Neuroscience 585 (2025) 133–143

Contents lists available at ScienceDirect

Neuroscience

journal homepage: www.elsevier.com/locate/nsc

# AlzFormer: Video-based space-time attention model for early diagnosis of Alzheimer's disease

Taymaz Akan [a,b], Sara Akan [c], Sait Alp [d], Christina Raye Ledbetter [e], Mohammad Alfrad Nobel Bhuiyan [a,*], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] Department of Medicine, LSU Health Shreveport, Shreveport, LA, USA
[b] Department of Software Engineering, Faculty of Engineering, Istanbul Topkapı University, Istanbul, Turkey
[c] Department of Computer Engineering, Faculty of Engineering, Istanbul Galata University, Istanbul, Turkey
[d] Department of Artificial Intelligence Engineering, Trabzon 61335, Turkey
[e] Department of Neurosurgery, LSU Health Shreveport, Shreveport, LA, USA

ARTICLE INFO

*Keywords:*
Alzheimer's disease
TimeSformer
Spatiotemporal
Attention
Deep learning

ABSTRACT

Early and accurate Alzheimer's disease (AD) diagnosis is critical for effective intervention, but it is still challenging due to neurodegeneration's slow and complex progression. Recent studies in brain imaging analysis have highlighted the crucial roles of deep learning techniques in computer-assisted interventions for diagnosing brain diseases. In this study, we propose AlzFormer, a novel deep learning framework based on a space–time attention mechanism, for multiclass classification of AD, MCI, and CN individuals using structural MRI scans. Unlike conventional deep learning models, we used spatiotemporal self-attention to model inter-slice continuity by treating T1-weighted MRI volumes as sequential inputs, where slices correspond to video frames. Our model was fine-tuned and evaluated using 1.5 T MRI scans from the ADNI dataset. To ensure the anatomical consistency of all the MRI data, All MRI volumes were pre-processed with skull stripping and spatial normalization to MNI space. AlzFormer achieved an overall accuracy of 94 % on the test set, with balanced class-wise F1-scores (AD: 0.94, MCI: 0.99, CN: 0.98) and a macro-average AUC of 0.98. We also utilized attention map analysis to identify clinically significant patterns, particularly emphasizing subcortical structures and medial temporal regions implicated in AD. These findings demonstrate the potential of transformer-based architectures for robust and interpretable classification of brain disorders using structural MRI.

## Introduction

Dementia diagnostics are complex and require a significant amount of time following the onset of the first clinical symptoms. The average time required for the process is 2.8 years for late-onset dementia and 4.4 years for young-onset dementia (Van Vliet et al., 2013). Alzheimer's disease (AD), the most common form of dementia, is a progressive and irreversible neurodegenerative condition marked by gradual deterioration in memory and cognitive function. Neuropathological changes related to AD are now believed to begin up to two decades prior to the emergence of clinical symptoms (Alp et al., 2024; Gelir et al., 2024). Early diagnosis of cognitive status is essential for the management and treatment of AD. Advancements in brain imaging techniques like Magnetic Resonance Imaging (MRI) have significantly impacted the diagnosis and prognosis of brain disorders in computer-assisted interventions (Fan et al., 2008). Therefore, many computer-aided approaches have been developed to enable more precise and efficient AD diagnosis (Jain et al., 2019; Ebrahimighahnavieh et al., 2020; Ebrahimi et al., 2021; Alinsaif et al., 2021; Loddo et al., 2022; Khojaste-Sarakhsi et al., 2022; Lin et al., 2023; Rangaraju et al., 2024; Chua et al., 2025).

* Corresponding author at: Division of Clinical Informatics, Department of Medicine, Louisiana State University Health Sciences Center, PO Box 33932, Shreveport, LA 71130-3932, USA.
*E-mail address:* Nobel.Bhuiyan@lsuhs.edu (M.A. Nobel Bhuiyan).

[1] The longitudinal data used in preparation for this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.neuroscience.2025.08.062
Received 9 May 2025; Accepted 30 August 2025
Available online 3 September 2025

**Table 1**
The details of data collections.

| | Set | # MRI Scans | NC | MCI | AD | Male | Female | Age (years) |
|---|---|---|---|---|---|---|---|---|
| ADNI1: Complete 1Yr 1.5 T (82 %) | Train | 1951 | 600 | 946 | 405 | 1341 | 953 | 75 ± 6.6 |
| ADNI1: Complete 1Yr 1.5 T (18 %) | Test | 342 | 105 | 166 | 71 | | | |

Machine learning provides a structured approach for the automatic classification of brain disorders by learning complex and subtle patterns from high-dimensional neuroimaging data. From a technical standpoint, this task typically involves three key stages: pre-processing, feature extraction, and classifier learning. These components are crucial for effectively analyzing brain imaging data and supporting clinical diagnosis (Suk et al., 2017). On the other hand, deep learning (DL) eliminates the need for manual feature engineering by learning task-specific representations directly from the raw input data. Rather than depending on predefined features, DL models extract informative patterns during training in a label-guided, end-to-end manner. As a result, the learned features are directly optimized for the classification objective, which makes DL especially effective for complex and high-dimensional modalities such as brain MRI.

MRI is a complex modality that requires various strategies to manage and analyze volumetric brain data. These include voxel-based (Armañanzas et al., 2017; Basaia et al., 2019; Solana-Lavalle and Rosas-Romero, 2021; Nemoto et al., 2021; Estudillo-Romero et al., 2022; Guan et al., 2022), region-of-interest (ROI)-based (Shinde et al., 2019; Ahmed et al., 2020; Feng et al., 2022; Shi et al., 2022), patch-based (Goenka and Tiwari, 2022; Liu et al., 2023; Huang and Qiu, 2024), and slice-based (Ebrahimi et al., 2021; Zhang et al., 2022; Sharma et al., 2022; Avram et al., 2024; Akan et al., 2023) methods. Voxel-based approaches analyze the entire brain volume but involve high computational costs and may lose crucial contextual information. ROI-based methods focus on specific brain regions typically affected by neurodegenerative diseases (NDDs), but they may overlook abnormalities in other areas. Patch-based methods examine small segments of brain images, which can help detect early signs of disease with fine-grained detail. Slice-based approaches divide volumetric brain data into two-dimensional slices, but they often fail to capture spatial dependencies between adjacent slices, which are crucial for understanding disease progression. Despite these limitations, each method contributes valuable insights into brain pathology.

Nevertheless, one of the overarching challenges in MRI-based brain analysis remains the high dimensionality of the data, contrasted with the relatively small number of available labeled samples. While transfer learning is often used to address this limitation, most pre-trained models are based on 2D natural images and do not capture the spatial continuity present in 3D brain MRI. Slice-based approaches frequently rely on such 2D models, which limits their ability to model volumetric context. As a result, the transferability of these models to medical imaging tasks remains constrained. To overcome this, we propose leveraging pre-trained video models originally designed to handle sequential and spatiotemporal information and fine-tuning them on MRI data. By treating the MRI volume as a sequence of slices, these models can exploit inter-slice dependencies in a way that aligns more naturally with the structure of brain imaging data.

Following the trends and successes in medical image analysis and deep learning, convolutional neural networks (CNN) have become increasingly popular in recent years. Still, they have not been shown to outperform conventional classifiers significantly. Most CNN studies perform no to minimal pre-processing of structural MRI scans as input for their classifier. In contrast, others employ more extensive pre-processing strategies that have proven successful for conventional classifiers, such as gray matter (GM) density maps. Although CNNs are designed to extract high-level features from raw imaging data, dedicated pre-processing that enhances disease-related features may improve the learning process for complex tasks, reducing model complexity and allowing for a more stable learning process. It is still unclear whether CNNs outperform conventional classifiers in AD classification and whether they benefit from extensive MRI pre-processing (Bron et al., 2021).

Recent studies, however, have shown that using DL models trained on MRI volumes can improve AD classification accuracy. E.g., (Lian et al., 2020) proposed a hierarchical CNN designed to potentially learn both local and global features from 3D MRI volumes while achieving 90 % accuracy in AD vs. CN classification. (Backstrom et al., 2018) used a 3D ConvNet and emphasized the importance of patient-specific data splitting to prevent data leakage. Their findings revealed a performance gap of up to 10 % between random and subject-level splits, emphasizing the need for rigorous evaluation protocols. Wen et al. (Wen et al., 2020) further reported that nearly half of the published deep learning models for AD had data leakage, which was primarily caused by improper longitudinal data partitioning.

Recently, Transformer-based architectures have emerged as an alternative to CNNs for vision tasks, offering several advantages for modeling long-range dependencies and sequential information. Transformers have less restrictive inductive biases compared to CNNs. This broadens the range of functions they can represent (Cordonnier et al., 2019; Zhao et al., 2020) and makes them more appropriate for large-scale datasets requiring fewer strong inductive priors. Second, convolutional kernels have been designed to capture short-range spatio-temporal information; however, they cannot model dependencies that expand beyond the receptive field. Unlike CNNs, which primarily rely on local receptive fields, Transformers utilize attention mechanisms to capture global context, making them well-suited for processing medical imaging data with spatial or temporal structure. In particular, video-based Transformers can model sequential relationships across MRI slices, providing a natural way to exploit inter-slice continuity without requiring full 3D convolutions. Finally, transformers are faster to train than 3D convolutional networks, can achieve significantly better test efficiency (with a slight loss in accuracy), and can be used on far longer video clips (i.e., more MRI slices). Motivated by these properties, we explore using a pre-trained video Transformer model fine-tuned for the classification of AD, MCI, and CN using slice-based MRI data.

Recent studies have begun to explore the potential of Transformer models for neuroimaging-based classification tasks. E.g., Liu et al. (Liu et al., 2023) proposed the Multi-Modal Mixing Transformer (3MT), a cascaded attention architecture that amalgamates 3D MRI with as many as 12 clinical features. During training, 3MT uses a modality dropout mechanism, which allows it to remain robust in the face of missing data. The model attained cutting-edge results in both AD classification and mild cognitive impairment conversion tasks, with generalization validated on the AIBL dataset without the need for retraining. Additional studies, including (Kushol et al., 2022), have investigated dual-transformer methodologies functioning in spatial and frequency domains, thereby substantiating the efficacy of Transformers for AD-related imaging tasks.

## Materials and method

### *Study population*

We used two datasets from the Alzheimer's Disease Neuroimaging Initiative (ADNI; adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership led by Principal Investigator Michael W. Weiner, MD. Its primary objective is to determine whether clinical and

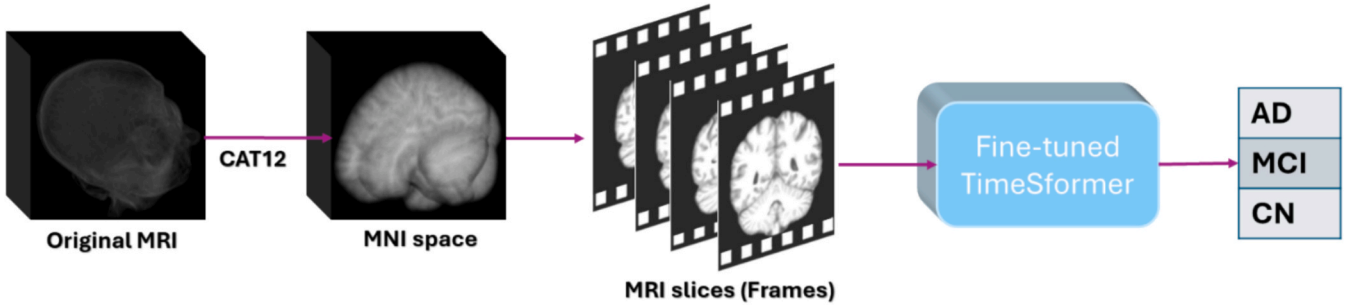


**Fig. 1.** Overview of the proposed AD classification pipeline. T1-weighted MRI scans are pre-processed and spatially normalized to a standard template. The processed volumes are then converted into sequential slices and treated as video-like inputs to a fine-tuned TimeSformer model, which performs multiclass classification into AD, MCI, and CN categories.

**Table 2**
Classification performance of AlzFormer on the test set ("ADNI1: Complete 3Yr 3 T"). Reported metrics include precision, recall, and F1-score for each class: AD, MCI, and CN. Overall accuracy, macro average, and weighted average scores are also provided.

| Classes | Precision | Recall | F1- score | Support |
|---|---|---|---|---|
| 0 (AD) | 0.96 | 0.94 | 0.95 | 71 |
| 1 (MCI) | 0.99 | 0.99 | 0.99 | 105 |
| 2 (CN) | 0.97 | 0.98 | 0.97 | 166 |
| Accuracy | | | 0.97 | 342 |
| Macro avg | 0.97 | 0.97 | 0.97 | 342 |
| Weighted avg | 0.97 | 0.97 | 0.97 | 342 |

**Table 3**
Performance comparison of different architectures for Alzheimer's disease classification.

| Architecture | ACC | Precision | Recall | F-score |
|---|---|---|---|---|
| CNN-Bi-LSTM (Akan et al., 2023) | 0.90 | 0.90 | 0.90 | 0.89 |
| CNN-Transformer (Alp et al., 2024) | 0.90 | 0.90 | 0.90 | 0.89 |
| ViT-Bi-LSTM (Akan et al., 2023) | 0.89 | 0.90 | 0.89 | 0.89 |
| ViT-Transformer (Alp et al., 2024) | 0.91 | 0.92 | 0.91 | 0.90 |
| **AlzFormer** | **0.97** | **0.97** | **0.97** | **0.97** |

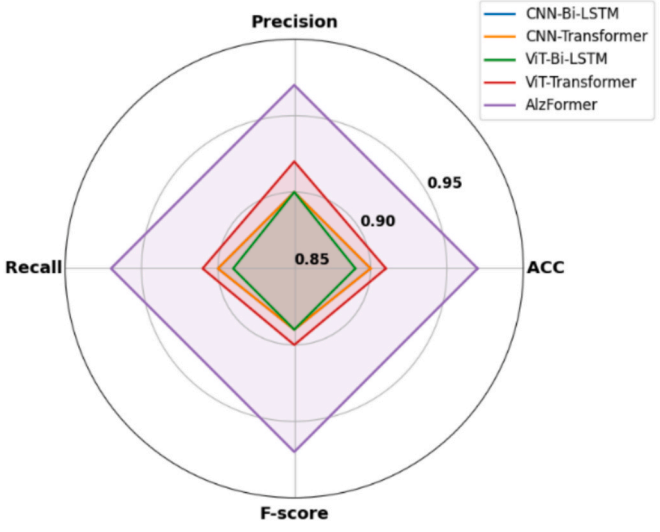


**Fig. 2.** Radar plot comparing the performance of baseline models and the proposed AlzFormer architecture across four evaluation metrics: Accuracy, Precision, Recall, and F-score.

neuropsychological assessments, serial magnetic resonance imaging (MRI), positron emission tomography (PET), and other biological markers can be integrated to track the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For the most up-to-date information, visit https://www.adni-info.org.

Our model was trained and evaluated using T1-weighted MRI scans from the "ADNI1: Complete 1Yr 1.5 T" data collection. The training set comprised 1951 scans (82 %), including 600 cognitively normal (NC), 946 with mild cognitive impairment (MCI), and 405 with Alzheimer's disease (AD). The test set consisted of 342 scans (18 %) with 105 NC, 166 MCI, and 71 AD cases. Across both sets, participants had a mean age of 75 ± 6.6 years, with a balanced sex distribution: 1341 male and 953 female subjects. Further demographic details are summarized in Table 1.

*Pre-processing*

Several steps are involved in the processing and analysis of sMRI data. The images must be aligned, registered to a standard template like MNI space, segmented into various tissue types like gray matter, white matter, and cerebrospinal fluid, and then subjected to voxel-based analyses (See Fig. 1). Since image registration helps standardize MRI scans using a fixed-size template, it is essential to guarantee these images' spatial alignment.

To standardize the coordinates and align all of the images in a common reference frame, we used the Montreal Neurological Institute (MNI) space. To wrap MRI scans into the MNI-152 space, skull stripping, normalization, and image registration were carried out using CAT-12 (https://neuro-jena.github.io/cat), a toolbox extension of SPM12 (Penny et al., 2011) designed for structural MRI pre-processing. The selection was also limited to the central 32 slices containing significant slices. The rest of the slices were excluded due to background information (None-brain area) unrelated to brain tissue.

*Classification*

Video classification involves learning spatial and temporal patterns by analyzing visual features within individual frames (spatial) and changes across frames over time (temporal). TimeSformer (Bertasius et al., 2021) is a convolution-free approach to video classification based on self-attention over space and time. We fine-tuned TimeSformer to perform classification on MRI volume for AD diagnosis. In our context, each T1-weighted brain MRI volume is treated as a temporal sequence of 2D coronal slices, where each slice corresponds to a single "frame." Our method, AlzFormer, enables direct spatiotemporal feature learning from sequences of MRI slice-level patches, capturing both intra-slice spatial patterns and inter-slice temporal dependencies critical for modeling progressive neurodegeneration.

Video classification involves learning spatial and temporal patterns by analyzing visual features within individual frames (spatial) and changes across frames over time (temporal). TimeSformer is a

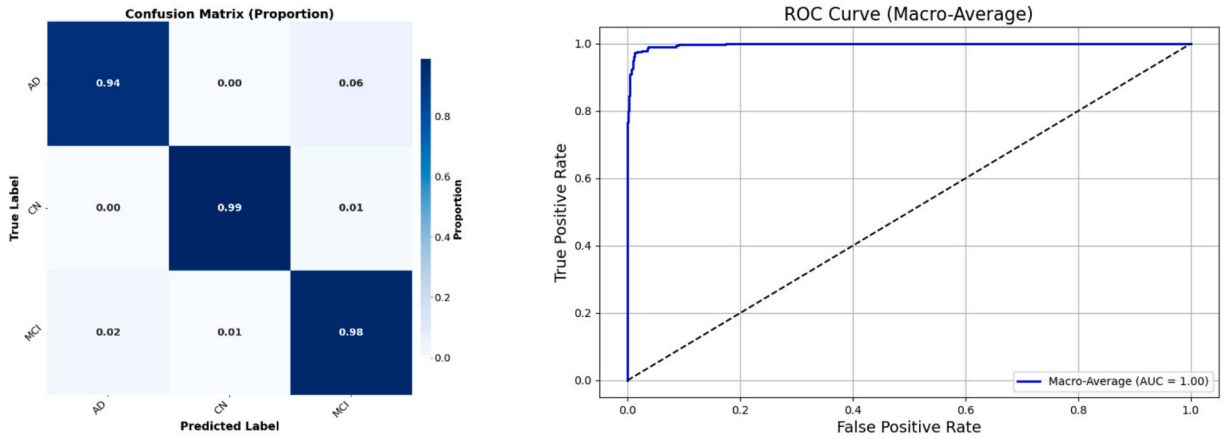


**Fig. 3.** **(a)** Normalized confusion matrix showing the proportion of predicted versus true labels on the test set. **(b)** Macro-averaged ROC curve for the classification of AD, MCI, and CN on the test set. The model achieved a macro-average AUC of 1, indicating high discriminative performance across all classes.

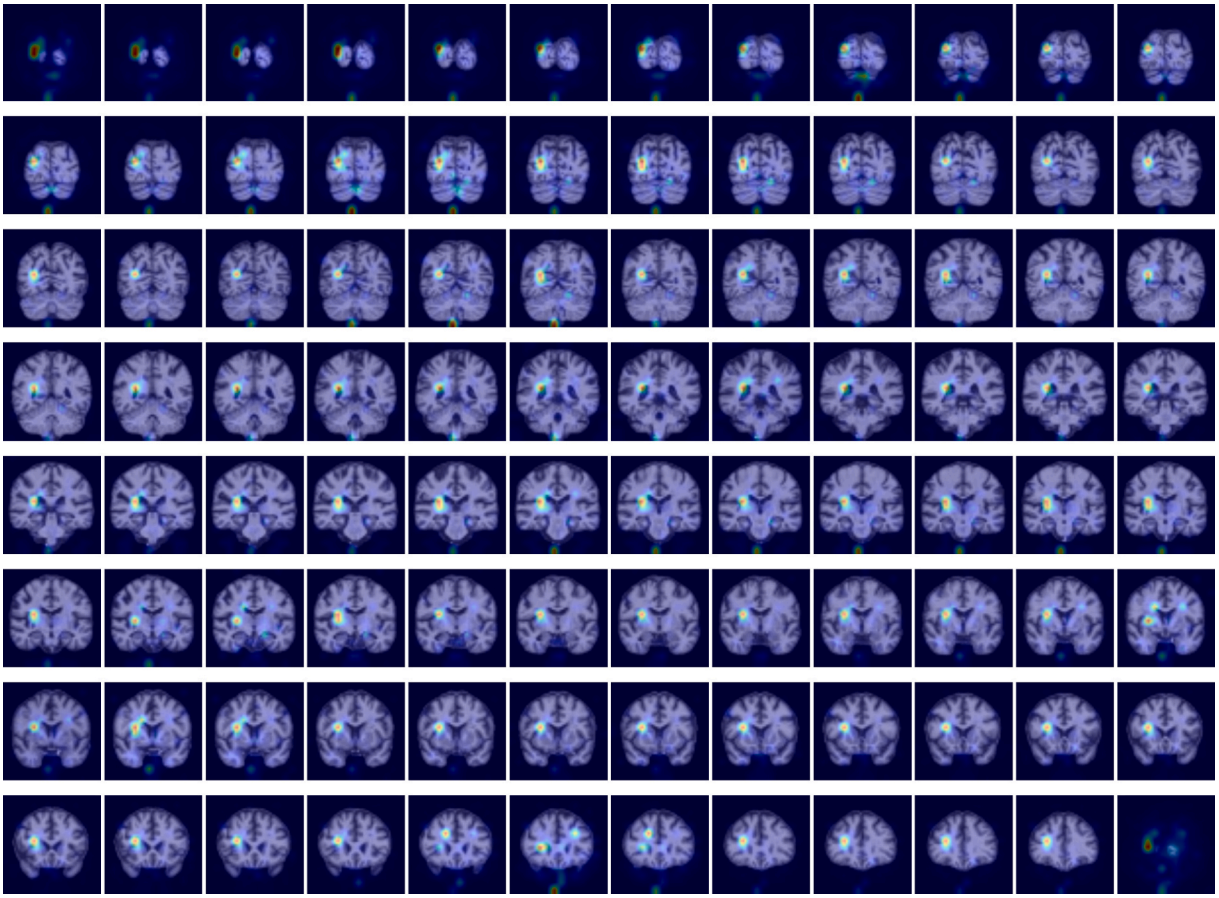

**Fig. 4.** Attention maps for a representative Alzheimer's disease subject.

convolution-free approach to video classification based on self-attention over space and time. We fine-tuned TimeSformer to perform classification on MRI volume for AD diagnosis. In our context, each T1-weighted brain MRI volume is treated as a temporal sequence of 2D coronal slices, where each slice corresponds to a single "frame." Our method, AlzFormer, enables direct spatiotemporal feature learning from sequences of MRI slice-level patches, capturing both intra-slice spatial patterns and inter-slice temporal dependencies critical for modeling progressive neurodegeneration.

Each MRI volume was sliced into 2D coronal images and tokenized into non-overlapping patches (e.g., 16 × 16 pixels). These patches were flattened and linearly projected into embeddings, followed by the addition of learnable spatial and temporal positional encodings. The resulting patch embeddings were passed through the transformer layers of TimeSformer, where self-attention is computed separately across spatial and temporal dimensions. This design allows the model to capture local anatomical features and their dependencies across slices jointly.

We initialized the model with weights pre-trained on the Kinetics-400 (Kay et al., 2017) video dataset for training. We replaced the final classification layer to perform multiclass classification across three diagnostic categories: AD, MCI, and CN. We used cross-entropy loss

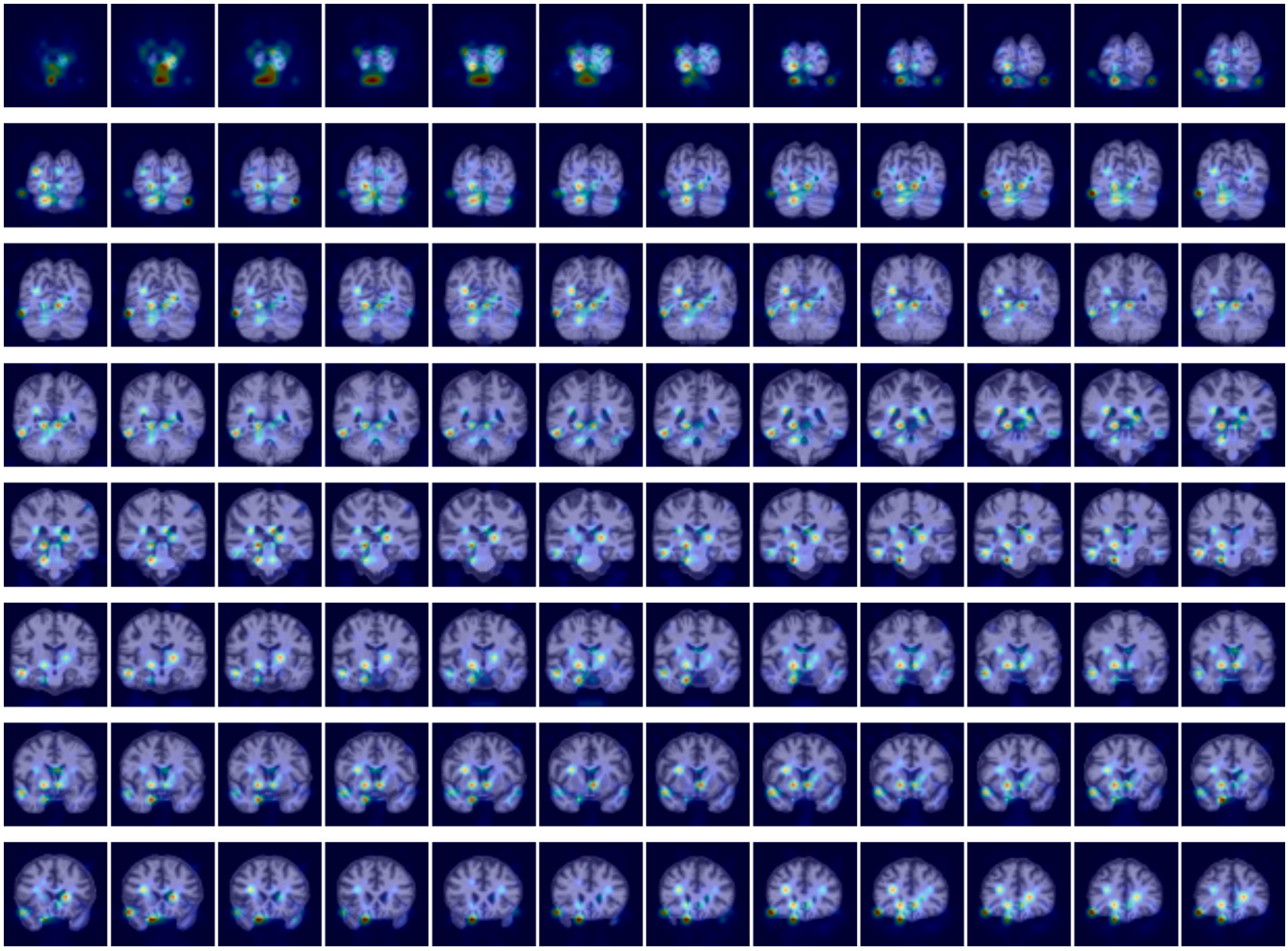

**Fig. 5.** Attention maps for a representative mild cognitive impairment subject.

function. Only the last transformer layer (layer 11) and the final classification head were unfrozen during fine-tuning. At the same time, the rest of the model remained frozen to preserve general visual representations learned from large-scale video data. The complete TimeSformer model contains approximately 121 million parameters, but we restricted training to 10,044,675 trainable parameters corresponding only to the last transformer layer and the classification head. The training was conducted using the AdamW optimizer with a learning rate 1e-4 and a weight decay of 1e-4. In addition, we utilized a Cosine Annealing Learning Rate Scheduler to gradually reduce the learning rate following a cosine function and improve convergence. In addition, the best-performing model (based on validation accuracy) was checkpointed for downstream use. An overview of the proposed classification pipeline using the fine-tuned TimeSformer model is illustrated in Fig. 1.

## Result

### *Classification performance*

Classification performance was quantified by the classification accuracy, precision, recall f1-score, area under the curve (AUC), and accuracy. The models were optimized on the training set; the test set was applied only to assess the final model.

On the test set, AlzFormer achieved an overall accuracy of 97 %. As shown in Table 2, and confusion matrix Fig. 3 **(a)**, the F1-scores for the three classes were 0.94 for AD, 0.99 for MCI, and 0.98 for cognitively normal (CN). The macro and weighted-average F1-scores were both 0.97, indicating balanced performance through the classes. Notably, no CN cases were misclassified as AD, and vice versa, indicating strong class separability between healthy controls and patients with dementia.

We computed the macro-averaged receiver operating characteristic (ROC) curve to assess the model's discriminative performance across all classes. As illustrated in Fig. 3 **(b)**, the model achieved a high AUC of 1, indicating excellent capability to distinguish between the three diagnostic categories. The steep rise in the curve close to the origin indicates a low false positive rate and high overall sensitivity across the classes.

We further compared the performance of a couple of baseline models, including CNN-BiLSTM, CNN-Transformer, ViT-BiLSTM, and ViT-Transformer, with our proposed model, AlzFormer. As shown in Table 3, AlzFormer outperforms all other models, achieving the highest accuracy, precision, recall, and F-score, demonstrating its effectiveness in Alzheimer's disease classification. Moreover, the spider plot (See Fig. 2) visually illustrates that **AlzFormer** consistently outperforms all baseline models, achieving superior scores in every metric, which further confirms the quantitative results presented in Table 3.

Based on overall performance across all metrics, AlzFormer ranks first, with ViT-Transformer coming in second with 0.91 accuracy and 0.90 F1-score. With identical scores (0.90 accuracy and 0.89 F1-score), CNN-BiLSTM and CNN-Transformer are tied for third place, while ViT-BiLSTM ranks lowest with slightly lower recall and F1-score (0.89). This ranking underscores the benefit of integrating spatial and temporal attention in AlzFormer to achieve precise and resilient MRI-based classification.

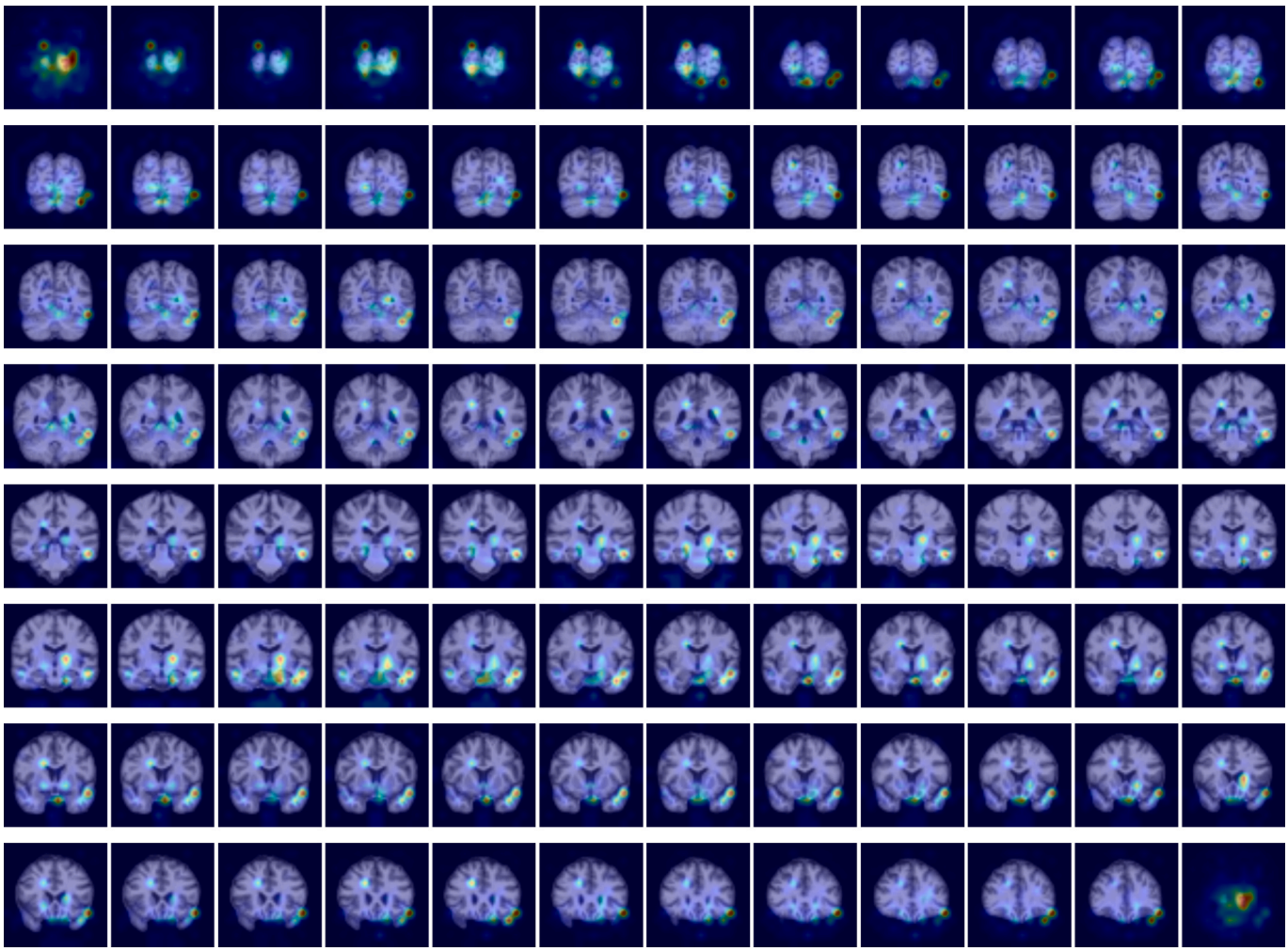

**Fig. 6.** Attention maps for a representative cognitively normal subject.

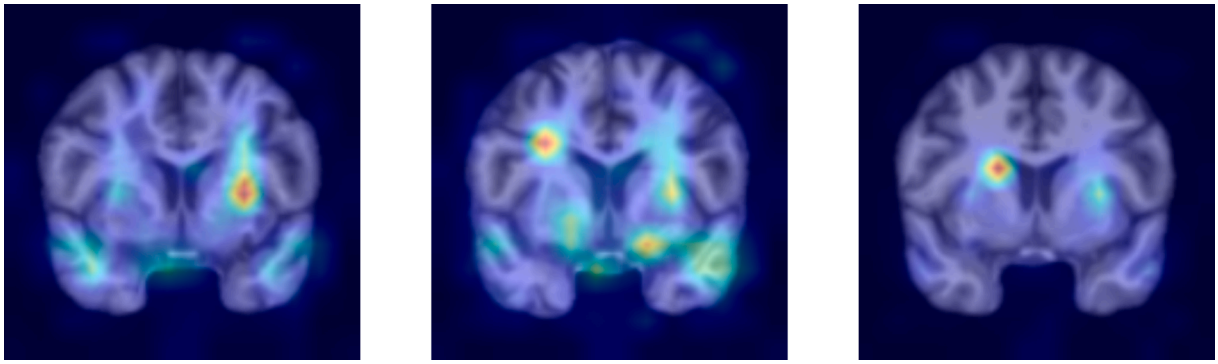

**Fig. 7.** Attention heatmaps at the 85th slice for representative CN, MCI, and AD subjects, respectively.

*Model analysis and representation*

We visualized class-specific attention maps over the input MRI slices for CN, MCI, and AD cases to better understand the model's decision-making process. As shown in Figs. 4–6, the TimeSformer model attends to distinct brain regions depending on the diagnostic category.

As illustrated in Fig. 4, the model's attention in a representative AD subject primarily focuses on the medial temporal lobe, particularly the hippocampus and Para hippocampal gyrus, across central coronal slices. Additional regions receiving attention include the putamen, thalamus, and posterior cingulate cortex, indicating the model's sensitivity to both limbic and subcortical structures affected by AD. These patterns are consistent with established neuropathological progression and highlight the model's ability to capture clinically relevant features.

As shown in Fig. 5, the model's attention in a representative MCI subject is more diffusely distributed than AD, with a prominent focus on medial temporal structures, including the hippocampus and the putamen, thalamus, and posterior cingulate cortex. These activations reflect early-stage alterations commonly associated with MCI, indicating that the model captures subtle changes in limbic and subcortical regions

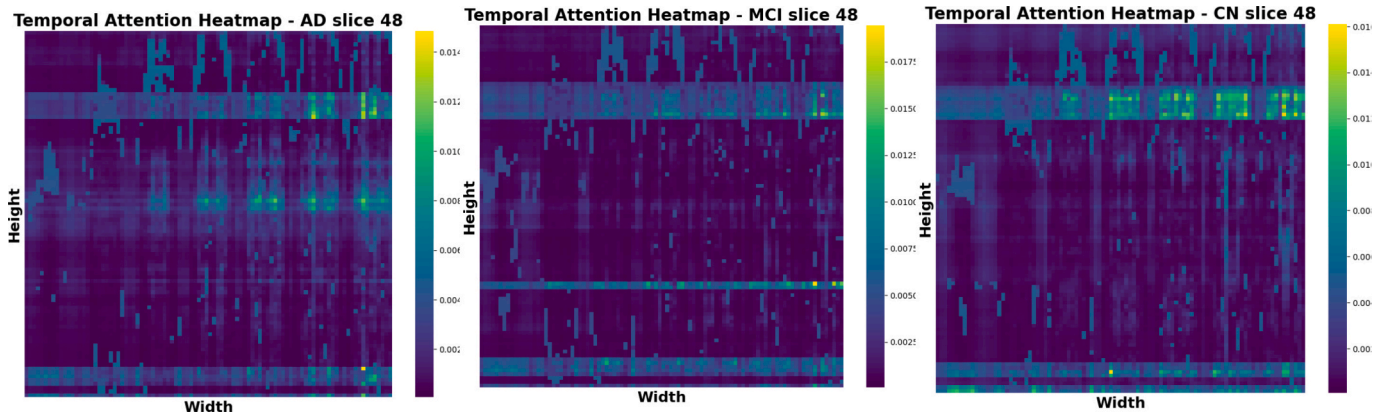


**Fig. 8.** Temporal attention heatmaps for a representative sagittal slice (slice 48) across CN, MCI, and AD subjects.

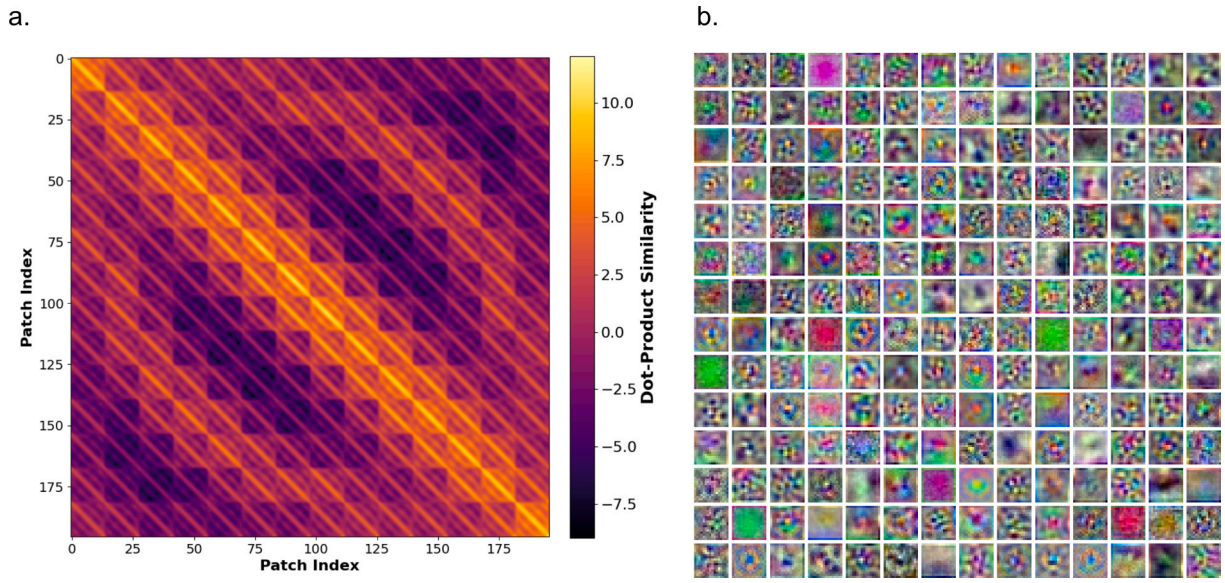


**Fig. 9.** **(a)** Dot-product similarity between learned positional embeddings of the AlzFormer model. A clear diagonal pattern suggests that the model captures spatial locality, with nearby patches having more similar positional encodings. **(b)** Visualization of the learned projection filters from the input embedding layer of the TimeSformer model.

before progressing to AD.

As shown in Fig. 6, the model's attention in a representative cognitively normal subject appears more dispersed and less concentrated in canonical Alzheimer's-related regions. Mild activations are observed in areas such as the putamen, precuneus, and occipital cortex but without the strong or consistent focus seen in MCI or AD cases. This pattern suggests the model recognizes the absence of pathological hallmarks and distributes attention more broadly, potentially reflecting the lower discriminative need in the absence of disease.

It is important to note that while the attention heatmaps frequently highlight regions such as the hippocampus and thalamus—structures commonly implicated in AD—they are **not disease-specific**. Similar structural changes may also occur during normal aging or in other neurodegenerative conditions. The attention maps simply reflect the regions that the model found statistically useful for classification within the given data distribution, rather than exclusive AD biomarkers. This distinction is critical for proper interpretation of model explanations.

To have better visual comparison across clinical groups, Fig. 7 presents attention heatmaps at the 48$^{th}$ slice from three different representative subjects corresponding to the CN, MCI, and AD classes. This fixed-slice view allows direct assessment of how the model distributes spatial attention at the same anatomical location. The medial temporal region, which is commonly impacted by structural degeneration, is the primary focus of attention in the AD subject. The MCI subject shows more diffuse and bilateral attention, reflecting intermediate anatomical changes, while the CN subject exhibits low-intensity and spatially uniform attention. These differences show that the model adjusts spatial attention to group-specific structural MRI anatomical features.

Fig. 8 illustrates the temporal attention heatmaps at sagittal slice 48 for AD, MCI, and CN. The CN subject's attention seems to be low-intensity and diffuse, suggesting that there is minimal focus in all three spatial dimensions. In contrast, the MCI subject shows slightly more structured yet sparse activation, particularly along horizontal bands. Notably, the AD subject exhibits more focused and intense temporal attention distributed along consistent spatial regions. This implies that the model has acquired the ability to selectively focus on temporally informative regions, which may be indicative of pathological progression as captured by slice-level dynamics.

Fig. 11. Chord diagrams illustrate feature similarity patterns after PCA-based dimensionality reduction for each diagnostic group: (a) AD, (b) MCI, and (c) CN. Each diagram visualizes pairwise cosine similarity between principal components of extracted features. 60 components were selected using PCA to retain the most informative dimensions of the feature space. The AD subject exhibits more focused and localized

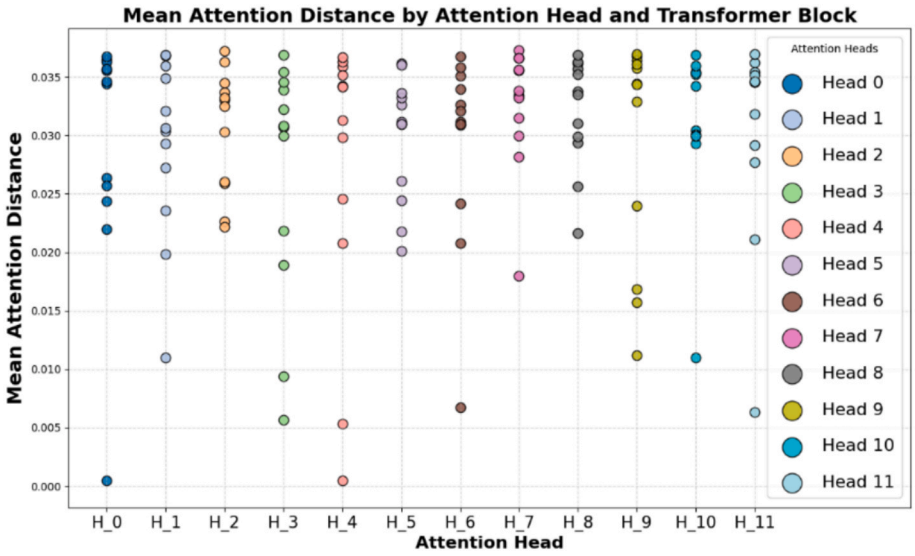


**Fig. 10.** Mean attention distance across attention heads and transformer blocks in the AlzFormer model. Each dot represents the average spatial distance a head attends to, revealing the transition from local to global attention patterns throughout the network.

feature interactions, while MCI and CN subjects demonstrate broader, more diffuse connectivity patterns, reflecting differing representational structures across diagnostic stages.

Transformers are inherently permutation-invariant and thus require positional information to model spatial or sequential relationships in the input data. In TimeSformer, this is achieved through learned positional embeddings. We visualized the similarity between positional embeddings by computing pairwise dot-product similarity across patch indices to understand how spatial structure is captured. In Fig. 9 **(a)**, the similarity matrix's diagonal structure indicates the positional encoding's spatial locality. Patches that are close together in the original image tend to have more similar positional embeddings, allowing the model to preserve relative spatial information even without convolutions.

The non-overlapping patches from MRI slices were used in the TimeSformer model to flatten each patch. We then applied a linear projection to convert the patches into an embedding. While these projections are not convolutional filters in the classical sense, visualizing them reveals structured and interpretable patterns. Fig. 9 **(b)** shows that many projection weights resemble edge detectors, circular patterns, or textured gradients—like the early-layer kernels of CNNs. This indicated that, despite lacking explicit convolutions, TimeSformer still captured localized spatial structure and provided informative representations for brain disorder classification.

We analyzed the mean attention distance (MAD) across attention heads and transformer blocks to investigate the spatial behavior of the model. The attention model's local and global patterns were measured using MAD. For each token, the geometric distance to all other tokens is weighted by the corresponding attention scores and then averaged across all tokens. As shown in Fig. 10, we compute the MAD for each attention head across all transformer blocks of AlzFormer. The early layers from nearby tokens identified localized attention, and global attention patterns were identified using deeper layers. This gradual receptive field expansion captures fine-grained local features and global context across the MRI slices.

## Discussion

This work aimed to explore using video-based deep learning models to diagnose AD. More specifically, we examine whether video modeling approaches that treat sequential MRI scans as video frames can effectively capture spatiotemporal patterns for classifying individuals into AD, MCI, and CN categories. The model was trained and evaluated on 1.5 Tesla MRI scan data from the public ADNI dataset. AlzFormer achieved 94 % accuracy on a test set, with F1-scores of 0.94 for AD, 0.99 for MCI, and 0.99 for CN, indicating balanced performance across the classes, with macro and weighted-average F1-scores of 0.97.

We compared AlzFormer with four state-of-the-art models: CNN-BiLSTM, CNN-Transformer, ViT-BiLSTM, and ViT-Transformer in order to evaluate its effectiveness. These models were consistently outperformed by AlzFormer in all evaluation metrics. It was the foremost model in terms of classification accuracy, precision, recall, and F1-score, with ViT-Transformer following as the second-best model. This comparative analysis highlights the enhanced efficacy of AlzFormer in modeling 3D brain MRI data, according to its spatiotemporal attention mechanism.

Although deep learning models have proved remarkable performance in brain disorder classification tasks, they are commonly called "black-box" systems because they are unable to interpret the results. This limitation raises concerns, particularly in clinical applications where understanding model decisions is essential. We addressed this problem by visualizing attention heatmaps for representative subjects from each diagnostic group across MRI slices (Figs. 4–6). These visualizations reveal which brain regions the model focuses on during prediction, revealing spatial patterns that influence classification outcomes. The heatmaps, which highlight discriminative regions, show that the model does not learn unquestioningly but rather captures disease-relevant features in the input. This promotes trust in model behavior and allows for neuroscientific validation of learned representations.

Transformers, particularly in video-based configurations, present potential for MRI-based classification. Using spatiotemporal self-attention, these models can learn inter-slice temporal relationships and intra-slice spatial features. This is an essential capability for capturing the progressive nature of neurodegenerative diseases. Our approach utilizes this architecture by treating MRI volumes as sequences of slices, allowing the model to exploit structural continuity without the complexity of full 3D convolutions. Additionally, using a pre-trained video model reduces reliance on extensive medical records and enables effective transfer learning from natural video domains. Regardless of the advantages of the transformer model, it requires a significant amount of time for computation, including extensive fine-tuning and pre-processing. Another limitation of our current model setup is it requires high-resolution MRI or MRI slices from 96 to form a meaningful spatiotemporal sequence. The model's performance may degrade in clinical settings with lower-resolution acquisitions or fewer slices of

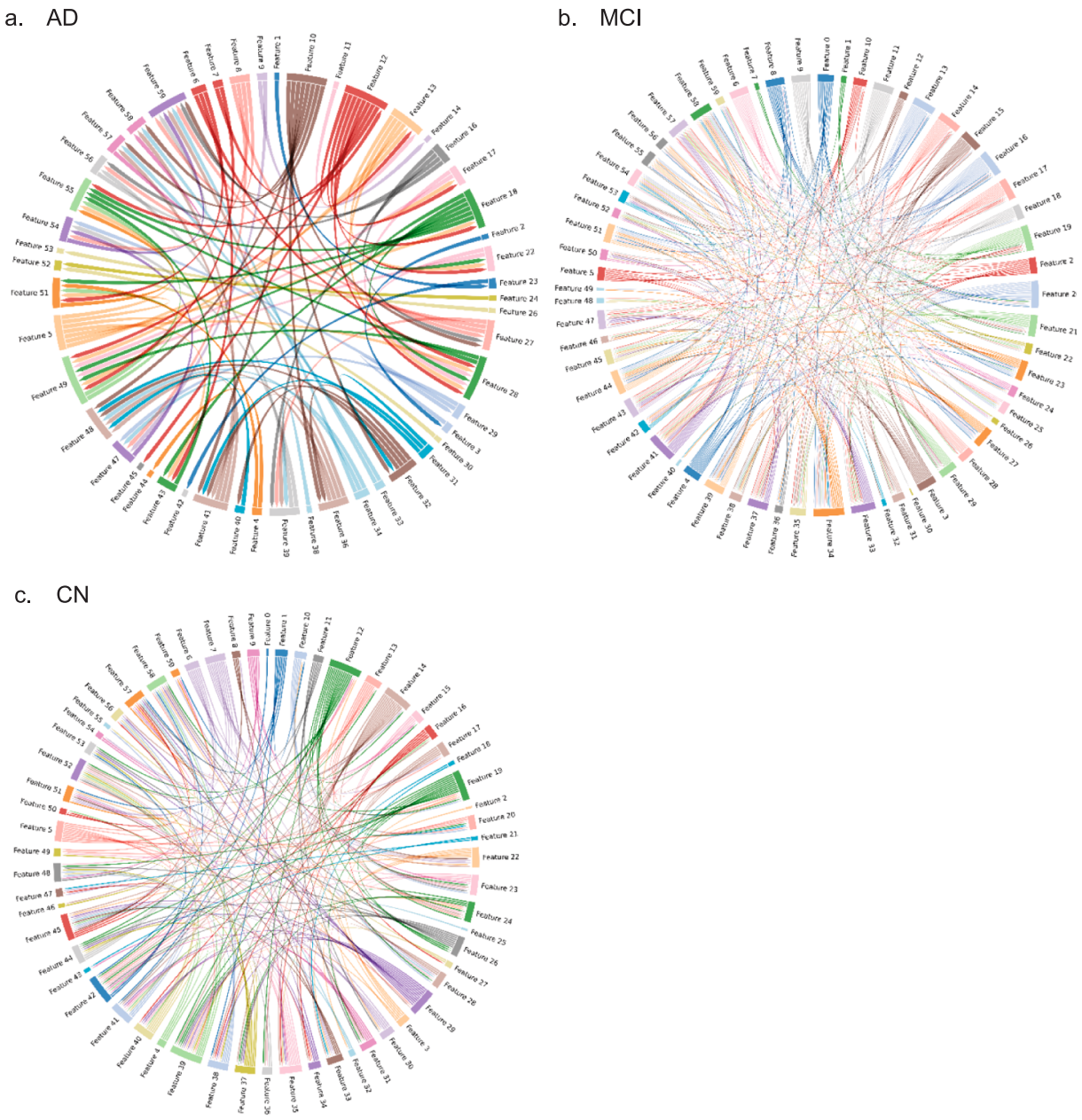


**Fig. 11.** Chord diagrams showing feature connectivity patterns for each diagnostic group: (a) AD, (b) MCI, and (c) CN. Each arc represents a feature pair with strong interactions captured by the attention mechanism.

MRI. However, pre-trained TimeSformer variants designed for shorter input sequences (e.g., 16 slices) offer potential compatibility.

Despite our proposed model's high performance, this study has a few limitations. First, our approach is solely based on structural MRI data and does not take into account multimodal inputs like clinical scores, cognitive assessments, or genetic biomarkers, which could improve diagnostic accuracy and clinical utility. Second, our model uses cross-sectional classification and does not account for disease progression, such as transitions from CN to MCI or MCI to AD. Modeling longitudinal changes necessitates follow-up imaging and progression labels, which are absent in our current dataset.

Future research should investigate the incorporation of multimodal data, encompassing cognitive assessments and genetic information, to enhance classification robustness and clinical significance. Furthermore, the integration of longitudinal MRI data would facilitate the creation of models capable of forecasting disease progression from CN to MCI and from MCI to AD. Progression-aware models may offer earlier and more tailored insights into disease trajectories, enhancing both diagnosis and prognosis in clinical environments.

## Conclusion

In this study, we introduced AlzFormer, a video-based deep learning framework that leverages spatiotemporal self-attention to classify Alzheimer's disease, mild cognitive impairment, and cognitively normal subjects using structural MRI. By treating MRI volumes as slice-level sequences and fine-tuning a pre-trained TimeSformer model, our approach effectively captured intra-slice spatial features and inter-slice dependencies. AlzFormer's overall accuracy on the test set was 94 %, with balanced class-wise F1-scores (AD: 0.94, MCI: 0.99, CN: 0.98) and an AUC of 1. This high classification performance on a test set demonstrates strong generalization under inter-scanner and inter-site variability. Furthermore, attention-based interpretability analyses revealed disease-relevant activation patterns, confirming that the model is

clinically relevant. For future work, we may use longitudinal data to model disease trajectories over time, as AD is, by nature, a progressive process. This would allow for early-stage diagnosis and tracking of Alzheimer's' progression.

## CRediT authorship contribution statement

**Taymaz Akan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Sara Akan:** Writing – review & editing, Writing – original draft. **Sait Alp:** Writing – review & editing, Writing – original draft. **Christina Raye Ledbetter:** Writing – review & editing. **Mohammad Alfrad Nobel Bhuiyan:** Writing – review & editing.

## Funding

This work was supported by an Institutional Development Award (IDeA) from the National Institutes of General Medical Sciences NIH under grant number P20GM121307 to MANB, NIH grants R01HL172970, R01HL145753, R01HL145753-01S1, and R01HL145753-03S1 to MSB; and Institutional Development Award (IDeA) from the National Institutes of General Medical Sciences of the NIH under grant number P20GM121307 and R01HL149264 to CGK. This project is also partially supported by Ike Muslow, MD, Endowed Chair in Healthcare Informatics of LSU Health Sciences Center Shreveport.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

Data collection and sharing for this project were funded by the Alzheimer's' Disease Neuroimaging Initiative (ADNI), the National Institutes of Health (Grant U01 AG024904), and the DOD ADNI Department of Defense (award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's' Association; Alzheimer's' Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Co.; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Co.; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development, LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research provides funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's' Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

### *Data availability*

The data are available on the ADNI website for download.

**Ethical Approval and Consent to participate:** Not applicable.

**Human Ethics:** Not applicable.

**Consent for publication:** Not applicable.

**Competing Interests:** The authors declare no potential competing interests.

### *Replication of results*

The codes and data used are available on request to enable the method proposed in the manuscript to be replicated by readers.

## References

Ahmed, S., Kim, B.C., Lee, K.H., Jung, H.Y., 2020. Ensemble of ROI-based convolutional neural network classifiers for staging the Alzheimer disease spectrum from magnetic resonance imaging. PLoS One 15 (12), e0242712. https://doi.org/10.1371/JOURNAL.PONE.0242712.

Akan, T., Alp, S., Bhuiyan, M.A.N., 2023. Vision Transformers and Bi-LSTM for Alzheimer's Disease Diagnosis from 3D MRI. In 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE) (pp. 530–535). IEEE.

Akan, T., Alp, S., & Bhuiyan, M.A.N., 2023. Vision Transformers and Bi-LSTM for Alzheimer's Disease Diagnosis from 3D MRI. In 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE) (pp. 530–535). IEEE.

Alinsaif, S., Lang, J., Initiative, A.D.N., 2021. 3D shearlet-based descriptors combined with deep features for the classification of Alzheimer's disease based on MRI data. Comput. Biol. Med. 138, 104879.

Alp, S., Akan, T., Bhuiyan, M.S., Disbrow, E.A., Conrad, S.A., Vanchiere, J.A., Bhuiyan, M.A.N., 2024. Joint transformer architecture in brain 3D MRI classification: its application in Alzheimer's disease classification. Sci. Rep. 14 (1), 8996.

Armañanzas, R., Iglesias, M., Morales, D.A., Alonso-Nanclares, L., 2017. Voxel-based diagnosis of Alzheimer's disease using classifier ensembles. IEEE J. Biomed. Health Inform. 21 (3), 778–784. https://doi.org/10.1109/JBHI.2016.2538559.

Avram, O., Durmus, B., Rakocz, N., Corradetti, G., An, U., Nittala, M.G., Halperin, E., 2024. Accurate prediction of disease-risk factors from volumetric medical scans by a deep vision model pre-trained with 2D scans. Nat. Biomed. Eng. 2024, 1–14. https://doi.org/10.1038/s41551-024-01257-9.

Backstrom, K., Nazari, M., Gu, I.Y.H., Jakola, A.S., 2018. An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images. Proceedings - International Symposium on Biomedical Imaging, 2018-April, 149–153. doi: 10.1109/ISBI.2018.8363543.

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., Filippi, M., 2019. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. NeuroImage: Clinical 21, 101645. https://doi.org/10.1016/J.NICL.2018.101645.

Bertasius, G., Wang, H., Torresani, L., 2021. Is space-time attention all you need for video understanding? In ICML (Vol. 2, p. 4).

Bron, E.E., Klein, S., Papma, J.M., Jiskoot, L.C., Venkatraghavan, V., Linders, J., Claassen, J.A.H.R., 2021. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. NeuroImage: Clinical 31, 102712.

Chua, J., Li, C., Antochi, F., Toma, E., Wong, D., Tan, B., Chen, C.L., 2025. Utilizing deep learning to predict Alzheimer's disease and mild cognitive impairment with optical coherence tomography. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 17 (1), e70041.

Cordonnier, J.-B., Loukas, A., Jaggi, M., 2019. On the relationship between self-attention and convolutional layers. arXiv preprint arXiv:1911.03584.

Ebrahimi, A., Luo, S., Chiong, R., 2021. Deep sequence modelling for Alzheimer's disease detection using MRI. Comput. Biol. Med. 134, 104537. https://doi.org/10.1016/J.COMPBIOMED.2021.104537.

Ebrahimighahnavieh, M.A., Luo, S., Chiong, R., 2020. Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. Comput. Methods Programs Biomed. 187, 105242. https://doi.org/10.1016/J.CMPB.2019.105242.

Estudillo-Romero, A., Haegelen, C., Jannin, P., Baxter, J.S.H., 2022. Voxel-based diktiometry: Combining convolutional neural networks with voxel-based analysis and its application in diffusion tensor imaging for Parkinson's disease. Hum. Brain Mapp. 43 (16), 4835–4851. https://doi.org/10.1002/HBM.26009.

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., Initiative, A.D.N., 2008. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 39 (4), 1731–1743.

Feng, J., Zhang, S.W., Chen, L., 2022. Extracting ROI-based contourlet subband energy feature from the sMRI image for Alzheimer's disease classification. IEEE/ACM Trans. Comput. Biol. Bioinf. 19 (3), 1627–1639. https://doi.org/10.1109/TCBB.2021.3051177.

Gelir, F., Akan, T., Alp, S., Gecili, E., Bhuiyan, M.S., Disbrow, E.A., (ADNI), A. D. N. I., 2024. Machine learning approaches for predicting progression to Alzheimer's disease in patients with mild cognitive impairment. J. Med. Biol. Eng. 1–21.

Goenka, N., Tiwari, S., 2022. AlzVNet: a volumetric convolutional neural network for multiclass classification of Alzheimer's disease through multiple neuroimaging computational approaches. Biomed. Signal Process. Control 74, 103500. https://doi.org/10.1016/J.BSPC.2022.103500.

Guan, H., Wang, C., Cheng, J., Jing, J., Liu, T., 2022. A parallel attention-augmented bilinear network for early magnetic resonance imaging-based diagnosis of

Alzheimer's disease. Hum. Brain Mapp. 43 (2), 760–772. https://doi.org/10.1002/HBM.25685.
Huang, F., Qiu, A., 2024. Ensemble vision transformer for dementia diagnosis. IEEE J. Biomed. Health Inform. https://doi.org/10.1109/JBHI.2024.3412812.
Jain, R., Jain, N., Aggarwal, A., Hemanth, D.J., 2019. Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images. Cogn. Syst. Res. 57, 147–159. https://doi.org/10.1016/J.COGSYS.2018.12.015.
Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., … Natsev, P., 2017. The kinetics human action video dataset. arXiv preprint arXiv: 1705.06950.
Khojaste-Sarakhsi, M., Haghighi, S.S., Ghomi, S.M.T.F., Marchiori, E., 2022. Deep learning for Alzheimer's disease diagnosis: a survey. Artif. Intell. Med. 130, 102332. https://doi.org/10.1016/J.ARTMED.2022.102332.
Kushol, R., Masoumzadeh, A., Huo, D., Kalra, S., Yang, Y.H., 2022. Addformer: Alzheimer's disease detection from structural mri using fusion transformer. Proceedings - International Symposium on Biomedical Imaging, 2022-March. doi: 10.1109/ISBI52829.2022.9761421.
Lian, C., Liu, M., Zhang, J., Shen, D., 2020. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE Trans. Pattern Anal. Mach. Intell. 42 (4), 880–893. https://doi.org/10.1109/TPAMI.2018.2889096.
Lin, Y., Li, X., Zhang, Y., Tang, J., 2023. Attention-based Efficient Classification for 3D MRI image of Alzheimer's Disease. In: In Proceedings of the 2023 6th International Conference on Sensors, Signal and Image Processing, pp. 34–39.
Liu, L., Liu, S., Zhang, L., To, X.V., Nasrallah, F., Chandra, S.S., 2023. Cascaded multi-modal mixing transformers for Alzheimer's disease classification with incomplete data. Neuroimage 277, 120267. https://doi.org/10.1016/J.NEUROIMAGE.2023.120267.
Liu, F., Yuan, S., Li, W., Xu, Q., Sheng, B., 2023. Patch-based deep multi-modal learning framework for Alzheimer's disease diagnosis using multi-view neuroimaging. Biomed. Signal Process. Control 80, 104400. https://doi.org/10.1016/J.BSPC.2022.104400.
Loddo, A., Buttau, S., Di Ruberto, C., 2022. Deep learning based pipelines for Alzheimer's disease diagnosis: a comparative study and a novel deep-ensemble method. Comput. Biol. Med. 141, 105032. https://doi.org/10.1016/J.COMPBIOMED.2021.105032.
Nemoto, K., Sakaguchi, H., Kasai, W., Hotta, M., Kamei, R., Noguchi, T., Asada, T., 2021. Differentiating dementia with lewy bodies and Alzheimer's disease by deep learning to structural MRI. J. Neuroimaging 31 (3), 579–587. https://doi.org/10.1111/JON.12835.
Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., 2011. Statistical parametric mapping: the analysis of functional brain images. Elsevier.
Rangaraju, B., Chinnadurai, T., Natarajan, S., Raja, V., 2024. Dual attention aware octave convolution network for early-stage Alzheimer's disease detection. Information Technology and Control 53 (1), 302–316.
Sharma, R., Goel, T., Tanveer, M., Murugan, R., 2022. FDN-ADNet: Fuzzy LS-TWSVM based deep learning network for prognosis of the Alzheimer's disease using the sagittal plane of MRI scans. Appl. Soft Comput. 115, 108099. https://doi.org/10.1016/J.ASOC.2021.108099.
Shi, D., Yao, X., Li, Y., Zhang, H., Wang, G., Wang, S., Ren, K., 2022. Classification of Parkinson's disease using a region-of-interest- and resting-state functional magnetic resonance imaging-based radiomics approach. Brain Imaging Behav. 16 (5), 2150–2163. https://doi.org/10.1007/S11682-022-00685-Y/FIGURES/6.
Shinde, S., Prasad, S., Saboo, Y., Kaushick, R., Saini, J., Pal, P.K., Ingalhalikar, M., 2019. Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI. NeuroImage: Clinical 22, 101748. https://doi.org/10.1016/J.NICL.2019.101748.
Solana-Lavalle, G., Rosas-Romero, R., 2021. Classification of PPMI MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease. Comput. Methods Programs Biomed. 198, 105793. https://doi.org/10.1016/J.CMPB.2020.105793.
Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A.D.N., 2017. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med. Image Anal. 37, 101–113.
Van Vliet, D., De Vugt, M.E., Bakker, C., Pijnenburg, Y.A.L., Vernooij-Dassen, M., Koopmans, R., Verhey, F.R.J., 2013. Time to diagnosis in young-onset dementia as compared with late-onset dementia. Psychol. Med. 43 (2), 423–432.
Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Colliot, O., 2020. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Med. Image Anal. 63, 101694. https://doi.org/10.1016/J.MEDIA.2020.101694.
Zhang, Y., Teng, Q., Liu, Y., Liu, Y., He, X., 2022. Diagnosis of Alzheimer's disease based on regional attention with sMRI gray matter slices. J. Neurosci. Methods 365, 109376. https://doi.org/10.1016/J.JNEUMETH.2021.109376.
Zhao, H., Jia, J., Koltun, V., 2020. Exploring self-attention for image recognition. In: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10076–10085.