ISSN 2007-9737

Multimodal Deep Learning Fusion Strategies for Alzheimer's Disease Classification

Ayrton Santos, Claudia I. Gonzalez*, Mario Garcia

Tijuana Institute of Technology/TECNM,
Mexico

cgonzalez@tectijuana.mx, ayrton.santos, mario@tectijuana.edu.mx

Abstract. This study explores multimodal deep learning models for diagnosing Alzheimer's disease (AD), integrating up to eight data modalities, including clinical, imaging, and genetic information, using the OASIS-3 dataset. Three fusion strategies (early, late, and intermediate) are implemented and compared to effectively combine heterogeneous data. Three case studies are considered: (1) binary classification using magnetic resonance imaging (MRI) and the Clinical Dementia Rating (CDR) scale to distinguish between cognitively normal individuals and those with AD dementia; (2) binary classification using eight modalities (MRI, CDR, CENTILOID, FAQ, NPI-Q, D1, C1, and B8), improving predictive accuracy and robustness; and (3) multiclass classification with the same eight modalities to predict cognitive normality, uncertain dementia, or other dementia subtypes. Experimental results show that multimodal models consistently outperform unimodal approaches, demonstrating superior classification performance and greater resilience to uncertainty. Despite challenges in model interpretability and dataset balancing, these findings underscore the potential of multimodal fusion strategies to improve computer-assisted diagnosis of Alzheimer's disease.

Keywords. Alzheimer's disease classification; multimodal deep neural networks; neurodegenerative diseases; multimodal CNN.

1 Introduction

Convolutional neural networks (CNNs) have been shown to be highly effective for analyzing complex biomedical information. A representative example is the study presented in [33], where CNNs were used to classify EEG signals into different neurological states, achieving high levels of accuracy in the classification of various brain-related conditions.

Multimodal deep learning [9] is an emerging approach that enables the simultaneous processing of different types of data such as text, images, audio and sensors. New trends in generative Artificial Intelligence AI have broken the monomodality barrier, also at a commercial level, since we are increasingly seeing more models on the market that seek to work with specific tasks while maintaining the multimodal essence.

Multimodal learning [26], for most humans, senses such as hearing, smell, taste, and sight are of great help in perceiving our environment. Multimodal means simultaneously combining different data from different senses or sensors. Based on this, researchers have been inspired to integrate multiple modalities into deep learning models. While some models are limited to numerical inputs and outputs, multimodal approaches offer the advantage of processing unstructured data, such as images or text. In the state of the art there are some multimodal deep learning approaches.

The research presented in [6] states that the world is multimodal, as we can see objects, feel textures, hear sounds, and experience flavors. For AI algorithms to achieve better results, it is first necessary to focus on how to capture and summarize multimodal data.

In [34], multimodality is used to predict endometrial cancer. They implement a multimodal model that combines clinical and visual information. Their architecture was designed to

consider multimodality, thus considering data that, with a unimodal approach, would normally be ignored. The authors note that the integration of three modalities provided them with superior performance, surpassing the clinical standard.

In [36], multimodality is implemented by attempting to match images with text. That is, it is necessary to match the text description with the images. The architecture proposes the use of three blocks; the first block is a convolutional neural network that extracts high-level features; the second block, a recurrent neural network to represent the semantic content of the text; the third, a multi-core module that allows fusing representations; and the fourth, a visual transformer that captures the global and local details of the images.

In [18] the authors analyzed how multimodality helps in the classification or mapping of remote sensing and land cover images, due to the high demands of urban planning, forest monitoring, soil analysis and possibly natural disaster management.

In the research presented in [38], the usefulness of developing multimodal deep learning models to predict a wide range of diseases or conditions using CT scans, MRIs, ultrasounds, and microscopy is analyzed. After feature extraction, a special multimodal fusion technique is employed to integrate the information. The model was evaluated in three main areas: classification, lesion localization, and generation of clinical descriptions. This innovative multimodal model proposes deep fusion between medical images and clinical reports, the neural network that extracts features from images, and the bidirectional network that facilitates the analysis of clinical reports. The study is carried out using a large database of medical images of different diseases.

In [28], the authors present a study where the combination of different drugs has proven to be effective for the treatment of cancer. The objective is to design new drugs and cancer therapies. Computational methods use synergy to predict effective drug combinations. The architecture is roughly composed of four main subnetworks: a protein interaction network, a network that extracts drug features, a drug-protein interaction network, and finally, a synergy prediction network.

In [20], multimodal deep learning is proposed to predict cardiovascular diseases. The premise of this study is to combine two types of data: clinical risk factors such as cholesterol, age, and blood pressure, among others. Retinal photographs are used, which are cheaper than cardiac CT scans. Retinal photographs are noninvasive and can show abnormalities. The study is relevant because it combines two types of data: images of blood vessels.

Neurodegenerative diseases, such as Alzheimer's, constitute a public health problem among the older adult population worldwide, wreaking havoc in areas such as the economic situation of family members and the loss of patients39; quality of life. Medicine has made significant advances in their practice. While there is currently no cure, improvements have been made in treatments to delay the onset of symptoms. Despite these advances, diagnosing these diseases is laborious and requires a series of diagnostic tests. In some cases, unfortunately, the disease is detected only after severe deterioration has already occurred. Early detection of neurodegenerative diseases is essential. This research is being conducted to explore and demonstrate how the development of predictive artificial intelligence models can improve diagnostic effectiveness. The aim of this work is to develop multimodal deep learning models for AD diagnosis, integrating up to eight data modalities, including clinical, imaging, and genetic information, using the OASIS-3 dataset [19]. The primary objective is to develop and implement a convolutional neural network (CNN)-based architecture capable of accurately classifying AD.

To this end, three fusion strategies, namely early, intermediate and late fusion, are implemented and compared, with the objective of identifying the advantages of each approach in effectively combining diverse inputs.

Three case studies are considered to evaluate model performance. The first focuses on binary classification with early fusion architecture using MRI images and Clinical Dementia Rating (CDR) to distinguish cognitively normal individuals from

AD dementia. The second case employs the same input modalities but utilizes a late fusion approach. The third case extends the binary classification task to eight modalities (MRI, CDR, CENTILOID, FAQ, NPI-Q, D1, C1, and B8) with an intermediate fusion strategy, aiming to enhance predictive accuracy and model robustness. The fourth case involves multiclass classification using the same eight modalities to predict cognitively normal, uncertain dementia, or other dementia subtypes, also using intermediate fusion to effectively integrate heterogeneous data sources. All data were extracted from the OASIS-3 repository after careful analysis, cleaning, and preprocessing.

Model performance is assessed using several metrics, including accuracy, precision, recall (sensitivity), specificity, and the area under the ROC curve (AUC). In general, this work is structured as follows: Section 2 reviews the state of the art in machine learning models applied to Alzheimer's disease diagnosis, including research using multimodal datasets, the datasets themselves, and multimodal fusion architectures in deep learning. Section 3 describes the proposed approach and methodology, detailing the overall workflow, the proposed architectures, the dataset employed, preprocessing procedures, training parameters, and evaluation metrics used for validation. Section 4 presents the experimental results, while Section 5 provides a discussion of these results. Finally, Section 6 offers conclusions and outlines directions for future work.

2 State of the Art

In recent years, the development of technology, computer vision, neural networks, and pattern recognition has allowed these tools to expand into other fields of research. Artificial neural networks and artificial intelligence are likely to be the future for solving numerous health problems. In this case, AI, with its ability to recognize patterns, can help predict which patients might be prone to developing a neurodegenerative disease. Below are a series of studies focusing on the use of machine learning and deep learning techniques applied to the diagnosis or classification of Alzheimer's disease.

2.1 Machine Learning Models Applied to the Diagnosis of Alzheimer's Disease

In [11], a dementia prediction system using machine learning was proposed through a project called OASIS (Open Access Series of Imaging Studies), provided by the Alzheimer's Research Center at the University of Washington. These data are preprocessed, and a data transformation is performed to create a suitable set for training the model. Some of the main approaches used are AdaBoost, Decision Trees, Extra Tree, Gradient Boost, K-Nearest Neighbor, Logistic Regression, Naive Bayes (NB), Random Forests, and Support Vector Machines to combine features. In addition, techniques such as the Minimal Absolute Selection and Shrinkage Operator (LASSO) were applied.

For this model, cross-sectional MRI images for young, middle-aged, and older adults, as well as longitudinal MRI images, were used. The dataset consists of 150 patients between the ages of 60 and 96. The best result was achieved with the Support Vector Machine (SVM) algorithm, using all features, achieving the highest accuracy, equivalent to 96.77%.

In [16], the implementation and comparison of Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) were proposed to optimize the hyperparameters of convolutional neural networks. These were applied to the classification of Alzheimer's disease using a dataset of magnetic resonance imaging scans. The study concludes that the approach based on PSO and CNN achieved high precision in classification, highlighting the importance of metaheuristics in improving model performance.

In [8], a 3D convolutional neural network is used to predict, based on MRIs, using biomarkers. The study consisted of 762 elderly patients: 459 healthy and 67 with mild cognitive impairment. The results showed that 236 patients had dementia, and the accuracy obtained was 76%.

In [21], a multimodal feature selection of 3 types for Alzheimer's detection was presented: MRI images, PET scans, and cerebrospinal fluid analysis. Multimodal learning involves using different types of data and learning the best from that combined information, be it text, audio, or

video. The problem is that common algorithms work with only one type of data at a time, which means that not all available information can always be used.

In [13], the work focuses on Alzheimer's disease; early prediction of Alzheimer's disease and mild cognitive impairment is performed using brain MRI and machine learning. Two datasets were used: the Open Access Imaging Study Series (OASIS) and the Alzheimer's Disease Neuroimaging Initiative (ADNI). Some of the algorithms used included support vector machines, decision trees, random forests, extremely random trees, linear discriminant analysis, logistic regression, and logistic regression with stochastic gradient descent. The best results were obtained with the random forests and extremely random trees algorithms.

Multiple statistical tools and learning algorithms are explored in [29]. Machine learning for the diagnosis of Alzheimer's disease in people over 75 years of age. As the technology improved, it was applied to medical image classification. The samples were split into 75% and 25% and used to train the algorithms, which run on powerful GPUs. Using convolutional neural networks, the optimized architecture called OViTAD was employed. These analysis sequences obtained average results of 94.32% and 97.88% for the fMRI and MRI sequences, respectively.

In [30], the AlzheimerNet framework is presented, a classifier based on convolutional neural networks, designed to identify all stages of Alzheimer's disease, with a group of healthy individuals and another of affected individuals, using the ADNI MRI dataset. This was achieved by applying preliminary preprocessing and data augmentation techniques. The implemented models were VGG16, MobileNetV2, AlexNet, ResNet50 and InceptionV3. Interestingly, InceptionV3 achieved high accuracy and was modified to create AlzheimerNet with an RMSprop optimizer; the achieved accuracy was 98.67%.

In [5], a model called Inception V3 was proposed, enhanced with transfer learning, and used to analyze MRI data and classify Alzheimer's patients and healthy individuals. A brightness adjustment was applied as part of the preprocessing to augment the dataset and improve accuracy. A technique called SMOTE was used to balance the classes. After preprocessing, the data was divided into validation and training sets, achieving an accuracy of 87.69% on the OASIS dataset, which has been widely used in previous studies.

In the proposal presented in [2], the research addresses feature selection, which produces better prediction performance than key gene sets. The five genes identified using the LASSO and Ridge methods achieved an area under the curve of 0.979. This work demonstrates how, with a small number of genes, it is possible to distinguish Alzheimer's disease from healthy controls with high accuracy. The dataset used comes from the Gene Expression Omnibus (GEO), filtered for Alzheimer's disease, and provides genetic data from brain tissue of healthy and affected patients within the same age range. Random methods, support vector machines, and convolutional neural networks were used.

An improved AdaBoost classification technique for diagnosis is presented in [31]. It detects Alzheimer's disease from MRI images, reducing construction time. The weak classifier uses the PSO algorithm, replacing the exhaustive search with an optimized PSO-based search in the weak classifier, which consists of a "decision tree." The results show that the PCA-PSO-AdaBoost combination is more effective than other methods, outperforming PCA-AdaBoost and PSO-SVM-Cuckoo-AdaBoost by 16.47%.

In [14] a study is proposed that compares the performance of three machine learning algorithms: Random Forest, Gradient Boosting and eXtreme Gradient Boosting, using multimodal biomarkers of subjects with mild cognitive impairment (MCI) obtained from the ADNI database. The prediction rate is related to the nature of the data, such as neuropsychological tests, Alzheimer's-related proteins, cerebrospinal fluid and MRI, among others. Electronic MRI data alone presented a lower accuracy (0.79), but multimodal data, which combine clinical and biological measures, achieved a higher accuracy (0.90).

2.2 Research Works Implementing a Multimodal Dataset

In the research presented in [13], [30], [31] and [14], the authors use a multimodal methodology. On the other hand, there are several exclusively multimodal studies, such as [4], which propose a multimodal model to improve the detection of neurodegenerative diseases. This model uses various advanced methods to analyze various data, such as time-frequency analysis, electromagnetic resonance, and genetic data.

These features allow for accurate diagnosis, with a 10% increase in accuracy and a 2.9% reduction in time. This study focused on a wide range of neurodegenerative diseases. The MC-RVAE model, designed for multimodal management of Alzheimer's disease [23], works with MRI and cognitive scores. This model was trained with synthetic and ADNU data with approximately 3000 epochs. Nearest neighbor models, random forests, and cluster factor analysis were used. The results were compared using a flexible and scalable ANOVA model for each task. In [22], a generative model is proposed that simultaneously estimates the continuous progression of biomarkers for different types of rare neurodegenerative diseases.

It uses transfer learning. Its most notable feature is that it includes a function called agnostic units, which represent dysfunctions in specific brain regions and are common to all diseases. In [10], a study is presented on a multimodal dataset, applying independent denoising autoencoders for each modality. In this work, a new metric for working with states and patches is presented, creating a classification framework and training a deep neural network (DNN) with supervised learning. The classification is reported to be promising and offers improvements over traditional statistical metrics, such as NMI and LCC, in terms of correct correspondence matching.

2.3 Multimodal Dataset Used in the State of the Art

Datasets are fundamental, as they form the basis for model development, training, and evaluation. These datasets provide the information needed to extract patterns and perform meaningful analyses.

Table 1. Summary of Multimodal Datasets in Neuroscience

Dataset	Description	Ref
OASIS	MRIs, clinical data	[37]
ADNI	MRI/PET, biomarkers	[3]
UK Biobank	MRI/CT, genomics	[32]
GEO	Gene expression	[25]
PPMI	clinical+biomarkers	[27]
MDS-UPDRS	Parkinson's assessment	[24]

Table 1 presents the most commonly used datasets in scientific research related to the detection of Alzheimer's and Parkinson's diseases.

2.4 Multimodal Fusion Architectures in Deep Learning

Multimodal deep learning is a subfield of artificial intelligence that focuses on developing models capable of learning and integrating information from multiple modalities or data sources, such as text, images, audio, video, and sensor signals. Unlike unimodal learning, which processes a single type of input, multimodal deep learning seeks to capture complementary and correlated information across different modalities to build richer, more robust, and contextual representations. This integration is achieved through deep neural architectures that extract, align, and fuse modality-specific features at various levels of abstraction (early, intermediate, or late fusion). These models enable improved performance in complex tasks, such as medical diagnosis, emotion recognition, or autonomic perception, where understanding the relationships between different types of data is essential. The selection of a fusion technique, in most cases, is based more on the researcher's intuition than on a systematic and rigorous evaluation of its performance. The architecture of a multimodal model can be classified according to the timing or stage at which data from different modalities are integrated. Broadly speaking, there are three widely used fusion approaches: early fusion, intermediate fusion, and late fusion, which are briefly described below [12]:

Table 2. Steps for Early Fusion

Step	Phase	Description
1.	Data Collection	Obtain multimodal data
2.	Preprocessing	Clean and normalize
3.	Feature Extraction	Modality-specific encoding
4.	Feature Fusion	Combine representations
5.	Joint Learning	Learn correlations
6.	Model Training	End-to-end training
7.	Evaluation	Performance assessment

1. Early Fusion: This type of fusion is characterized by combining information from one or more modalities during the initial stages, usually at the feature level [12]. In this fusion, the features extracted from each modality are combined before the classification layer. Each modality passes through its own encoder (CNN for images, MLP for tabular data), and the resulting embeddings are concatenated into a single multimodal feature vector. This fused vector is then fed into the final fully-connected layers for classification [35]. The steps for Early Fusion [7] are described in Table 2.

2.4.1 Late Fusion

In late fusion, each modality is processed independently through its own complete network (encoder + classifier). Each branch produces a single prediction, which is combined at the end using averaging, weighted averaging, or voting to generate the final prediction. Figure 2.3 shows a diagram of what early and late fusion might look like [35]. The steps to process the Late Fusion in Multimodal Deep Learning [15] are explained in Table 3.

2.4.2 Intermediate Fusion

This type of fusion achieves a balance, integrating data as it combines it, allowing for effective results using the specific features of each modality. Intermediate fusion stands out for the robustness of the models, as they preserve the distinctive characteristics of each method. In intermediate fusion, models can reveal patterns that might be overlooked by other techniques [1]. In [17], intermediate fusion processes features from the modalities separately, combines them, and then processes them before issuing a verdict.

Table 3. Steps for Late Fusion in Multimodal Deep Learning

Step	Phase	Description
1.	Data Collection	Obtain multimodal data
2.	Preprocessing	Modality-specific preparation
3.	Feature Extraction	Independent encoders
4.	Prediction	Individual outputs
5.	Fusion	Combine predictions
6.	Decision	Final classification
7.	Evaluation	Performance assessment

After combining the features, additional steps are performed, such as sending them through a convolutional neural network, to finally make a decision. The fundamental difference with respect to other fusions lies in the timing of the decision to fuse. The Steps for Intermediate Fusion [17] in Multimodal Deep Learning are described in Table 4.

3 Methodology

In this Section, the methodological process followed for the development of multimodal deep learning models applied to Alzheimer's disease classification is described. The general workflow, proposed architectures, utilized dataset, preprocessing procedures, training parameters, and evaluation metrics employed to validate the results are presented.

3.1 Model Architecture

This Section presents the general workflow of the methodology. The flow integrates the different stages, from the acquisition and preprocessing

Table 4. Steps for Intermediate Fusion in Multimodal Deep Learning

Step	Phase	Description
1.	Data Collection	Obtain multimodal data
2.	Preprocessing	Modality-specific preparation
3.	Feature Extraction	Independent encoders
4.	Intermediate Output	Individual representations
5.	Fusion	Combine intermediate features
6.	Joint Learning	Shared representation learning
7.	Prediction	Downstream tasks
8.	Training	End-to-end optimization
9.	Evaluation	Performance analysis

of multimodal data to the implementation and evaluation of the deep learning models.

3.1.1 Multimodal Deep Learning with Early Fusion

Figure 1 presents the general process of the multimodal deep learning model with early fusion designed for the prediction of neurodegenerative diseases, specifically for the classification of Alzheimer's disease, using two data modalities: MRI and CDR. The following subsections describe in detail the stages that comprise this workflow:

— **Data Input:** CSV files containing tabular information (CDR) and images (MRI) provided by OASIS-ID are loaded.

— **Preprocessing:** For the images, the black background is cropped, each channel is equalized, images are resized to 128×128, converted to tensors, and normalized. For the tabular data, relevant numerical columns are selected, and missing values are filled with 0.

— **Neural Networks (and Feature Extraction):** The CDR branch processes the tabular vector as a sequence and extracts the most relevant information. The MRI branch applies several layers to the slice to detect visual patterns (edges, textures, structures) and produces an image feature vector.

— **Early Fusion:** The summaries from CDR and MRI are merged into a single combined vector. Training: The model is trained to optimize this combined classifier.

— **Classification:** The system outputs a binary prediction (Cognitively Normal vs. AD Dementia).

3.1.2 Multimodal Deep Learning with Late Fusion

Figure 2 illustrates a deep learning architecture for Alzheimer's disease classification in which the information from the two modalities, MRI images and CDR data, is combined at the end of the network process (late fusion). The workflow of the process is described below.

— **Data Input**: As in early fusion, the process begins with two separate branches: one for MRI images and another for CDR data, also matched by OASIS-ID.

Preprocessing: For tabular data, the CSV is read, missing values are filled with 0, and relevant numerical columns are selected. For images, grayscale conversion and resizing to 128×128 are applied.

— **Neural Networks**:

– **CDR Branch (tabular data)**: The clinical variable vector for each patient (CDR) is processed through a small network that detects patterns, producing a compact numerical vector capturing the most relevant information. This summary is used to generate a partial prediction from the tabular network.

– **MRI Branch (images)**: Brain slices (MRI) are processed through layers that detect visual features such as edges, textures, and structures, reducing the image to a numerical summary. This vector is used to generate a partial prediction from the image branch.

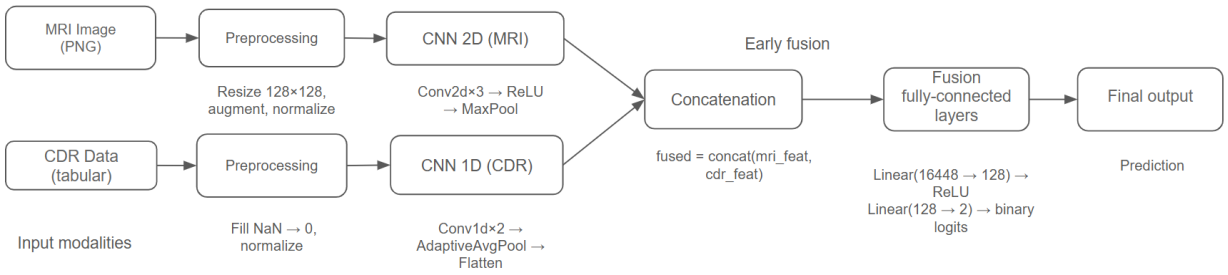

Fig. 1. Multimodal Deep Learning with Early Fusion

— **Feature Extraction**: The intermediate outputs of each branch are the compact vectors summarizing the most important information from each modality (CDR or MRI). In the final workflow, these feature vectors feed the layers that generate predictions during the late fusion stage.

— **Late Fusion**: Predictions (logits) from both modalities (CDR and MRI) are concatenated. The combined prediction vector is passed through an additional dense layer, which acts as a classifier that learns the best way to weight and combine the individual predictions from CDR and MRI for the final outcome.

— **Classification**: The process concludes with the final label prediction (Cognitively Normal or AD Dementia), completing the classification workflow.

3.1.3 Multimodal Deep Learning with Intermediate Fusion

Figure 3 presents a deep learning architecture adapted for the use of multiple data modalities. Its main feature is intermediate fusion, whereby the different modalities are integrated at an intermediate stage of the model, allowing a balance between the initial independence of each information source and their subsequent combination. The following steps outline the workflow of this process:

— **Data Input**: The process begins with multiple data sources, divided into two branches: one handling magnetic resonance imaging (MRI) and the other handling tabular data. The tabular branch includes clinical and biomarker data, described as follows:

  – **CDR (Clinical Dementia Rating)**: Clinical assessment scale.
  – **CENTILOID**: A PET imaging biomarker for amyloid plaque density.
  – **FAQ (Functional Activities Questionnaire)**: Questionnaire on functional activities.
  – **NPI-Q (Neuropsychiatric Inventory Questionnaire)**: Questionnaire on neuropsychiatric symptoms.
  – **D1**: Global clinical diagnosis.
  – **C1**: Specific cognitive tests.
  – **B8**: Physical and neurological findings.

— **Data Preprocessing**: MRI images undergo automatic black-background cropping, channel-wise equalization, and resizing to 132×132. Tabular data are matched using OASIS-ID. Relevant columns are selected, and missing values are filled with zeros.

— **Neural Networks**: The MRI branch uses a VGG-like network with three convolutional blocks to extract image features. The tabular branch employs a multilayer perceptron with several linear layers interleaved with BatchNorm, LeakyReLU, and Dropout for robust feature extraction.

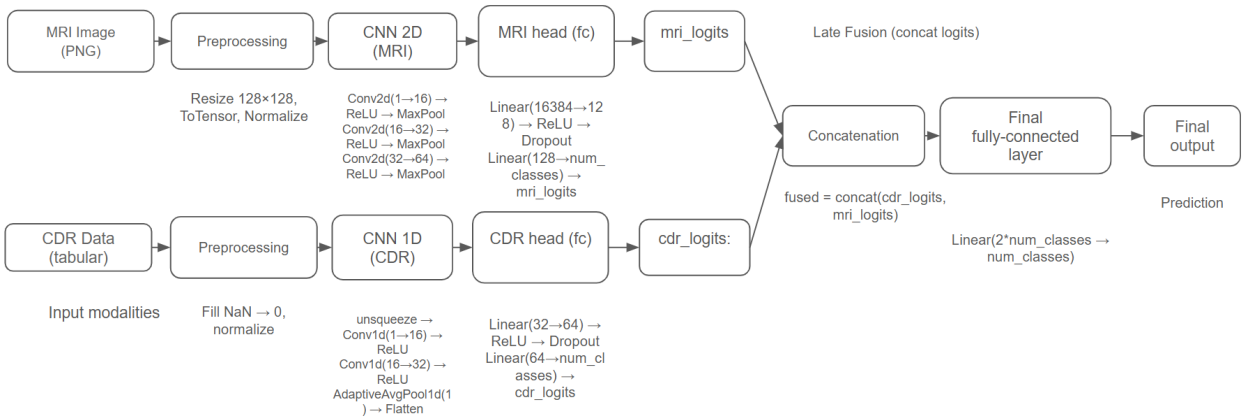

Fig. 2. Multimodal Deep Learning with Late Fusion

— **Feature Extraction**: Each branch produces a compact summary capturing the most relevant information from its respective input (images or tabular data).

— **Intermediate Fusion**: The processed feature vectors from both MRI and tabular branches are concatenated. Unlike early fusion (which combines raw features) or late fusion (which combines predictions), intermediate fusion merges refined features from each branch. The resulting vector is fed into fully connected layers, which learn cross-modal representations, performing true "fusion" at a higher, more abstract level.

— **Output Layer (Classification)**: The output of the fully connected fusion layers is passed to a final classification layer. In the first architecture, the model performs binary classification (Cognitively Normal vs. Dementia). In the second architecture, a multiclass model classifies the samples into Cognitively Normal, Uncertain Dementia, or Other Dementia.

3.2 Multimodal Dataset

OASIS-3 is a multimodal longitudinal dataset specializing in Alzheimer's disease (Table 5). It contains information on 1,378 patients, aged between 42 and 95 years, collected over 30 years through various studies at the Knight Alzheimer's Disease Research Center at Washington University in St. Louis. The dataset includes raw MRI and PET (Positron Emission Tomography) scans, as well as clinical and cognitive assessments, structured into more than 2,800 MRI sessions (T1w, FLAIR, ASL, DTI, among others), and more than 2,100 PET sessions with different tracers (PIB, AV45, FDG, and Tau). Additionally, it features a preprocessed section using FreeSurfer, ideal for researchers who do not wish to manually clean raw MRI data. Thanks to this tool, scientists and engineers can develop highly accurate artificial intelligence models [19].

Table 5. Dataset Summary

Aspect	Description
Cohort	1,378 (755 CN, 622 CI); 42–95 yrs
Data	Random IDs; normalized dates
MRI	2,842 (T1w, T2w, FLAIR, ASL)
PET	2,157 (PIB/AV45, FDG)
Fusion	Multi-modal (MRI+PET+clinical)

Data Preparation Access to the dataset was first requested from the corresponding data repository. Upon approval, approximately 800 GB of raw data were downloaded, comprising medical evaluations, clinical and family histories, computed tomography (CT) scans, and magnetic resonance imaging (MRI) data. The initial phase of preprocessing

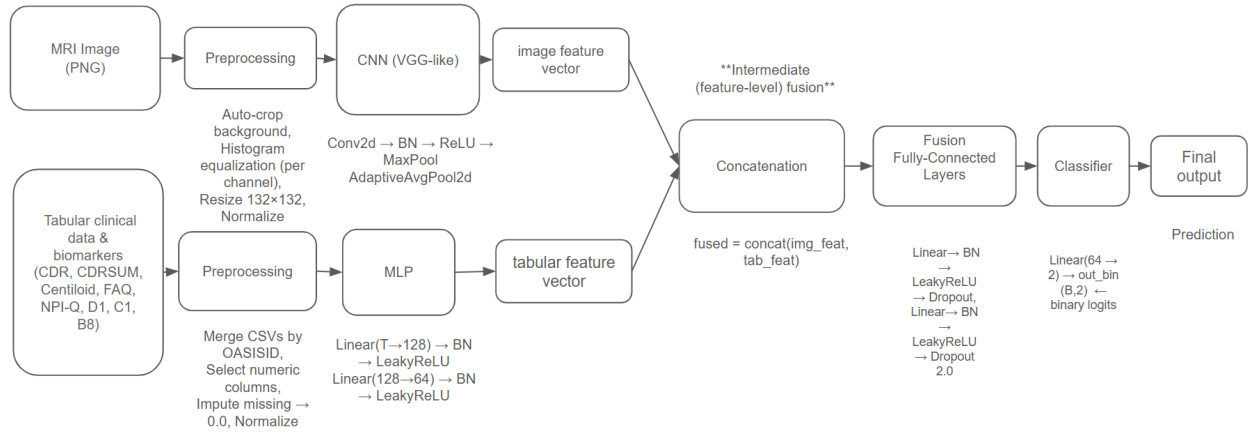

Fig. 3. Multimodal Deep Learning with Intermediate Fusion

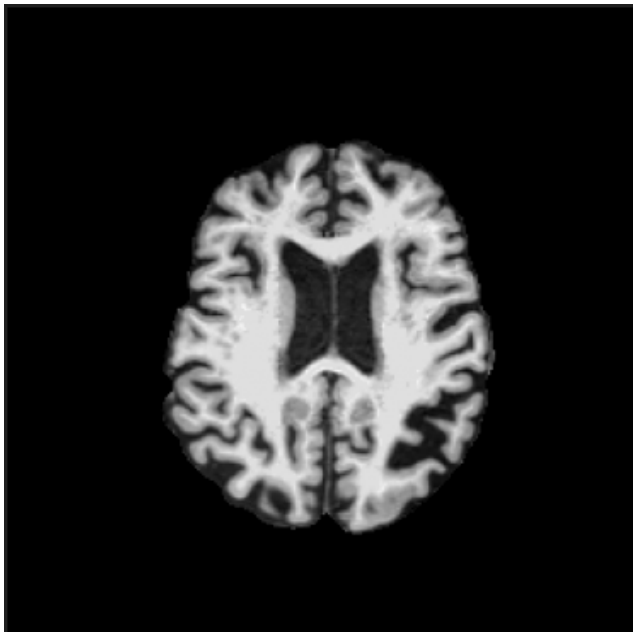

Fig. 4. MRI of a brain with AD Dementia

Fig. 5. CDR volumetric brain data

focused specifically on the MRI modality, given its relevance for structural brain analysis in Alzheimer's disease.

1. **Data Preprocessing: MRI and CDR**

   The initial stage of processing focused on cleaning the MRI images. Fortunately, the team providing the dataset included a specific section with images in a format compatible with the FreeSurfer tool, which facilitates their handling. Notably, these MRI images were already pre-segmented, meaning the brain structures were isolated from surrounding tissues (e.g., bone or non-neural organs), thus facilitating more accurate pattern recognition.

   Using this dataset, 26 axial slices of the

brain were selected, as these sections include the hippocampus, a key region for the early detection of Alzheimer's disease. Subsequently, a script was used to convert the images to grayscale and resize them to 128 × 128 pixels. For the processing of the tabular CDR data, each patient's most recent diagnosis was considered to determine the presence or absence of Alzheimer's disease. Using this information, a script was executed to match the MRI images with their corresponding CDR values, thus generating the dataset required to train the convolutional neural network. The preprocessing described above produced a dataset containing 1,248 MRI and CDR records, which were divided into 70% for training, 15% for validation, and 15% for testing. This dataset is binary and consists of two classes (Cognitively Normal and AD Dementia), representing whether the patient is healthy or has Alzheimer's disease. Figure 4 illustrates an MRI image of a patient with AD Dementia. Figure 5 presents a representation of the tabular CDR data.

2. **Preprocessing of 8 Data Modalities (MRI, CDR, Centiloid, FAQ, NPI-Q, D1, C1, and B8)**

   For this preprocessing, a second dataset was created. Similarly, MRI images were preprocessed and matched with the corresponding tabular data. For the images, cropping was applied, histogram equalization was performed, and they were resized to 132 × 132 pixels; pixel values were normalized, and simple augmentations were applied, including horizontal flipping and rotation of ±10°. For the tabular data, different CSV files were merged using OASIS-ID; missing values were filled with 0, and the tabular data were normalized to improve model stability and performance. This process ensures that the tabular data are clean and correctly linked so that the multimodal model can effectively process them. From this preprocessing, two datasets were generated. The first corresponds to a binary classification with the classes Cognitively Normal and AD Dementia; this dataset contains 127 samples per class, totaling 254 samples, with 70allocated for training, 15% for validation, and 15% for testing. The second dataset is intended for multiclass classification, including three categories: Cognitively Normal, Uncertain Dementia, and Other Dementia. This dataset contains 127 samples per class, totaling 381 samples, with 70% used for training, 15% for validation, and 15% for testing. Table 6 describes the eight modalities obtained from the OASIS-3 dataset.

Table 6. Data Modalities for Dementia Classification

Modality	Description
MRI Images	Brain structure
CDR / CDR-SB	Dementia severity scale
Centiloid	Amyloid load measure
FAQ	Daily functioning
NPI-Q	Behavioral symptoms
Clinical Diagnosis	Final diagnosis

3.3 Training Details

The hyperparameters and configurations for each of the multimodal networks were determined through multiple experiments, manually adjusting the number of epochs, filters, activation functions, and batch size. The details of the parameters used are presented in Table 7 for the early fusion architecture, Table 8 for late fusion, Table 9 for binary intermediate fusion, and Table 10 for multiclass intermediate fusion.

3.4 Evaluation Metrics

This section describes the evaluation metrics considered to measure the capabilities of classification models. The evaluation is based on the confusion matrix, which contains four key measures:

— True Positives (TP): Correctly predicted positive cases.

— True Negatives (TN): Correctly predicted negative cases.

Table 7. Parameters and Configuration for Early Fusion (CDR + MRI)

Parameter	Value	Description
Batch size	32	Batch size used for training/validation
Learning rate	1×10^{-3}	Initial learning rate for Adam
Epochs	100	Number of epochs in training
CNN (tabular)	Conv1D → Pool → FC	Extracts numerical summary from CDR data (dim = 64)
CNN (MRI)	Conv2D ×3 + MaxPool → Flatten (16384)	Extracts visual summary from MRI slice (dim = 16,384)
Early Fusion	Concatenation → FC(128) → FC(2)	Combines both summaries and produces binary logits
Output	Logits (B, 2) → Softmax → Probabilities	Final classification: Normal / AD Dementia
Loss / Opt.	CrossEntropyLoss; Adam (wd 1×10^{-4})	Training function and optimizer
Metrics	accuracy, precision, recall, specificity, AUC	Calculated per epoch

Table 8. Parameters and Configuration for Late Fusion (CDR + MRI)

Parameter	Value	Description
Batch size	32	Batch size used for training/validation
Learning rate	1×10^{-3}	Initial learning rate for Adam
Epochs	100	Number of epochs in training
CNN (tabular)	Conv1D (1→16→32) → AdaptiveAvgPool1d → Flatten → FC(32→64) → Dropout → Linear(64→C)	Generates tabular logits (CDR logits) of size num_classes.
CNN (MRI)	Conv2D (1→16→32→64) con MaxPool×3 → Flatten → FC(64*16*16→128) → Linear(128→C)	Generates image logits (MRI logits) of size num_classes.
Late Fusion	Concatenation (CDR logits + MRI logits) → Linear (2*C → C)	Combines the partial predictions from each branch and produces the final output
Output	Logits (B, C) → Softmax → Probabilities	Final classification: Cognitively Normal / AD Dementia
Loss / Opt.	CrossEntropyLoss; Adam	Training function and optimizer
Metrics	accuracy, precision, recall, specificity, AUC	Calculated per epoch

— False Positives (FP): Negative cases incorrectly predicted as positive.

— False Negatives (FN): Positive cases incorrectly predicted as negative.

1. **Accuracy**: measures the overall direction of the classifier and is defined as the proportion of correctly predicted instances relative to the total number of instances. Values range from 0 to 1, with 1 indicating perfect classification. Accuracy is expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

2. **Precision**: quantifies the model's ability to avoid false positives. It represents the proportion of true positives among all positive predictions. Important when the cost of false positives is high (e.g., spam detection where legitimate emails should not be marked as

Table 9. Parameters and Configuration for Intermediate Fusion (Binary, 8 Modalities)

Parameter	Value	Description
Batch size	32	Batch size used for training/validation
Learning rate	1×10^{-3}	Initial learning rate for Adam
Epochs	100	Number of epochs in training
CNN (MRI)	Conv2D 3→16→32→64 + AdaptiveAvg-Pool2d((1,1)) → Flatten	Extracts visual features
MLP (tabular)	Linear(→128)→BN →MLP: 128→128→64 (BatchNorm, LeakyReLU, Dropout)	Processes tabular data and produces tabular features (dimension = 64)
Intermediate Fusion	concat(img feat, tab feat) → FC(→128→64)	Fuses image and tabular representations before the binary classifier
Output	Binary output (2 multimodal classes)	Binary output for classification (Cognitively Normal/Dementia)
Loss / Opt.	CrossEntropyLoss; Adam	Training function and optimizer
Metrics	accuracy, precision, recall, specificity, AUC	Calculated per epoch

Table 10. Parameters and Configuration for Intermediate Fusion (8 Modalities, Multiclass)

Parameter	Value	Description
Batch size	32	Batch size used for training/validation
Learning rate	1×10^{-3}	Initial learning rate for Adam
Epochs	100	Number of epochs in training
CNN (MRI)	Conv2D: 3→16→32→64 + BN + AdaptiveAvg-Pool2d((1,1))	Extracts visual features
MLP (tabular)	Linear(→128)→BN →Dropout → Linear(→128)→... → salida 64	Processes tabular data and produces tabular features (dimension = 64)
Intermediate Fusion	concat(img feat, tab feat) → FC(128) → FC(64)	Fuses representations before the final classifier
Output	Output (3 multimodal classes)	Output for classification Cognitively Normal / AD Dementia
Loss / Opt.	CrossEntropyLoss; Adam	Training function and optimizer
Metrics	accuracy, precision, recall, specificity, AUC	Calculated per epoch

spam). This is expressed by:

$$Precision = \frac{TP}{TP + FP}. \quad (2)$$

3. **Recall (Sensitivity)**: Sensitivity, also known as completeness or true positive rate, measures the model's ability to identify all relevant positive cases. It is essential when failure to detect positive cases is costly (for example, when diagnosing diseases where false negatives can have serious consequences). This is calculated by:

$$Recall = \frac{TP}{TP + FN}. \quad (3)$$

4. **Specificity**: Specificity measures the model's ability to correctly identify negative cases, representing the true negative rate: It is essential when it is important to correctly identify negative cases (e.g., screening tests where false alarms must be minimized). This

is expressed as follows:

$$Specificity = \frac{TN}{TN + FP}. \quad (4)$$

5. **AUC (Area Under the ROC Curve)**: The Receiver Operating Characteristic (ROC) curve plots the true positive rate (Sensitivity) versus the false positive rate (1 - Specificity) at different classification thresholds. The AUC provides an aggregate measure of performance across all possible classification thresholds. Provides a single measure of the model's overall performance. AUC = 0.5 indicates random prediction, while AUC = 1.0 represents perfect classification Models with an AUC of 0.9 are considered excellent, AUC of 0.8 are considered good, and AUC of 0.7 are considered acceptable. The AUC is calculated as follows:

$$AUC = \int_0^1 TPR(FPR)d(FPR), \quad (5)$$

where TPR is the True Positive Rate (Sensitivity) and FPR is the False Positive Rate (1 - Specificity).

4 Results

This section presents the experiments performed. Training was carried out on a computer with an Intel Core i9-13900KF processor, with two combined RAM modules to-taling 128 GB of RAM, and an NVIDIA GeForce RTX 4080 SUPER graphics card with 16 GB of GDDR6X VRAM. The PyTorch framework was used, along with the Python OS module for working with system files, NumPy for mathematical operations, Pandas for managing tabular data, and PIL for working with images. Before selecting these final ar-chitectures, preliminary tests were performed with different configurations to identify which offered the highest accuracy. For the evaluation of the experiments, the following metrics were considered: accuracy, precision, sensitivity, specificity, and area under the curve (AUC), previously explained in Section 3.4.

4.1 Case Study 1: Early Fusion and Late Fusion on Multimodal Alzheimer's Classification

In this case study, the architectures illustrated in Figure 1 for early fusion and Figure 2 for late fusion was employed, using MRI and Clinical Dementia Rating (CDR) data as input modalities. Details on the dataset configuration are provided in Section 3.2.1. The model was designed to perform a binary classification (Cognitively Normal and AD De-mentia), predicting whether the tested subject had a diagnosis of Alzheimer's disease. To evaluate performance and robustness, 30 independent experiments were performed for each fusion strategy (early fusion and late fusion) using the training parameters specified in Tables 7 and 8. The total training time for the early fusion model was approximately 44 hours, while the late fusion model required approximately 39 hours to complete training. Tables 11 and 12 present the results of the 30 experiments conducted using the binary early fusion and late fusion architectures, respectively.

4.2 Case Study 2: Multimodal Intermediate Fusion for Multi-Class and Binary Alzheimer's Classification

In this case study, the architecture depicted in Figure 3 was employed, incorporating eight input modalities: MRI, Clinical Dementia Rating (CDR), Centiloid values, Functional Activities Questionnaire (FAQ), Neuropsychiatric Inventory Questionnaire (NPI-Q), clini-cal diagnoses, cognitive assessments, and neurological examination results. The configu-ration of the dataset used is summarized in Section 3.2.1. Two experimental setups were conducted using this architecture: 1) Binary classifica-tion with eight input modalities. 2) Multi-class classification (three output classes) with the same set of eight modalities. For both experiments, an intermediate fusion strategy was applied. The training parame-ters used for the Binary and Multi-class classification are detailed in Table 9 and 10 re-spectively. The results of the 30 independent runs for each classification task are reported in Tables 13 (binary classification) and 14 (three-class classification), respectively.

Table 11. Results of the model using early fusion

Avg. Acc	Avg. Prec	Avg. Sen	Avg. Spec	Avg. AUC
0.9912	0.9920	0.9822	0.9796	0.9984
0.9931	0.9908	0.9882	0.9763	0.9984
0.9929	0.9926	0.9951	0.9789	0.9985
0.9928	0.9935	0.9938	0.9832	0.9985
0.9921	0.9957	0.9853	0.9894	0.9985
0.9922	0.9888	0.9934	0.9695	0.9981
0.9912	0.9928	0.9943	0.9797	0.9986
0.9924	0.9950	0.9950	0.9874	0.9982
0.9924	0.9940	0.9913	0.9846	0.9985
0.9908	0.9926	0.9771	0.9791	0.9972
0.9919	0.9914	0.9878	0.9782	0.9984
0.9910	0.9919	0.9948	0.9773	0.9984
0.9922	0.9939	0.9946	0.9839	0.9984
0.9918	0.9947	0.9897	0.9867	0.9985
0.9912	0.9923	0.9973	0.9790	0.9984
0.9907	0.9941	0.9889	0.9836	0.9985
0.9901	0.9948	0.9967	0.9868	0.9985
0.9925	0.9911	0.9991	0.9735	0.9984
0.9912	0.9938	0.9946	0.9842	0.9984
0.9923	0.9928	0.9914	0.9814	0.9983
0.9934	0.9947	0.9867	0.9864	0.9987
0.9886	0.9942	0.9960	0.9820	0.9979
0.9927	0.9938	0.9875	0.9841	0.9987
0.9916	0.9915	0.9932	0.9764	0.9976
0.9930	0.9938	0.9930	0.9846	0.9983
0.9890	0.9955	0.9901	0.9890	0.9986
0.9916	0.9918	0.9845	0.9788	0.9984
0.9930	0.9889	0.9931	0.9697	0.9985
0.9909	0.9902	0.9957	0.9708	0.9985
0.9920	0.9944	0.9897	0.9863	0.9985

Table 12. Results of the model using late fusion

Avg. Acc	Avg. Prec	Avg. Sen	Avg. Spec	Avg. AUC
0.9881	0.9900	0.9941	0.9730	0.9967
0.9930	0.9959	0.9948	0.9883	0.9973
0.9934	0.9962	0.9954	0.9884	0.9974
0.9893	0.9918	0.9936	0.9782	0.9967
0.9843	0.9841	0.9949	0.9577	0.9962
0.9882	0.9915	0.9925	0.9772	0.9965
0.9742	0.9804	0.9844	0.9485	0.9925
0.9879	0.9889	0.9949	0.9700	0.9967
0.9864	0.9877	0.9937	0.9677	0.9962
0.9890	0.9904	0.9947	0.9747	0.9972
0.9912	0.9931	0.9951	0.9813	0.9974
0.9905	0.9932	0.9943	0.9810	0.9971
0.9923	0.9947	0.9954	0.9842	0.9968
0.9923	0.9953	0.9946	0.9864	0.9968
0.9901	0.9920	0.9948	0.9783	0.9965
0.9886	0.9902	0.9945	0.9735	0.9970
0.9864	0.9897	0.9911	0.9742	0.9962
0.9874	0.9910	0.9918	0.9761	0.9972
0.9905	0.9924	0.9951	0.9789	0.9971
0.9918	0.9945	0.9951	0.9835	0.9966
0.9926	0.9956	0.9944	0.9880	0.9969
0.9839	0.9839	0.9942	0.9576	0.9954
0.9930	0.9959	0.9950	0.9879	0.9972
0.9907	0.9927	0.9951	0.9795	0.9968
0.9892	0.9911	0.9943	0.9764	0.9968
0.9897	0.9921	0.9941	0.9787	0.9964
0.9886	0.9910	0.9942	0.9742	0.9960
0.9883	0.9908	0.9937	0.9746	0.9966
0.9872	0.9903	0.9928	0.9730	0.9959
0.9833	0.9832	0.9945	0.9548	0.9957

4.3 Comparative and Statistical Analysis

This section presents a comparative and statistical analysis of the performance achieved by the different architectures implementing various fusion strategies, aiming to quantitatively identify the best results and provide a solid evidence-based rationale. Three main fusion strategies are evaluated: early fusion, intermediate fusion, and late fusion.

4.3.1 Statistical Analysis Between Early Fusion and Late Fusion Using the Z-Test

A statistical comparison of the performance between early fusion and late fusion was conducted. The evaluation metric used was accuracy from the experiments presented in Tables 11 and 12. A Z-test was applied to determine significant differences between the two architectures, testing the hypothesis that early fusion outperforms late fusion.

Null hypothesis (H_0): There is no significant difference, or the performance of early fusion is inferior or equal to that of late fusion:

$$H_0 : \mu_{early} \leq \mu_{late}.$$

Alternative hypothesis (H_a): Early fusion is significantly superior to late fusion in terms of performance:

$$H_a : \mu_{early} > \mu_{late}.$$

A Z-test for two independent samples was applied, and the results are summarized in Table 15. A Z-statistic of 2.43738 and a p-value of 0.01479 were obtained. Since the p-value (0.01479) is less than the significance level $\alpha = 0.05$, the null hypothesis (H_0) is rejected. This indicates that early fusion is statistically significant and superior to late fusion in terms of accuracy.

Table 13. Results of the model using intermediate fusion for binary classification

Avg. Acc	Avg. Prec	Avg. Sen	Avg. Spec	Avg. AUC
0.9406	0.9430	0.9396	0.9417	0.9726
0.9331	0.9378	0.9263	0.9399	0.9686
0.9312	0.9386	0.9225	0.9399	0.9661
0.9303	0.9346	0.9320	0.9287	0.9669
0.9350	0.9398	0.9287	0.9412	0.9693
0.9360	0.9415	0.9255	0.9465	0.9683
0.9370	0.9446	0.9286	0.9455	0.9695
0.9353	0.9427	0.9267	0.9439	0.9660
0.9304	0.9344	0.9278	0.9331	0.9636
0.9375	0.9447	0.9276	0.9474	0.9705
0.9357	0.9376	0.9350	0.9365	0.9665
0.9353	0.9383	0.9310	0.9396	0.9674
0.9379	0.9407	0.9352	0.9406	0.9709
0.9418	0.9467	0.9356	0.9480	0.9725
0.9336	0.9364	0.9323	0.9349	0.9694
0.9334	0.9409	0.9234	0.9434	0.9693
0.9335	0.9393	0.9309	0.9361	0.9676
0.9252	0.9312	0.9143	0.9361	0.9621
0.9346	0.9410	0.9287	0.9406	0.9695
0.9305	0.9376	0.9207	0.9403	0.9652
0.9361	0.9444	0.9233	0.9490	0.9699
0.9371	0.9411	0.9299	0.9443	0.9691
0.9381	0.9429	0.9326	0.9436	0.9718
0.9388	0.9440	0.9309	0.9466	0.9734
0.9294	0.9333	0.9249	0.9339	0.9668
0.9341	0.9412	0.9245	0.9436	0.9673
0.9291	0.9360	0.9177	0.9405	0.9670
0.9346	0.9386	0.9294	0.9398	0.9689
0.9292	0.9349	0.9149	0.9435	0.9641
0.9359	0.9427	0.9210	0.9508	0.9690

Table 14. Results of the model using intermediate fusion for three-class classification

Avg. Acc	Avg. Prec	Avg. Sen	Avg. Spec	Avg. AUC
0.9076	0.9151	0.9033	0.9118	0.9575
0.9085	0.9225	0.8877	0.9293	0.9599
0.9078	0.9173	0.8970	0.9187	0.9595
0.9076	0.9193	0.8994	0.9158	0.9617
0.9056	0.9183	0.8883	0.9228	0.9567
0.9079	0.9201	0.8909	0.9250	0.9608
0.9089	0.9219	0.8912	0.9266	0.9595
0.9060	0.9190	0.8957	0.9163	0.9582
0.9043	0.9153	0.8864	0.9221	0.9578
0.9064	0.9164	0.8948	0.9180	0.9604
0.9100	0.9218	0.8944	0.9255	0.9611
0.9033	0.9168	0.8886	0.9180	0.9584
0.9041	0.9177	0.8863	0.9220	0.9587
0.9096	0.9220	0.8913	0.9280	0.9593
0.9043	0.9170	0.8883	0.9202	0.9549
0.9015	0.9123	0.8948	0.9083	0.9538
0.9033	0.9182	0.8817	0.9248	0.9555
0.9070	0.9195	0.8923	0.9217	0.9565
0.9044	0.9192	0.8850	0.9239	0.9571
0.9070	0.9179	0.8938	0.9203	0.9563
0.9041	0.9131	0.8949	0.9134	0.9586
0.9064	0.9214	0.8841	0.9287	0.9591
0.9068	0.9185	0.8922	0.9213	0.9599
0.9075	0.9165	0.8984	0.9166	0.9599
0.9087	0.9216	0.8896	0.9279	0.9591
0.9102	0.9212	0.8974	0.9231	0.9599
0.9135	0.9257	0.8932	0.9337	0.9634
0.9038	0.9143	0.8912	0.9165	0.9577
0.9094	0.9224	0.8904	0.9284	0.9610
0.9031	0.9124	0.8934	0.9128	0.9554

4.3.2 Statistical Analysis Between Early Fusion and Intermediate Fusion Using the Z-Test

This section evaluates whether the early fusion architecture provides significantly better performance than intermediate fusion. A one-tailed Z-test is used, with accuracy as the evaluation metric.

For the statistical test, the mean accuracy values reported in Tables 11 and 13 are considered, comparing binary classification results between early fusion and intermediate fusion. The details of the Z-test are described below.

Null hypothesis (H_0): There is no significant difference, or the performance of early fusion is inferior or equal to that of intermediate fusion:

$$H_0 : \mu_{early} \leq \mu_{intermediate}.$$

Alternative hypothesis (H_a): Early fusion is significantly superior to intermediate fusion in terms of performance:

$$H_a : \mu_{early} > \mu_{intermediate}.$$

The results of the Z-test are presented in Table 16, with a Z-statistic of 69.10353 and a p-value of 0.00000. Since the p-value (0.00000) is much smaller than $\alpha = 0.05$, the null hypothesis (H_0) is rejected. This indicates that early fusion is statistically significant and superior to intermediate fusion.

4.3.3 Comparative Analysis of All Models

Table 17 presents the average results obtained from Tables 11 to 14. The models com-pared include: a unimodal binary classification architecture with CDR input data; a uni-modal binary classification architecture with MRI input data; a multimodal early fusion architecture for binary classification using CDR and MRI; a multimodal late fusion archi-tecture for binary classification using CDR and MRI; a multimodal intermediate fusion architecture for binary classification with eight input modalities (MRI + CDR + Centiloid +

Table 15. Statistical analysis using Early Fusion and Late Fusion

Metric	Early Fusion	Late Fusion
Mean	0.99291	0.99099
Standard Deviation	0.00181	0.00393
Z-score	2.43738	

Table 16. Statistical analysis using Early Fusion and Intermediate Fusion

Metric	Early Fusion	Intermediate Fusion
Mean	0.99291	0.93968
Standard Deviation	0.00181	0.00381
Z-score	69.10353	

Table 17. Comparative Summary of the Performance of the Implemented Models

Model	Acc	Prec	Sen	Spe	AUC
Unimodal (CDR)	0.9930	0.9948	0.9959	0.9852	0.9988
Unimodal (MRI)	0.7744	0.6582	0.3761	0.9259	0.6510
Early Fusion	0.9918	0.9929	0.9916	0.9811	0.9984
Late Fusion	0.9887	0.9910	0.9939	0.9755	0.9965
Inter. Fusion (2C)	0.9343	0.9397	0.9274	0.9413	0.9683
Inter. Fusion (3C)	0.9066	0.9185	0.8919	0.9214	0.9586

FAQ + NPI-Q + D1 + C1 + B8); and a multimodal intermediate fusion architecture for mul-ticlass classification with eight input modalities (MRI + CDR + Centiloid + FAQ + NPI-Q + D1 + C1 + B8).

5 Discussion of Results

In this study, six classification architectures were compared using different data modalities and fusion strategies. The evaluated models included unimodal architectures (CDR and MRI) and multimodal architectures with early, late, and intermediate fusion, for both binary and multiclass classification, using up to eight input modalities (MRI, CDR, Centiloid, FAQ, NPI-Q, D1, C1, and B8). The average results for the metrics (accuracy, precision, sensitivity, specificity, and AUC) are presented in Tables 11 to 14 and were statistically analyzed using z-tests to compare the performance of the different fusion strategies. The main findings are discussed below:

1. Differences in performance between early and late fusion for binary classification: Early fusion (CDR+MRI) demonstrated slightly better performance than late fusion across all evaluated metrics, with an average accuracy of 0.9918 versus 0.9887, precision 0.9929 versus 0.9910, sensitivity 0.9916 versus 0.9939, specificity 0.9811 versus 0.9755, and AUC 0.9984 versus 0.9965.

Although these differences are modest, they consistently indicate improved overall predictive ability. Statistical analysis using a z-test confirmed the superiority of early fusion ($z = 2.43738$, $p = 0.01479 < 0.05$), providing evidence that integrating all modalities from the beginning of the model is statistically advantageous. This approach allows the network to learn joint relationships between features, fully exploiting the combined information to recognize complex patterns across modalities. In contrast, late fusion offers the benefit of allowing each submodel to specialize in its specific modality, which reduces interference between heterogeneous data types and simplifies training when modalities vary significantly. Notably, highly informative modalities such as CDR benefit most from early fusion, whereas modalities with different scales or higher noise levels may be better managed through late fusion, where specialization mitigates interference effects.

2. Impact of intermediate fusion on binary vs. multiclass tasks for binary classification, intermediate fusion with eight modalities achieved an average accuracy of 0.9343, lower than early (0.9918) and late fusion (0.9887), but higher than unimodal MRI (0.7744). In multiclass classification, performance decreased further (accuracy 0.9066), reflecting the greater task complexity and class distribution. Despite this, intermediate fusion offers the advantage of capturing complex interactions between modalities at intermediate layers, which is particularly useful when data contain nonlinear relationships or dependencies that cannot be directly exploited by early or late fusion. This gradual integration approach is especially relevant for large multimodal datasets, as it facilitates

the representation of intricate patterns across multiple modalities.

3. Impact of early vs. intermediate fusion in binary tasks The z-test comparing early and intermediate fusion for binary classification yielded $p = 0.0000$ (< 0.005), indicating that early fusion is significantly superior. This emphasizes that, for binary tasks, full integration of modalities from the start maximizes predictive capability compared to gradual feature combination.

4. Comparison of all models Unimodal models showed uneven performance, with CDR alone achieving excellent results (accuracy 0.9930) while MRI alone was limited (accuracy 0.7744), highlighting that not all modalities provide equal discriminative power. Multimodal models improved overall performance, with early fusion proving the most effective strategy for binary classification. Incorporating multiple modalities (eight inputs) using intermediate fusion improved performance compared to MRI alone but did not surpass the effectiveness of early fusion with CDR+MRI, indicating that increasing the number of modalities alone is insufficient and that the integration strategy is critical. Multimodal approaches enhance model robustness by integrating complementary information and capturing complex inter-modality relationships, whereas unimodal approaches are simpler and require less data and computation but are limited if the modality lacks sufficient discriminative power. Early fusion is ideal when modalities are complementary and full integration is desired, late fusion is advantageous when modalities are heterogeneous and submodel specialization is needed, and intermediate fusion is particularly useful for exploring complex interactions, especially in multiclass tasks or when using multiple modalities.

The results demonstrate that both the fusion strategy and the quality of modalities significantly influence performance. For binary classification with highly informative modalities such as CDR and MRI, early fusion is the most effective. Intermediate fusion emerges as a promising alternative for modeling complex interactions among multiple modalities, particularly in multiclass tasks.

6 Conclusions

This study presents and evaluates three multimodal deep learning architectures including early fusion, late fusion, and intermediate fusion for Alzheimer's disease (AD) classification. The experimental results demonstrated that early fusion consistently achieved the highest predictive performance in binary classification tasks, outperforming both late and intermediate fusion approaches in terms of accuracy, precision, and AUC, as confirmed by statistical analysis. Intermediate fusion, while beneficial for capturing complex inter-modal interactions in multiclass tasks, showed lower performance for binary classification, likely due to the challenges of integrating multiple modalities, some with incomplete data. Overall, multimodal models outperformed unimodal approaches, highlighting the importance of incorporating complementary data sources to improve predictive robustness.

Moreover, multiclass classification proved more challenging than binary classification, partly due to limited sample sizes for other dementia subtypes, emphasizing the need for larger and more balanced datasets.

The study underscores the importance of multimodality in enhancing model robustness and predictive power, particularly for early detection of AD. Nonetheless, challenges remain in handling missing data, optimizing architectures, and improving model interpretability. For future work, several directions are proposed: exploring advanced fusion strategies such as attention-based or adaptive fusion to dynamically weight modality contributions; expanding datasets by including more patients, additional modalities, or data from other repositories; evaluating models in real clinical settings with feedback from medical experts; developing user-friendly interfaces for clinical data input and model interaction; and extending the methodology to other neurodegenerative diseases, cerebrovascular conditions, or brain tumors.

Additionally, incorporating interpretability techniques and strategies for handling missing modalities will be essential to enhance clinical applicability and reliability. Ultimately, a robust multimodal deep learning system capable of analyzing MRI

and complementary clinical data could become a valuable tool for rapid and accurate identification of neurological disorders, potentially improving patient outcomes and supporting early intervention.

Acknowledgments We thank the TECNM/Tijuana Institute of Technology and SECIHTI for financial support through grant CF-2023-I-555.

References

1. **Agostinho, L., Ricardo, N., Pereira, M., Hiolle, A., Pinto, A. (2022).** A practical survey on visual odometry for autonomous driving in challenging scenarios and conditions. IEEE Access, Vol. 10, pp. 72182–72205. DOI: 10.1109/ACCESS.2022.3188990.
2. **Alamro, H., Thafar, M., Albaradei, S., Alzahrani, A. (2023).** Exploiting machine learning models to identify novel Alzheimer's disease biomarkers and potential targets. Scientific Reports, Vol. 13, pp. 4979. DOI: 10.1038/s41598-023-30904-5.
3. **Alzheimer's Disease Neuroimaging Initiative (2025).** Adni. https://adni.loni.usc.edu/. Accessed: 28 June 2025.
4. **Anand, R., Priyan, T., Brahmam, M., Balusamy, B., Benedetto, F. (2024).** Imnmagn: Integrative multimodal approach for enhanced detection of neurodegenerative diseases using fusion of multidomain analysis with graph networks. IEEE Access, Vol. 12, pp. 73095–73112. DOI: 10.1109/ACCESS.2024.3403860.
5. **Ansi, R., Gowtham, N., Ramachandran, S., Praneeth, S. (2023).** Revolutionizing Alzheimer's disease prediction using inceptionv3 in deep learning. 2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, Coimbatore, India, pp. 1155–1160. DOI: 10.1109/ICECA58529.2023.10395534.
6. **Baltrušaitis, T., Ahuja, C., Morency, L.-P. (2019).** Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 41, No. 2, pp. 423–443. DOI: 10.1109/TPAMI.2018.2798607.
7. **Boulahia, S., Amamra, A., Madi, M., Dornaika, F. (2021).** Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Machine Vision and Applications, Vol. 32, pp. 121. DOI: 10.1007/s00138-021-01249-8.
8. **Chattopadhyay, T., Ozarkar, S. S., Buwa, K., Thomopoulos, S. I., Thompson, P. M., Alzheimer's Disease Neuroimaging Initiative (2023).** Predicting brain amyloid positivity from t1 weighted brain mri and mri-derived gray matter, white matter and csf maps using transfer learning on 3d cnns. bioRxiv. DOI: 10.1101/2023.02.15.528705. Preprint.
9. **Cheng, J., Huang, C., Zhang, J., Wu, B., Zhang, W., Liu, X., Zhang, J., Tang, Y., Zhou, H., Zhang, Q., Gu, M., Dong, J., Zhang, X. (2024).** Multimodal deep learning using on-chip diffractive optics with in situ training capability. Nature Communications, Vol. 15, pp. 6189. DOI: 10.1038/s41467-024-50677-3.
10. **Cheng, X., Zhang, L., Zheng, Y. (2016).** Deep similarity learning for multimodal medical images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Vol. 6, No. 3, pp. 248–252. DOI: 10.1080/21681163.2015.1135299.
11. **Dhakal, S., Azam, S., Hasib, K., Karim, A., Jonkman, M., Al Haque, A. (2023).** Dementia prediction using machine learning. Procedia Computer Science, Vol. 219, pp. 1297–1308. DOI: 10.1016/j.procs.2023.01.414.
12. **Dietz, S., Altstidl, T., Zanca, D., Eskofier, B., Nguyen, A. (2024).** How intermodal interaction affects the performance of deep multimodal fusion for mixed-type time series. arXiv. DOI: 10.48550/arXiv.2406.15098.
13. **Diogo, V. S., Ferreira, H. A., Prata, D., Alzheimer's Disease Neuroimaging Initiative (2022).** Early diagnosis of Alzheimer's disease using machine learning: A multi-diagnostic, generalizable approach.

Alzheimer's Research & Therapy, Vol. 14, pp. 107. DOI: 10.1186/s13195-022-01047-y.

14. **Franciotti, R., Nardini, D., Russo, M., Onofrj, M., Sensi, S. (2023).** Comparison of machine learning-based approaches to predict the conversion to Alzheimer's disease from mild cognitive impairment. Neuroscience, Vol. 514, pp. 143–152. DOI: 10.1016/j.neuroscience.2023.01.029.

15. **Gadzicki, K., Khamsehashari, R., Zetzsche, C. (2020).** Early vs. late fusion in multimodal convolutional neural networks. 2020 IEEE 23rd International Conference on Information Fusion (FUSION), IEEE, Rustenburg, South Africa, pp. 1–6. DOI: 10.23919/FUSION45008.2020.9190246.

16. **Gonzalez, C. I. (2025).** Designing optimal cnns architectures using metaheuristic algorithms applied to the classification of Alzheimer's disease. Computación y Sistemas, Vol. 29, No. 1, pp. 179–189. DOI: 10.13053/CyS-29-1-5512.

17. **Guarrasi, V., Aksu, F., Caruso, C., Di Feola, F., Rofena, A., Ruffini, F., Soda, P. (2025).** A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. Image and Vision Computing, Vol. 158, pp. 105509. DOI: 10.1016/j.imavis.2025.105509.

18. **Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B. (2020).** More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Transactions on Geoscience and Remote Sensing, Vol. 59, No. 5, pp. 4340–4354. DOI: 10.1109/TGRS.2020.3016820.

19. **LaMontagne, P., Benzinger, T., Morris, J., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A., Raichle, M., Cruchaga, C., Marcus, D. (2019).** Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. medRxiv. DOI: 10.1101/2019.12.13.1901490.

20. **Lee, Y. C., Cha, J., Shim, I., Park, W.-Y., Kang, S. W., Lim, D. H., Won, H.-H. (2023).** Multimodal deep learning of fundus abnormalities and traditional risk factors for cardiovascular risk prediction. NPJ Digital Medicine, Vol. 6, pp. 14. DOI: 10.1038/s41746-023-00748-4.

21. **Li, J., Xu, H., Yu, H., Jiang, Z., Zhu, L. (2022).** Multi-modal feature selection with anchor graph for Alzheimer's disease. Frontiers in Neuroscience, Vol. 16, pp. 1036244. DOI: 10.3389/fnins.2022.1036244.

22. **Marinescu, R., Lorenzi, M., Blumberg, S., Young, A., Planell-Morell, P., Oxtoby, N., Eshaghi, A., Yong, K., Crutch, S., Golland, P., Alexander, D. (2019).** Disease knowledge transfer across neurodegenerative diseases. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019, Springer, Cham, Switzerland, Vol. 11765, pp. 860–868. DOI: 10.1007/978-3-030-32245-8$_{95}$.

23. **Martí-Juan, G., Lorenzi, M., Piella, G. (2023).** MC-RVAE: Multi-channel recurrent variational autoencoder for multimodal Alzheimer's disease progression modelling. NeuroImage, Vol. 268, pp. 119892. DOI: 10.1016/j.neuroimage.2023.119892.

24. **Movement Disorders Society (2025).** Movement disorders society unified Parkinson's disease rating scale (UPDRS). https://www.movementdisorders.org/. Accessed: 28 June 2025.

25. **NCBI**, **.** Gene expression omnibus (geo). https://www.ncbi.nlm.nih.gov/geo/. Accessed: 31 October 2025.

26. **Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. (2011).** Multimodal deep learning. Proceedings of the 28th International Conference on Machine Learning (ICML-11), Omnipress, Madison, WI, USA.

27. **Parkinson's Progression Markers Initiative (2025).** Parkinson's progression markers initiative (PPMI). https://www.ppmi-info.org/.

28. **Rafiei, F., Zeraati, H., Abbasi, K., Ghasemi, J., Parsaeian, M., Masoudi-Nejad, A. (2023).** Deeptrasynergy: Drug combinations using multimodal deep learning with transformers. Bioinformatics. DOI: 10.1093/bioinformatics/btad438.

29. **Sarraf, S., DeSouza, D., Anderson, J., Tofighi, G., Alzheimer's Disease Neuroimaging Initiative (2017).** Deepad: Alzheimer's disease classification via deep convolutional neural networks using mri and fmri. bioRxiv. DOI: 10.1101/070441. Preprint.

30. **Shamrat, F. M. J. M., Akter, S., Azam, S., Karim, A., Ghosh, P., Tasnim, Z., Hasib, K. M., De Boer, F., Ahmed, K. (2023).** Alzheimernet: An effective deep learning-based proposition for Alzheimer's disease stages classification from functional brain changes in magnetic resonance images. IEEE Access, Vol. 11, pp. 16376–16395. DOI: 10.1109/ACCESS.2023.3244952.

31. **Thangavel, S., Selvaraj, S. (2023).** Machine learning model and cuckoo search in a modular system to identify Alzheimer's disease from mri scan images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Vol. 11, No. 6, pp. 1753–1761. DOI: 10.1080/21681163.2023.2187239.

32. **UK Biobank (2025).** Uk biobank: A major resource for health research. https://www.ukbiobank.ac.uk/. Accessed: 28 June 2025.

33. **Villazana, S., Montilla, G., Eblen, A., Maldonado, C. (2021).** Detección de señales eeg epilépticas utilizando redes convolucionales basada en la transformada synchrosqueezing acolchada. Computación y Sistemas, Vol. 25, No. 2, pp. 269–286. DOI: 10.13053/CyS-25-2-3461. Epileptic Signal Detection Using Quilted Synchrosqueezing Transform Based Convolutional Neural Networks.

34. **Volinsky-Fremond, S., Horeweg, N., et al. (2024).** Prediction of recurrence risk in endometrial cancer with multimodal deep learning. Nature Medicine, Vol. 30, pp. 1962–1973. DOI: 10.1038/s41591-024-02993-w.

35. **Wadekar, S., Chaurasia, A., Chadha, A., Culurciello, E. (2024).** The evolution of multimodal model architectures. arXiv. DOI: 10.48550/arXiv.2405.17927.

36. **Wang, J., Zhang, H., Zhong, Y., Liang, Y., Ji, R., Cang, Y. (2024).** Advanced multimodal deep learning architecture for image-text matching. 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI), IEEE, Changchun, China, pp. 1185–1191. DOI: 10.1109/ICETCI61221.2024.10594167.

37. **Washington University in St. Louis (2025).** Open access series of imaging studies (OASIS). https://sites.wustl.edu/oasisbrains/. Accessed: 28 June 2025.

38. **Yao, Z., Lin, F., Chai, S., He, W., Dai, L., Fei, X. (2024).** Integrating medical imaging and clinical reports using multimodal deep learning for advanced disease analysis. 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE), IEEE, pp. 1217–1223. DOI: 10.1109/ICSECE61636.2024.10729527.

Article received on 31/10/2025; accepted on 15/12/2025.
**Corresponding author is Claudia I. Gonzalez.*