

ORIGINAL RESEARCH

Open access

Multi-Task Learning and Attention-Enhanced 3D CNNs for Alzheimer's disease Classification

Nora Schneider^{1*}, Tobias Frank¹, Andreas Müller², Christoph Meier³

Abstract

In recent years, deep learning has emerged as a promising framework for identifying Alzheimer's disease (AD) within clinical neuroimaging. Nonetheless, conventional architectures frequently fail to optimize feature extraction from medical scans and underutilize auxiliary data resources vital for diagnostic categorization. To resolve these challenges, this paper introduces a deep three-dimensional convolutional neural network architecture that integrates attention mechanisms with multi-task learning. An enhanced primary C3D network is deployed to generate initial, coarse low-level feature maps. This design implements a novel convolutional block tailored to capture the specific structural properties of magnetic resonance imaging (MRI) data, along with a secondary block configured to compute localized attention weights for distinct pixel coordinates within the feature map, which are subsequently combined with the map's output. Thereafter, a series of fully connected layers implements multi-task learning, producing three distinct outputs that encompass the primary classification objective. During the training phase, the two auxiliary outputs leverage backpropagation to refine and enhance the primary classification performance. Empirical results demonstrate that this proposed framework outperforms existing state-of-the-art models for AD detection, yielding superior classification accuracy and enhanced performance metrics on the Alzheimer's disease Neuroimaging Initiative dataset. These outcomes illustrate a compelling direction for prospective diagnostic classification research.

Keywords Deep learning, Alzheimer's disease (AD), Magnetic resonance imaging (MRI), Multi-task learning

*Correspondence:

Nora Schneider

nora.schneider@natfak.uni-freiburg.de

¹ Department of Neuroimaging and Multi-Task Learning, Faculty of Medicine, University of Freiburg, Freiburg, Germany

² Department of Alzheimer's Disease Classification and Attention Mechanisms, Faculty of Medicine, Heidelberg University, Heidelberg, Germany

³ Department of 3D CNN Architectures for Medical Imaging, Faculty of Engineering, University of Hohenheim, Stuttgart, Germany

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative pathology that presents substantial diagnostic difficulties during its initial presentation. Although AD represents the leading etiology of dementia, the syndrome can also be precipitated by alternative pathological conditions. Because differentiating among specific dementia subtypes is frequently complicated, establishing a precise estimation of Alzheimer's dementia prevalence remains problematic. It is widely recognized that AD constitutes approximately two-thirds of all clinical dementia diagnoses. Projections derived from mathematical modeling of future dementia rates indicate that an expanding proportion of the population will be affected, driven by shifting variables such as rising life expectancy, evolving mortality patterns, and the prevalence of cardiovascular diseases. The traditional framework for diagnosing AD via magnetic resonance imaging (MRI) evaluations is labor-intensive and inherently subjective. Consequently, the development of a more streamlined and objective early diagnostic methodology is imperative. Deep learning (DL) strategies applied to MRI datasets have shown great potential

for categorizing AD patients; however, maximizing accuracy and minimizing technical constraints remain critical objectives. To reliably differentiate between AD, mild cognitive impairment (MCI), and healthy control (HC) cohorts using MRI data, this investigation introduces a DL model built upon 3D convolutional neural networks (CNNs). This system is designed to heighten the precision of AD diagnoses while overcoming the inherent boundaries of traditional machine learning (ML) paradigms. Ultimately, this research seeks to advance effective, unbiased diagnostic frameworks for AD and clarify the utility of DL methodologies in neurodiagnostic applications.

To address the shortcomings of conventional ML techniques in diagnosing AD, we introduce an end-to-end classification framework driven by a deep three-dimensional CNN. This architecture integrates an attention strategy to maximize its capacity for feature extraction. Furthermore, it incorporates twin auxiliary subtasks—specifically, Clinical Dementia Rating (CDR) scale score regression and Mini-Mental State Examination (MMSE) score regression—to optimize the primary AD classification performance and facilitate multi-task learning. The CDR Scoring Table offers descriptive reference points to assist clinicians in assigning appropriate scores based on clinical interviews and diagnostic judgment. This table is applied in tandem with CDR Worksheets. Utilizing a dedicated scoring algorithm or standard scoring protocols, a global CDR Score can be derived alongside individual domain ratings. Because the CDR is a copyrighted diagnostic instrument, its utilization requires an appropriate license. The Global CDR Score is employed to define and monitor the progression of a patient's cognitive impairment or dementia status, where 0 indicates healthy cognition, 0.5 signifies questionable or very mild dementia, 1 represents mild dementia, 2 denotes moderate dementia, and 3 indicates severe dementia.

The Mini-Mental State Examination (MMSE) serves as a brief cognitive screening tool that offers a rapid, dependable measurement of an individual's cognitive status. It is frequently applied to identify cognitive deficits, quantify impairment severity, and track cognitive changes over time. In both clinical practice and research settings, the MMSE is the most ubiquitously deployed brief cognitive assessment. The tool evaluates multiple cognitive dimensions, including language production, visuospatial construction, orientation, repetition, verbal recall, attention, and calculation. The proposed pipeline begins with data standardization to achieve data enhancement. Next, an attention mechanism is incorporated, and the 3D convolutional network is executed to extract features contextualized by attention weights, thereby capturing the most informative spatial layout information for the classification task. Finally, to refine the performance of the primary classification task, multiple fully connected (FC) layers undergo iterative backpropagation to simultaneously resolve the classification output, the CDR score, the MMSE score, and the two auxiliary subtasks.

Convolutional neural networks are capable of processing image data by compressing large volumes of raw information into compact representations while effectively maintaining the structural traits of the original image. Mirroring the architecture of standard neural networks, CNNs are comprised of input, hidden, and output layers. Following convolutional operations, the high-dimensional properties of the scan are managed by partitioning the feature matrix into distinct localized blocks to calculate their maximum or average values. This step reduces spatial dimensionality, accelerates network processing, and suppresses overfitting. Because the dimensional features of the scan remain highly complex after convolution, the feature matrix is segmented into separate blocks to compute their maximum or average values, which assists in minimizing feature dimensionality, accelerating processing speed, and preventing overfitting. The final layer of the CNN model, the fully connected layer, generates the definitive class probabilities by aggregating all localized features and the individual channel feature matrices into unified vector representations.

As demonstrated by both ablation and comparative testing, the proposed architecture functions efficiently on the AD classification task, showing that the DL model successfully isolates critical information from medical scans. Comparative experiments confirm that this strategy achieves superior performance in AD classification relative to conventional ML methodologies. This framework may aid the development of more efficient and objective clinical pipelines for AD diagnosis. Before the early 2000s, the sole definitive technique for validating an AD diagnosis was a post-mortem autopsy. Today, laboratory assays and neuroimaging modalities are widely accessible, allowing clinicians and researchers to observe biological expressions of the pathology—or biomarkers—in living patients. At present, blood tests should be used as a complement to other diagnostic procedures rather than as a standalone diagnostic solution for dementia. However, these advanced diagnostic protocols are not yet universally accessible. Research teams funded by the National

Institute on Aging continue to investigate pathways toward more precise, less invasive, and faster methodologies for diagnosing Alzheimer's.

A variety of domestic and international researchers have pursued classification studies leveraging MRI scans from individuals with AD. Conventional ML techniques typically depend on manual or semi-manual feature engineering to establish a baseline for classification. Varying feature sets generate divergent classification performances, and the requirements for classifier selection alter accordingly. For example, literature [1, 2] coupled the volumes of the hippocampus and the entorhinal cortex as input features, using Support Vector Machines (SVM) to execute AD classification; similarly, literature [3] introduced a multi-stage classification protocol for AD detection where a Gaussian Naive Bayesian classifier is utilized in the initial stage. Following the initial classification results, specific samples are then passed to an SVM or KNN classifier for refined reclassification. These older classification techniques risk losing vital data during the feature selection stage, presenting substantial constraints. In recent years, a succession of highly sophisticated and optimized network architectures, including Alex-Net [4], VGGNet [5], and GoogLeNet [6], have been introduced.

Although these aforementioned strategies have yielded strong results in AD classification, they fail to maximize the utility of all data relevant to AD diagnostics. First, while multiple 3D image classification algorithms exist within deep learning, the computational burden of 3D convolutional network processing is exceptionally high. Consequently, most AD classification workflows still rely on 2D convolutional learning applied to individual 3D image slices, which causes a loss of critical three-dimensional structural data. Second, clinical diagnosis of AD depends not only on MRI scans but also necessitates alternative auxiliary indicators. However, these additional clinical metrics are currently neglected in the majority of AD classification tasks, leading to an underutilization of informational resources. Alzheimer's disease remains a time-dependent, deteriorating neurological condition that is difficult to detect in its initial stages. The traditional method of diagnosing AD via MRI analysis is labor-intensive and subjective. Thus, a more efficient and objective approach to early diagnosis is essential. While DL methods applied to MRI scans have demonstrated great promise for categorizing AD patients, improvements in accuracy and the reduction of technical constraints remain imperative.

To address the aforementioned challenges, this paper presents an end-to-end framework for AD classification founded upon a deep three-dimensional CNN (CNN). This method integrates an attention mechanism to amplify feature extraction proficiency and introduces two complementary subtasks: Clinical Dementia Rating (CDR) Scale (CDR) score regression and Mini-Mental State Examination (MMSE) score regression. These auxiliary pathways serve to optimize primary AD classification performance and enable multi-task learning. Within this pipeline, data standardization is first applied to achieve data enhancement. Following this, the attention architecture is introduced, and the 3D convolutional network is deployed to yield feature representations embedded with attention weights, thereby isolating the most informative spatial layout attributes for classification objectives. To refine the output generated by the primary classification task, multiple fully connected (FC) layers carry out iterative backpropagation, ultimately resolving the definitive classification results, the CDR scores, the MMSE scores, and the dual auxiliary subtasks. Ablation and comparative experiments confirm that the proposed methodology exhibits strong performance on the AD classification task, operating as an effective DL model capable of isolating clinical insights from neuroimaging scans. This study presents a DL model for AD diagnosis using neuroimaging features that can select features automatically and discover the optimal network architecture. The overarching goal is to fill the existing research gap by introducing a streamlined and highly capable DL model for AD diagnosis that outperforms current state-of-the-art architectures, thereby supporting the early prevention of AD and reducing clinical workloads.

Related Work

The introduction of ResNet [7] fundamentally advanced the progress of DL, establishing it as a widely adopted framework within the domain of medical image classification [8]. Nevertheless, a substantial architectural discrepancy exists between networks optimized for medical images versus those designed for natural scene classification. Natural image classification typically necessitates deeper network topologies due to the extensive variety of category targets.

Conversely, medical imaging models generally target highly specific pathological classifications (such as identifying tumors) and must place a greater priority on granular, detailed localized information. Grounded in DL principles, research in literature [9] modified the layer configuration of VGG16 to realize an enhanced AD classification outcome. Similarly, the authors of literature [10] utilized standard 3D convolutional architectures alongside residual convolutional blocks to extract features from MRI scans; based on a lightweight engineering principle, they conducted numerous empirical trials to determine the optimal network depth and the specific number of channels per layer. In other work, literature [11–13] constructed a novel 3D convolutional architecture designated as HadNet, utilizing Bayesian optimization to fine-tune the model's hyperparameters.

Evaluating high-dimensional feature spaces forms the foundational core of the diagnostic techniques applied in AD classification, a process that is frequently susceptible to the curse of dimensionality. Feature selection operates as an effective strategy to reduce feature volume by systematically isolating and removing non-informative attributes. Current AD classification frameworks employ a variety of feature selection strategies, including multi-task feature selection, group lasso, and principal component analysis. Within multi-modal architectures, class labels are commonly predicted by executing a direct combination of these isolated traits from each respective modality. When utilizing a classical multi-modal AD fusion methodology, these diverse modalities are typically combined via linear integration. To properly characterize and isolate the contextual relevance across disparate data streams, sample significance analysis plays a vital role within the multi-modal fusion process.

The existing literature reviews a diverse selection of DL models proposed for diagnosing AD, including convolutional neural networks, recurrent neural networks (RNNs), and autoencoders. These prior studies analyze the distinct advantages and structural limitations of each framework while emphasizing the ongoing challenges tied to feature extraction and selection. Driven by the rapid expansion of the internet, organizations can now aggregate vast datasets from the Internet of Things (IoT), social media streams, and alternative heterogeneous sources in a wide array of formats. This continuous, accelerated expansion of data volume creates large-scale datasets characterized by massive dimensionality. In the modern era of big data, feature selection techniques have become increasingly critical because they help mitigate dimensionality friction and prevent overfitting throughout the learning process. Given that certain algorithmic designs are fundamentally better suited to specific data profiles than others, it is advisable to implement all applicable models and systematically evaluate their comparative outcomes. Hyperparameter tuning represents the iterative process of isolating the most optimal value for each specific hyperparameter to elevate overall model accuracy. Successfully calibrating these variables demands a comprehensive understanding of their underlying definitions and their specific influence on the global network architecture—a process that can be replicated across multiple top-performing models. Furthermore, ensemble techniques represent the most heavily utilized strategy among winning solutions in data science competitions. This methodology achieves superior predictive outputs by simply integrating the independent predictions of multiple individual models. This aggregation is typically executed using either boosting or bagging (Bootstrap Aggregating). Finally, cross-validation stands as one of the most vital components of data modeling workflows. Before finalizing an architecture, it is standard practice to validate the model's generalizability on an independent sample that was strictly withheld from the training phase.

In Ref. [14], an intelligent, personalized framework is introduced that leverages ML techniques to predict and prevent mental health conditions by integrating multi-source features, including biometric data, lifestyle habits, behavioral patterns, and social media footprints. Machine learning is recognized for introducing a vast array of distinct learning paradigms and methodologies, with supervised and unsupervised learning serving as two ubiquitous examples. Supervised learning utilizes feedback from labeled datasets to predict specific target outcomes, making it uniquely suited for addressing classification and regression tasks. The core objective of this learning paradigm is to interpret data with respect to designated metrics. By contrast, unsupervised learning acts independently of external labels, seeking to map and interpret internal data structures on its own. Unsupervised learning operates without predefined training benchmarks or external structural rules. To resolve specialized problem sets, individual classifiers are often meticulously combined and structured through ensemble learning processes. Ensemble learning is generally implemented to maximize global model performance or to reduce the risk of selecting an underperforming individual network. Additionally, neural networks

and deep learning have recently gained widespread favor among ML methodologies due to their superior capability to resolve complex problems across fields like image identification, audio recognition, and natural language processing. By mimicking the functional mechanisms of biological neural networks within the brain, these computational techniques enable systems to learn directly from observational datasets.

In Ref. [15], the authors introduce a DL architecture called 3ACL designed to classify brain tumors using 3D MRI data, demonstrating strong accuracy metrics that outperform contemporary models. By converting high-dimensional 3D MRI arrays into a 2D space using the 3t2FTS-v2 protocol, the system produces highly discriminative 2D-1D image mappings embedded with increased diagnostic significance. In alternative methods like the 3t2FTS or 3t2FTS-v1 protocols, first-order statistics (FOS) traits are calculated to construct 2D-1D image representations that act as a unique identifier for a 3D voxel. However, these older transition techniques restrict their scope exclusively to FOS traits. They fail to evaluate supplementary feature extraction frameworks or normalization methods that could potentially generate more robust 2D-1D images by introducing distinct, discriminatory data. It is apparent that the core inspiration and structural design of this methodology can be successfully extended to analyze both 3D brain tissue structures and localized tumor volumes. Moreover, all iterations of these frameworks can be applied to investigate 3D-defined neurological disorders like Alzheimer's disease as well as oncological malignancies. Consequently, both versions can be leveraged to activate traditional ML frameworks (such as deep learning and transfer learning) that are built upon 2D analytical evaluations.

The investigation presented in Ref. [16] delivers a comprehensive analysis of ML algorithms tailored for the early identification of mild cognitive impairment and AD, highlighting encouraging outcomes across a variety of research initiatives. During the initial phases of the pathology, both pharmacological and nonpharmacological interventions show high efficacy in mitigating cognitive and behavioral symptoms. Within this developmental stage, it is critical to conceptualize AD as a continuous spectrum; individuals exhibiting MCI who may subsequently transition to AD dementia already harbor the underlying pathology, even if their overt cognitive symptoms have not yet fully materialized. This reality highlights the urgent necessity of successfully differentiating between MCI patients who will eventually convert to AD dementia and those who will remain clinically stable.

According to Ref. [17], a deep learning model is introduced that surpasses existing algorithms by learning an underlying latent feature representation of data to optimize AD classification. Convolutional adversarial autoencoders are presented in Ref. [18] as a specialized technique to enhance classification accuracy over existing frameworks while minimizing data variability in multi-center AD classification tasks. Utilizing a hybrid framework that merges self-supervised and semi-supervised learning strategies, the authors of [19] introduce an approach for medical image classification that outperforms contemporary models and establishes high accuracy rates. In Ref. [20], the authors advocate for the use of convolutional autoencoders as a powerful DL technique to map the intricate structural pathology of AD and classify the disease with high precision. The model introduced in Ref. [21] provides an architecture for AD classification that achieves excellent accuracy rates and outperforms historical algorithms by imputing missing data arrays within multimodal brain imagery. The research presented in Ref. [12] introduces an ensemble model for AD categorization that aggregates multiple deep networks, yielding a higher degree of accuracy than current standalone algorithms. Furthermore, the study in Ref. [13] presents an ensemble composed of deep neural networks trained via transfer learning for AD classification, demonstrating performance that surpasses existing models.

Recent DL configurations optimized for AD diagnosis using neuroimaging features include 3D RSA-DNN [22], dual attention multi-instance DL [23], NMF-TDNet [24], 3D CNN-based multichannel contrastive learning [25], a lightweight framework [26], AlzheimerNet [27], DEMNET [28], multi-stream CNN [29], the weakly supervised and multimodal SDPN [30], and a deep multi-task multi-channel learning framework [31]. The diagnostic accuracy recorded across these various models spanned a range from 85.6% to 92.96%.

A multitude of research studies have explored the synergy between deep learning models and neuroimaging modalities for diagnosing AD and mild cognitive impairment (MCI). Li et al. [32] introduced a deep sparse autoencoder network that demonstrated strong performance when fusing structural MRI scans with positron emission tomography data. PET

functions as a minimally invasive, functional imaging technology that utilizes positron-emitting radiotracers to capture critical insights regarding human biology and localized biochemistry. Because it operates as an inherently three-dimensional (3D) imaging approach, it requires detector arrays featuring high detection efficiencies alongside enhanced spatial resolution. Mehta *et al.* [33] engineered a cascaded DL framework that achieved optimized performance levels on medical imaging repositories, including the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Faisal and Kwon [34] formulated an automated identification system driven by a deep CNN model that yielded strong performance metrics when validated on the ADNI-2 dataset. Gamal *et al.* [35] proposed an ensemble of 3D CNN models designed for the early detection of AD. Finally, Liu *et al.* [36] demonstrated an accuracy rate of 85.6% on the ADNI dataset by leveraging a multimodal neuroimaging feature learning strategy tailored for the multiclass diagnosis of AD. Together, these collective discoveries confirm that the integration of DL architectures and neuroimaging methods holds immense potential for the early identification and personalized therapeutic management of progressive brain disorders such as Alzheimer's and Parkinson's.

For the prediction of MCI conversion, relying on a single-modal neuroimage may prove insufficient as it captures merely a localized component of the structural changes associated with cerebral atrophy. Conversely, multi-modal neuroimaging provides supplementary data points that can be integrated to illuminate the synergistic relationships among distinct neuroimaging modalities. Multi-modal fusion, which encompasses early, late, and hybrid fusion strategies, represents a prominent frontier within multi-modal ML. Multimodal learning framework principles are frequently applied across fields such as image categorization and image registration. The academic interest in multi-modal fusion arises from two primary benefits. First, the observation of identical pathological phenomena across multiple modalities can generate highly robust predictive outcomes. Second, to enhance the precision of classification results, complementary informative features can be isolated from diverse modalities through the utilization of Sparse Autoencoders. In particular, deep learning frameworks—which have recently driven substantial breakthroughs in AI technology—have demonstrated highly positive effects when applied to genomic medicine and healthcare environments. The objective of computer-aided AI approaches, such as DL architectures, is to facilitate data-driven algorithms capable of comprehensively improving the diagnostic accuracy of AD through the analysis of neuroimaging and/or genetic data. For instance, by leveraging neuroimaging inputs like MRI data, CNNs have been effectively integrated into DL models to classify AD. Because these encouraging methodologies can be seamlessly extended to heterogeneous data types, including neuroimaging and genomics, DL models have been projected to assume a pivotal role in the diagnosis and prospective prediction of AD in upcoming investigations.

The progressive neurological impairment characteristic of AD profoundly disrupts memory, linguistic capabilities, and emotional stability. The precise underlying etiology of AD remains elusive, and traditional clinical diagnostic pathways remain subjective and labor-intensive. To prevent AD from advancing to severe stages and to halt pathological progression, early identification is critical. Although MRI represents a cornerstone modality in diagnosing Alzheimer's disease, traditional ML methodologies applied to MRI classification suffer from constrained feature selection mechanisms, leading to inconsistent classification performance. Deep learning architectures have recently yielded highly encouraging outcomes in AD diagnosis; nonetheless, optimizing their accuracy and overall performance remains a critical objective. The existing literature details various DL configurations proposed for AD diagnosis, outlining their distinct structural advantages alongside persistent technical challenges. Multiple contemporary investigations have analyzed the application of DL and neuroimaging techniques for identifying AD and mild cognitive impairment (MCI). These architectures achieved diagnostic accuracies ranging from 85.6% to 92.96%. This paper seeks to resolve the current deficit of efficient AD diagnostic methodologies by introducing a DL framework capable of automated feature selection and autonomous network structure optimization. The overarching goal is to advance early AD prevention strategies and alleviate clinical workloads. Broadly, this review underscores the profound capacity of DL and neuroimaging workflows to enable personalized therapeutic strategies for neurodegenerative disorders like AD and MCI.

Alzheimer's disease represents the leading etiology of dementia, accounting for 60%–80% of all diagnosed cases. The clinical trajectory of Alzheimer's disease initiates with mild cognitive impairment (MCI) and systematically deteriorates into a neurodegenerative form of dementia. This pathology actively impairs cognitive processing, damages neuronal

networks, induces memory deficits, and severely compromises the execution of basic daily activities. Structurally, Alzheimer's disease is classified as a multifaceted, progressive neurological brain disorder. Notably, individuals diagnosed with MCI exhibit a significantly higher susceptibility to developing AD relative to the general population. At present, the absence of a definitive disease-modifying therapy for AD constitutes the primary obstacle confronting researchers in this discipline. Nevertheless, contemporary therapeutic interventions can mitigate symptom severity or delay the manifestation of new clinical signs. Consequently, it is imperative to identify AD during its early, prodromal phase. Achieving precise and early identification of AD via computer-aided diagnostic instruments serves to lower the substantial economic burdens associated with patient care—costs that are projected to increase dramatically. Traditional ML techniques applied to early AD diagnosis frequently depend on two primary feature extraction paradigms: region of interest (ROI)-based features and voxel-based features.

While an extensive body of literature has focused on deploying DL and neuroimaging frameworks for diagnosing AD and mild cognitive impairment (MCI), a notable gap persists regarding the integration of heterogeneous features—such as biometric data, lifestyle factors, behavioral patterns, and social media footprints—to engineer personalized intelligent systems for predicting and preventing mental health issues. Machine learning constitutes an AI strategy that leverages a diverse array of methodologies to guide algorithmic learning. Within the healthcare sector, supervised learning, unsupervised learning, and DL represent the three most frequently implemented learning paradigms. Additional ML frameworks include reinforcement learning and semi-supervised learning, the latter of which merges elements of both supervised and unsupervised learning. In these frameworks, the algorithm operates as an agent within an interactive environment, learning iteratively through a process of error commission and reward acquisition based on performance milestones. Rather than differentiating rigidly between samples and source populations, ML techniques isolate descriptive informational patterns within data arrays to project clinical outcomes for individual patients. While the descriptive components of standard statistics share similarities with ML, the inferential dimension—the core of statistical science—remains distinct because it relies strictly on sample data to formulate generalized conclusions regarding the wider population from which the sample was drawn. Modern ML approaches surpass conventional statistical techniques due to their capacity to map complex, non-linear, high-dimensional interactions that can significantly enhance predictive power. Furthermore, given the exceptional accuracy rates achieved by lightweight DL frameworks and multimodal SDPNs in alternative medical image classification tasks, their diagnostic utility must be thoroughly investigated for AD detection. Finally, because deep multi-task and multi-channel learning architectures have demonstrated strong performance across other medical image processing domains, their potential for optimizing AD diagnostics warrants detailed evaluation.

Proposed Work

The presented study addresses the current lack of a highly effective methodology for diagnosing AD. Traditional AD diagnosis reliant on medical neuroimaging is fundamentally subjective and time-consuming. Conventional ML techniques are subject to clear limitations, such as manual feature selection protocols that risk discarding vital structural data. Although Alzheimer's disease stands as the most prevalent form of dementia, clinical medicine currently lacks treatments capable of arresting its progression. This scarcity of effective therapeutic interventions is directly linked to a deficit of reliable disease endpoints and definitive biomarkers. Implementing ML models to analyze EEG recordings represents one strategy to overcome the constraints inherent to current diagnostic modalities. However, persistent challenges include the necessity for substantial sample sizes, precise EEG source localization utilizing high-density acquisition arrays, a lack of automation, and the subjective nature of artifact removal. Another complicating factor is the utilization of sub-optimal feature selection within black-box ML frameworks like deep neural networks. To bridge these gaps, this paper proposes a DL methodology for AD diagnosis that leverages neuroimaging properties to autonomously select informative features and isolate the optimal network topology. By introducing a DL model for AD diagnosis that outperforms existing state-of-the-art architectures, this work aims to resolve current academic limitations, support the early prevention of AD, and mitigate diagnostic burdens on clinicians.

The structural workflow of the proposed AD classification methodology is illustrated in **Figure 1**. The architecture principally comprises a foundational network module, an attention module, and a multi-task learning module. The primary network module is engineered for feature extraction and integrates a total of 5 convolutional blocks (Block). The attention module is embedded directly within the midsection of the primary network to amplify the model's capacity to isolate sufficient visual data features. Concurrently, the multi-task learning module is integrated at the terminus of the primary network to enrich the framework with data derived from two supplementary tasks. Convolutional blocks serve as the fundamental structural units of a CNN, which is utilized for image recognition tasks. These blocks consist of one or more convolutional layers designed to isolate and extract key characteristics from the input image. To compress the spatial dimensions of the generated feature maps while maintaining critical informational attributes, one or more pooling layers are frequently deployed downstream from the convolutional layers. Throughout the operation of these convolutional layers, dedicated filters are swept across the input image to isolate features such as edges, textures, and geometric shapes. Subsequently, the feature maps undergo downsampling via the application of pooling layers to the convolutional outputs. By minimizing the spatial dimensions of these feature maps, the computational efficiency of the network is significantly enhanced, and the propensity for overfitting is reduced. To generate a definitive prediction or classification for the image, one or more fully connected layers are sequentially applied to the outputs of the pooling layers. In the initial phases of a CNN architecture, convolutional blocks are utilized when the primary goal centers on feature extraction from the raw input image. These blocks are frequently replicated multiple times throughout the network, with each sequential block capturing increasingly granular structural detail from the output of the preceding layer. This hierarchical organization enables the network to learn progressively detailed features, thereby advancing its overall image recognition capabilities.

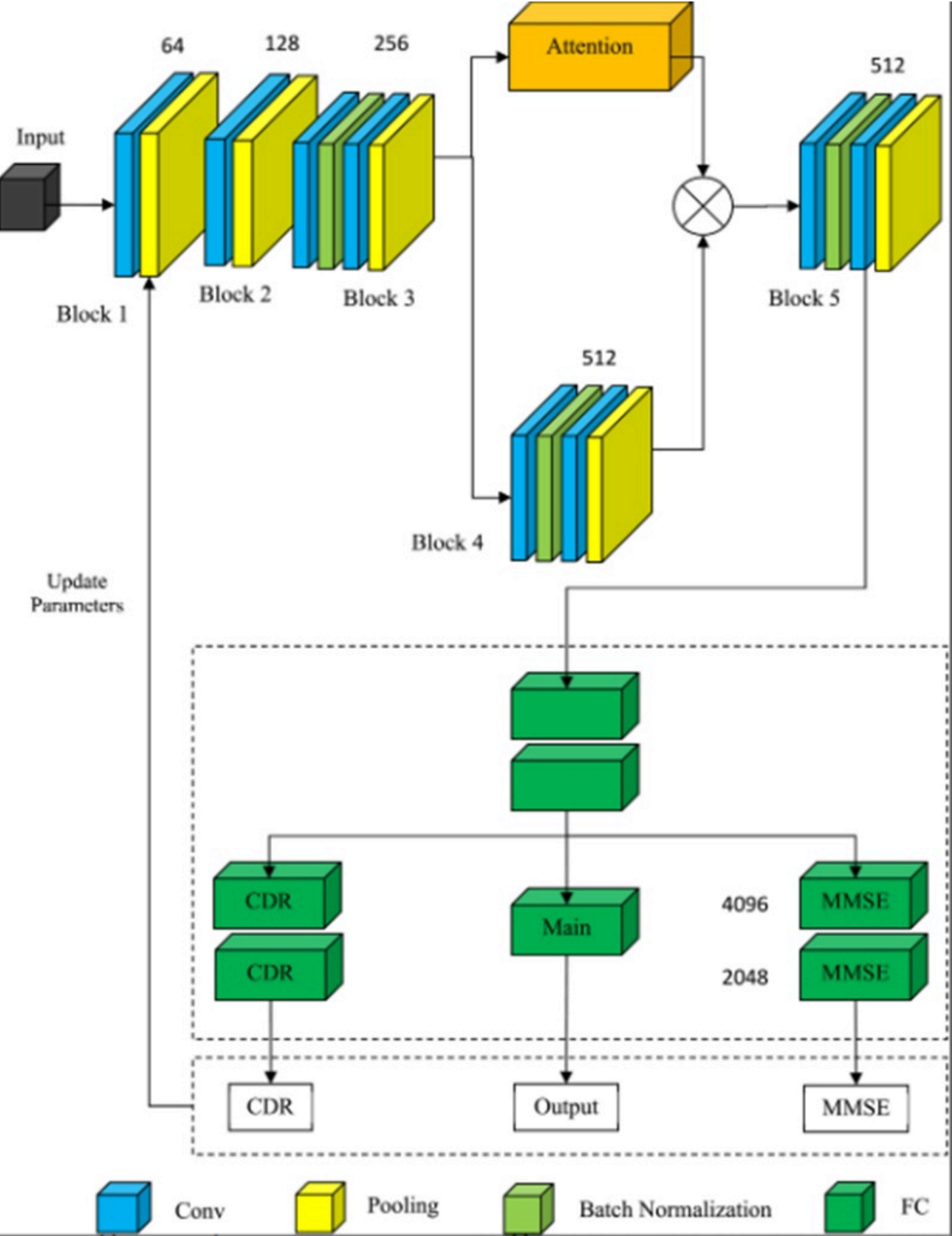


Figure 1. 3D convolutional network based on the attention mechanism and multi-task learning.

Basic network

Brain MRI scans are inherently three-dimensional data arrays. A large volume of historical research has relied heavily on the two-dimensional slice extraction of brain MRI volumes; nonetheless, the clinical preprocessing required to isolate 2D slice profiles is labor-intensive and highly prone to discarding critical structural features. To reliably categorize AD, MCI, and HC cohorts using MRI data, this investigation introduces a DL model built upon 3D CNNs. This architecture is designed to heighten the precision of AD diagnoses while overcoming the inherent boundaries of traditional machine learning (ML) paradigms. Ultimately, this research seeks to advance effective, unbiased diagnostic frameworks for AD and clarify the utility of DL methodologies in neurodiagnostic applications. The presented study addresses the current lack of a highly effective methodology for diagnosing AD. Traditional AD diagnosis reliant on medical neuroimaging is fundamentally subjective and time-consuming. Conventional ML techniques are subject to clear limitations, such as manual feature selection protocols that risk discarding vital structural data. The study proposes a DL model for AD diagnosis using neuroimaging features that can select features automatically and discover the optimal network architecture. The overarching goal is to fill the existing research gap by introducing a streamlined and highly capable DL model for AD diagnosis that outperforms current state-of-the-art architectures, thereby supporting the early prevention of AD and reducing clinical workloads.

Drawing upon insights from established research paradigms, this paper constructs a three-dimensional computational volume. The architectural concept of three-dimensional convolutional networks was pioneered in literature [37, 38] to address the demands of human action analysis and kinematic recognition. In comparison to two-dimensional convolutional operations, three-dimensional convolutions exhibit a superior capacity to capture sequential configurations and preserve a larger density of features. In this framework, the C3D network structure [24, 25, 37] is selected as the primary backbone network to retrieve the essential characteristics of the neuroimaging data. The C3D architecture functions as a classical, general-purpose framework that was initially deployed for behavior recognition and video feature extraction—marking it as a foundational network model in that domain. Due to the massive computational overhead associated with processing 3D MRI scans and the stark divergence between clinical neuroimaging repositories and standard action recognition datasets, this paper adapts the network topology to facilitate a more rapid and precise deployment of C3D for AD classification tasks. The definitive network configuration is prescribed in Table 1.

Table 1. Structure of improved C3D network.

Structural units	Improve network
Block 1	$3 \times 3 \times 3$, 64, Conv1a
	$1 \times 2 \times 2$, Pool
Block 2	$3 \times 3 \times 3$, 128, Conv2a
	$2 \times 2 \times 2$, Pool
Block 3	$3 \times 3 \times 3$, 256, Conv3a
	256, batch normalisation
	$3 \times 3 \times 3$, 256, Conv3b
	$2 \times 2 \times 2$, Pool
Block 4	$3 \times 3 \times 3$, 512, Conv4a
	512, batch normalisation
	$3 \times 3 \times 3$, 512, Conv4b

Block 5	2 × 2 × 2, Pool
	3 × 3 × 3, 512, Conv5a
	512, batch normalisation
	3 × 3 × 3, 512, Conv5b
	2 × 2 × 2, Pool

The primary backbone architecture preserves five distinct convolutional blocks derived from the baseline C3D model, where each block incorporates a downsampling pooling layer designed to compress the image scale while expanding the network's receptive field. Empirical investigations detailed in literature [37] confirm that within three-dimensional convolutional frameworks, utilizing a uniform, compact convolution kernel size of $3 \times 3 \times 3$ across all convolutional layers yields the most optimized performance outcomes; consequently, every convolutional layer within the proposed primary network implements a $3 \times 3 \times 3$ kernel configuration. The initial pooling layer features a unique size and stride length specified at $1 \times 2 \times 2$. Conversely, the dimensional properties and step length of all subsequent pooling layers are uniformly set to $2 \times 2 \times 2$. Under this configuration, if the input array dimensions for Block 1 are represented as (N, C, D, H, W), the resulting output dimensions generated by the terminal Block 5 will resolve to (N, C', D/16, H/32, W/32), where N denotes the batch count, C and C' represent the respective channel counts, and D, H, W express the three-dimensional spatial coordinates of the image. Because the sample volume of medical datasets is typically highly constrained, the integration of a batch normalization layer (Batch Normalization, BN) is exceptionally well-suited for processing clinical image profiles. Its primary function centers on accelerating network convergence speeds and suppressing the onset of overfitting. Accordingly, this paper embeds a BN layer within the primary network architecture, significantly augmenting global network performance while compressing training time-frames. Following the execution of Block 5, a fully connected layer aggregates the pooled features to compute the probability distributions across the different target categories, subsequently nominating the class associated with the maximum probability score as the final classification result.

To facilitate feature extraction across the three independent subtasks and the principal classification target, dedicated convolutional layers are systematically deployed. A dedicated merging layer designed to synthesize all learned features from the auxiliary subtasks is positioned immediately downstream from the dense layer. Prior to interacting with the final output layers, all diagnostic tasks are routed through two fully connected layers. To maximize performance generalization across unseen testing data, this multi-task learning paradigm leverages the statistical relationships of associated tasks. The hard parameter-sharing technique represents a highly prevalent methodology for learning multiple distinct tasks simultaneously, operating via a singular unified architecture that shares a common subset of inner parameters to introduce an inductive bias during the training phase. This approach has attracted substantial interest across academic and corporate research sectors due to its conceptual simplicity, its capacity to enhance model generalisation, and its efficiency in lowering computational overhead. However, determining the optimal protocol for mixing the gradients of disparate tasks to enable synchronized learning presents a noted challenge, given that individual tasks frequently establish conflicting objective trajectories. To resolve this friction, we apply principles of multi-objective optimization, introducing a methodology that accounts for the temporal behavior of task gradients to construct a dynamic bias that fine-tunes the weight distribution of each individual task during backpropagation. To ensure that this synchronized learning process consistently guides the network toward the maximum performance threshold across all combined tasks, the algorithm dynamically allocates greater optimization focus toward tasks that exhibit divergence or that failed to capture performance benefits during preceding training iterations. Consequently, we empirically confirm that the proposed strategy surpasses contemporary state-of-the-art models when tasked with learning contradictory objectives.

Attention module

The implementation of attention mechanisms has expanded dramatically across contemporary DL domains, yielding highly favorable outcomes within speech recognition, natural language processing, and image processing environments [14, 26]. Speech recognition functions as an algorithmic capability that empowers a software application to translate spoken linguistic expressions into written text formats. This domain is alternatively referred to as automatic speech recognition, computer speech recognition, or speech-to-text technology. While a variety of speech recognition software and hardware configurations exist, the most sophisticated modern frameworks rely fundamentally on AI and ML principles. To properly decode and process spoken language structures, these models synthesize rules of grammar, syntax patterns, signal compositions, and the acoustic architecture of audio and vocal signals. Ideally, these systems learn iteratively through ongoing operational deployment, refining their responsive profiles with each subsequent interaction. The most advanced systems further allow enterprises to customize and calibrate the underlying technology to match highly specific operational parameters, ranging from corporate brand recognition requirements to localized linguistic variations and vocal idiosyncrasies.

Specifically, natural language processing (NLP) represents a specialized subfield of computer science and artificial intelligence (AI) that focuses on engineering computational systems capable of understanding written and spoken human languages in a manner that closely mirrors human comprehension. NLP, often intersecting with computational linguistics, integrates rule-based modeling of human language structures with advanced statistical, ML, and DL network architectures. Operating in tandem, these technical frameworks empower computers to analyze human communication delivered via textual or acoustic data streams, thereby "understanding" the comprehensive meaning of the expression, including the speaker's or writer's core intent and underlying emotional tone. NLP serves as the core computational engine driving algorithms that execute machine translation between distinct languages, process spoken verbal commands, and rapidly synthesize massive repositories of unorganized text data, even within real-time environments.

Digital image processing involves translating a raw visual image into a digital format and applying specialized computational operations to extract meaningful, structured information. When implementing designated signal processing techniques, the underlying image processing architecture typically handles all image inputs as 2D signals. The fundamental core of an image processing framework is a general-purpose computing environment, which can scale in architecture from standard personal computers up to high-performance supercomputing clusters. To satisfy demanding performance benchmarks, specially engineered hardware configurations are occasionally deployed in highly specialized applications. The baseline workflow of image processing initiates with image acquisition. Alternatively referred to as the preprocessing phase within the broader image processing pipeline, this step involves capturing the target visual data from a hardware-based input source. Following acquisition, the process of emphasizing and highlighting specific structural attributes of interest hidden within an unrefined scan is designated as image enhancement.

The primary objective of an attention mechanism is to dynamically isolate and select the most contextually relevant data components out of a vast volume of input data, closely mimicking the biological visual attention systems found in the human brain. Once the training phase is complete, the calculated attention weights become fixed during the testing cycle; however, because the input testing samples vary, the final attention distribution shifts accordingly. The absolute value of the computational force is sequentially integrated, meaning that the attention map generated for each specific sample is highly customized. Depending on the operational mode of expression, attention mechanisms are broadly categorized as either hard or soft attention frameworks [16, 28, 29]. Hard attention operates on a binary selection principle where a specific spatial coordinate is either strictly prioritized or entirely excluded, framing the selection as a 0/1 problem. Conversely, soft attention allocates continuous variations of focus across different regions or feature channels, utilizing a spectrum of values bounded between 0 and 1 to express the precise intensity of attention assigned to each distinct zone. Soft attention frameworks are inherently differentiable, allowing the model to compute gradients and resolve the optimal weights of the attention module via standard backpropagation. The attention module developed in this study is engineered using a spatial domain attention methodology founded upon soft attention principles. This spatial domain

attention framework operates on the premise that the contextual data residing within alternative channel domains dictates the exact distribution of attention mapping across the various coordinates of the image.

A multi-dimensional tensor containing data values strictly bounded within the range of [0, 255] can be conceptually managed as a digital image. When deploying a textual sequence for classification, translation, or alternative NLP workflows, the original text-based Transformer architecture utilizes those localized words as its primary input token stream. To adapt the core Transformer framework for the Vision Transformer (ViT) architecture, we implement the minimal necessary structural modifications required to transition the model from processing discrete words to operating directly on raw image blocks. We then systematically evaluate the degree to which the model can autonomously map visual structures without relying on hardcoded inductive biases. Because ViT lacks an inherent, pre-programmed understanding of the relative spatial positions of individual patches within a scan, or even the basic reality that an image possesses a 2D structure, it must extract this essential structural knowledge directly from the training dataset and preserve it within specialized position embeddings. When validated on the ImageNet repository, where the ViT framework is initially trained, it secures a top-1 accuracy score of 77.9%. By comparison, the top-performing conventional CNN architectures trained on ImageNet without supplementary data streams currently achieve an accuracy of 85.8%; thus, while the initial ViT outcomes are highly respectable for an introductory paradigm shift, they remain noticeably below the established state of the art. Due to its lack of built-in structural assumptions regarding image properties, ViT exhibits a tendency to overfit the ImageNet task despite the implementation of mitigating strategies such as strict regularization protocols. To rigorously analyze the impact of dataset scaling on global model performance, we train the ViT framework on ImageNet-21k (comprising 14 million distinct images across 21 thousand separate categories) and the JFT repository (encompassing 300 million images across 18 thousand classes). We subsequently compare these performance outcomes against a state-of-the-art CNN architecture designated as Big Transfer (BiT) trained on the identical data repositories. As observed in prior benchmarks, when restricted to standard ImageNet training (1M images), the ViT model severely underperforms in comparison to its convolutional counterpart (BiT).

As illustrated in **Figure 2**, beyond the standard convolutional layers, pooling layers, and BN blocks that constitute the primary backbone network, the integrated Attention module additionally incorporates a normalization layer designed to constrain the definitive output range. Convolutional layers represent the fundamental building blocks of all CNN architectures. The process of convolution itself is a straightforward mathematical operation involving the application of a localized filter to an input array to generate a corresponding activation map. By continuously sweeping the identical filter matrix across an input array, such as a clinical photograph, the model constructs a feature map that reflects the exact locations and relative intensities of a detected visual feature within the input space. The innovative nature of convolutional neural networks lies in their capability to adhere to the strict requirements of a specific predictive modeling task, such as automated image classification, while simultaneously and automatically learning a massive array of parallel filters customized to the underlying training dataset. Consequently, input photographs can have highly specialized characteristics isolated that are not easily captured through alternative methods. Convolutional neural networks, or CNNs, represent a specialized category of neural network architectures designed explicitly for interacting with two-dimensional image arrays, though they can be adapted to process one-dimensional and three-dimensional datasets with equal utility. This layer executes a specific procedure known as "convolution". Within the context of a CNN framework, convolution operates as a linear process that entails multiplying a designated set of weights by the input data array, mirroring the functional mechanism of a traditional neural network layer. Given that this methodology was originally engineered to manage two-dimensional inputs, a two-dimensional array of weights, formally designated as a filter or a kernel, is systematically multiplied by an intersecting array of incoming input data.

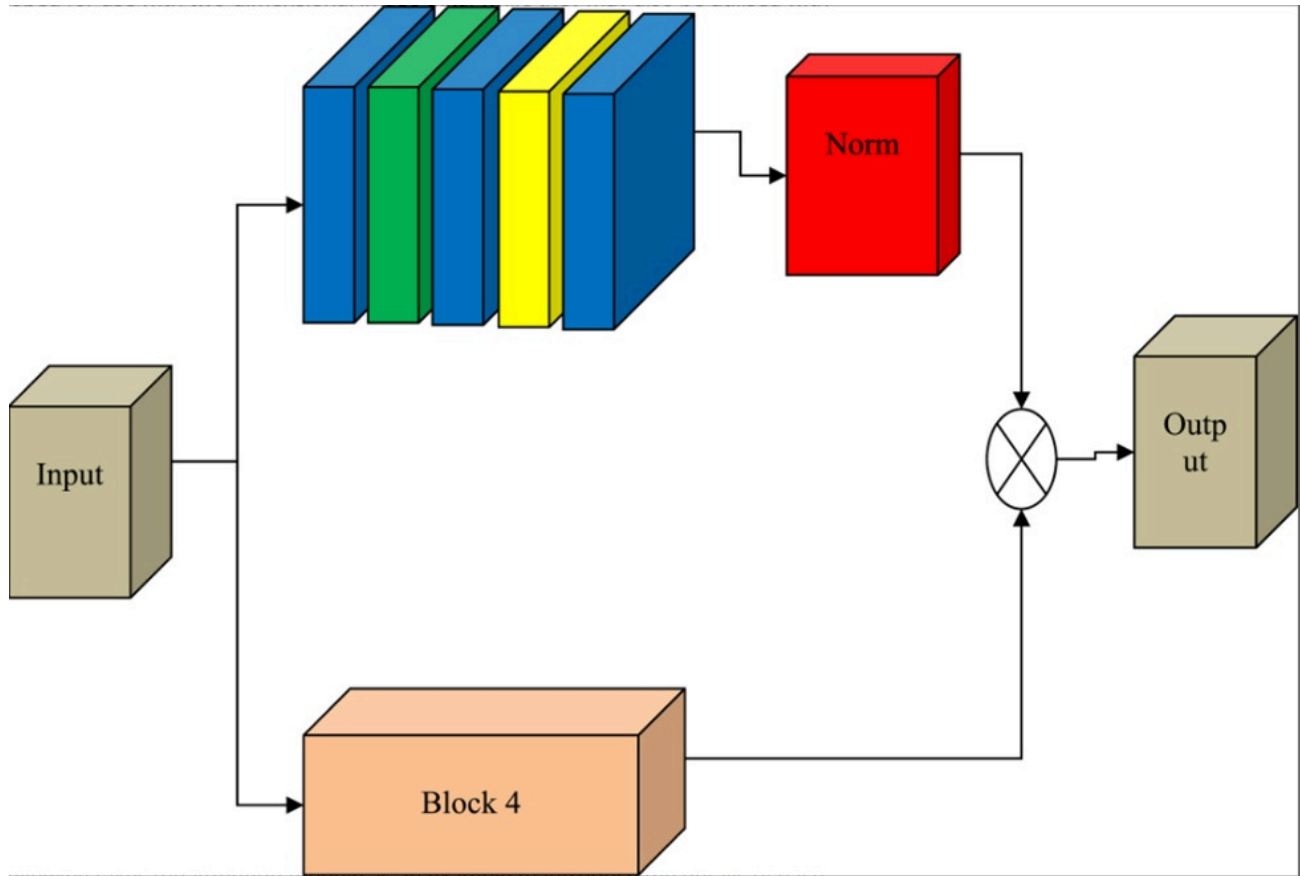


Figure 2. Attention module.

The subsequent section details the specific operational mechanics of each individual network layer, utilizing the Attention module as a representative example. Let F_3 denote the output tensor generated by Block 3 in **Figure 1**. This tensor F_3 is routed simultaneously into both the Attention module and Block 4. Within the Attention module, the tensor first undergoes a convolutional operation, the mathematical expression for which is formalized in **Eq. 1**:

$$F_{Att-l}^i = \sum_{j \in M^N} F_N^j * K_{Att-l}^{ij} + b_{Att-l}^i \quad (1)$$

Where F_{Att-l}^i represents the i -th feature map corresponding to the n -th layer within the Attention module, F_N^j signifies the j -th feature map delivered as an output from the preceding Block N , and K_{Att-l}^{ij} defines the localized convolution kernel mapping F_{Att-l}^i and F_N^j within the n -th layer of the Attention architecture. The operator $*$ denotes the standard convolution operation, b_{Att-l}^i represents the associated bias term, and M^N reflects the total index set of feature maps produced by Block N .

Because convolutional layers continuously update their internal parameters throughout the training sequence, the statistical distribution of subsequent internal input data frequently experiences destabilizing shifts. To mitigate this phenomenon, a Batch Normalization (BN) layer is introduced to standardize the intermediate feature representations across the network. The foundational calculation for the BN transformation is expressed as follows:

$$y_i = \gamma \left(\frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta \quad (2)$$

In this formulation, x_i acts as the raw input data vector, y_i represents the normalized output representation, μ defines the calculated mean, and σ^2 denotes the data variance. The variables γ and β operate as learnable reconstruction parameters optimized along the channel dimension to ensure that the core properties of the learned feature distribution are preserved. Prior to executing the activation function, the forward propagation sequence incorporating the BN layer is computed via the following expression:

$$F_{ATT-2} = f(BN(F_{ATT-1})) \quad (3)$$

Where $f(\cdot)$ represents the Rectified Linear Unit (ReLU) activation function, $BN(\cdot)$ denotes the normalization function executed by the BN layer, F_{Att_1} signifies the output tensor yielding from the prior convolution operation, and F_{Att_2} maps the definitive state post-normalization and activation.

Applying the principles of Eq. 1, the subsequent convolutional layer yields an intermediate output F_{Att_3} , which is directly passed into the pooling layer. This framework implements a max-pooling paradigm, isolating the peak value within a designated spatial region as the representative output. This operation serves to reduce spatial resolution while expanding the model's overall receptive field, according to the following mathematical definition:

$$F_{Att-4}^i = w_{Att-4}^i \cdot \text{down} \left(f \left(F_{Att-3}^i \right) \right) + b_{Att-4}^i \quad (4)$$

Where w_{Att-4}^i characterizes the pooling parameter associated with the i -th feature map, $\text{down}(\cdot)$ defines the specific downsampling pooling function, and F_{Att-3}^i denotes the i -th feature map received from the preceding layer. Following the pooling stage, the tensor transforms into F_{Att_5} via a convolutional operation matching Eq. 1, which is then forwarded to the normalization layer. The explicit objective of this normalization layer is to map the outputs of the preceding layer into a bounded probability distribution, formulated as follows:

$$F_{Att}^i = \frac{F_{Att-5}^i - \min \left(F_{latten} \left(F_{Att-5}^i \right) \right)}{\text{sum} \left(F_{Att-5}^i \right)} \quad (5)$$

Where F_{Att}^i signifies the i -th feature map produced by the terminal stage of the Attention module, and F_{Att-5}^i corresponds to the i -th feature map emerging from the prior convolution. The operator $\text{Flatten}(\cdot)$ acts as a flattening function that reshapes a multi-dimensional tensor into a unified one-dimensional array, $\min(\cdot)$ computes the minimum value across each row of the current feature map, and $\text{sum}(\cdot)$ executes a summation across each respective row.

Utilizing the mathematical workflows outlined above, the output tensor F_4 from Block 4 is isolated in an identical manner. Given an input batch size and spatial layout of (b, c, d, h, w) , the spatial dimensions of F_4 resolve to $(b, 2*c, d/2, h/2, w/2)$, whereas the definitive output dimensions of the Attention module are constrained to $(b, 1, d/2, h/2, w/2)$. The downstream input array fed into Block 5 is established by calculating the product of these two distinct tensors, as formalized in Eq. 6:

$$F = F_{Att} \cdot F_4, i = 1, 2, \dots, 2 * c \quad (6)$$

More explicitly, the computed output of the Attention module is systematically multiplied by each individual feature map produced by Block 4, thereby introducing the spatial attention weights into the core architectural pipeline.

Multi-task learning module

Multi-task learning frameworks are engineered to transfer learned knowledge elements from a primary task across to complementary task spaces, leveraging the interconnected informational structures embedded within multiple distinct objectives to improve global learning precision. The underlying premise requires that all parallel subtasks maintain a high degree of statistical relevance. Furthermore, by training these tasks simultaneously, multi-task learning networks facilitate parameter sharing [17, 30, 31], which considerably bolsters the model's overall capacity for generalizability. Within deep learning paradigms, multi-task architectures typically manifest as either soft parameter-sharing networks or hard parameter-sharing networks. In soft parameter-sharing configurations, each separate task maintains its own independent model and parameter set; conversely, hard parameter-sharing mechanics enforce a shared hidden layer architecture across all objectives, restricting unique parameter assignments exclusively to the individual terminal output layers of each specific task. In a direct comparison, the hard parameter-sharing mechanism demands a vastly reduced parameter volume, which effectively suppresses model overfitting and renders it uniquely ideal for small-scale clinical datasets.

This study implements a multi-task learning architecture grounded in hard parameter-sharing principles to achieve automated AD classification. The core task is defined as AD classification, while the secondary auxiliary subtasks are designated as CDR score regression and MMSE score regression. Clinically, evaluating a patient's degree of cognitive decline culminates in a comprehensive global score; the MMSE stands as one of the most trusted and widespread screening instruments for cognitive impairment. Because both diagnostic metrics strongly correlate with AD progression patterns, they function effectively as auxiliary training objectives.

The implementation strategy for this multi-task framework is detailed within the Multi-task module in [Figure 1](#). The final classification probabilities, alongside the predicted CDR and MMSE scores, are resolved through the integration of fully connected layers. The baseline calculation governing a fully connected layer is structured as follows:

$$y^i = f(w^i \cdot (\text{Flatten}(x^i)) + b^i) \quad (7)$$

Where x^i represents the i -th incoming feature map from the input stream, and w^i defines the specific weights optimized for the fully connected layer. The network shares two consecutive dense layers featuring channel dimensions of 4096 and 2048, respectively. A dropout layer is additionally embedded within this sequence to randomly deactivate neurons at a fixed probability, suppressing overfitting tendencies while accelerating training throughput. The primary classification branch terminates with a fully connected layer containing 2 channels, whereas the auxiliary CDR and MMSE pipelines each branch into two subsequent dense layers. These three distinct tasks undergo synchronized training, sharing all baseline parameters prior to the network split, and updating the global model parameters via backpropagation, with the primary AD classification task serving as the definitive clinical output.

The global objective function synthesizes the cross-entropy classification loss derived from the true clinical labels and predicted classifications with the regression losses from the auxiliary CDR and MMSE tasks. The global loss function is computed using [Eq. 8](#):

$$\text{Loss} = L_{\text{main}} + \alpha L_{\text{CDR}} + \lambda L_{\text{MMSE}} \quad (8)$$

Where L_{main} represents the cross-entropy loss tracking AD classification, L_{CDR} signifies the regression loss for the CDR task, L_{MMSE} denotes the regression loss for the MMSE task, and α and λ serve as loss balancing coefficients.

While L_{main} is evaluated via a standard cross-entropy loss function, both L_{CDR} and L_{MMSE} are calculated utilizing the Huber loss function, the mathematical expression for which is given below:

$$L_{\delta}(I_i, y_i) = \frac{1}{N} \sum_{i=1}^N \begin{cases} \frac{1}{2}(I_i - y_i)^2, |I_i - y_i| \leq \delta \\ \delta|I_i - y_i| - \frac{1}{2}\delta^2, |I_i - y_i| > \delta \end{cases} \quad (9)$$

Where I_i characterizes the true ground-truth label of the i -th sample, y_i represents the model's predicted scalar value, N denotes the total number of samples within the cohort, and δ operates as a tuning hyperparameter.

Dataset and Preprocessing

Experimental data

The empirical data utilized in this study is drawn from the public Alzheimer's Disease Neuroimaging Initiative (ADNI) repository, made available by Michael W. Weiner. The comprehensive ADNI dataset encompasses four distinct successive sub-cohorts: ADNI-1, ADNI-GO, ADNI-2, and ADNI-3. This specific investigation into AD categorization leverages baseline structural MRI data from the ADNI-1 cohort, which was acquired between the years 2004 and 2009 utilizing a standardized 1.5 T imaging protocol. The designated experimental dataset comprises 638 distinct patient scans, which were systematically partitioned into a training set, a validation set, and an independent testing set, accounting for 70%, 10%, and 20% of the aggregate data volume, respectively.

Image preprocessing

An illustrative example of an structural MRI volume from the ADNI dataset is provided in **Figure 3**. Moving from left to right, the visualization displays the Coronal, Sagittal, and Axial planes of the MRI scan, which correspond to the anterior-posterior, left-right, and superior-inferior cross-sections, respectively. Because subjects can exhibit varied head orientations during scan acquisition—such as tilting downwards, leaning laterally, or extending backwards—the raw MRI volumes often manifest at non-standardized angles. Furthermore, unrefined brain MRI scans contain extensive non-cerebral structures (such as the cranium) that offer no diagnostic value for AD classification; these non-brain regions introduce substantial noise parameters that degrade classification performance.

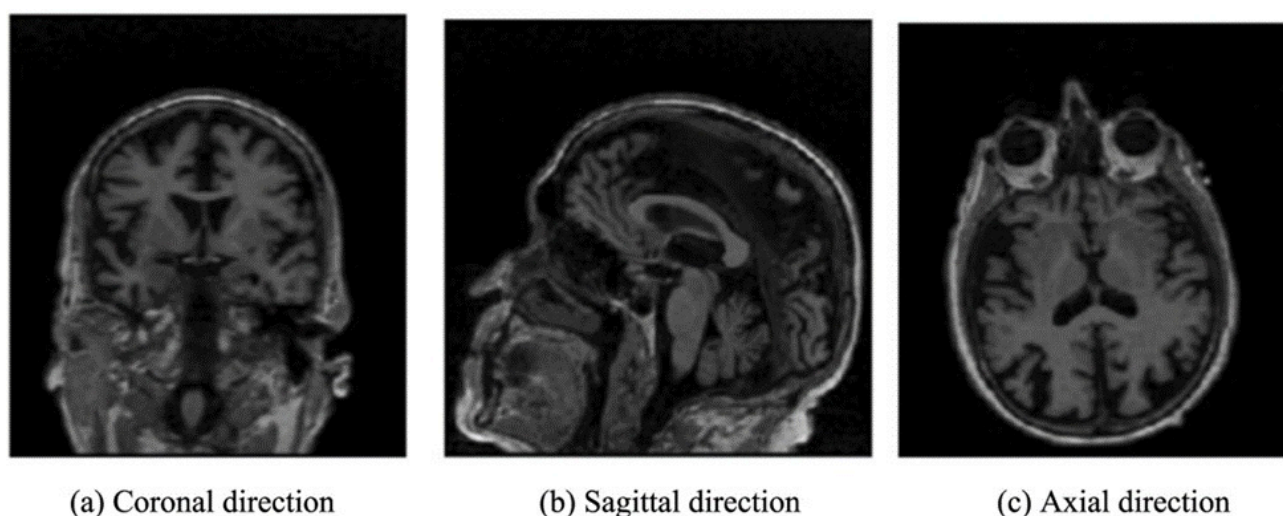


Figure 3. Magnetic resonance imaging (MRI) image example.

To resolve these technical data variations, the complete preprocessing workflow is illustrated in **Figure 4**, using the Sagittal plane as a representative example, and is executed through the specific sequential steps detailed below.

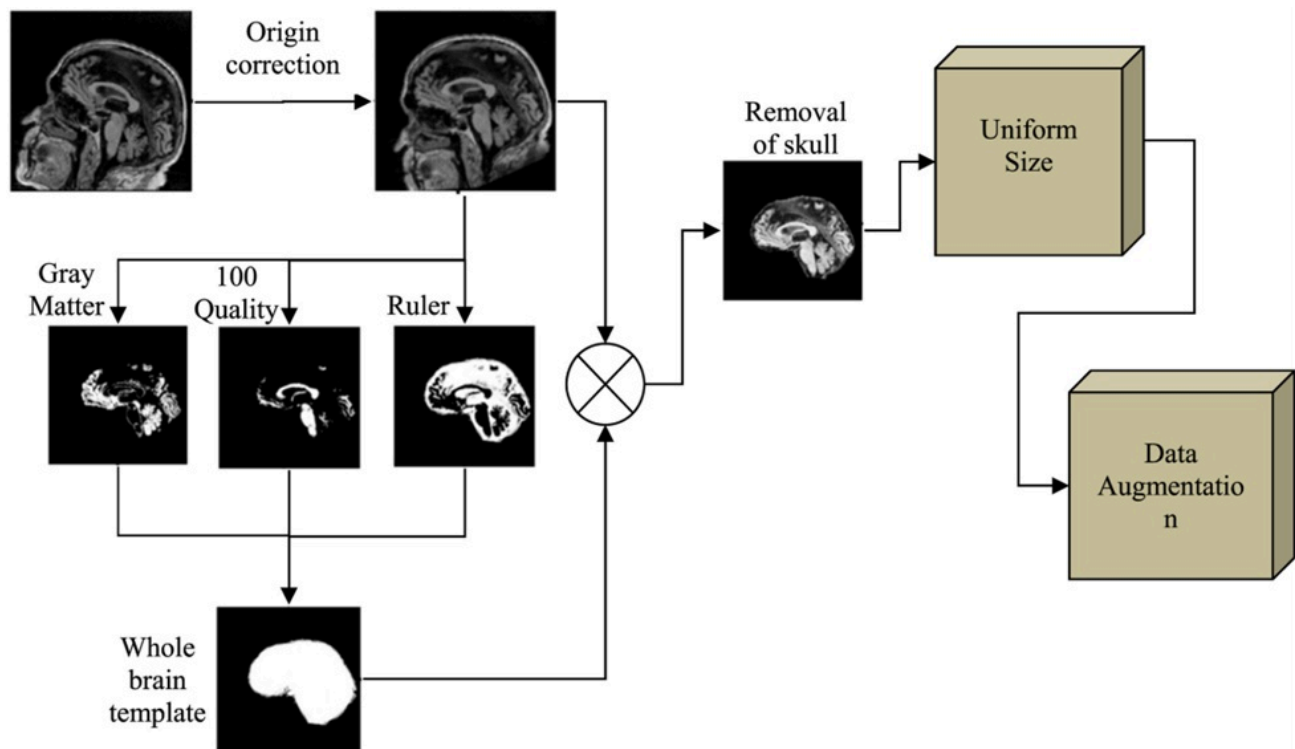


Figure 4. Preprocessing flow.

- **Step 1. Origin correction:** The Statistical Parametric Mapping (SPM) toolbox within the MATLAB environment is deployed to standardize the coordinate origin of each brain MRI volume via the AC-PC alignment method, where AC denotes the anterior commissure and PC represents the posterior commissure. Operationally, the vertical axis of the digital crosshair is aligned to pass directly through the midline of the coronal and axial planes, forcing the primary intersection point to coincide with the AC spatial coordinate as closely as possible. Subsequently, the horizontal axis within the sagittal view is manipulated to intersect both the AC and PC markers, thereby completing the coordinate origin correction.
- **Step 2. Remove the skull:** The skull stripping pipeline is executed across three distinct phases. First, the origin-corrected MRI volume is processed via the SPM segmentation tool to partition the scan into three tissue classes: gray matter, white matter, and alternative cerebrospinal fluids. These three segmented tissue tracks are then directly aggregated to generate a global binary brain mask. Finally, a element-wise dot multiplication is computed between this binary tissue template and the origin-corrected raw MRI volume, isolating the cerebral structure and removing the skull.
- **Step 3. Unify the size:** The raw MRI scans within the shared dataset exhibit highly heterogeneous dimensions, including formats such as $256 \times 256 \times 166$ and $192 \times 192 \times 160$. To alleviate the computational strain imposed on the 3D CNN architecture, the skull-stripped brain volumes undergo spatial downsampling. This process yields a highly uniform final tensor size of $79 \times 95 \times 79$ across all subjects.
- **Step 4. Data augmentation:** To address data scarcity constraints and expand the limited volume of experimental training data, data augmentation is achieved by applying stochastic noise injection, random spatial mirroring, and intensity normalization protocols directly to the preprocessed MRI volumes.

Experimentation and Analysis

Evaluation criteria

To perform a rigorous quantitative assessment of the proposed methodology, the framework implements accuracy (Accuracy, Acc), sensitivity (Sensitivity, Sens), and specificity (Specificity, Spec)—the most widely established metrics within neuroimaging classification literature—as the primary evaluation indicators on the ADNI dataset. The formal mathematical equations governing these criteria are expressed as follows:

$$A_{\alpha} = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

$$Sens = \frac{TP}{TP + FN} \quad (11)$$

$$Spec = \frac{TN}{FP + TN} \quad (12)$$

Where FP signifies a false positive error, representing the volume of healthy control scans incorrectly categorized as pathological; TP denotes a true positive outcome, indicating the volume of validation patients correctly identified with the disease; TN represents a true negative evaluation, tracking the percentage of accurately categorized healthy control subjects; and FN refers to a false negative error rate, measuring the number of pathological samples incorrectly flagged as normal. To ensure balanced evaluation, a consistent distribution of patient scans was allocated across the validation cohorts.

Experimental setup and parameter analysis

In this investigation, TensorFlow [18] was implemented as the deep learning framework. Following initial pre-training on the UCF101 action recognition dataset, the convolutional layer weights of the C3D model were utilized to initialize the parameters of the matching network layer. Optimization was performed using the adaptive moment estimation (Adam) technique over 10,000 total training epochs. The initial learning rate was established at 0.0001, and was systematically scaled down to 1/20 of its contemporary value at every 25-epoch interval. The training batch size was set to 8 samples per iteration, while the hyperparameters δCDR and $\delta MMSE$ from Eq. 9 were designated as 0.5 and 7.0, respectively. The loss balance weights for the auxiliary subtasks of CDR regression and MMSE regression were denoted by α and λ , respectively, in Eq. 8. Given that the primary objective of this architecture is AD classification, the values for α and λ are conventionally restricted to the range [0, 1]. In the context of the HC experiment, λ was initially held constant at 0, while α was adjusted incrementally by steps of 0.1. Figure 5 illustrates the classification performance on the validation set across these distinct values of α post-training. The empirical results indicate that the classification of AD versus HC achieves optimal performance when α equals 1.0; consequently, α was fixed at 1.0 for all subsequent multi-task learning evaluations involving the AD versus HC cohorts.

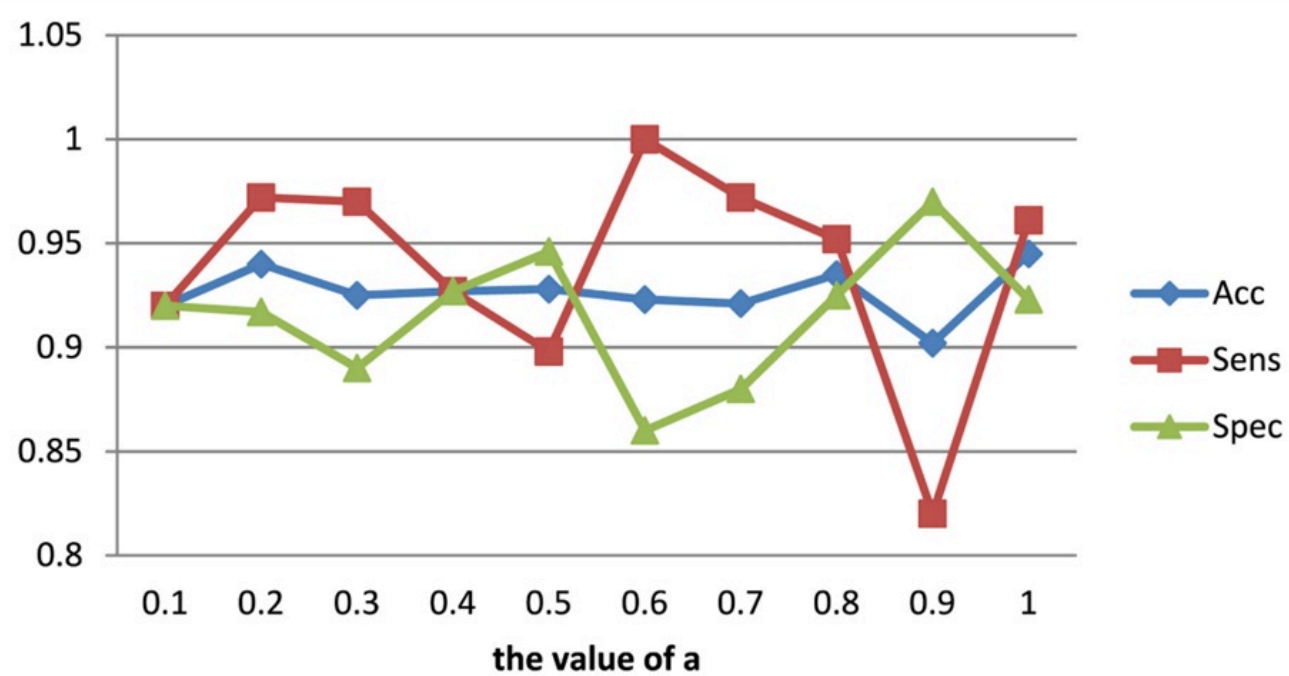


Figure 5. Performance of different values of α (AD versus HC).

Building upon the previous experimental phase, the hyperparameter α was fixed at 1.0 while the λ parameter was fine-tuned within the AD versus HC classification framework. Adjustments to λ were likewise conducted at 0.1 increments. As demonstrated in **Figure 6**, the evaluation achieves peak classification efficacy when $\lambda = 0.3$. Accordingly, for the AD versus HC trials, α and λ were finalized at 1.0 and 0.3, respectively.

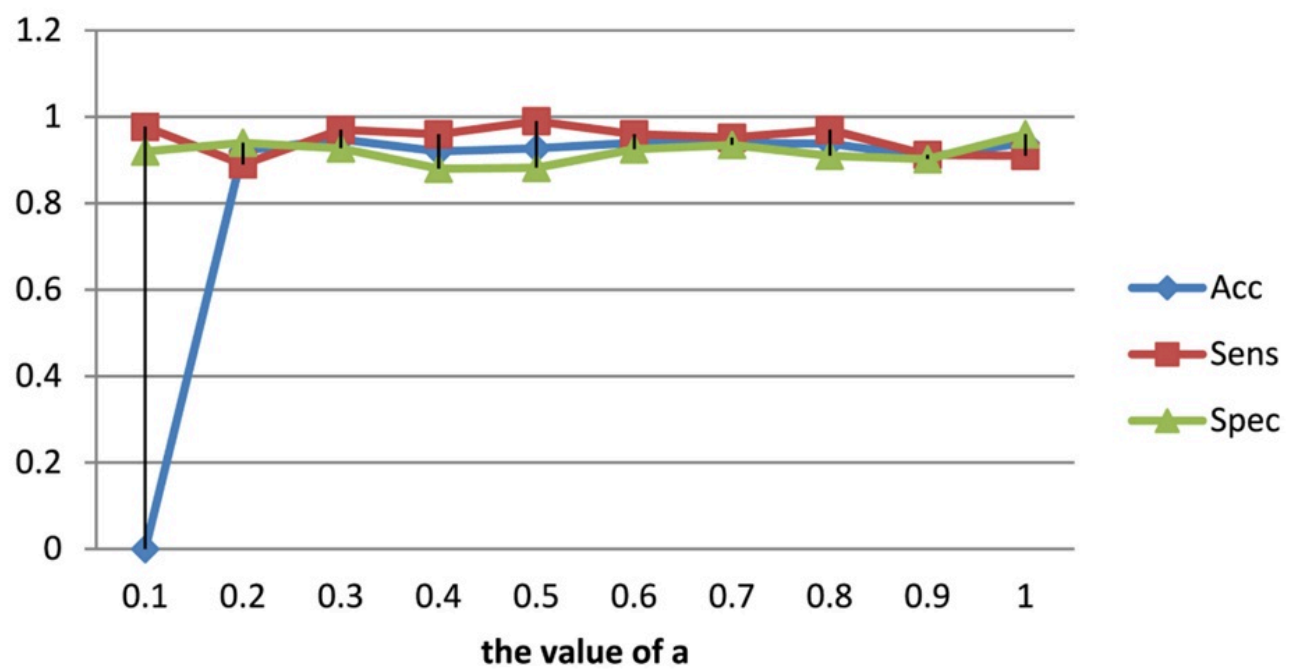


Figure 6. Performance of different values of λ (AD versus HC).

Adopting an identical methodology for the MCI versus HC cohort, the optimal configuration for hyperparameters α and λ was determined via **Figures 7 and 8** to be 0.2 and 0.6, respectively. These derived values for α and λ were strictly maintained across all subsequent multi-task learning experiments.

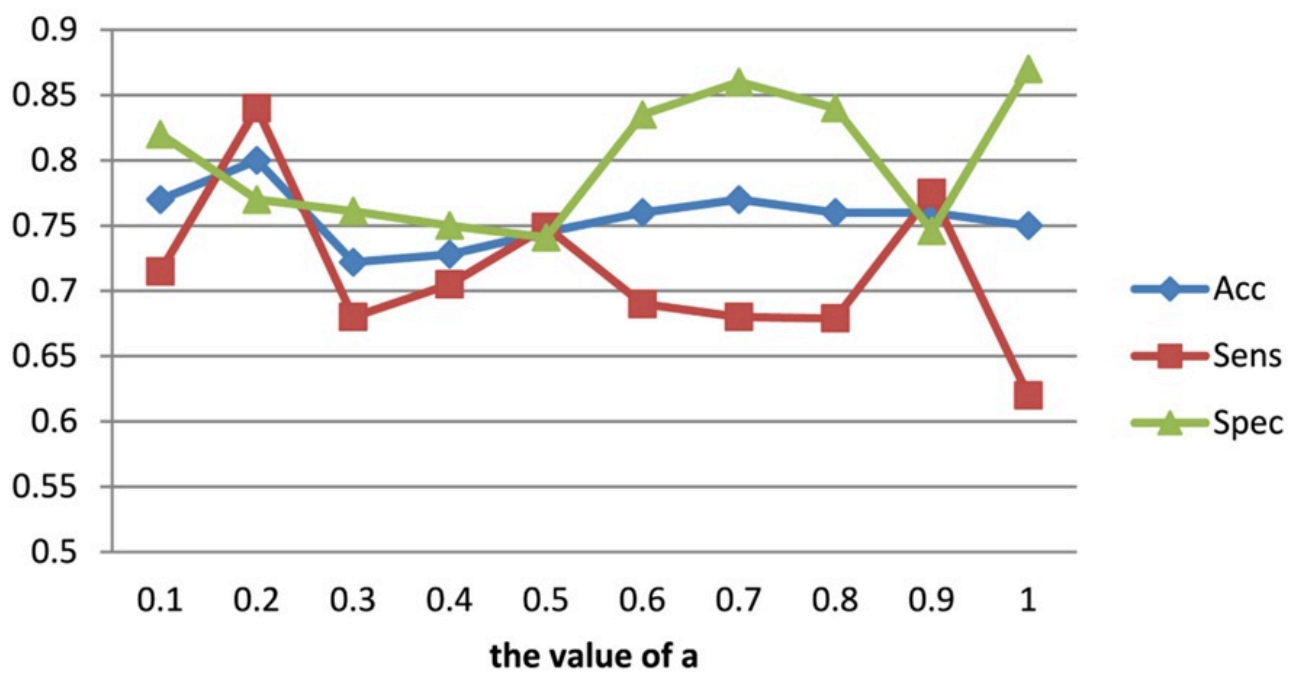


Figure 7. Performance of different values of α (MCI versus HC).

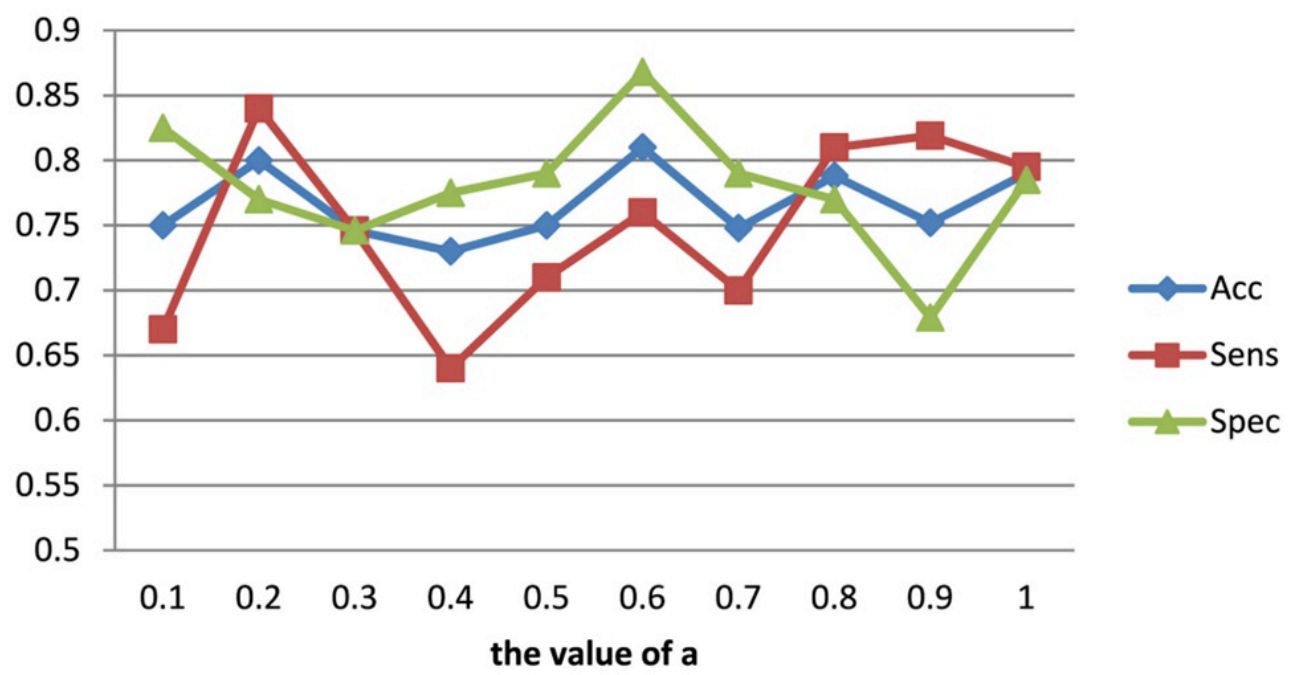


Figure 8. Performance of different values of λ (MCI versus HC).

Experimental comparison

An ablation study was structured to validate the practical contribution of each component within the proposed methodology, with the comparative outcomes summarized in [Table 2](#).

Table 2. Comparison of ablation experiments.

Model	Acc	
	MCI versus HC	AD versus HC
C3D	0.5000	0.8721
C3D + BN	0.7273	0.8983
C3D + BN + Attention	0.7386	0.9186
C3D + BN + Multitask	0.7642	0.9157
Presented work	0.8097	0.9448

Across the complete dataset, classification accuracy within the MCI versus HC trials is consistently lower than that observed in the AD versus HC trials. This discrepancy emerges because brain atrophy patterns in MCI patients exhibit greater structural similarity to HC subjects than to advanced AD patients, thereby introducing greater diagnostic ambiguity. The C3D + BN configuration denotes the integration of a batch normalization layer into the baseline C3D network. As shown in [Table 2](#), this integration yields a notable performance enhancement in the MCI versus HC tasks. Prior to incorporating the BN layer, the network struggled to differentiate between MCI and HC instances, yielding an accuracy rate of approximately 0.5, which is statistically equivalent to random chance. This baseline limitation likely stems from the fact that the original C3D network was pre-trained on the UCF101 dataset, which consists of human action videos that diverge substantially from medical image classification features. Consequently, when training the standard C3D network on brain MRI scans, the limited sample sizes of the MCI and HC cohorts induced rapid overfitting, which inflated test set errors. The alternative C3D + BN + Attention model represents the targeted addition of an attention mechanism directly to the batch-normalized baseline.

The empirical data in [Table 2](#) indicate that the attention mechanism primarily enhances classification accuracy within the AD versus HC trials. Conversely, the C3D + BN + Multitask framework, which isolates the effects of multi-task learning, provides a greater performance boost to the MCI versus HC tasks than to the AD versus HC tasks, where the attention mechanism remains more influential. When both components are deployed simultaneously (the proposed algorithm), the accuracy improves by 7% for AD versus HC classification relative to the baseline C3D network. Furthermore, the accuracy for MCI versus HC classification increases substantially from 0.5 to 0.8097. These metrics validate the combined utility of the BN layer, multi-task learning strategies, and the attention mechanism in optimizing AD classification performance.

[Table 3](#) evaluates the classification performance of the proposed framework against several recent benchmark studies. For instance, the approach in literature [\[19\]](#) implements a recurrent gated unit within an RNN architecture to extract cross-sectional slice features for subsequent classification. The method described in literature [\[20\]](#) incorporates ensemble learning techniques into a convolutional network to establish a multi-slice ensemble classification framework. Meanwhile, literature [\[21\]](#) evaluates morphological variations of the hippocampus utilizing a 3D densely connected convolutional network for AD detection. The comparative metrics in [Table 3](#) reveal that the proposed algorithm possesses a distinct performance advantage over the first two slice-based methodologies in both AD versus HC and MCI versus HC classifications. Specifically, classification accuracy improved by 3%–13% and 2% across the respective cohorts, while

diagnostic sensitivity increased by 6% and 5.5%. This performance gap exists because the former two frameworks rely strictly on 2D slice data; even though their slice selection methods show marked improvements over traditional techniques that simply extract central slices from sequential intervals, the retained features remain inherently incomplete, which compromises classification accuracy.

Table 3. Comparison of classification performance.

Model	MCI versus HC			AD versus HC		
	Spec	Sens	Acc	Spec	Sens	Acc
Literature [19] algorithm	0.8000	0.7808	0.7886	0.9100	0.9140	0.9119
Literature [20] algorithm	-	-	-	-	-	0.8103
Literature [21] algorithm	0.6996	0.7727	0.7464	0.9372	0.9063	0.9229
Presented work	0.7670	0.8352	0.8097	0.9250	0.9722	0.9448

The third benchmark method also leverages a three-dimensional convolutional network; however, it diverges from the proposed algorithm in its capacity to classify MCI versus HC instances. In these specific trials, the benchmark method underperforms, trailing the proposed algorithm by margins of 6% in accuracy, 6% in sensitivity, and 7% in specificity. This disparity highlights the proposed algorithm's capacity to exploit the clinical correlations between CDR scales, MMSE scores, and primary AD categories, alongside its ability to isolate highly relevant spatial features from 3D MRI scans via the attention mechanism. Notably, when compared across all three baseline methods, the specificity of the proposed algorithm is somewhat lower. Specificity essentially dictates the proportion of true HC subjects who are correctly identified as negative for the disease. Because false-positive patients typically receive secondary follow-up assessments that rectify initial misdiagnoses, sacrificing a marginal degree of specificity is an acceptable trade-off to secure higher diagnostic sensitivity.

Conclusion

To resolve the clinical and technical challenges outlined above, this paper introduces an end-to-end AD classification methodology constructed around a deep 3D convolutional neural network (CNN). The framework integrates an attention mechanism to refine feature extraction alongside two auxiliary regression subtasks—specifically, CDR Scale score regression and MMSE score regression—to guide the multi-task learning process and optimize overall AD classification accuracy. By combining the 3D convolutional network with attention allocation, the model successfully isolates features weighted by regional importance, capturing the structural spatial info most critical to classification performance. Ultimately, this deep 3D CNN algorithm offers an effective means of mining medical image features without neglecting complementary clinical metrics.

Validation experiments performed on the ADNI dataset confirm that the proposed model yields robust AD classification performance and outpaces current state-of-the-art algorithms. The underlying data was sourced from the public repository of the Alzheimer’s Disease Neuroimaging Initiative (ADNI), which comprises four successive sub-phases: ADNI-1, ADNI-GO, ADNI-2, and ADNI-3. This research specifically utilizes 1.5 T structural MRI data from the ADNI-1 cohort collected between 2004 and 2009. The complete experimental dataset of 638 samples was partitioned into training, validation, and testing sets using a 70%, 10%, and 20% ratio, respectively. Despite these positive outcomes, the proposed algorithm still exhibits diminished performance in MCI versus HC trials relative to AD versus HC evaluations. Future work will focus on mapping the distinct structural features of MCI pathology by integrating multi-scale spatial data, implementing adversarial training samples, and refining the classification loss functions to better manage class imbalances during training phases.

Acknowledgements

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through General Research Project under grant number (GRP/75/44).

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 08 Apr 2023

Revised: 04 Jun 2023

Accepted: 12 Jul 2023

Published online: 15 September 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Shojaei S, Abadeh MS, Momeni Z. An evolutionary explainable deep learning approach for Alzheimer's MRI classification. *Expert Syst Appl*. 2023;220:119709.

<https://doi.org/10.1016/j.eswa.2023.119709>.

Subramanyam Rallabandi VP, Seetharaman K. Deep learning-based classification of healthy aging controls, mild cognitive impairment and Alzheimer's disease using fusion of MRI-PET imaging. *Biomed Signal Process Control*. 2023;80(Pt 1):104312.

<https://doi.org/10.1016/j.bspc.2022.104312>.

Alorf A, Khan MUG. Multi-label classification of Alzheimer's disease stages from resting-state fMRI-based correlation connectivity data and deep learning. *Comput Biol Med*. 2022;151(Pt A):106240.

<https://doi.org/10.1016/j.compbiomed.2022.106240>.

El-Geneedy M, Moustafa HED, Khalifa F, Khater H, Abdelhalim E. An MRI-based deep learning approach for accurate detection of Alzheimer's disease. *Alex Eng J*. 2023;63:211-21.

<https://doi.org/10.1016/j.aej.2022.07.062>.

Han R, Liu Z, Chen CLP. Multi-scale 3D convolution feature-based broad learning system for Alzheimer's disease diagnosis via MRI images. *Appl Soft Comput*. 2022;120:108660.
<https://doi.org/10.1016/j.asoc.2022.108660>.

Sharma S, Guleria K, Tiwari S, Kumar S. A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer disease using MRI scans. *Measurement (Lond)*. 2022;24:100506.
<https://doi.org/10.1016/j.measen.2022.100506>.

Liu F, Yuan S, Li W, Xu Q, Sheng B. Patch-based deep multi-modal learning framework for Alzheimer's disease diagnosis using multi-view neuroimaging. *Biomed Signal Process Control*. 2023;80(Pt 2):104400.
<https://doi.org/10.1016/j.bspc.2022.104400>.

Liang W, Zhang K, Cao P, Liu X, Yang J, Zaiane O. Rethinking modeling Alzheimer's disease progression from a multi-task learning perspective with deep recurrent neural network. *Comput Biol Med*. 2021;138:104935.
<https://doi.org/10.1016/j.compbiomed.2021.104935>.

Liu Y, Mazumdar S, Bath PA. An unsupervised learning approach to diagnosing Alzheimer's disease using brain magnetic resonance imaging scans. *Int J Med Inform*. 2023;173:105027.

Sharma R, Goel T, Tanveer M, Murugan R. 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*. 2022;115:108099.
<https://doi.org/10.1016/j.asoc.2021.108099>.

Loddo A, Buttau S, Di Ruberto C. Deep learning based pipelines for Alzheimer's disease diagnosis: a comparative study and a novel deep-ensemble method. *Comput Biol Med*. 2022;141:105032.

Choi JY, Lee B. Combining of multiple deep networks via ensemble generalization loss, based on MRI images, for Alzheimer's disease classification. *IEEE Signal Process Lett*. 2020;27:206-10.

Tanveer M, Rashid AH, Ganaie MAG, Reza M, Razzak I, Hua KL. Classification of Alzheimer's disease using ensemble of deep neural networks trained through transfer learning. *IEEE J Biomed Health Inform*. 2022;26(4):1453-63.
<https://doi.org/10.1109/JBHI.2021.3083274>.

Hu Z, Wang Z, Jin Y, Hou W. VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer's disease prediction. *Comput Methods Programs Biomed*. 2023;229:107291.
<https://doi.org/10.1016/j.cmpb.2022.107291>.

Demir F, Akbulut Y, Taşcı B, Demir K. Improving brain tumor classification performance with an effective approach based on new deep learning model named 3ACL from 3D MRI data. *Biomed Signal Process Control*. 2023;81:104424.
<https://doi.org/10.1016/j.bspc.2022.104424>.

Muhammed Niyas KP, Thiyagarajan P. A systematic review on early prediction of mild cognitive impairment to Alzheimer's using machine learning algorithms. *Int J Intell Netw*. 2023;4:74-88.

Dong A, Zhang G, Liu J, Wei Z. Latent feature representation learning for Alzheimer's disease classification. *Comput Biol Med*. 2022;150:106116.
<https://doi.org/10.1016/j.compbiomed.2022.106116>.

Cobbinah BM, Sorg C, Yang Q, Ternblom A, Zheng C, Han W, et al. Reducing variations in multi-center Alzheimer's disease classification with convolutional adversarial autoencoder. *Med Image Anal*. 2022;82:102585.
<https://doi.org/10.1016/j.media.2022.102585>.

Wang Y, Song D, Wang W, Rao S, Wang X, Wang M. Self-supervised learning and semi-supervised learning for multi-sequence medical image classification. *Neurocomputing*. 2022;513:383-94.
<https://doi.org/10.1016/j.neucom.2022.09.097>.

Martinez-Murcia FJ, Ortiz A, Gorriz JM, Ramirez J, Castillo-Barnes D. Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders. *IEEE J Biomed Health Inform*. 2020;24(1):17-26.
<https://doi.org/10.1109/JBHI.2019.2914970>.

Gao X, Shi F, Shen D, Liu M. Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in Alzheimer's disease. *IEEE J Biomed Health Inform*. 2022;26(1):36-43.
<https://doi.org/10.1109/JBHI.2021.3097721>.

Zhu W, Sun L, Huang J, Han L, Zhang D. Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. *IEEE Trans Med Imaging*. 2021;40(9):2354-66.
<https://doi.org/10.1109/TMI.2021.3077079>.

Lao H, Zhang X. Regression and classification of Alzheimer's disease diagnosis using NMF-TDNet features from 3D brain MR image. *IEEE J Biomed Health Inform*. 2022;26(3):1103-15.

Li J, Wei Y, Wang C, Hu Q, Liu Y, Xu L. 3-D CNN-based multichannel contrastive learning for Alzheimer's disease automatic diagnosis. *IEEE Trans Instrum Meas*. 2022;71:1-11. Art. no. 5008411.

Rashid AH, Gupta A, Gupta J, Tanveer M. Biceph-Net: a robust and lightweight framework for the diagnosis of Alzheimer's disease using 2D-MRI scans and deep similarity learning. *IEEE J Biomed Health Inform*. 2023;27(3):1205-13.
<https://doi.org/10.1109/JBHI.2022.3174033>.

Shamrat FMJM, Akter S, Azam S, Karim A, Ghosh P, Tasnim Z. AlzheimerNet: an effective deep learning based proposition for Alzheimer's disease stages classification from functional brain changes in magnetic resonance images. *IEEE Access*. 2023;11:16376-95.
<https://doi.org/10.1109/ACCESS.2023.3244952>.

Murugan S, Chandran V, Sumithra MG, Gao XZ. DEMNET: a deep learning model for early diagnosis of Alzheimer diseases and dementia from MR images. *IEEE Access*. 2021;9:90319-29.
<https://doi.org/10.1109/ACCESS.2021.3090474>.

Ashtari-Majlan M, Seifi A, Dehshibi MM. A multi-stream convolutional neural network for classification of progressive MCI in Alzheimer's disease using structural MRI images. *IEEE J Biomed Health Inform*. 2022;26(8):3918-26.
<https://doi.org/10.1109/JBHI.2022.3155705>.

Liu M, Zhang J, Lian C, Shen D. Weakly supervised deep learning for brain disease prognosis using MRI and incomplete clinical scores. *IEEE Trans Cybern*. 2020;50(7):3381-92.
<https://doi.org/10.1109/TCYB.2019.2904186>.

Shi J, Zheng X, Li Y, Zhang Q, Ying S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. *IEEE J Biomed Health Inform*. 2018;22(1):173-83.
<https://doi.org/10.1109/JBHI.2017.2655720>.

Liu M, Zhang J, Adeli E, Shen D. Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. *IEEE Trans Biomed Eng*. 2019;66(5):1195-206.
<https://doi.org/10.1109/TBME.2018.2869989>.

Li F, Tran L, Thung KH, Ji S, Shen D, Li J. A robust deep model for improved classification of AD/MCI patients. *IEEE J Biomed Health Inform*. 2015;19(5):1610-6.
<https://doi.org/10.1109/JBHI.2015.2429556>.

Mehta R, Christinck T, Nair T, Bussy A, Premasiri S, Costantino M, et al. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. *IEEE Trans Med Imaging*. 2022;41(2):360-73.
<https://doi.org/10.1109/TMI.2021.3114097>.

Faisal FUR, Kwon GR. Automated detection of Alzheimer's disease and mild cognitive impairment using whole brain MRI. *IEEE Access*. 2022;10:65055-66.

Gamal A, Elattar M, Selim S. Automatic early diagnosis of Alzheimer's disease using 3D deep ensemble approach. *IEEE Access*. 2022;10:115974-87.

Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Trans Biomed Eng*. 2015;62(4):1132-40.
<https://doi.org/10.1109/TBME.2014.2372011>.

Zhao Y, Wang X, Che T, Bao G, Li S. Multi-task deep learning for medical image computing and analysis: a review. *Comput Biol Med*. 2023;153:106496.

Zhang X, Han L, Zhu W, Sun L, Zhang D. An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE J Biomed Health Inform*. 2022;26(11):5289-97.
<https://doi.org/10.1109/JBHI.2021.3066832>.