A Multimodal Approach Integrating 3D-CNN and LSTM for Prediction of Temporal Progression of Alzheimer's Disease

Durga Supriya HL[1], Swetha Mary Thomas[2], Sowmya Kamath S.[3]
Healthcare Analytics and Language Engineering (HALE) Lab,
Dept. of Information Technology, National Institute of Technology Karnataka, Surathkal,
Srinivasnagar P.O., Mangaluru 575025 India
Email: sowmyakamath@nitk.edu.in

***Abstract*—Alzheimer's Disease (AD) poses a substantial healthcare challenge marked by cognitive decline and a lack of definitive treatments. As the global population ages, the prevalence of AD escalates, underscoring the urgency for more advanced diagnostic techniques. Current single-modality methods have limitations, emphasizing the critical need for early detection and precise diagnosis to facilitate timely interventions and the development of effective therapies. We propose a novel multimodal medical diagnostic framework for AD employing a hybrid deep learning model. This framework integrates a 3D Convolutional Neural Network (CNN) to extract detailed intra-slice features from MRI volumes and a Long Short-Term Memory (LSTM) network to capture intricate inter-sequence patterns indicative of AD progression. By leveraging longitudinal MRI data alongside biomarkers and cognitive scores, our framework significantly enhances diagnostic accuracy, particularly in the early stages of the disease. We incorporate Grad-CAM to enhance the interpretability of MRI-based diagnoses by highlighting relevant brain regions. Through the fusion of spatial, temporal features, cognitive scores, and demographic information, our approach effectively classifies patients into Alzheimer's Disease (AD) or Cognitive Normal (CN) categories. Our multimodal approach achieves a remarkable accuracy of 92.65%, outperforming state-of-the-art works by a margin of 6%.**

Keywords: **Multi-modal Diagnosis, Alzheimer's Disease, Classification, Explainable AI**

I. INTRODUCTION

Alzheimer's Disease (AD) represents an increasingly significant public health challenge worldwide as the aging population grows. The neuro-degenerative disorder severely impacts cognitive functions, resulting in memory loss, impaired reasoning, language deficits, and an eventual inability to perform daily tasks. Unfortunately, diagnosing AD is fraught with challenges. Current clinical methods heavily rely on neuropsychological assessments, patient history, and caregiver interviews, which are subjective and prone to error. Moreover, definitive confirmation typically requires post-mortem examination, making early and accurate diagnosis elusive during a patient's lifetime. The reliance on clinical symptoms alone leaves room for misdiagnosis, delays in intervention, and missed opportunities for early treatment. Despite extensive research into the underlying biological mechanisms of AD, innovative diagnostic approaches are urgently needed to overcome these hurdles.

AD diagnosis relies on various types of data, each providing unique insights into the condition. Clinical data, obtained through patient interviews and examinations, offers crucial information on symptom progression and cognitive decline. Neuroimaging techniques, such as MRI and PET scans, reveal structural and functional brain abnormalities, including atrophy and abnormal protein deposits. Cognitive assessments provide quantitative measures of cognitive function over time. However, most existing works typically focus on utilizing a single type of data like MRI or PET scans to visualize structural and functional changes in the brain associated with AD, while ignoring other data sources. Integrating these diverse data sources improves diagnostic accuracy, monitors disease progression, and identifies potential therapeutic targets. Advancements in technology and interdisciplinary approaches hold promise for enhancing Alzheimer's diagnosis and treatment strategies, ultimately improving patient outcomes and quality of life.

To address the complexities of AD diagnosis, our solution is to adopt a multi-modal approach centered on deep learning techniques and magnetic resonance imaging (MRI) data. We aim to leverage demographic features and cognitive scores alongside MRI data to enhance the analytical capabilities of our model. Recognizing the importance of longitudinal data, our approach unfolds incrementally, gradually introducing additional modalities and longitudinal data to assess the model's evolving performance. This iterative process encompasses the integration of demographic features and cognitive scores with MRI data, culminating in a comprehensive model that harnesses the strengths of all three modalities. Through this approach, we aim to develop a robust diagnostic tool capable of accurately detecting AD progression, thereby contributing to advancements in neurodegenerative disease diagnosis and management.

Another significant requirement in medical applications is the importance of explainability and interpretability. While deep learning models such as convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) models offer

promising avenues for improved diagnostic accuracy, their black-box nature raises concerns about interpretability and trustworthiness in clinical settings. To address these concerns, we incorporate Class Activation Mapping mechanisms for providing visual explanations of model decisions. This emphasis on transparency and interpretability not only enhances clinicians' understanding of the underlying features driving diagnostic predictions but also instills confidence in the model's reliability, aiding in decision-making processes.

The remainder of this article is organized into several key sections aimed at comprehensively exploring the proposed multimodal approach for predicting the temporal progression of Alzheimer's disease. In Section II, a thorough review of related works is presented, providing insights into existing methodologies and gaps in current research. Following this, the details of dataset collection and preprocessing techniques employed to ensure data quality and consistency are presented in Section III. This section also covers the process of integrating 3D-CNN and LSTM networks for capturing temporal dependencies within the data. In Section IV, the experimental results for various fusion techniques and combinations of data modalities investigated are presented, followed by concluding remarks and future work.

II. Related Work

Existing works on AD prediction and diagnosis encompass a wide range of approaches, reflecting the complexity of the disease and advancements in technology. Researchers have explored various methodologies, from traditional machine learning to advanced deep learning models, often integrating diverse data sources such as clinical records and neuro-imaging data. Emphasis has been placed on incorporating temporal dynamics and disease progression information, as well as on enhancing interpretability through techniques like explainable artificial intelligence (XAI). This diverse array of approaches aims to improve early detection, personalized treatment, and overall understanding of AD pathology.

El-Sappagh et al. [1] conducted a study comparing five machine learning algorithms to predict Alzheimer's disease progression using data from the ADNI database. They enhanced the models by incorporating time-series features such as comorbidities and medication history, aiming to capture the disease's evolving nature over time. Their optimized models demonstrated that the effectiveness of Random Forest in leveraging temporal features to accurately predict Alzheimer's disease progression. By considering dynamic factors alongside static variables, the models gain a more nuanced understanding of disease trajectories, potentially enabling earlier detection and personalized treatment strategies. Their study underscores the importance of advanced machine learning techniques and diverse data integration in Alzheimer's prediction. Moscoso et al. [2] investigated the utility of MRI volumetric measures for early detection of AD in Mild Cognitive Impairment (MCI) patients. They found that stable MCI patients often progress to AD, challenging the adequacy of non-disease training examples. By combining MRI data from the hippocampus and entorhinal cortex, they enhanced prediction accuracy over time. Their study highlights the importance of refining diagnostic approaches for detecting AD in its early stages, especially considering the dynamic nature of disease progression. The findings underscore the significance of neuroanatomical markers associated with AD pathology in improving predictive models.

Zhang et al. [3] proposed a novel multimodal deep learning framework for mental disorder detection, integrating facial expressions, gestures, and verbal content to improve accuracy. Their approach effectively identified bipolar disorder and depression, suggesting potential for generalization across various mental health conditions. The integration of facial expressions, gestures, and verbal content allows for a holistic assessment, capturing subtle nuances indicative of mental health issues. The deep learning architecture enables automatic feature extraction from complex multimodal data, helping in early detection and intervention.

Khagi et al. [4] conducted a study on deep learning techniques for diagnosing AD using imaging modalities, specifically exploring 3D CNN architectures. They introduced divNet, a novel approach designed to improve classification performance by modifying reception areas. They evaluated divNet's effectiveness in terms of memory usage, parameter count, and classification accuracy. Tu et al. [5] introduced a multimodal model tailored for AD diagnosis that integrates geometric algebra-based feature extension and influence degree-based filtration with an Artificial Neural Network (ANN). By combining these diverse techniques, their model aims to capture intricate patterns and relationships within multimodal data, providing a more comprehensive understanding of Alzheimer's disease pathology. They reported high accuracy in diagnosing both AD and MCI, surpassing the performance of existing state-of-the-art methods.

Viswan et al. [6] emphasized the importance of explainability in Artificial Intelligence (AI), particularly within deep neural networks. They investigated visual explanation methods for convolutional neural networks (CNNs) and evaluated techniques such as Grad-CAM and Ablation-CAM. Through various techniques such as LIME for textual data, SHAP for numeric explanations, and rule-based systems, researchers have gained valuable insights. Textual analysis using XAI methods has revealed distinctive linguistic patterns in medical transcripts, aiding in the classification of dementia patients. Numeric explanations provided by SHAP have quantified predictor importance, enhancing understanding of model decision-making and predictor stability. Rule-based explanations generated by XAI methods have simplified decision rules, facilitating early identification and diagnosis of cognitive decline. Visual explanations, particularly heatmaps generated by LRP, GradCAM, and SHAP, have helped clinicians pinpoint critical features in medical images, such as MRI scans, leading to improved diagnosis and treatment planning.

Avelar et al. [7] investigated the complexity of AD diagnosis and progression through a multilayer network approach,

addressing the challenge posed by its multifactorial and heterogeneous nature. Using data from 490 subjects (147 cognitively normal, 287 with mild cognitive impairment, and 56 with AD) sourced from ADNI, the study integrates various biomarkers including structural MRI, amyloid-β, PET, cerebrospinal fluid, cognition, and genetics. Employing multilayer community detection, the model accurately identifies AD cases and predicts future AD progression with approximately 90% accuracy, even in cases misdiagnosed clinically. Notably, the framework successfully distinguishes subtypes among MCI participants transitioning to AD or reverting to cognitive normalcy. The study's findings highlight the efficacy of the multilayer approach in capturing the heterogeneity of AD, offering valuable insights into the disease's progression and providing a comprehensive understanding of its multidisciplinary profiles. Shaker et al. [8] proposed a novel deep learning model for early detection of AD progression using multimodal time series data. The model integrates stacked convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) architectures to jointly predict AD multiclass progression and four cognitive scores regression tasks. Evaluation on Alzheimer's Disease Neuroimaging Initiative (ADNI) data involving 1536 subjects demonstrates state-of-the-art performance, highlighting the model's effectiveness in analyzing heterogeneous temporal data and accurately predicting patients' future status.

The extensive review of prominent research works on AD prediction and diagnosis revealed several limitations and research gaps. Firstly, while many studies have explored various methodologies and data sources, there remains a challenge in integrating disparate data modalities effectively. Additionally, despite efforts to incorporate temporal dynamics, the predictive models could not fully capture the evolving nature of the disease over time. While fusion strategies have been explored, there is still a need to investigate a wider range of fusion techniques to optimize predictive performance. Our work aims to address these limitations by leveraging 3D MRI scans, which provide detailed anatomical information crucial for early AD detection. By considering temporal features, the model can capture disease progression dynamics more accurately. Furthermore, exploring different fusion strategies and combinations of data modalities will enhance the model's ability to extract comprehensive information from heterogeneous data sources, potentially leading to more robust and accurate predictions of AD onset and progression. The focus of this work is to develop a deep learning-based framework for the enable early prediction of AD progression utilizing multimodal data comprising MRI images, demographic features, and cognitive scores. We also aim to ensure model explainability using visual attention based models to provide insights into the regions of interest within the MRI images that contribute to the classification decisions.

III. Materials and Methods

A. Dataset Specifics

The Alzheimer's Disease Neuroimaging Initiative (ADNI) [9] provides a comprehensive repository of neuroimaging, clinical, genetic, and biochemical data, which was utilized in this study. The dataset incorporates three distinct modalities to offer a multifaceted perspective on Alzheimer's disease progression. The first modality comprises demographic data, encompassing three fundamental features – *gender*, *age*, and *education*. These factors furnish essential contextual information for characterizing the study cohort. The second modality integrates cognitive scores (CS) and selected critical biomarkers, including *APOE4* (Apolipoprotein E4), *ADAS13* (Alzheimer's Disease Assessment Scale - Cognitive Subscale 13), *FAQ* (Functional Activities Questionnaire), *MMSE* (Mini-Mental State Examination), and various *RAVLT* (Rey Auditory Verbal Learning Test) scores. These biomarkers play a crucial role in assessing cognitive decline and disease progression. The third modality involves 3D MRI neuroimaging collected at five specific time steps – *baseline* (BL), *month 12* (M12), *month 24* (M24), *month 36* (M36), *month 48* (M48) visits. This imaging data provides valuable insights into structural changes in the brain over time, contributing to a more comprehensive understanding of AD.

The dataset provides data of a total of 204 patients. As seen in Fig. 1, among these, 18 patients transitioned from Cognitively Normal (CN) to Alzheimer's Disease (AD). The data collection process involved obtaining scans from five distinct visits: baseline (BL), 12 months (M12), 24 months (M24), 36 months (M36) and 48 months (M48). Additionally, due to the limited number of patients diagnosed with Alzheimer's Disease, we considered and recategorized *Mildly Cognitively Impaired* (MCI) patients as AD, to ensure a well-defined cohort for analysis.

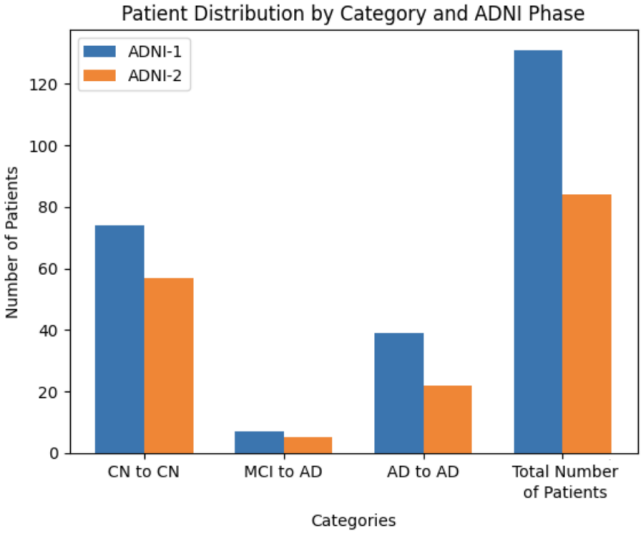

Fig. 1: Patient Distribution

We used 3T T1-weighted anatomical sequences which were recorded using the volumetric 3D MPRAGE protocol with a Sagittal acquisition type, ensuring high-quality imaging. The MRI scans had a voxel resolution of $1 \times 1 \times 1$ mm, providing detailed spatial information. Each MRI scan consisted of sagittal axis slices which could be downloaded as individual

DICOM files. To streamline the data and make it compatible for further analysis, a crucial step involved the conversion of these DICOM images into a consolidated format. Specifically, all the sagittal axis slices of a single MRI scan were transformed into a single nii.gz (Neuroimaging Informatics Technology Initiative - compressed) file.

B. Data Preprocessing

A preprocessing pipeline was designed to eliminate extraneous information, correct intensity variations, align volumes to a common anatomical space, focus analyses on relevant brain structures, and streamline the dataset for more efficient processing. Each image was subjected to a set of uniform analyses, ensuring consistent and accurate comparisons across diverse MRI scans in the development of a multimodal AD diagnosis system. The preprocessing pipeline was automated to process the input images, through a series of python scripts.

Firstly, a process of reorientation to a standard space was performed, for adjusting image orientation or flipping left and right. These operations are integral to the foundational steps of preprocessing, ensuring consistent voxel processing and interpretation across diverse software and systems. In our specific case, we employed FSLEyes [10] to visualize the data, revealing instances where images had undergone a 180° rotation. To rectify this, we utilized the fslreorient2std tool [10] from FSL to standardize the orientation of these images before proceeding with further processing (Fig. 2). Next, bias field correction (BFC) was performed to address brightness issues in MRI scans resulting from various factors, such as the patient's positioning in the MRI scanner or the scanner's version. These brightness issues manifest as low-frequency, smoothed, and undesirable signals within the scanner, impacting the overall image quality. If left uncorrected, these nonhomogeneities can adversely affect subsequent processing steps, including skull stripping. We employed the N4 BFC algorithm [11] from the advanced normalization tools (ANTs). This utilizes an improved B-spline fitting routine allowing for multiple resolutions in the correction process, thereby enhancing the effectiveness of bias field correction in optimizing image quality (Fig. 3).

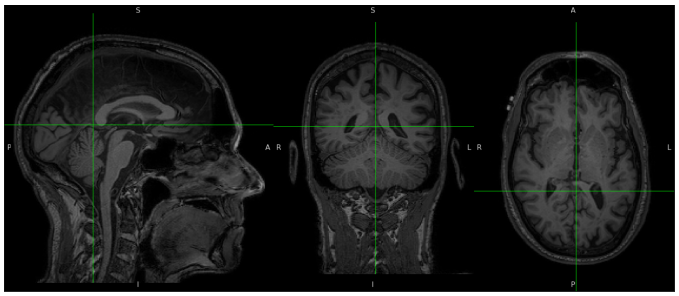

Fig. 2: MRI after reorientation

Another critical step is the MNI152 standard template registration, used for aligning images based on brain structures, facilitating the comparison of different MRI scans. In our study, all the MRI scans were aligned to be compatible with the

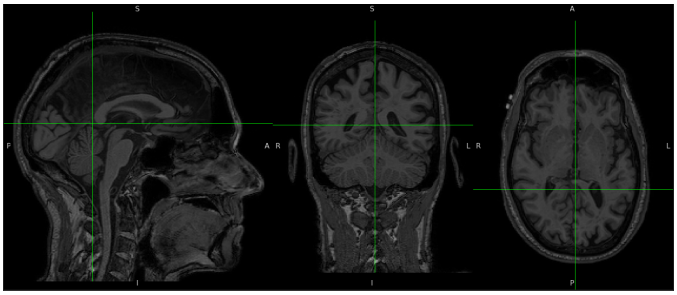

Fig. 3: MRI after bias field correction

MNI152_T1_1mm.nii.gz template. The affine transformation performed ensures alignment without deforming the images, involving fundamental transformation steps like rescaling, rotation, translation, and shearing. The FLIRT tool in FSL was utilized for the registration, with correlation ratios serving as the similarity metric. This approach successfully standardized all MRIs to the MNI152 template space [12] as shown in Fig. 5, ensuring uniformity and compatibility by bringing all images to the same dimensions (182, 218, 182). As a final step, skull stripping was performed, to isolate the brain tissue from non-brain tissue in MRI images. This process is essential for eliminating extraneous information, such as residual neck voxels, which can act as noise and contribute to the high dimensionality of training data, complicating the subsequent classification task. We employed the Brain Extraction Tool (BET2) [13] from FSL to perform skull stripping as shown in Fig. 6. This tool is effective in accurately calculating a brain mask, allowing for the removal of remaining body parts and noise from the brain region.

C. Spatiotemporal Feature Extraction from MRI scans

Following the loading of the nii.gz files using the nibabel library [14] and their conversion into numpy arrays, a crucial step in the data preparation process involved reshaping the image data for each patient into 5D tensors. This transformation is essential to make the data suitable for input into the machine learning model. The tensors were structured with dimensions (5, 1, 182, 218, 182), aligning precisely with the expected input shape for the model. This reshaping process ensured that the spatial information from the neuroimaging data was appropriately organized for subsequent analysis. Additionally, considering the categorical nature of the labels associated with the MRI scans, where CN is represented as 0 and AD as 1, the data was appropriately labeled to facilitate supervised learning.

The 3D CNN model designed for the classification task of Alzheimer's Disease utilizes a hierarchical architecture tailored for processing three-dimensional data. The model comprises four convolutional blocks, each contributing to the extraction of increasingly complex features from the input MRI data. The initial convolutional block incorporates two convolutional layers with ReLU activation functions, followed by 3D max-pooling and dropout for regularization. This block is crucial for capturing low-level features. Subsequent convolutional blocks, each comprising two convolutional layers,

build upon this foundation to extract hierarchical features. The progressive increase in the number of channels in these layers enables the model to discern more intricate representations. To enhance flexibility and prevent overfitting, dropout layers are strategically placed after each max-pooling operation. This is especially important in medical imaging tasks where data may be limited, necessitating robust feature learning. The final convolutional block is succeeded by a flatten layer, transforming the output of the convolutional layers into a one-dimensional tensor. The 3D CNN processes the 5 visits' MRI images and extracts spatial features which are then fed into the LSTM layer as shown in Fig. 4. This sequential integration of 3D CNN and LSTM allows the model to leverage both spatial and temporal information from the MRI volumes and cross-sectional biomarkers, enhancing the overall diagnostic accuracy and stability, particularly in the early stages of AD.

TABLE I: Cognitive Scores and Demographic information

Sl.no	Variable	Description
1	ADAS-13	13-item AD assessment scale
2	FDG	Fluorodeoxyglucose
3	TAU	A Protein
4	PTAU	A Protein
5	CDRSB	Clinical Dementia Rating
6	MMSE	Mini-Mental State Examination
7	RAVLT Immediate	Rey Auditory Verbal Learning Test
8	RAVLT Learning	Rey Auditory Verbal Learning Test
9	RAVLT Forgetting	Rey Auditory Verbal Learning Test
10	RAVLT % Forgetting	Rey Auditory Verbal Learning Test
11	FAQ	Functional Activities Questionnaire
12	MOCA	Montreal Cognitive Assessment
13	Hippocampus	Hippocampus Volume
14	AGE	Age
15	PTGENDER	Gender
16	PTEDUCAT	Education
17	APOE4	No. of £4 alleles of APOE

D. Cognitive, Demographic and Hippocampus data:

In this work, the diverse information available in the ADNI dataset through three distinct modalities was harnessed. The primary modality involves 3D MRI neuroimaging, allowing us to delve into detailed neural imaging for comprehensive analysis as described before. The second modality encompasses demographic data, featuring key attributes such as gender, age, and education. This demographic information enriches our understanding of the patient profile, contributing valuable context to the diagnostic process. The third modality incorporates a Cognitive Score (CS) along with selected critical biomarkers. This includes 13 features, such as APOE4, ADAS13, FAQ, MMSE, and 4 RAVLT scores which are summarized in Table I. Leveraging this diverse set of biomarkers allows for a more holistic assessment of AD, considering both genetic predispositions and cognitive performance.

We conducted a meticulous preprocessing of the ADNI dataset to extract pertinent information aligned with our chosen modalities. Focusing specifically on patients corresponding to the MRI image data, we curated a dataset that encompassed demographic features (gender, age, and education) and a cognitive score along with critical biomarkers. To ensure uniformity in our dataset, we transformed categorical data into numerical categories. Additionally, we performed normalization on selected columns to address variations in data ranges. This preprocessing step was crucial for optimizing the performance of our subsequent machine learning model. The curated dataset was then divided into training and test sets, with an 80-20% split, to facilitate robust model training and evaluation. To seamlessly integrate our data into the PyTorch framework, we converted them into PyTorch tensors, ensuring compatibility with the neural network architecture. We designed a 2-layered fully connected neural network with 32 and 16 neurons, respectively. This configuration was chosen to strike a balance between model complexity and computational efficiency. Subsequently, the tensor vectors were passed through the neural network, leveraging its capacity to capture intricate patterns within the data.

E. Information Fusion:

In the final stage of the multimodal AD diagnosis framework, the output of the last layer of the Deep Neural Network

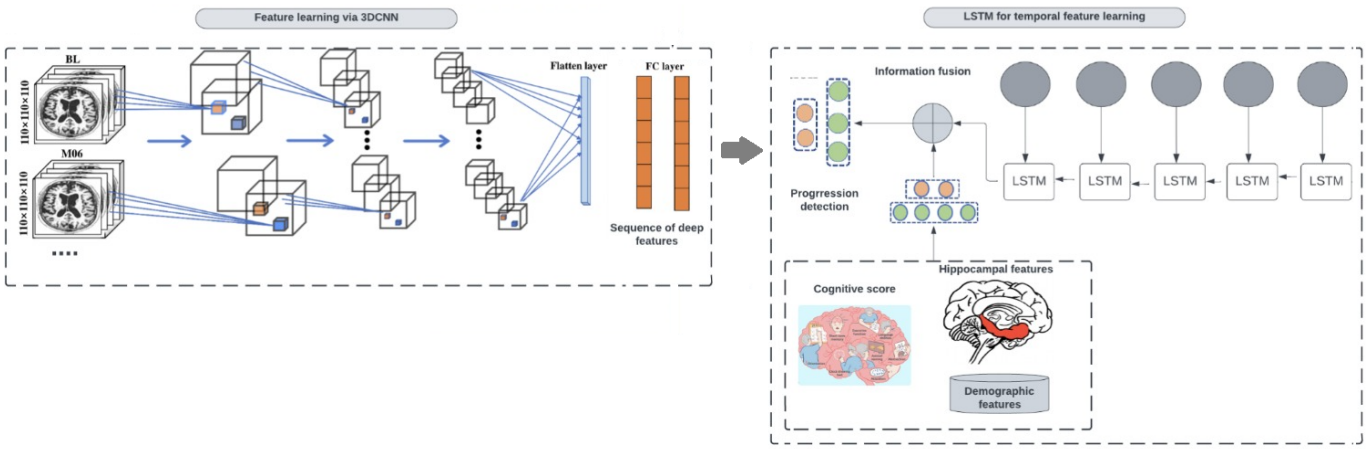

Fig. 4: Process of extracting spatiotemporal features

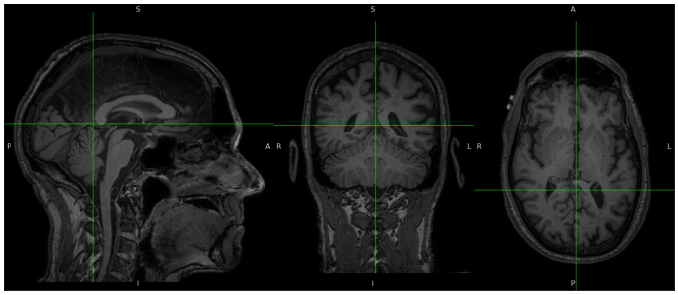

Fig. 5: MRI after MNI152 registration

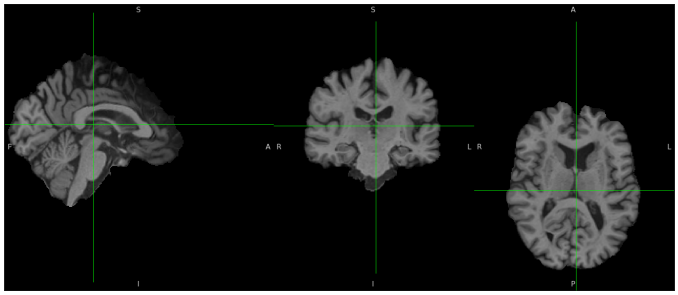

Fig. 6: MRI after skull stripping

(DNN), consisting of 16 feature vectors, is concatenated with the output of the LSTM. This concatenated information is then processed by a subsequent network composed of two dense layers with 128 and 64 hidden units respectively, and 1 dense output later with 2 units. The first two dense layers contribute to hierarchical feature learning and representation, while the last layer serves as the output layer for the classification task. The final layer outputs probabilities between 0 and 1, describing the likelihood of a patient progressing to Alzheimer's Disease or remaining in a Cognitive Normal state.

F. Model Explainability:

Class Acitvation Mapping (CAM) is a technique used for explaining the decisions made by deep learning models, in image classification tasks. In the context of AD diagnosis using MRI scans, Grad-CAM [15] helps to highlight the regions of the brain that the model considers important for making its prediction. The process begins with a pre-trained deep learning model that has been trained to classify MRI scans as either healthy or Alzheimer-affected. Grad-CAM specifically focuses on the final convolutional layer of this model. This layer captures high-level features that are crucial for making the classification decision, making it a suitable starting point for analysis. It analyzes the gradients flowing into the final convolutional layer. These gradients represent the rate of change of the loss function concerning the model's parameters.

Specifically, Grad-CAM computes the gradient of the predicted class score concerning the feature maps of the last convolutional layer. By analyzing these gradients, it determines how much each pixel in the feature maps influences the final prediction. Next, it averages the gradients across the channels within each neuron in the last convolutional layer. This results in feature importance maps that highlight which parts of the feature maps are most relevant for the classification task. These importance maps are initially low-resolution, so Grad-CAM up-samples them to match the dimensions of the input MRI scan. For a specific class, such as AD, Grad-CAM calculates importance scores for each feature map. These scores reflect how much each feature map contributes to the prediction of that class. By combining these importance scores with the feature map activations, Grad-CAM generates a heatmap as shown in Fig. 9. Brighter regions in the heatmap correspond to features deemed crucial for the Alzheimer's prediction, potentially indicating areas of the brain exhibiting atrophy or other disease signatures. Healthcare professionals can use this heatmap to understand which regions of the brain are influencing the model's prediction of Alzheimer's disease, aiding in their diagnosis and understanding of the model's behavior.

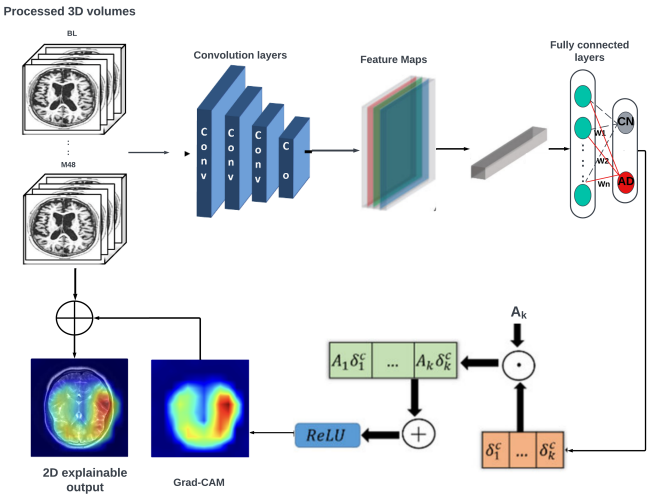

Fig. 7: Model Explainability

IV. Experimental Results and Analysis

The framework was implemented using the PyTorch library, running on a GPU with 12 GB RAM. Training of the 3D-CNN-BRNN model was conducted end-to-end, utilizing the Adam optimizer for parameter optimization. After extensive exploration of hyperparameter configurations, the training adopted a learning rate of 0.00027. Stratified cross-validation (CV) was employed in each fold to ensure balanced batches during training. The number of epochs was set at 100, corresponding to the point where the loss reached its minimum value in each fold during the 5-fold CV process. We performed multiple experiments to evaluate the efficacy of our approach for Alzheimer's detection leveraging multimodal data sources. These experiments involved investigating various combinations of the 3 modalities - MRI data, cognitive scores, and demographic information, different fusion strategies such as early and late fusion, and comparing our results with state-of-the-art works.

Standard metrics like accuracy, precision, recall, and F1-score were used for assessing the performance of the proposed approach. Accuracy measures the overall correctness of the model's predictions, which is the ratio of correctly classified instances to the total instances in the dataset. This is computed as per Eq. (1), where, TP_{AD} denotes true positives for Alzheimer's disease, TN_{CN} denotes true negatives for cognitively normal, FP_{CN} denotes false positives for cognitively normal, and FN_{AD} denotes false negatives for Alzheimer's disease.

$$Accuracy = \frac{TP_{AD} + TN_{CN}}{TP_{AD} + TN_{CN} + FP_{CN} + FN_{AD}} \quad (1)$$

Precision quantifies the accuracy of positive predictions made by the model, indicating the proportion of correctly predicted AD instances out of all instances predicted as AD (Eq. 2). Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify AD instances, indicating the proportion of correctly predicted AD instances out of all actual AD instances (Eq. 4). The F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives. It is particularly useful when the class distribution is imbalanced (Eq. 4).

$$\text{Precision} = \frac{TP_{AD}}{TP_{AD} + FP_{AD}} \quad (2)$$

$$\text{Recall} = \frac{TP_{AD}}{TP_{AD} + FN_{CN}} \quad (3)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Table II presents the evaluation metrics for different combinations of modalities used in our Alzheimer's detection model, including MRI (Magnetic Resonance Imaging), CS (Cognitive Scores), and Demographics. The results indicate that combining all three modalities together yields a better performance in terms of accuracy, precision, recall, and F1 score compared to individual modalities. By combining these modalities, the model can leverage a more comprehensive set of features, capturing both structural and biochemical changes alongside demographic risk factors. Consequently, the performance of the model improves significantly when all three modalities are integrated, highlighting the importance of multimodal approaches in Alzheimer's detection using deep learning models.

TABLE II: Performance evaluation of different combinations of modalities.

Data Modality	Accuracy	Precision	Recall	F1-score
MRI	83.34	0.8511	0.8386	0.8448
MRI + Demographic	85.66	0.8529	84.75	0.8502
MRI + CS	89.12	0.8777	0.8687	0.8782
MRI + CS + Demographic	92.65	0.9047	0.9139	0.9093

For assessing the performance of models when features from multiple modalities are to be utilized, we used two feature fusion approaches – Early Fusion and Late Fusion. Early Fusion involves combining the features from different modalities at the input level, whereas Late Fusion integrates the predictions from the model trained on each modality. In Table III, the performance metrics are presented for both Early Fusion and Late Fusion using the combination of all three modalitiesThe results show that both fusion strategies perform well, with Late Fusion slightly outperforming Early Fusion across all metrics, including Accuracy, Precision, Recall, and F1 Score. By processing predictions from individual modality-specific models and combining them at the decision-making level, late fusion enables the model to leverage the specialized information extracted by each modality. This approach allows for a more comprehensive consideration of the diverse features captured by different modalities, leading to improved performance. Late fusion also offers flexibility in decision making, adaptively weighing the contributions of each modality's prediction based on their reliability. Additionally, late fusion helps mitigate overfitting by preventing the model from relying too heavily on a single modality or feature representation, resulting in more robust and generalizable predictions.

TABLE III: Performance evaluation of different fusion strategies

Fusion technique	Accuracy	Precision	Recall	F1-score
Early Fusion	89.57	0.8866	0.8857	0.8861
Late fusion	92.65	0.9047	0.9239	0.9093

Table IV compares the performance metrics of the proposed methodology with state-of-the-art works which are described in Section II. The proposed methodology demonstrates a slightly superior performance across all metrics compared to the base paper, indicating the effectiveness of the developed Alzheimer's detection model. Additionally, the Confusion Matrix (Fig. 8) can be referenced to gain insights into the model's classification performance for each class. With 37 true positives, the model effectively identifies individuals with Alzheimer's disease, demonstrating its capability to accurately classify positive cases. Additionally, the high count of 152 true negatives indicates the model's proficiency in correctly recognizing individuals without Alzheimer's.

To enable model explainability, the Grad-CAM model was adapted for the visualization of the prediction regarding the progression of Alzheimer's Disease. This was generated based on five prior MRI scans, which yielded insightful results (depicted in Fig. 9). The heatmaps generated by Grad-CAM provided a clear indication of the regions within the brain that significantly influenced the model's decision-making process. By highlighting specific areas of activation or importance within the MRI scans, the visualization offered valuable insights into the neural correlates associated with Alzheimer's progression. This approach not only elucidated the areas of the brain crucial for accurate predictions but also contributed to a better understanding of the underlying neuropathological processes involved in the disease progression. Overall, the Grad-CAM based visual attention plots enhanced the interpretability

TABLE IV: Benchmarking the proposed approach with State-of-the-art works

Work	Techniques used	Modality	Accuracy	Precision	Recall	F1-score
Shaker et al. [8]	CNN-BiLSTM	6 timesteps of MRI, cognitive scores, demographic data	90.8	-*	-*	-*
Liangxiu et al. [16]	Residual Self-Attention Deep Neural Network + Grad-CAM	MRI and demographic data	91	-*	-*	-*
Tong et al. [17]	Deep learning model	MRI	86	0.86	0.87	0.86
Our Work	3DCNN-LSTM + Grad-CAM	5 timesteps MRI, demographic data, cognitive scores	**92.65**	0.9047	0.9139	0.9093

Not reported

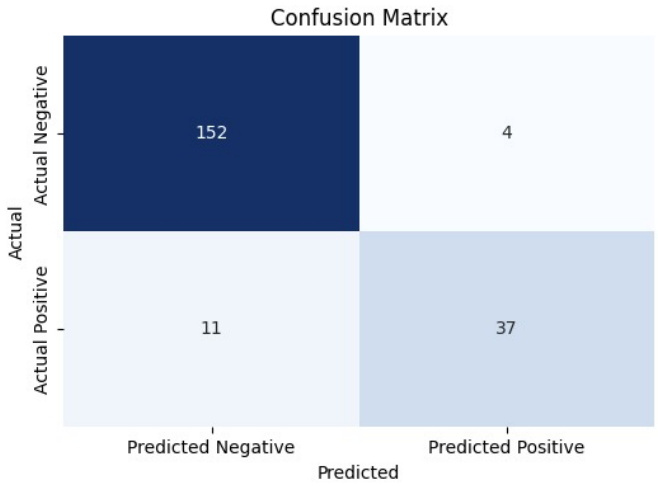

Fig. 8: Confusion Matrix for the proposed approach

of the model's predictions, shedding light on the intricate relationship between neuro-imaging data and Alzheimer's disease progression.

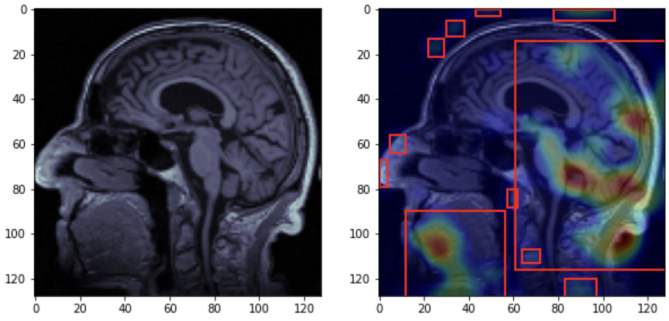

Fig. 9: Visualization of heatmap overlayed on a slice of the MRI scan

V. CONCLUSION & FUTURE WORK

In this article, a comprehensive approach to Alzheimer's prediction utilizing a multimodal framework integrating MRI scans, cognitive scores, and demographic features across multiple patient visits was presented. Through the utilization of a 3D CNN model for spatial feature extraction from MRI data followed by LSTM for temporal feature extraction, combined with cognitive and demographic data, we achieved promising results in distinguishing between Alzheimer's and cognitively normal individuals, outperforming state-of-the-art works by a margin of upto 6%. The incorporation of Grad-CAM for heatmap visualization facilitated the identification of crucial regions contributing to predictions, enhancing interpretability. Furthermore, early and late fusion strategies were investigated, which revealed that late fusion yields superior performance compared to early fusion, underscoring the importance of integrating information at later stages. Moreover, a comparative analysis demonstrated the efficacy of leveraging all three modalities—MRI, cognitive scores, and demographic features—highlighting the significance of a holistic approach in Alzheimer's prediction.

As part of extended work, we intend to subject our approach to a detailed validation study with subject matter experts to verify the model's accuracy while also prioritizing transparency and interpretability. Our future endeavors involve meticulous examination of explainability techniques, aiming not only to provide precise predictions but also to offer transparent insights into the decision-making processes of the deep learning model. This collaborative approach is crucial for real-world applications, particularly in sensitive domains like healthcare, and for gaining trust in the model's outputs and facilitating its potential integration into clinical decision-making processes.

REFERENCES

[1] S. El-Sappagh, H. Saleh, R. Sahal, T. Abuhmed, S. R. Islam, F. Ali, and E. Amer, "Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data," in *Future Generation Computer Systems*, Feb. 2021, pp. 2680–699. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X20329824

[2] A. Moscoso, J. Silva-Rodríguez, J. M. Aldrey, J. Cortés, A. Fernández-Ferreiro, N. Gómez-Lado, Álvaro Ruibal, and P. Aguiar, "Prediction of alzheimer's disease dementia with mri beyond the short-term: Implications for the design of predictive models," in *NeuroImage: Clinical*, 2019, p. 101837. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213158219301871

[3] Z. Zhang, W. Lin, M. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 344–350.

[4] B. Khagi and G.-R. Kwon, "3d cnn design for the classification of alzheimer's disease using brain mri and pet," *IEEE Access*, vol. 8, pp. 217 830–217 847, 2020.

[5] Y. Tu, S. Lin, J. Qiao, Y. Zhuang, and P. Zhang, "Alzheimer's disease diagnosis via multimodal feature fusion," *Computers in Biology and Medicine*, vol. 148, p. 105901, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001048252200645X

[6] M. M. K. S. . F. H. Vimbi Viswan, Noushath Shaffi, "Explainable artificial intelligence in alzheimer's disease classification: A systematic review," *Cognitive Computation*, vol. 16, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s12559-023-10192-x#citeas

[7] B. Avelar-Pereira, M. Belloy, R. O'Hara *et al.*, "Decoding the heterogeneity of alzheimer's disease diagnosis and progression using multilayer networks," *Molecular Psychiatry*, vol. 28, pp. 2423–2432, 2023. [Online]. Available: https://doi.org/10.1038/s41380-022-01886-z

[8] S. El-Sappagh, T. Abuhmed, S. M. R. Islam, and K. Kwak, "Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data," *Neurocomputing*, 06 2020.

[9] "Alzheimer's Disease Neuroimaging Initiative (ADNI) database." [Online]. Available: https://adni.loni.usc.edu/

[10] P. McCarthy, "FSLeyes," Jul. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3937147

[11] "Advanced Normalization Tools." [Online]. Available: http://stnava.github.io/ANTs

[12] "MNI Atlases - FslWiki." [Online]. Available: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases

[13] "BET - FslWiki - Skull Stripping." [Online]. Available: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET

[14] "NiBabel." [Online]. Available: https://nipy.org/nibabel/

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[16] X. Zhang, L. Han, W. Zhu, L. Sun, and D. Zhang, "An explainable 3d residual self-attention deep neural network for joint atrophy localization and alzheimer's disease diagnosis using structural mri," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5289–5297, 2022.

[17] J. Venugopalan, L. Tong, H. Hassanzadeh *et al.*, "Multimodal deep learning models for early detection of alzheimer's disease stage," *Scientific Reports*, vol. 11, p. 3254, 2021. [Online]. Available: https://doi.org/10.1038/s41598-020-74399-w