allowing later layers to adapt to the nuances of Alzheimer’s classification. The architecture includes:

a. Feature Extraction:

DenseNet169 is employed as the backbone; early layers are frozen, while later layers remain trainable.

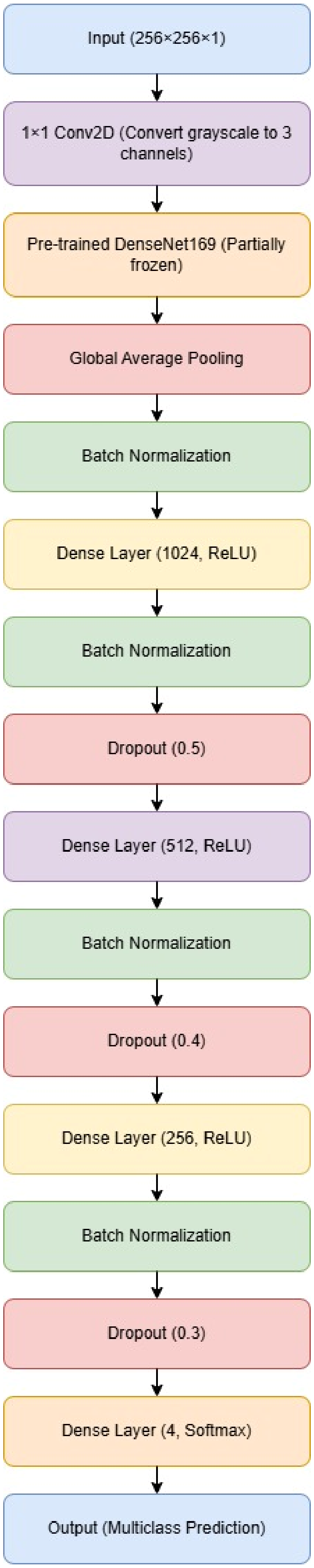
b. Global Average Pooling:

Similar to the binary model, global average pooling is used to transform feature maps into a fixed-length vector representation.

c. Fully Connected Layers:

This vector is passed through a series of dense layers with increasing capacity. Specifically, the network incorporates dense layers with 1024, 512, and 256 units, each

`



followed by batch normalization and dropout layers (with dropout rates of 0.5, 0.4, and 0.3, respectively).

d. Output Layer:

The final layer consists of four neurons corresponding to the four classes, with a softmax activation function that outputs a probability distribution over the classes.

Loss Function and Optimization

The multiclass model employs the categorical cross-entropy loss function:

$$L = -\Sigma [y_i \cdot \log(p_i)]$$

where y_i is the one-hot encoded true label and p_i is the predicted probability for class i, summed over all classes. The model is optimized using the Adam optimizer with an initial learning rate of 0.0002. Similar to the binary model, adaptive learning rate scheduling and early stopping are applied. Data augmentation (rotations, translations, zoom, brightness adjustments) further enhances model generalization. Class weights are calculated (with a modified weighting strategy to emphasize classes with subtle differences) using:

$$w_i = N / (C \cdot n_i)$$

where C equals 4 in this case.

Training and Evaluation

The multiclass model is trained for up to 40 epochs, and its performance is monitored using validation accuracy, loss curves, and per-class metrics (accuracy, precision, recall, and F1-score). The confusion matrix is analyzed to identify misclassification patterns—specifically, the model exhibited challenges distinguishing between “No Impairment” and “Very Mild Impairment,” with a significant number of misclassifications observed in these categories.

`

In summary, both models employ rigorous preprocessing, transfer learning via DenseNet architectures, and advanced optimization strategies to address the challenges inherent in Alzheimer’s MRI classification. The binary model achieves high performance with near-perfect classification metrics, while the multiclass model, though promising, requires further refinement to handle subtle inter-class differences effectively.

System Architecture:

The overall system architecture integrates multiple modules—from data ingestion and preprocessing to model training and evaluation—into a unified pipeline that supports both binary and multiclass Alzheimer’s MRI classification tasks. The architecture is designed to efficiently process MRI scans and extract relevant features using pre-trained deep learning models, while incorporating techniques such as data augmentation, normalization, and class weighting to improve performance and generalization.

Data Ingestion

MRI scans are collected and organized into structured directories. For the binary classification task, images are sorted into two folders ("Impairment" and "No Impairment"), and for the multiclass task, they are divided equally among four categories ("Mild Impairment," "Moderate Impairment," "No Impairment," and "Very Mild Impairment"). This structured organization facilitates automated loading and preprocessing.

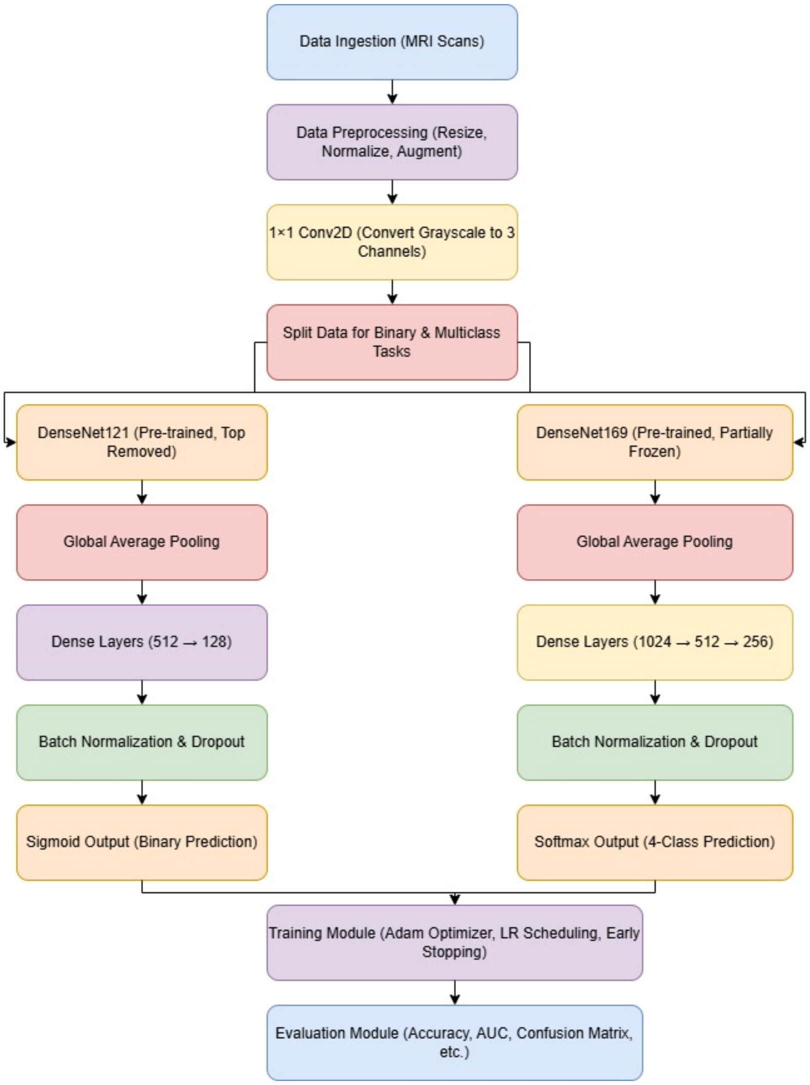
Preprocessing and Conversion

Each image is resized to match the model requirements—224×224 pixels for the binary model and 256×256 pixels for the multiclass model. Min–max normalization is applied to scale pixel intensities into the [0, 1] range using the formula:

$$x\_norm = (x - x\_min) / (x\_max - x\_min)$$

Data augmentation techniques, such as rotations, translations, zooming, and brightness adjustments, are employed to enhance the diversity of the training data and reduce overfitting. Additionally, because the MRI scans are in grayscale, a 1×1 convolution layer is applied to convert single-channel images into three-channel inputs, making them compatible with the pre-trained DenseNet models.

`


Data Ingestion (MRI Scans)
Data Preprocessing (Resize, Normalize, Augment)
1×1 Conv2D (Convert Grayscale to 3 Channels)
Split Data for Binary & Multiclass Tasks
DenseNet121 (Pre-trained, Top Removed)
DenseNet169 (Pre-trained, Partially Frozen)
Global Average Pooling
Global Average Pooling
Dense Layers (512 → 128)
Dense Layers (1024 → 512 → 256)
Batch Normalization & Dropout
Batch Normalization & Dropout
Sigmoid Output (Binary Prediction)
Softmax Output (4-Class Prediction)
Training Module (Adam Optimizer, LR Scheduling, Early Stopping)
Evaluation Module (Accuracy, AUC, Confusion Matrix, etc.)

Dual-Branch Pipeline

After preprocessing, the data is directed into two parallel branches:

a. Binary Branch:

– Utilizes DenseNet121 (with its top layers removed) as a feature extractor.

– Global average pooling aggregates spatial feature maps into a fixed-length vector.

– This vector is processed by fully connected layers with batch normalization and dropout, culminating in a final sigmoid-activated output that provides a binary prediction.

– The binary model is optimized using binary cross-entropy loss, where:

`

$$L = –(1/N)\ \Sigma\ [y \cdot \log(p) + (1 – y) \cdot \log(1 – p)]$$

b. Multiclass Branch:

– Employs DenseNet169 with partial layer freezing to retain robust pre-trained features while fine-tuning later layers.

– Global average pooling is used to convert feature maps into a vector representation.

– A series of dense layers with decreasing units (1024, 512, and 256) are applied, with batch normalization and dropout, leading to a final softmax layer that outputs a probability distribution over the four classes.

– The multiclass model uses categorical cross-entropy loss defined as:

$$L = – \Sigma\ [y\_i \cdot \log(p\_i)]$$

Training Module

Both branches are trained using the Adam optimizer with an adaptive learning rate schedule and early stopping mechanisms based on validation loss. Class weighting is applied in both models to compensate for any imbalances in the dataset. The training module is responsible for iteratively updating model weights and monitoring performance metrics such as accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC).

Evaluation Module

Post-training, the Evaluation Module assesses the performance of both models using a variety of metrics. For the binary model, a detailed classification report and confusion matrix confirm near-perfect performance. In the multiclass model, per-class accuracy and confusion matrices are analyzed to identify areas of misclassification and challenges in distinguishing subtle differences between classes.

Both branches are trained using the Adam optimizer with tailored learning rate schedules, early stopping, and class weighting to counteract class imbalance. The system's Training Module coordinates these optimizations, while the Evaluation Module computes performance metrics (accuracy, precision, recall, F1-score, AUC, and confusion matrices) for comprehensive model assessment.

`

Experimental results

This section presents a comprehensive analysis of the performance of the two developed models. The results are detailed separately for the binary classification model and the multiclass classification model. Each section includes descriptions of the dataset, training behavior, evaluation metrics, and insights derived from confusion matrices and classification reports.

A. Binary Classification Results

Dataset Overview

• The binary dataset comprises 5,120 MRI images equally divided into two classes: “Impairment” (2,560 images) and “No Impairment” (2,560 images).

• Images are resized to 224×224 pixels and normalized using min–max normalization:

$$x\_norm = (x - x\_min) / (x\_max - x\_min)$$

• To adapt grayscale images for the DenseNet121 model, a 1×1 convolution layer is applied to convert them into three-channel inputs.

Training Performance

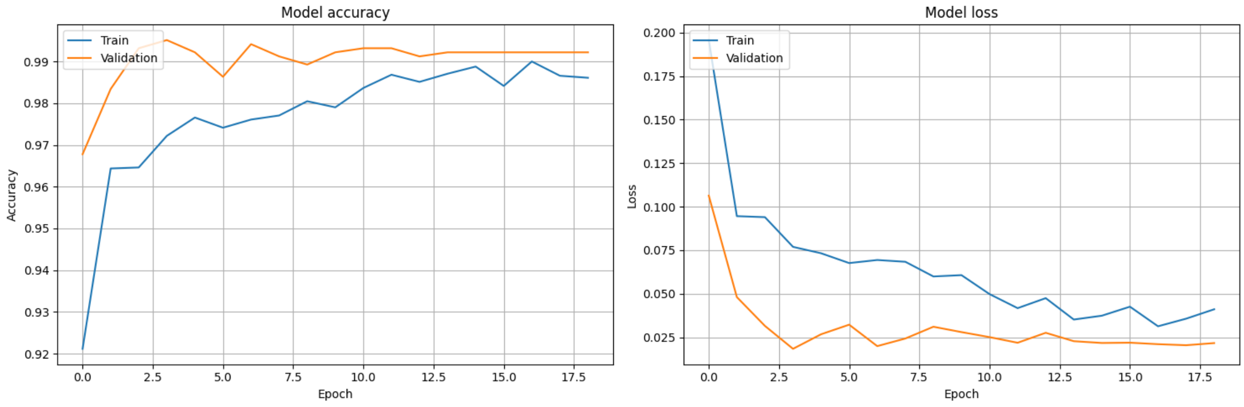
• The binary model, built on DenseNet121, demonstrated rapid convergence during training. Early epochs showed a consistent rise in both training and validation accuracy while the loss decreased steadily.

• The final training accuracy reached approximately 98.6%, with the validation accuracy approaching 99.2%.

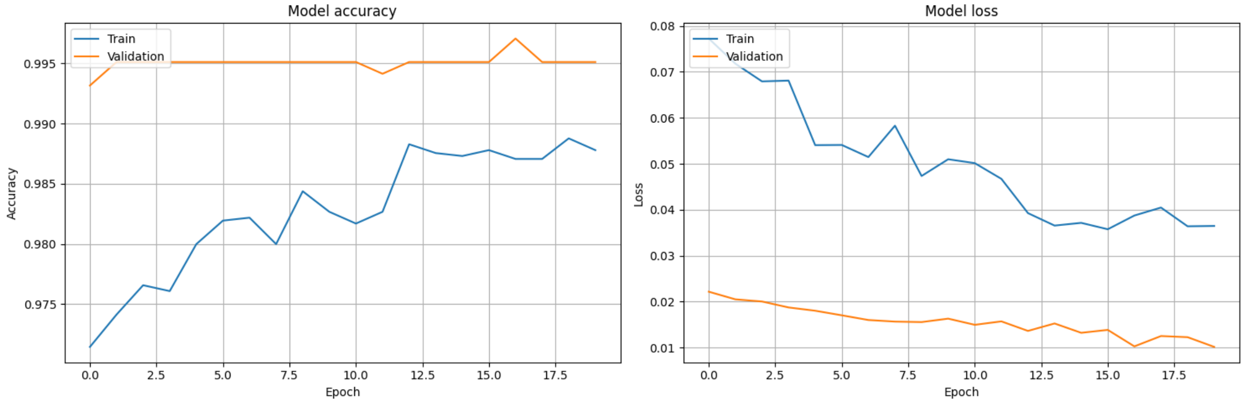
• Loss values stabilized at around 0.0411 on the training set and 0.0217 on the validation set.

• Fine-tuning by unfreezing selected layers of the DenseNet121 model further improved performance, with the best fine-tuned model achieving a validation accuracy close to 99.5% and a reduction in loss (training loss ~0.0365, validation loss ~0.0102).

Before fine tuning



After fine tuning



Evaluation Metrics

- The binary cross-entropy loss function, defined as

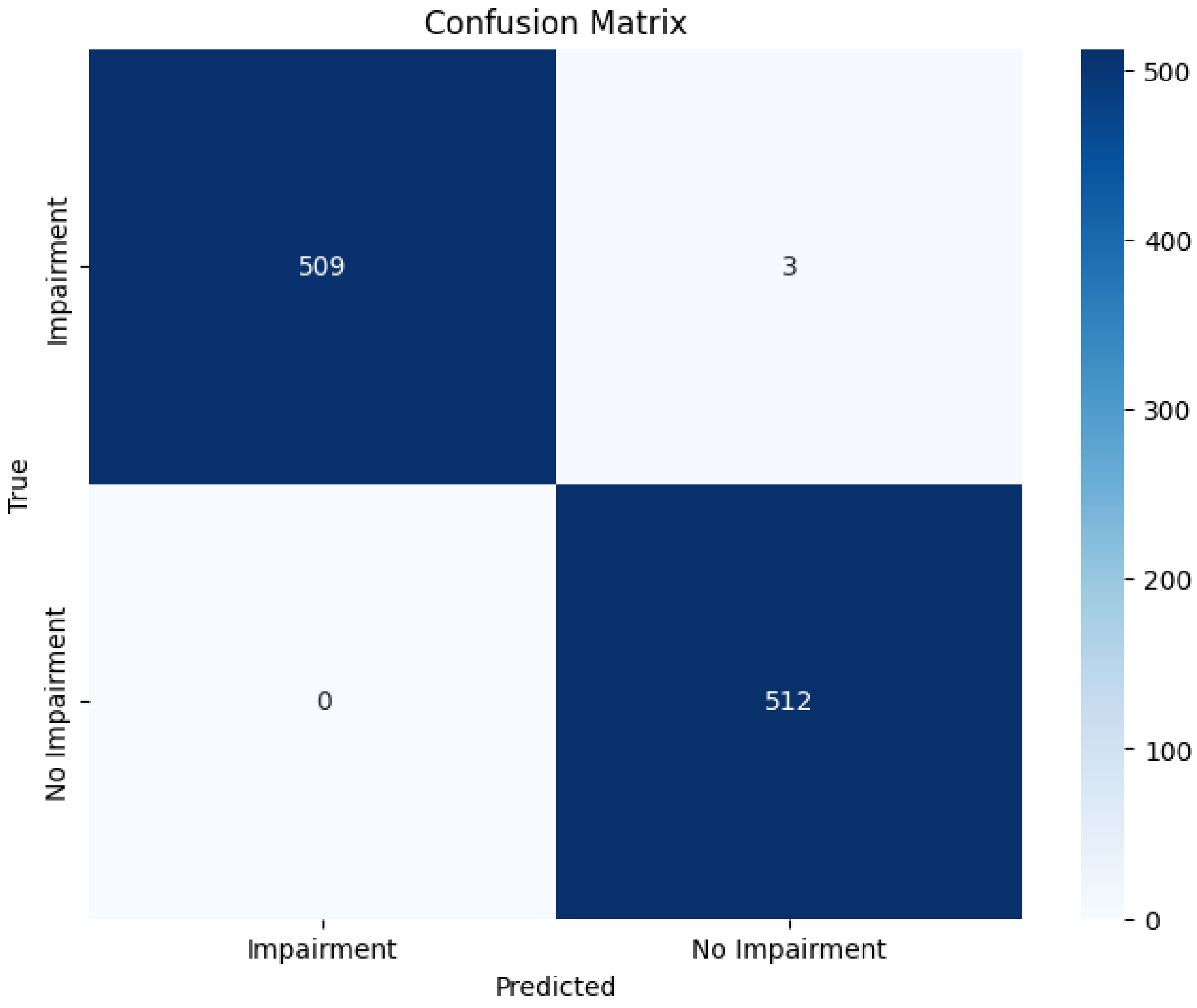
$$L = -(1/N)\ \Sigma\ [y \cdot \log(p) + (1 - y) \cdot \log(1 - p)],$$

provided a reliable measure of prediction error.

- Detailed classification reports revealed that both classes exhibited precision, recall, and F1-scores near 1.00, indicating almost perfect discrimination between “Impairment” and “No Impairment.”

- The confusion matrix confirmed that misclassifications were minimal, with the majority of samples correctly classified, thereby validating the model’s robustness and reliability.

`



Additional Observations

• Data augmentation techniques (rotations, translations, zoom, brightness adjustments) and class weighting (w_i = N / (C · n_i), where C=2) were essential to counteract any overfitting tendencies and ensure balanced learning.

• The high performance of the binary model suggests that the combination of DenseNet121-based feature extraction and rigorous preprocessing can serve as a highly effective screening tool in a clinical setting.

`

B. Multiclass Classification Results

Dataset Overview

• The multiclass dataset contains 10,240 MRI images, evenly distributed among four classes: “Mild Impairment,” “Moderate Impairment,” “No Impairment,” and “Very Mild Impairment” (2,560 images per class).

• Images are resized to 256×256 pixels, which helps in capturing finer anatomical details.

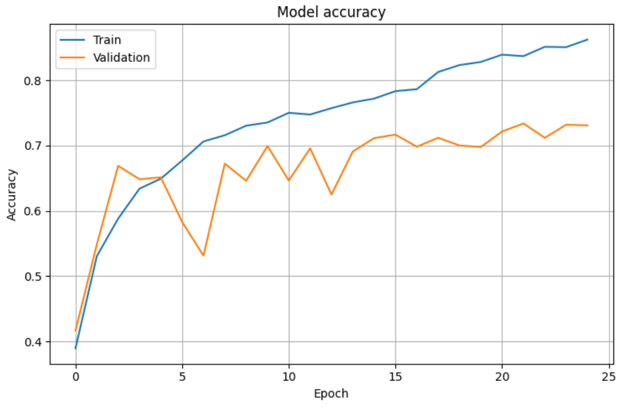
• Similar to the binary task, min–max normalization is applied, and a 1×1 convolution layer converts grayscale images into three-channel inputs, making them suitable for the DenseNet169 model.

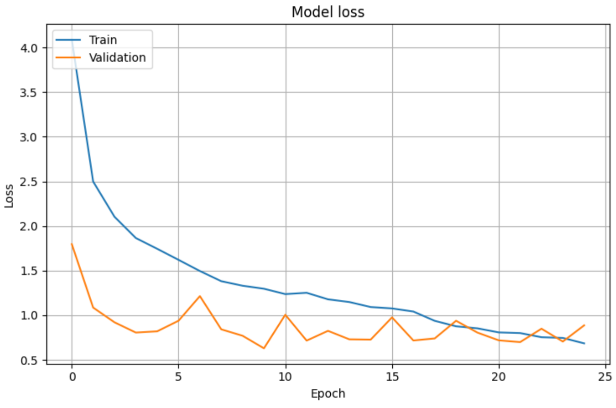
Training Performance

• The multiclass model, based on DenseNet169 with partial layer freezing, was trained for up to 40 epochs. The training process involved extensive data augmentation and a carefully tuned learning rate schedule (starting at 0.0002).

• Although the model showed steady improvements during the initial epochs, the validation accuracy plateaued at around 73.3%.

• Loss values decreased consistently during training; however, the validation loss remained relatively higher compared to the binary model, reflecting the increased complexity of differentiating between four closely related classes.





`

Per-Class Analysis and Evaluation

• The multiclass model's overall validation accuracy was approximately 73.3%, but per-class performance varied considerably:

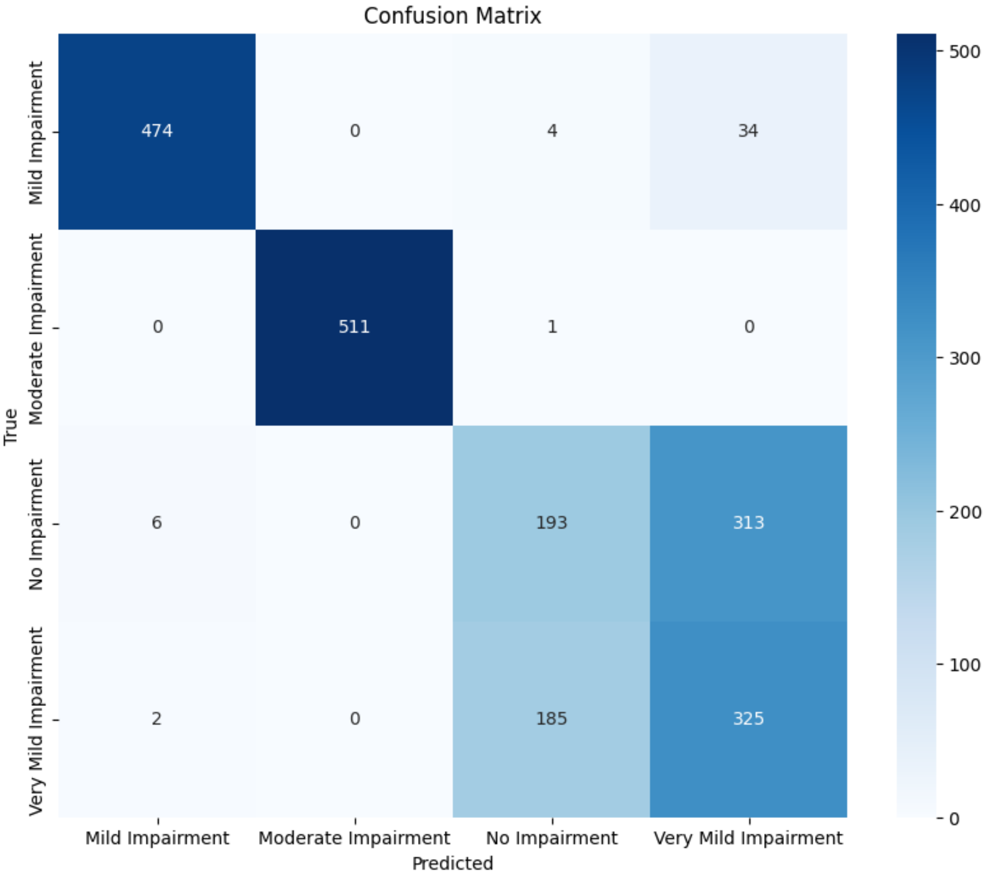
> – "Moderate Impairment" achieved very high accuracy (≈ 99.8%), demonstrating that the model could capture pronounced pathological features.
>
> – "Mild Impairment" performed well with an accuracy of around 92.6%.
>
> – Conversely, "No Impairment" and "Very Mild Impairment" yielded lower accuracies of approximately 37.7% and 63.5%, respectively.

• The categorical cross-entropy loss function, defined as

$$L = -\Sigma [y_i \cdot \log(p_i)],$$

was used to compute the discrepancy between the predicted and true class distributions.

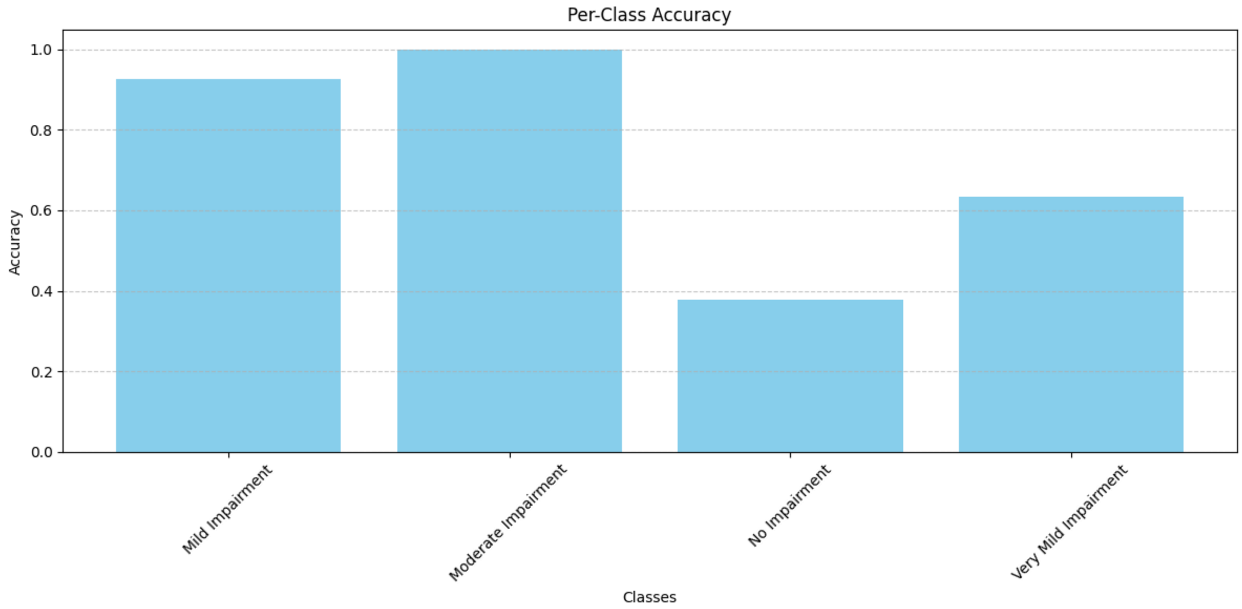
• The confusion matrix revealed significant misclassifications, particularly within the "No Impairment" class, where 319 samples were incorrectly classified. This suggests that the



True \ Predicted	Mild Impairment	Moderate Impairment	No Impairment	Very Mild Impairment
Mild Impairment	474	0	4	34
Moderate Impairment	0	511	1	0
No Impairment	6	0	193	313
Very Mild Impairment	2	0	185	325

subtle differences between “No Impairment” and “Very Mild Impairment” remain challenging for the model.

• Detailed precision, recall, and F1-score metrics further underscore that while some classes are robustly detected, others suffer from lower predictive performance, indicating the need for additional refinement in feature extraction and class separation.



Additional Observations

• Class weighting in the multiclass scenario was adapted to emphasize classes with subtle differences, using the formula:

w_i = N / (C · n_i),   with C = 4.

• The application of data augmentation—through slight rotations, shifts, zoom, and brightness variation—helped improve generalization but was insufficient to completely overcome the intrinsic challenge of class overlap.

• The fluctuations in the training curves for the multiclass model suggest that the model might benefit from further optimization, such as experimenting with more complex architectures, advanced loss functions (e.g., focal loss), or integrating additional clinical features to support the imaging data.

Overall, the experimental results demonstrate that the binary classification model achieves near-perfect performance, making it highly reliable for a screening application. In contrast, the multiclass model, despite its promising performance in certain categories, reveals significant challenges in differentiating subtle impairment levels. These findings highlight the potential of deep learning in Alzheimer's MRI classification and point to specific areas for future improvement, such as enhanced feature extraction methods and more sophisticated optimization strategies.

Discussion on Challenges and Optimization

- The lower performance for certain classes in the multiclass task underscores the difficulty in capturing nuanced features in MRI scans when the differences are subtle.

- While the overall validation accuracy of approximately 73.3% is promising, the model's uneven performance across classes indicates that additional refinement—such as further data augmentation, improved feature extraction methods, or advanced loss weighting strategies—may be required.

- The training curves for multiclass classification, which show fluctuations in both accuracy and loss, suggest that more sophisticated optimization techniques could further enhance model stability and performance.

Overall, the experimental results demonstrate that the binary classification model achieves near-perfect performance, making it suitable for reliably distinguishing between impaired and non-impaired subjects. In contrast, while the multiclass model shows promise, it also reveals clear areas for further investigation and optimization, particularly in handling the more challenging subtle distinctions between certain impairment levels. These findings provide valuable insights into the effectiveness of transfer learning and the need for targeted improvements in complex medical image analysis tasks.

`

Discussion

The experimental results highlight notable differences in performance between the binary and multiclass classification models, offering valuable insights into the strengths and limitations of each approach.

The binary classification model achieved exceptional performance, with training and validation accuracies exceeding 98% and 99%, respectively. This near-perfect performance is reflected in the extremely low loss values and classification metrics—precision, recall, and F1-scores all approach 1.0. The success of the binary model can be attributed to several factors. First, the problem of distinguishing between “Impairment” and “No Impairment” presents a clearer decision boundary compared to the multiclass scenario. The use of DenseNet121 as a feature extractor, combined with rigorous preprocessing (resizing, normalization, and data augmentation), effectively captured the salient features in the MRI scans. Furthermore, the application of a 1×1 convolution layer ensured compatibility with the pre-trained network by converting grayscale images to three-channel inputs, preserving essential image details.

In contrast, the multiclass classification model, while promising, exhibited significant challenges. With an overall validation accuracy of around 73.3%, the multiclass model struggled particularly with differentiating between classes that have subtle differences—specifically between “No Impairment” and “Very Mild Impairment.” The DenseNet169-based model, which was chosen for its higher capacity, did show high accuracy for classes with more pronounced features, such as “Moderate Impairment” (approximately 99.8%) and “Mild Impairment” (around 92.6%). However, the lower accuracies for “No Impairment” (approximately 37.7%) and “Very Mild Impairment” (approximately 63.5%) suggest that the network had difficulty discerning the nuanced differences between these categories. The confusion matrix analysis further confirmed that a considerable number of samples from the “No Impairment” class were misclassified, pointing to the intrinsic challenge of subtle feature differentiation in medical imaging.

Several factors may contribute to the performance disparity between the two models. The binary model benefits from a simpler decision space and a balanced dataset, which makes the classification task more straightforward. Conversely, the multiclass model faces a more complex classification task, where the overlap in imaging characteristics between classes can lead to ambiguity in the model's predictions. Despite the use of data augmentation and class weighting strategies, these techniques may not be sufficient to overcome the inherent challenge of capturing fine-grained differences in MRI features.

Additionally, the training curves for the multiclass model displayed fluctuations in both accuracy and loss, suggesting that the model might be sensitive to the learning rate schedule and the amount of training data available. The partial freezing of DenseNet169's layers helped preserve useful pre-trained features but may also have limited the network's ability to fully adapt to the unique characteristics of the Alzheimer's MRI data.

In summary, while the binary classification model demonstrates that deep learning can be highly effective in screening for Alzheimer's-related impairment, the multiclass model reveals the need for further refinement when dealing with subtle gradations of disease severity. Future work should focus on exploring more advanced network architectures, alternative loss functions, and potentially integrating multimodal data to improve the multiclass model's performance. These insights underscore the importance of tailoring deep learning approaches to the specific challenges of medical image analysis, where even minor differences in features can have significant clinical implications.

Conclusion and Future Work

Conclusion: This project demonstrates the potential of deep learning approaches in the classification of Alzheimer’s disease using MRI scans. The binary classification model, based on DenseNet121, achieved near-perfect performance—with training and validation accuracies exceeding 98% and 99%, respectively—and minimal misclassification as evidenced by its low loss values and near-unity precision, recall, and F1-scores. These results validate the model’s robustness and highlight its suitability as a screening tool for distinguishing between “Impairment” and “No Impairment.” In contrast, the multiclass model built on DenseNet169, although promising, achieved an overall validation accuracy of approximately 73.3%. Detailed analysis revealed that while the model performed exceptionally well in identifying Moderate and Mild Impairment, it struggled with subtle distinctions between “No Impairment” and “Very Mild Impairment.” These challenges underscore the inherent complexity of differentiating among closely related classes in medical imaging.

Future Work: Building upon these findings, future research should explore several avenues to improve multiclass performance and broaden the clinical applicability of the models. First, more advanced or hybrid network architectures could be investigated to enhance feature extraction, especially for capturing subtle variations in MRI scans. The incorporation of alternative loss functions, such as focal loss, may better address the class imbalance and the challenge of differentiating borderline cases. Moreover, integrating multimodal data—including clinical assessments, neuropsychological scores, and genetic information—could enrich the feature space and potentially improve diagnostic accuracy. Additional data augmentation techniques and ensemble methods could further bolster model robustness. Finally, prospective clinical studies and validation on larger, more diverse datasets are essential to assess generalizability and refine these models for real-world diagnostic workflows.

`

References

[1] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700-4708).

[2] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A Survey on Deep Learning in Medical Image Analysis. Medical Image Analysis, 42, 60-88.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

[4] Suk, H.-I., Lee, S.-W., & Shen, D. (2014). Hierarchical Feature Representation and Multimodal Fusion with Deep Learning for AD/MCI Diagnosis. NeuroImage, 101, 569-582.

[5] Alzheimer's Disease Neuroimaging Initiative (ADNI).