Applied Soft Computing Journal 162 (2024) 111749

Contents lists available at ScienceDirect

Applied Soft Computing

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

# Information fusion-based Bayesian optimized heterogeneous deep ensemble model based on longitudinal neuroimaging data

Nasir Rahim [a], Shaker El-Sappagh [a,b,c], Haytham Rizk [d], Omar Amin El-serafy [d], Tamer Abuhmed [a,*]

[a] Information Laboratory (InfoLab), Department of Computer Science and Engineering, College of Computing and Informatics, Sungkyunkwan University, Suwon 16419, South Korea
[b] Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt
[c] Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Banha 13518, Egypt
[d] Neurology Department, Faculty of Medicine, Cairo University, 11562, Egypt

## HIGHLIGHTS

- Heterogenous deep ensemble model for detecting AD progression using longitudinal data.
- Lightweight autoencoder followed by PCA for energy compact deep feature representation.
- Bayesian optimization on various time series models (GRU, LSTM, 1D-CNN).
- Explainable AI approach representing AD's progression across various time steps.
- Model's robustness was tested on an independent dataset, assessing its generalizability.

## ARTICLE INFO

*Keywords:*
AD progression detection
Ensemble network
Multimodal information fusion
Longitudinal data analysis

## ABSTRACT

The fusion of multimodal longitudinal data is difficult but crucial for enhancing the accuracy of deep learning models for disease identification and helps provide tailored and patient-centric decisions. This study explores the fusion of multimodal data to detect the progression of Alzheimer's disease (AD) using ensemble learning. We propose a heterogeneous ensemble framework of Bayesian-optimized time-series deep learning models to identify progressive deterioration of brain damage. Experimental results show that the heterogeneous ensemble of three models with patient's temporal data outperforms all other variants of ensemble models by achieving an average performance of 95% for accuracy. We also propose a novel explainability approach, which enables domain experts and practitioners to better comprehend the model's final decision. The visual explainability of infected brain regions and the model's robustness is evaluated by our two medical domain experts showing its promising use in real medical environment. To evaluate the model's generalizability and robustness, our optimized model is tested on a dataset with different distribution. The experiments demonstrate that the proposed model, which was trained on ADNI data, exhibits reliable generalization to NACC data with an average precision of 90%, recall of 91%, F1-score of 89%, AUC of 88%, and accuracy of 88%.

## 1. Introduction

Alzheimer's disease (AD) is the most prevalent chronic degenerative brain disease worldwide. It is primarily associated with people over 60 years of age [1]. The World Health Organization reports that 50 million people suffer from AD, with estimates suggesting that this number will triple by 2050 [2]. Currently, there is no cure for AD at this time, and existing treatments can only slow its progression. Early detection of AD is essential for improving treatment outcomes as the disease becomes less treatable when detected in its late stages [3]. Mild cognitive impairment (MCI) is often considered a transitional stage between normal brain function and AD. Recent studies indicate that 10–15% of people with MCI progress to AD annually. Several machine learning (ML) techniques aim to differentiate between the stages of AD, including

* Corresponding author.
*E-mail address:* tamer@skku.edu (T. Abuhmed).

https://doi.org/10.1016/j.asoc.2024.111749
Received 4 December 2023; Received in revised form 2 April 2024; Accepted 3 May 2024
Available online 15 May 2024

cognitive impairments such as stable mild cognitive impairment (sMCI) and progressive mild cognitive impairment (pMCI) [4]. However, these techniques have several limitations, such as the limited number of data samples, single modality of data, and baseline data only. Such diagnostic approaches have an increased chance of misdiagnosing the actual disease owing to the lack of sufficient information [5,6,7].

Neuroimaging, especially magnetic resonance imaging (MRI), is very effective in diagnosing diseases related to human cognition, especially neurodegenerative diseases, such as AD [6,8]. MRI is a non-invasive, highly effective method for brain imaging, offering precise depictions of brain size and structure and allowing for the identification of minor abnormalities. With the emergence of deep learning (DL) algorithms in the field of computer vision, many researchers have successfully implemented deep neural network (DNN)-based algorithms in the medical domain using neuroimaging modalities and have achieved prominent results. These algorithms process raw data, such as neuroimaging, and generate high-level representational features "on the fly" related to the specified disease during the learning process [9]. For instance, Jiang et al. [10] proposed a hybrid convolutional neural network-support vector machine (CNN-SVM) model for the early detection of MCI and AD. They utilized the MRI modalities of 70 MCI and 50 normal control (NC) subjects and trained a visual geometry group (VGG) model to extract feature embeddings. Then, the LASO feature selector was employed to perform feature selection and train an SVM model with selected features for the binary classification (MCI vs. AD) task. Thanh et al. [11] used resting state functional MRI (rsFMRI) data from South Korean patients to train and evaluate a 3D-CNN to differentiate between healthy controls and patients with AD. They generated 3D maps from rsFMRI volumes and used these maps to train the 3D-CNN and predict MMSE scores. They further optimized the obtained feature set using the least absolute shrinkage method and evaluated ML algorithms such as support vector regressors and tree regressors to improve the model's performance. Zhang et al. [12] proposed a novel lightweight binary classifier for CN and AD classification based on whole-brain 3D MRI volumes. They focused on the axial plane of the MRI volume, preprocessed it with spatial normalization, skull stripping, and measurement of stationary wavelet entropy and used a single-layered neural network with a particle swarm optimizer for parameter tuning. Lu et al. [13] proposed a DNN model based on multiscale PDG-PET to identify metabolic changes caused by AD pathology and differentiate them from those observed in NCs. The authors also combined multiple classifiers with varying validation configurations, making their approach more stable and robust. Most diagnostic studies mentioned earlier in the AD domain mainly focused on a single modality of neuroimaging data.

Many studies have proven that integrating multimodal data can enhance a model's performance in the disease identification process and help provide tailored and patient-centric decisions [14,15]. Multimodal data, such as neuropsychological battery results, cognitive scores, demographics, and neuroimaging data in the AD domain, have been shown to significantly improve model performance and reduce the negative effects of noisy data [16]. In addition, owing to the multimodal diagnostic process, the resulting models are widely accepted in the medical environment [5,6,7,17]. Bi et al. [18] created a clustering evolutionary random forest method to address the problem of limited training data. They started by randomly selecting samples and combining the features to form an initial random forest. This approach includes the concept of clustering evolution, which eliminates redundant and irrelevant decision trees. In addition, the author conducted experiments on various ways of combining features and ultimately combined feature construction, sample classification, and feature selection to detect AD pathology. Xu et al. [15] used multimodal imaging data to classify patients with MCI and AD. They utilized three imaging modalities: MRI, FDG-PET, and Florbetapir PET scans. The combination of these modalities allows for a more comprehensive analysis of brain anatomy and metabolism, leading to a more accurate classification algorithm. According to Huang et al. [19], a combination of MRI and CSF modalities can help distinguish cognitively normal individuals from patients with MCI. Gray et al. [20] trained a random forest (RF) classifier to distinguish CN vs. MCI vs. AD using FDG-PET, MRI, CSF, and genetic features. Although the studies in the medical domain produced satisfactory results for identifying diseases, they only utilized data from a single time step and did not consider any follow-up data after the baseline visit. Furthermore, these studies failed to consider the time-series aspect of the data, which has the potential to reveal the impact of changes in sequential features over time and enhance the classification performance. In addition, the omission of subsequent time steps from the data bounds causes researchers to lose the most crucial information about disease progression [15,21].

The management and analysis of time-series data provides crucial information for the assessment of neurodegenerative diseases, particularly AD, which is a serious form of chronic cognitive impairment. Furthermore, it may prove to be challenging to differentiate between CN and AD when analyzing degenerative brain diseases based on baseline or single-visit data only [22]. Very limited research has been conducted based on time-series data for AD progression detection [16,23,24]. For instance, Chincarini et al. [25] employed time-series data analysis of the ADNI dataset to detect AD progression using longitudinal MRI. Four MRI volumes were used as input data for each patient: two volumes from the baseline visit, one volume taken at a 12-month follow-up, and one from a 24-month follow-up appointment. This study aimed to capture the morphometry of hippocampal subregions to monitor AD progression in brain tissues and was formulated as two binary classification tasks, namely, CN vs. AD and CN vs. MCI. The study yielded a 93% area under the ROC curve for the CN vs. AD task, and 88% for the CN vs. MCI task. Similar to this study, Moradi et al. [26] trained a semi-supervised ML algorithm using MRI data to predict the progression from MCI to AD in the following three years. El-Sappagh et al. [16] proposed multimodal multitask DL architectures to jointly predict multiple tasks simultaneously, such as the current status of the disease, critical cognitive scores, and the progression time of AD patients. The study was conducted based on time-series data of 1,537 patients to diagnose CN vs. AD and sMCI vs. pMCI patients. Moore et al. [27] investigated the correlation between pairs of data points at varying time steps using a conventional random forest (RF) model. A combination of demographic features, physical data from brain scans, and cognitive scores was investigated for AD prediction. The use of time-series multimodal data is highly encouraged by domain experts for the development of diagnostic systems. This is because it enhances the possibility of the development of a system that is accurate, stable, and intuitive from a medical perspective, as evident from the literature [28]. Abuhamed et al. [29] proposed a two-stage long short-term memory (LSTM) model to detect AD progression. In the first stage, patients' health status was classified as CN vs. MCI vs. AD, whereas in the second stage, a regression model specifying the conversion time for pMCI patients was predicted. El-Sappagh et al. [23] proposed a cost-effective AD progression detection technique using a conventional ML approach to predict the diagnosis of a patient from one of four categories: CN, AD, pMCI, and sMCI. A fusion of multimodal longitudinal data of comorbidities, medications, and cognitive scores was used to implement, evaluate, and optimize a group of five ML models. The goal was to detect the progression of AD at M48 by analyzing the longitudinal data of a patient at four time steps (i.e., baseline [BL], month 6 [M06], month 12 [M12], and month 18 [M18]).

The ensemble machine learning (EL) algorithms have demonstrated promising results in various clinical applications [30,5,31,32]. Despite the challenges faced while training an ML classifier in the medical domain, such as limited data availability and the complex nature of medical imaging data [33], EL algorithms have proven to be an effective solution. EL algorithms employ a group of decision-making systems or classifiers that apply various strategies to combine their decisions, leading to improved prediction of new data. These classifiers are typically trained together, and their decisions are then combined based on a

predefined strategy. In the medical domain, EL approaches view clinical decision making as a learning algorithm that searches for the best candidate for AD outcomes in a hypothesis space. In situations with a lack of sufficient data or medical expertise, learning algorithms and/or physicians may come up with many AD outcome hypotheses with comparable predictability levels. The combination of these classifiers or physicians enables the algorithm to make more accurate decisions and reduces the risk of selecting the wrong classifier or physician. Likewise, physicians, in general, have expertise in a particular pathology and, as a result, their diagnosis may lean toward what they are most familiar with. However, by constructing an ensemble of multiple classifiers or physicians, their decisions are combined and the risk of relying on the wrong classifier/physician is reduced. The ensemble of multiple algorithms or physicians leads to more accurate conclusions and may provide a better approximation of the true unknown outcome than any individual classifier or physician. Wo et al. [34] proposed an improved EL model to diagnose the severity of AD. This model combines three classifiers with both weighted and unweighted schemes. The authors first evaluated the importance of image-derived features for classification performance and then localized the top-rated features in brain regions related to AD progression. The primary imaging modality used in the experiments was 11 C-PIB PET; however, the authors acknowledge that the diversity of the base classifiers can be further explored. Loddo et al. [35] introduced an EL network comprising deep CNN models for AD diagnosis. The authors evaluated various CNN structures that were pretrained on the ImageNet dataset, including AlexNet [36], different variants of ResNet [37], and GoogleNet [38]. These models were then fine-tuned using the data from the ADNI, OASIS, and Kaggle MRI datasets. The most effective features were selected and utilized to train an ensemble of bagged tree models for AD diagnosis. El-Sappagh et al. [5] proposed an ensemble network of six ML classifiers for detecting AD progression. They utilized cognitive scores, medication history, and demographic features and tuned SVM, decision tree, random forest, KNN, MLP, and XGBoost classifiers. In this study, we compared the accuracy of an individual classifier with the overall accuracy achieved by a pool of ensemble frameworks. Sadat et al. [39] proposed an EL framework of six DL models, including VGG, different variants of InceptionNet, EfficientNet, and a custom DL model and evaluated them with various combinations of these classifiers using a weighted average technique. Their proposed framework was tested with the OASIS dataset, and they reported a 96% classification accuracy for early-stage diagnosis of AD.

A drawback of the existing EL-based techniques in the literature is that they are either based on classic ML-based classifiers with handcrafted features or DL-based techniques with single or multimodal data from the baseline step only. While existing literature has explored various methods for AD diagnosis, there remains a distinct lack of comprehensive models that effectively integrate longitudinal neuroimaging and cognitive data to capture the accurate progression of the disease over time. This research gap highlights the need for novel methodologies capable of leveraging multimodal data to provide accurate and clinically relevant insights into AD progression. To the best of our knowledge, none of the existing studies have utilized the feature extraction process from longitudinal 3D MRI fused with cognitive scores for detecting the progression of AD. In this study, we propose a novel DL framework that uses longitudinal MRI data from three timesteps (i.e., BL, M06, and M12) and fuses it with cognitive scores at the baseline visit to predict AD progression. We propose a lightweight convolutional autoencoder (CAE) that creates high-level feature embeddings for MRI slices at each time step. These high-level features from longitudinal MRI scans effectively model temporal dependencies in the data. Furthermore, to obtain a more energy-compact representation of the deep features, we applied the principal component analysis (PCA) technique to the deep features to remove redundant feature maps obtained from the encoder module and retain features with the highest variance. The obtained time-series features were fused with the patient cognitive scores to train the deep learning models (i.e., 1D CNN, LSTM, and GRU). A well-known Bayesian optimizer was used to optimize these baseline deep-learning classifiers.

Unlike previous studies that often focused solely on neuroimaging or cognitive scores, our methodology incorporates a multimodal perspective, offering a comprehensive understanding of AD progression. Furthermore, we employ heterogeneous ensemble networks to optimize model performance, surpassing the accuracy achieved by individual models and existing DL approaches [31,35,40]. This EL framework allows us to leverage the complementary strengths of different base classifiers, enhancing the robustness and generalizability of our model across diverse datasets and cohorts. The resulting models have been used to build several ensemble classifiers. We tested various combinations of homogenous and heterogeneous (i.e., two and three networks) ensemble models to build the best framework for AD progression detection. We considered 110 middle slices from the MRI volume of each patient at each time step in the evaluation phase and obtained class probabilities from the proposed EL model. The final probabilities obtained for 330 ($110 \times 3$) (BL, M06, and M12) MRI slices were averaged, where each probability represented a patient's health status, specifying whether the patient has had progressed to AD or remained cognitively normal. In summary, the key contributions of the proposed study are as follows:

- We proposed a heterogeneous deep ensemble model based on multimodal time series MRI data fused with cognitive scores to predict the disease progression three years later (i.e., at month 48 [M48]). We considered a larger portion of 3D MRI volume at three-time steps (110 middle slices from each volume) in the decision-making process. These slices are composed of the most critical brain tissues, such as the Hippocampus, Amygdala, and its subregions, which are mainly affected by AD.
- We propose an optimized lightweight CAE that creates high-level representational features for 2D MRI slices at each time step of the patient's longitudinal MRI volumes (i.e., BL, M06, and M12). The obtained feature maps were further compressed using PCA to achieve an energy-compact lower-dimensional representation of the input data. The evaluation results of the proposed CAE were compared with those of VGG-CAE and UNET-CAE and outperformed in all combinations of experiments.
- We further investigated the effect of multimodal data by fusing the patient cognitive scores from the baseline with the MRI data. The combination of these cross-sectional biomarkers and longitudinal MRI features was further used to optimize a list of time-series models, such as 1D CNN, LSTM, and GRU, using the Bayesian optimizer.
- We tested various combinations of ensemble networks (i.e., two- and three-base classifiers) to achieve the best ensemble framework for AD progression detection with the highest accuracy.
- A comprehensive analysis of the performance of the proposed model was performed using a real-world dataset from ADNI, with 1,692 MRI volumes. The data were longitudinal multimodal data from 3D MRI scans (i.e., BL, M06, and M12) in addition to 14 cognitive scores. The proposed framework was evaluated using a variety of experiments to 1) show the performance using a single modality, 2) show the impact of multimodality, 3) explore the role of temporal features, 4) show the effect of combining the decisions of homogenous EL models, and 5) show the effect of heterogeneous EL models on the overall performance of the proposed framework.
- Our results suggested that the heterogeneous EL model outperformed all other architectures and the existing literature of DL modeling by achieving 96%, 96%, 96%, 97%, and 95% for precision, recall, F1-score, AUC, and accuracy, respectively.
- An explainable approach was designed to visually demonstrate the progressive patterns of disease progression over the longitudinal time steps. This was accomplished by utilizing the guided grad-cam

technique to showcase the brain regions that are activated during the decision-making processes of the model.
- To evaluate the proposed model's generalizability across cohorts, we applied the Bayesian optimized trained model using ADNI data to the NACC test data. This allowed us to test the effectiveness of the model on an independently collected cohort and assess its ability to perform disease identification in new datasets.
- The proposed approach underwent meticulous evaluation by medical domain experts to ensure the alignment of its output with the disease progression detection process manually conducted by a physician. The whole ML pipeline has been guided and validated by the domain experts. Including humans in the loop of model optimization improves the trustworthiness of the resulting model. Domain experts thoroughly assessed and validated the accuracy and reliability of the model's predicting disease progression.
- The resulting assessments established a strong correlation between the proposed approach's outcomes and the diagnostic decisions made by medical professionals, thereby enhancing its clinical relevance and usability in real-world scenarios.

The remainder of this paper is organized as follows. In Section 2, the proposed material and methods are discussed. The proposed ensemble framework is presented in Section 3. Section 4 details the experimental setup, and Section 5 presents a discussion of the results obtained. Section 6 compares the study with current state-of-the-art research. Section 7 presents the proposed XAI method for visualizing the longitudinal progression of the disease, while Section 8 assesses the model's robustness on a new set of cohorts. The limitations of the study are addressed in Section 9, and the paper is concluded in Section 10 with suggestions for future directions for extending the study.

## 2. Materials and methods

This section discusses the essential building blocks in the design of a novel EL framework for detecting AD progression. The proposed study was conducted based on longitudinal 3D MRI at three time steps (BL, M06, and M12) and the patients' cognitive scores. We trained a lightweight convolutional autoencoder with 2D MRI slices that produced a highly representational feature at each time step. We further utilized the PCA technique on deep features to obtain a more energy-compact representation of the input features. Using the Bayesian optimization approach, these features were utilized to tune a group of time-series models (i.e., 1D CNN, LSTM, and GRU). We tested different combinations of homogenous and heterogeneous EL networks with multimodal data to investigate their effects on AD progression.

### 2.1. Convolutional autoencoder (CAE)

An autoencoder (AE) comprises two submodules: an encoder and decoder. Each module is a subnetwork that creates an input-output relationship. The main purpose of an AE is to represent high-dimensional feature information in a low-dimensional feature space. In contrast, a CAE is a better-suited neural network for image processing because it utilizes the full potential of CNNs to exploit images. In CAE, the local spatiality of the input location is preserved owing to the shared weights across all input locations. For a grayscale image input, the $i$th feature map is represented as

$$h^i = s\left(x * \quad W^i + b^i\right) \tag{1}$$

where b is a broadcasted bias across all feature maps, * represents the convolution operation between trainable parameters $W^i$ and feature maps $x$, and $s$ is an activation function.

In the decoding step, the bias term is applied to each latent map obtained from the encoder module, and reconstruction is performed as follows:

$$y = s\left(\sum_{i\in H}\left(h^i * \quad \widehat{W}^i + c\right)\right) \tag{2}$$

where $c$ represents the bias term, $h^i$ represents the latent feature maps from the encoder network, and $\widehat{W}^i$ shows a flip operation over both weight dimensions. The back-propagation method calculates the gradient of the error function for a given parameter set.

### 2.2. Principal component analysis (PCA)

PCA is a general-purpose tool for dimensionality reduction and data analysis [41]. PCA plays a critical role in a wide range of research fields, including pattern recognition, artificial intelligence, and data mining. The core idea of PCA is the linear projection of high-dimensional data samples into a low-dimensional space during the computational process. For instance, in a fine-grained image recognition task, the local and global features of an input image are considered to distinguish subordinate categories from entry-level categories. The list of common features includes a color histogram, texture features, and edge shapes. If all these features are considered in the computation process, hundreds or even thousands of feature vectors will be processed for a single image. There must be a method to extract the most important features from the available data, where the original data are replaced by a linear combination of these features. The main aim of PCA is to reduce the dimensions of the original image by extracting features with a significant variance and removing components with a low variance. Features with large variances allow us to keep a larger amount of information compact, while features with low variance are eliminated because they refer to the less important components of data. PCA is applied to the MRI data owing to its mathematical properties, such as the eigenvalue decomposition of the covariance matrix of the data ($\sum$).

Eq. 3 demonstrates the process of deriving principal components from multidimensional feature vectors through eigenvalue decomposition.

$$\sum = \quad A\Lambda A^T \tag{3}$$

The eigenvector matrix $A = (a_1, a_2, \ldots, a_N)$ and the calculated diagonal matrix $\Lambda$, composed of eigenvalues, are used in this calculation. By utilizing the first K eigenvectors of $A$, a transformed feature vector $z^i$can be obtained from an original feature vector element $xi$ (pixel of a 2D MRI slice) via Eq. 4:

$$\begin{aligned} z_i &= [z_1 \quad z_2 \quad \vdots z_k \quad] \\ &= [a_{11}\cdots \quad a_{1N}\vdots \quad \ddots \quad \vdots \quad a_{K1}\cdots \quad a_{KN}][x_1\vdots \quad x_N] \end{aligned} \tag{4}$$

Another noticeable aspect of Eq. 4 is that the transformed principal components have no correlation with each other, based on the properties of the eigenvectors. In this study, because we are dealing with $h \times w \times d$ dimensional feature maps from the encoder module, neighboring channels may have redundant feature maps. PCA can remove this correlation through dependency on the principal components. The variance explained through the first few principal components can be represented as

$$\frac{\sum_{i=1}^{k}\lambda_i}{\lambda_1 + \quad \lambda_2 + \quad \lambda_3 + \ldots \quad \lambda_N} \tag{5}$$

To achieve an efficient dimensionality reduction, it is important to determine the appropriate number of principal components. This is because the first principal component explains the highest variance in information, with the explained variance decreasing as more components are added.

### 2.3. 1D convolution neural network (1D CNN)

In the field of artificial neural networks, 2D CNNs are deep

feedforward networks modeled after the mammalian visual cortex. In contrast, compact 1D CNNs are designed to work with 1D signals and are optimized for applications that have limited labeled data and high signal variation from various sources (e.g., patients, devices, motors, or circuits). These networks consist of two types of layers:1) CNN layers, which perform both 1D convolution and subsampling, and 2) fully connected layers, also known as multilayer perceptron (MLP). The configuration of a 1D CNN is determined by several key factors, including the number of hidden CNN and MLP layers/neurons, size of the filter (kernel) in each CNN layer, subsampling factor in each CNN layer, and choice of pooling and activation operators. One of the main differences between 1D and 2D CNNs is that, while 2D CNNs use 2D matrices for kernels and feature maps, 1D CNNs utilize 1D arrays. The CNN layers in a 1D CNN are responsible for processing raw 1D data and learning to extract relevant features that are subsequently used in the classification task performed by the MLP layers. This integration of feature extraction and classification into a single process result in improved classification performance, which can be optimized. Furthermore, 1D CNNs have low computational complexity owing to the presence of only one computationally intensive operation, which is a series of 1D convolutions that are linear weighted sums of two 1D arrays. These operations can be performed in parallel for both forward and backward propagations. Fig. 1 illustrates the functional mechanism of a 1D CNN.

### 2.4. Long short-term memory (LSTM)

LSTM networks are an advanced version of RNNs, primarily designed to address the vanishing and exploding gradient problems [42]. A conventional LSTM cell consists of three gates that control the flow of information in the network, namely, the forget gate $f_t$, input gate $i_t$, and output gate $O\_t$. The gates are defined as follows:

$$f_t = \sigma(W_f * \quad X_t + \quad R_f * \quad h_{t-1} + \quad b_f)$$

$$i_t = \sigma(W_i * \quad X_t + \quad R_i * \quad h_{t-1} + \quad b_i)$$

$$O_t = \sigma(W_o * \quad X_t + \quad R_o * \quad h_{t-1} + \quad b_o)$$

where $X\_t$ is the input data, $h_{t-1}$ is the previous hidden state, W and R are matrices of trainable parameters, and b represents the vectors of the trainable biases. The sigmoid function σ is used to ensure that the values of these gates are between 0 and 1. The following equations show the mathematical form of the LSTM units.

$$C_t = \tanh(W_c * \quad X_t + \quad R_c * \quad h_{t-1} + \quad b_c)$$

$$C_t = \quad f_t * \quad C_{t-1} + \quad i_t * \quad C_t$$

$$h_t = \quad f_t * \tanh(C_t)$$

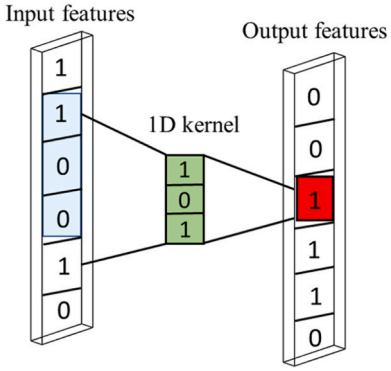


**Fig. 1.** Illustration of 1D CNN.

$$y_t = \sigma(W_y * \quad h_t + \quad b_y)$$

The candidate cell states $C_t$ are calculated based on the input data $X_t$ and previous hidden state $h_{t-1}$. The current cell state $C_t$ is determined by the forget gate $f_t$, previous cell state $C_{t-1}$, input gate $i_t$, and candidate cell state $C_t$. The Hadamard product Θ is used to elementwise multiply the matrices involved in these computations. Finally, the output $y_t$ is computed by applying the relevant weights $W_y$ and biases $b_y$ to the hidden state $h_t$.

### 2.5. Gated recurrent unit (GRU)

The LSTM network was first introduced in 1997 and has become widely recognized for its ability to effectively retain long-term dependencies in language processing. However, its complex design results in lengthy training processes. To address this, GRU networks were developed as a modified version of LSTMs with reduced complexity and improved accuracy [43]. GRUs are similar to LSTMs but have fewer parameters and two gates instead of three: update gate $U_t$and and reset gate $r_t$. The update gate regulates the state of the hidden layers in the model, whereas the reset gate determines how much of the past information is retained by determining certain aspects of the memory. The equations below show the GRU units, where $\dot{h}_t$ represents the candidate hidden state.

$$u_t = \sigma(W_u \quad x_t + \quad R_u \quad h_{t-1} + \quad b_u)$$

$$r_t = \sigma(W_r \quad x_t + \quad R_r \quad h_{t-1} + \quad b_r)$$

$$\dot{h}_t = \tanh(W_h \quad x_t + (r_t \quad \odot h_{t-1})R_h + \quad b_h)$$

$$h_t = (1 - \quad u_t) \quad \odot h_{t-1} + \quad u_t \quad \odot \quad \dot{h}_t$$

$$y_t = \sigma(W_y \quad h_t + \quad b_y)$$

### 2.6. Network optimization using a Bayesian optimizer

The main goal of hyper-parameter tuning is to select the optimum hyperparameter set for a particular machine learning algorithm to obtain the best results when tested on a test set [44]. The general expression of the optimization method for any ML-based algorithm can be represented by Eq. 6, where f is the performance, x is a hyperparameter setting, and $x_{opt}$ is the optimum choice.

$$x_{opt} = \quad argmax_x \quad _{\in X} \quad f(x) \tag{6}$$

Eq. 6 is a function of $x_{opt}$ that maximizes the $f(x)$ function: The components of $x_{opt}$, which denote the hyperparameter set, may be continuous real, integers, or categorical values, meaning that all name sets are discrete. The term $f(x)$ is an objective function that a model attempts to minimize on a validation dataset. The objective function is also known as a Gaussian Process (GP). $X$ represents the search space for the $x$ hyperparameter. The Bayesian optimization method is derived from Bayes' theorem, which uses prior knowledge to make inferences. Eq. 7 represents Bayes' theorem. In Bayes' theorem, the posterior distribution is directly proportional (∝) to the product

of the likelihood $P(Z|X)$ and a priori distribution $(P(X))$. In contrast, the prediction performed by $(P(X))$ is also known as "belief" [45].

$$P(X|Z) \propto P(Z|X)P(X) \tag{7}$$

The hyperparameter tuning process can be achieved in several ways, and the most commonly adopted methods are grid and random searches. These methods are generally suitable for tuning the hyperparameters of the model; however, the evaluation of hyperparameters during the tuning process does not provide any learning knowledge. Therefore, we turn to Bayesian optimization to overcome the disadvantages of grid and random search. During the Bayesian optimization process, a probability

model is first created that converts the values of the hyperparameters into the probabilities of obtaining a particular value of the objective function, also known as the score. Using these scores, the most promising values for the hyperparameters are selected for the evaluation of the objective function. In the proposed framework, we used a Bayesian optimizer to tune the hyperparameters of each time-series model (i.e., 1DCNN, LSTM, and GRU). In the testing phase, we first collected the probabilities from each time-series model generated for the entire set of 110 2D MRI slices fused with CSs and then averaged the resultant 330 probabilities to compute the final prediction for CN or AD patients. Fig. 2 in Section 3 explains the pipeline for information fusion and model optimization in the proposed framework.

### 2.7. Ensemble models

Ensemble learning (EL) involves utilizing multiple decision-making systems to improve the accuracy of new data predictions [46,31,32]. The three primary ensemble learning techniques are bagging, boosting, and stacking. These techniques can considerably enhance the generalization performance of the learning system. The most common methods for generating individual decision-making systems in an EL are heterogeneous and homogenous classifiers [47]. In the heterogeneous approach, different learning algorithms are applied to the same training data, whereas in the homogenous approach, the same learning algorithms are applied to different training sets. Strategies for combining results from individual decision-making systems include averaging, voting, and learning. The choice of method depends on the intended use of integrated learning. For example, if the goal is a regression task, the results from each system may be averaged or weighted. In the case of a classification task, the output probabilities from each classifier are voted on to obtain the final result [34]. There are two types of voting: absolute majority voting, in which more than half of the individual systems produce the same output, and relative majority voting, in which most systems produce a specific output. The result of relative majority voting is considered the final outcome of the integrated learning.

## 3. Proposed ensemble framework

As shown in Fig. 2, the proposed framework was designed to utilize the best input feature set to accurately detect AD progression. All MRI volumes were first pre-processed using a standard pre-processing pipeline. It has been widely shown in the literature that non-preprocessed volumes can significantly reduce the models' performance because of the existence of non-necessary skull regions in the training data and can act as noise [48]. After the preprocessing step, we divide the entire dataset into training and testing sets with an 80% and 20% split ratio. We then extract 110 2D coronal slices from the middle of the MRI volume and save it as a new 3D volume of size $h \times w \times$ 110 for each time step (i.e., BL, M06, and M12). It has been proven from the literature that the middle slices from MRI scans represent the most informative brain regions that are highly vulnerable to neurodegenerative diseases [24]. In addition, the coronal plane is known for its best representational view of specified brain tissues that are commonly affected by AD, such as the hippocampus and amygdala [49,50]. In this manner, only the most informative 2D slices from the 3D volume are processed by the proposed framework, which can capture the spatiotemporal features from the multimodal three-dimensional longitudinal nature of the training data. Our training set was composed of 451 patients with 1353 (451 $\times$ 3 time steps) MRI volumes, whereas the test set contained 113 patients with 339 (113 $\times$ 3 time steps) MRI volumes. Then, we trained an autoencoder with 2D MRI slices to learn the best representation of the input image in the latent feature space. Owing to the redundant feature maps in the feature set obtained from the encoder module, we further processed each feature vector using PCA to obtain a compact representation of the deep vector. The extracted feature vector for all three time steps is fused with CSs in late fusion manner and employed for optimizing the hyperparameter of each time series model (i.e., 1D CNN, LSTM, and GRU). A detailed discussion of the multimodal features fusion and optimized hyperparameters is presented in Table 2, Table 3, and Table 4. We further tested various combinations of homogenous and heterogeneous ensemble networks of time-series models with feature embeddings collected from different CAEs (i.e., VGG-CAE, UNET-CAE, and the proposed CAE). The main aim was to build the best combination of ensemble networks with the best set of feature embeddings.

### 3.1. Image preprocessing

Image preprocessing refers to the removal of non-essential information in RAW MRI volumes. These steps allow accurate comparisons of different-sized brain scans by domain experts and researchers. In this study, we preprocessed

the raw MRI volumes using a predefined pipeline [49,51,52]. The resulting volumes were subsequently subjected to visual inspection by our medical domain experts to validate their quality and accuracy. The results showed that model performance was significantly improved compared to using the raw MRI volumes, and the computation process was reduced as the data only contained disease-related information (brain tissue only, with no skull or neck regions in the MRI).

The preprocessing steps include:

1. Ensure the orientation of MRI: In the first step, we manually checked the orientation of each MRI volume by visualizing them using FSLView. This ensured that each MRI volume was correctly positioned according to the standard template. This step is crucial because discrepancies in orientation can lead to inaccuracies in subsequent processing steps, such as registration to a common template space. Ensuring consistent orientation across all MRI volumes enhances the accuracy and reliability of downstream analyses, facilitating meaningful comparisons and interpretations of the neuroimaging data [53].
2. In the second step, we performed bias field correction on each volume. This process was done using the N4BiasFieldCorrection package from the Advanced Normalization Tools (ANT) [54]. N4BiasFieldCorrection utilizes a non-parametric approach based on B-spline fitting to estimate and remove the bias field from each MRI volume. The algorithm iteratively estimates the bias field while simultaneously correcting for intensity inhomogeneity, resulting in improved image quality and more accurate subsequent processing steps [49].
3. In the third step, we conducted image registration to align each MRI volume with the MNI152 template space, a widely used standard anatomical reference in neuroimaging research. The registration process was performed using the FLIRT (FMRIB's Linear Image Registration Tool) package, a component of the FSL (FMRIB Software Library) suite, executed via the command prompt. The parameters specified in the FLIRT command were carefully chosen to optimize the registration process. Specifically, we configured FLIRT with 256 bins for histogram matching, allowing for robust intensity-based alignment between the MRI volumes and the template. Additionally, we employed 12 degrees of freedom to account for translation, rotation, and scaling transformations, providing flexibility to capture spatial variations across subjects. To interpolate voxel intensities during the registration process, we utilized spline interpolation, which preserves image detail while minimizing interpolation artifacts. This choice of interpolation method contributes to the accuracy of the registration, ensuring that anatomical structures are appropriately mapped onto the template space. Moreover, we employed correlation ratio as the cost function, which measures the similarity between intensities in the MRI volumes and the template. By maximizing this similarity metric, FLIRT optimized the registration parameters to achieve the best alignment between the individual MRI volumes and the MNI152 template [55].

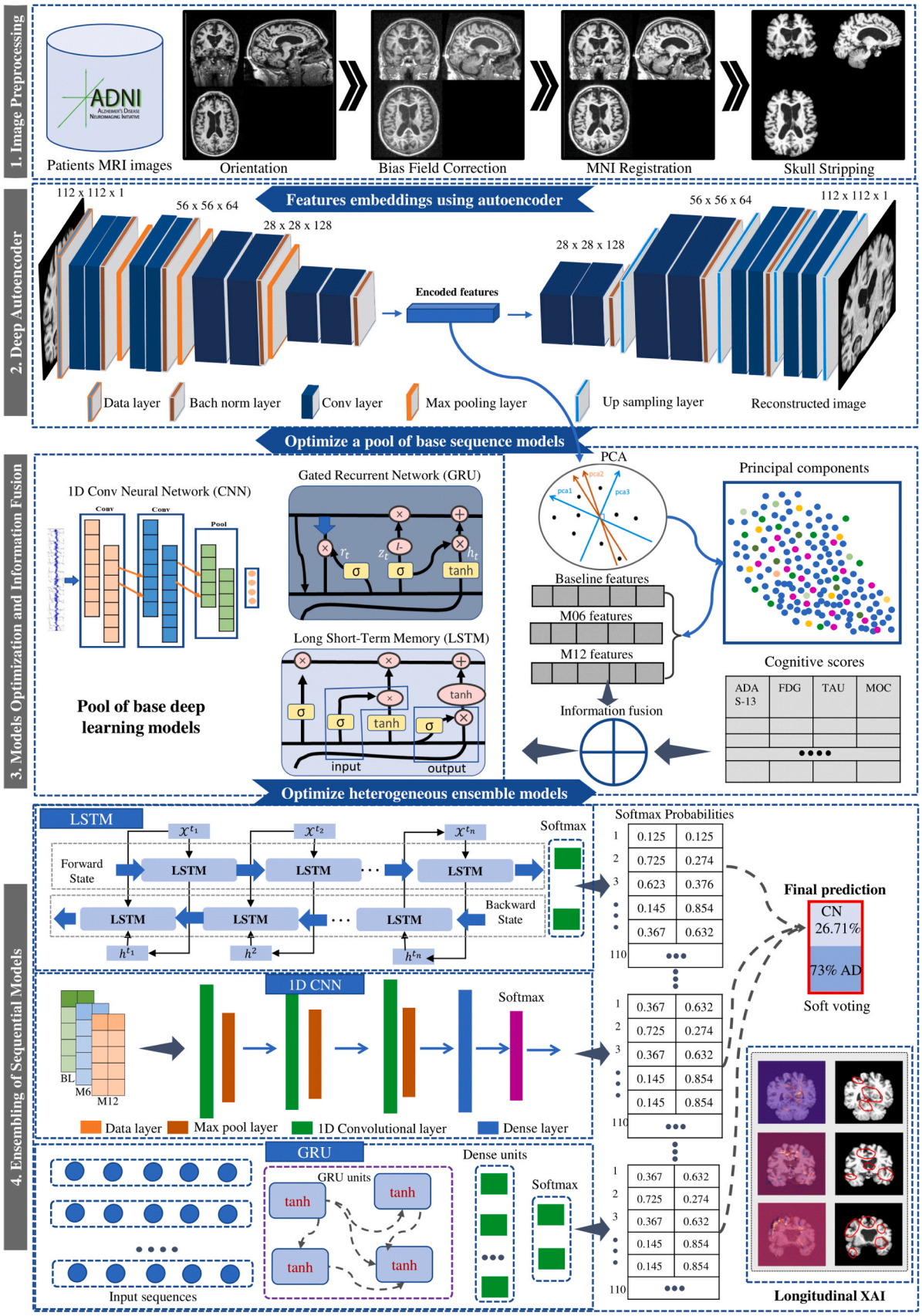


**Fig. 2.** Proposed framework for detection of AD progression.

4. Finally, we conducted skull stripping, a crucial preprocessing step aimed at isolating the brain tissue from non-brain structures, using the Brain Extraction Tool (BET2) [56] package within the FSL suite. In skull stripping, we specified optional parameters to guide the BET2 algorithm in accurately delineating the brain tissue. These parameters included a fractional intensity threshold, which determines the sensitivity of the algorithm in detecting brain boundaries. By setting the fractional intensity threshold to 0.5, we ensured that voxels with intensities lower than half of the maximum intensity within the image were considered as non-brain tissue, effectively delineating the brain from surrounding structures. By carefully selecting these parameters, we aimed to optimize the performance of the BET2 algorithm in accurately extracting the brain tissue while minimizing the inclusion of non-brain structures [57].

### 3.2. Deep convolutional autoencoder (CAE)

In this study, we proposed a lightweight CAE that produces high-level representational features to input images in a latent feature space. Compared with the conventional AE, the proposed framework is a significantly improved version of the unsupervised AE, in which it collaborates with CNN to extract features from the image. Instead of stacking deeper layers sequentially using a convolution + max pooling layer, we maintained the spatial dimensions of the input image in two consecutive convolution layers to preserve the local structure of the brain tissues. This strategy helps stabilize the training process and avoids corruption of the feature space, whereas the batch normalization layer prevents overfitting of the network. The encoder module comprises six convolution layers and four max-pooling layers. Max pooling is utilized after the second, fourth, and sixth convolution layers to preserve the spatial features of the input image. Instead of using max pooling after every convolution layer, we deployed it after two convolution layers, allowing the network to extract valuable information in the spatial dimension. The first two convolution layers each used 32 kernels of size $3 \times 3$, followed by batch normalization and max-pooling. The number of kernels increases in the deeper layers as the spatial dimension of the feature maps decreases, with 64 kernels in the third and fourth layers and 128 kernels in the fifth and sixth layers. At this point, we obtained HxWx128 feature maps, providing a latent representation of the input image. The obtained feature vector for the input 2D slice at each time step (i.e., BL, M06, and M12) was further processed through PCA to obtain a more compact representation of the feature vector with fewer dimensions. The decoder module of the proposed CAE comprises two convolutional layers of size $3 \times 3$ with 64 kernels, followed by a batch normalization layer and an up-sampling layer. This was followed by two more convolution layers, each with 32 kernels of size $3 \times 3$, along with a batch normalization layer and an upsampling layer. In the final layer, a single kernel of size $3 \times 3$ was applied to the obtained feature maps, resulting in a single channel (grayscale) 2D MRI slice. The rectified linear unit (ReLU) activation function was utilized in all hidden layers. The loss function used was binary cross-entropy, calculated between the reconstructed and labelled images (where the labelled image is the same input image used for model optimization during the training process). The proposed CAE was trained on $112 \times 112 \times 1$ grayscale images for 150 epochs using the Adam optimizer with a learning rate of 0.0001 and a batch size of 16. The arrangement of the convolution and up/down-sampling layers in the CAE architecture was determined through a trial-and-error process, involving extensive experimentation with various architectural designs. Table 1 lists the architectural design of the proposed CAE network. Fig. 3 shows the input and reconstructed images of the proposed lightweight CAEs.

### 3.3. Feature compression using PCA

To obtain a more energy-compact representation of the extracted deep features, we applied PCA [41] to the feature vector obtained from the encoder part of the proposed network. Because many filters in the deeper layers are highly correlated and potentially detect the same features in multiple feature maps, they make insignificant contributions to accuracy and unnecessarily increase the computational cost. To remove redundant features from the obtained feature maps, we applied PCA to the output feature maps of the encoder module. This process reduces the highly correlated insignificant and redundant maps to a compact feature set by retaining only the most useful information. We took the output feature maps from the encoder module ($14 \times 14 \times 128$), fed them to PCA, and produced a compact feature vector. To determine the best compact representation, we generated different vector sizes (i.e., 2048, 1024, 512, and 256). The same process was repeated for all the slices at each time step in the longitudinal dataset (BL, M06, and M12). For each PCA representation setting (i.e., 2048, 1024, 512, and 256), we fed the DL models with the PCA representation and found that the performance achieved with a 1024-dimensional feature set was not significantly different from the performance achieved with a 2048-dimensional feature vector. Furthermore, the performance achieved with the 512 and 256 feature sets showed degradation in the overall accuracy. Therefore, we used the

**Table 1**
Design of the proposed CAE.

| Layer ID | Layer Name | Number of Kernels | Kernel Size/ Stride | Stride & Padding | Output Size |
|---|---|---|---|---|---|
| 0 | Input | - | - | - | $112 \times 112 \times 1$ |
| 1 | Conv1 | 32 | $3 \times 3$ | 1, 1 | $112 \times 112 \times 32$ |
| 2 | Conv2 | 32 | $3 \times 3$ | 1, 1 | $112 \times 112 \times 32$ |
| 3 | Batch normalization | 32 | - | - | $112 \times 112 \times 32$ |
| 4 | Max pooling | - | $2 \times 2$ | 2, - | $56 \times 56 \times 32$ |
| 5 | Conv | 64 | $3 \times 3$ | 1, 1 | $56 \times 56 \times 64$ |
| 6 | Conv | 64 | $3 \times 3$ | 1, 1 | $56 \times 56 \times 64$ |
| 7 | Batch normalization | 64 | - | - | $56 \times 56 \times 64$ |
| 8 | Max pooling | - | $2 \times 2$ | 2, - | $28 \times 28 \times 64$ |
| 9 | Conv | 128 | $3 \times 3$ | 1, 1 | $28 \times 28 \times 128$ |
| 10 | Conv | 128 | $3 \times 3$ | 1, 1 | $28 \times 28 \times 128$ |
| 11 | Batch normalization | 128 | - | - | $28 \times 28 \times 128$ |
| 8 | Max pooling | - | $2 \times 2$ | 1, 1 | $14 \times 14 \times 128$ |
| 12 | Up sampling2d | - | $2 \times 2$ | 2, - | $28 \times 28 \times 128$ |
| 13 | Conv | 64 | $3 \times 3$ | 1, 1 | $28 \times 28 \times 64$ |
| 14 | Conv | 64 | $3 \times 3$ | 1, 1 | $28 \times 28 \times 64$ |
| 15 | Batch normalization | 64 | - | - | $28 \times 28 \times 64$ |
| 16 | Up sampling2d | - | $2 \times 2$ | 2, - | $56 \times 56 \times 64$ |
| 17 | Conv | 32 | $3 \times 3$ | 1, 1 | $56 \times 56 \times 32$ |
| 18 | Conv | 32 | $3 \times 3$ | 1, 1 | $56 \times 56 \times 32$ |
| 19 | Batch normalization | 32 | - | - | $56 \times 56 \times 32$ |
| 20 | Up sampling2d | - | $2 \times 2$ | 2, - | $112 \times 112 \times 32$ |
| 21 | Conv | 1 | $3 \times 3$ | 1, 1 | $112 \times 112 \times 1$ |

Padding: valid = 0, same = 1

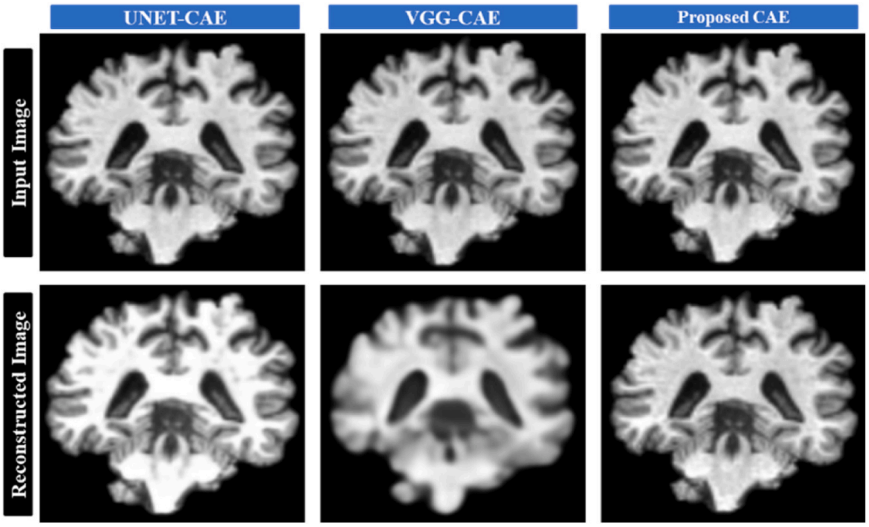


**Fig. 3.** Reconstructed images using different autoencoders.

1024-dimensional feature vector as the representational feature vector for each 2D MRI.

### 3.4. Bayesian optimization of time series models

In this study, we aimed to propose optimized models for time-series data analysis and then created an ensemble of these models to achieve ensemble-based AD progression detection. We optimized three deep learning-based time-series models using a Bayesian optimizer: 1D CNN, LSTM, and GRU. A Bayesian optimizer with five-fold cross-validation (CV) was employed for model optimization, and the mean precision, mean recall, mean F1 score, mean AUC, and mean accuracy for each optimized model were reported. Furthermore, the achieved accuracies were reported for a single time step, that is, baseline data only (BL), two time steps (BL∼M06), and three time steps (BL∼M12). The reported accuracies for every single time step in the training data show the effectiveness of adding more time steps to the performance of detecting AD progression. We also investigated and compared the accuracies achieved with embeddings obtained from VGG-CAE and UNET-CAE. In addition, we investigated the effect of multimodal data, that is, MRI-based embeddings fused with cognitive scores. Experiment 1 shows the performance output of Bayesian optimized models using a single modality (i.e., MRI embeddings). Experiment 2 shows the performance results of Bayesian optimized models using multimodality (i.e., MRI embeddings + cognitive scores).

*Bayesian optimization of 1D CNN:* We used a Bayesian optimizer to optimize various hyperparameters to design the proposed 1D CNN for progression detection. For the proposed framework of the 1D CNN, we optimized the number of convolutional layers, size of the convolutional kernels, number of dropouts, number of dense layers, number of dense units, and the learning rate. The search space for the proposed 1D CNN comprised four convolutional layers and three dense layers, followed by an output layer. The Bayesian optimization algorithm chooses the best hyperparameter set from the given parameter space. Table 2 presents the architectural design of the optimized 1D CNN. The search space column represents the range of kernels, size of kernels, and range of the dropout threshold, and the Bayesian optimizer chooses the best parameter sets from the search space column. Notably, we increase the range of learnable parameters in the deeper layer to learn more abstract representations of the input data. The Adam optimizer was used to train the network weight with the learning rate options set to 1e-2, 1e-3, and 1e-4. We set the step size to 32 in each layer, specifying the distance between two consecutive sample values in the given range. The network was regularized by deploying a batch normalization layer and dropout layer to avoid overfitting.

Furthermore, we utilized an early stopping callback to stop unnecessary training of the network if no improvement in the validation accuracy was achieved. The optimizer was run for 25 epochs with a batch

**Table 2**
Optimized 1D CNN architecture developed by the Bayesian optimizer.

| Layer ID | Layer Type (Name) | Search Space Range (Kernels/Kernel size/Dropout) | Number of Kernels/Dropout | Kernel Size/Stride | Output Size |
|---|---|---|---|---|---|
| 0 | Input | - | - | - | 3072 × 1 |
| 1 | Conv1D_1 | 96–128/3–5/- | 96/- | 5×1/2 | 1535 × 96 |
| 2 | Batch_Normalization_1 | -/-/- | 96/- | - | 1535 × 96 |
| 3 | Max_Pooling1d_1 | -/-/- | -/- | 2/2 | 767 × 96 |
| 4 | Conv1D_2 | 128–160/3–5/- | 128/- | 3×1/1 | 765 × 128 |
| 5 | Batch_Normalization_2 | -/-/- | 128/- | - | 765 × 128 |
| 6 | Max_Pooling1D_2 | -/-/- | -/- | 2/2 | 382 × 128 |
| 7 | Conv1D_3 | 128–192/3–5/- | 162/- | 3×1/1 | 380 × 162 |
| 8 | Batch_Normalization_3 | -/-/- | 162/- | - | 380 × 162 |
| 9 | Max_Pooling1D_3 | -/-/- | -/- | 2/2 | 190 × 162 |
| 10 | Dropout_1 | -/-/0.1–0.5 | -/0.3 | - | 190 × 162 |
| 11 | Flattening Layer | -/-/- | -/- | - | 30780 + 17 |
| 12 | Dense_1 | -/-/512–104 | -/- | - | 832 |
| 13 | Dropout_2 | -/-/0.1–0.5 | -/0.3 | - | 832 |
| 14 | Dense_2 | -/-/64–128 | - | - | 567 |
| 15 | Dropout_2 | -/-/0.1–0.5 | -/0.1 | - | 64 |
| 16 | SoftMax | -/-/- | -/- | - | 2 |

size of 16 and maximum number of trials of 120. We chose the maximum number of trials to 120 to produce a wide range of search spaces for the Bayesian optimizer to select the best parameter set for the proposed 1D CNN. The hyperparameter set returned by the Bayesian optimizer with the highest possible accuracy comprised three convolutional layers, followed by two dense layers and an output layer. The chosen learning rate was set to 1e-3. The architectural details of the optimized 1D CNN are presented in Table 2.

*Bayesian optimization of LSTM:* We also utilized an optimized version of the LSTM network that detects AD progression from the longitudinal feature embeddings of MRI fused with cognitive scores. A Bayesian optimizer was used to select the best hyperparameter set, including the number of LSTM layers, number of LSTM cells in each layer, activation function, dropout threshold, number of dense layers, number of dense units in each layer, and learning rate. The hyperparameter search space for the proposed LSTM network comprised five LSTM layers and three dense layers, followed by an output layer. Table 3 shows the architectural design of the LSTM network and the hyperparameter search space. The search space column represents the range of LSTM units, choice of activation function, and the range of the dropout threshold that the Bayesian optimizer will search from. The parameter-selected column represents the chosen parameters, whereas the regularization column represents the amount of dropout/L1 regularizers. The Adam optimizer was used to train the network weight with the learning rate options set to 1e-2, 1e-3, and 1e-4. The step size was set to 32, and early stopping callback was utilized to stop unnecessary training of the network. The optimizer was run for 50 epochs with a batch size of 16, and the maximum number of trials was set to 120. The best-returned hyperparameter set comprised four LSTM layers, followed by a single dense layer and an output layer. The chosen learning rate was set to 1e-4. Table 3 lists the optimized hyperparameters for the proposed LSTM network.

*Bayesian optimization of GRU:* In several studies, GRU performed better than LSTM networks because of its simpler architectural design and fewer trainable parameters. Therefore, in this study, we used an optimized version of the GRU network to detect AD progression using multimodal data (longitudinal MRI + CS). A Bayesian optimizer was used to select the best hyperparameter set for the proposed GRU network, including the number of GRU layers, number of GRU units in each layer, activation function, number of dropouts, number of dense layers, number of dense units in each layer, and learning rate. The hyperparameter search space for the proposed GRU network was composed of five GRU hidden layers and three dense layers, followed by an output layer. The Bayesian optimizer searches for the best hyperparameter set in the specified parameter space. The Adam optimizer was used for weight optimization with the learning rate options set to $1e-2$, $1e-3$, and $1e-4$. The step size was set to 32, and an early stopping callback was used to stop unnecessary network training. The input batch size was set to 16, and the number of epochs was set to 32, with a maximum trial of 120. Table 4 lists the optimized hyperparameters selected by the Bayesian optimizer. The proposed GRU network comprises three GRU layers and two dense layers followed by an output layer.

### 3.5. Ensemble of sequential models

In our proposed framework, we examined an ensemble of diverse combinations of Bayesian-optimized base classifiers for the detection of AD progression using longitudinal MRI. In addition, we examined the impact of incorporating multimodal data on disease identification. After optimizing each model using the Bayesian optimizer, we explored the prediction output of various combinations of ensemble models, such as 1D CNN + LSTM, 1D CNN + GRU, LSTM + GRU, and 1D CNN + LSTM + GRU. Longitudinal embeddings (i.e., BL, M06, and M12) fused with cognitive scores for 110 slices were passed through a pool of ensemble networks, and probabilities from the base classifiers were collected for each class. Subsequently, class-wise soft voting was applied to the collected probabilities and averaged to obtain each class's final probabilities. The maximum probability value specifies the target class label for input data representing a cognitively normal or progressive AD patient. Fig. 5 shows an example of multimodal data processed through the proposed EL framework for single-patient disease diagnosis.

## 4. Experimental setup

We tested our model using the GPU versions of Keras 2.0, Keras Tuner 1.2.0, and NVIDIA TITAN X GPU with 32 GB of memory and 128 GB of RAM. We carefully prepared the training data to prevent any type of data leakage. We repeated each experiment five times by employing five-fold CV and reported the average performance using evaluation metrics including mean precision, mean recall, mean F1-score, mean AUC, and mean accuracy.

### 4.1. Dataset

We utilized the dataset obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a widely used open-source platform for research purposes [58]. The ADNI was established in 2003 as a public-private partnership, with an initial capital budget of $60 million earmarked for a five-year period. The primary objective of the program was to investigate the feasibility of using serial MRI, PET, and other clinical assessments, biomarkers, and neuropsychological evaluations, to monitor the progression of MCI and detect early signs of neurodegenerative disease. Early diagnosis of AD progression using critical biomarkers would significantly benefit physicians and researchers working toward developing new therapies and improving the efficacy of

**Table 3**
Bayesian optimization of the LSTM model.

| Layer ID | Layer (Name/ Type) | Search Space Range (Units/Activation Function/ Dropout/L1) | Parameters Selected (LSTM/Dense/Activation Function) | Regularization (Dropout/L1) | Output Shape |
|---|---|---|---|---|---|
| 0 | Input | -/-/-/- | -/-/- | -/- | 3 × 1024 |
| 1 | LSTM_1 (LSTM) | 416–576/-/-/0.1–0.001 | 512/-/- | -/0.003 | 3 × 512 |
| 2 | Activation Function | -/ [Sigmoid, Tanh]/-/- | -/-/Tanh | -/- | 3 × 512 |
| 3 | LSTM_2 (LSTM) | 440–680/-/-/0.1–0.001 | 512/-/- | -/0.005 | 3 × 512 |
| 4 | Activation Function | -/ [Sigmoid, Tanh]/-/- | -/-/Tanh | -/- | 3 × 512 |
| 5 | LSTM_3 (LSTM) | -/440–680/-/- | 384/-/- | -/0.002 | 3 × 384 |
| 6 | Dropout_1 (Dropout) | -/0.1–0.5/-/- | -/-/- | 0.3/- | 3 × 384 |
| 7 | Flattening Layer | -/-/-/- | -/-/- | -/- | 1152 + 17 |
| 8 | Dense_1 (Dense) | 256–512/-/- | 416 | -/- | 416 |
| 9 | Dropout _2 (Dropout) | -/0.1–0.5/-/- | -/-/- | 0.2/- | 416 |
| 10 | Dense _3(Dense) | -/-/- | Softmax | - | 2 |

**Table 4**
Bayesian optimization of the GRU model.

| Layer ID | Layer (Name/ Type) | Search Space Range (Units/Activation Function/ Dropout/L1) | Parameters Selected (GRU/Dense/Activation Function) | Regularization (Dropout/ L1) | Output Shape |
|---|---|---|---|---|---|
| 0 | Input | -/-/-/- | -/-/- | -/- | 3 × 1024 |
| 1 | GRU_1(GRU) | 256–512/-/-/0.001–0.1 | 384/-/- | -/0.002 | 3 × 512 |
| 2 | Activation Function | -/ [Sigmoid, Tanh]/-/- | -/-/Tanh | - | 3 × 512 |
| 3 | GRU_2 (GRU) | 416–576/-/-/0.001–0.1 | -/512/- | -/0.006 | 3 × 384 |
| 4 | Activation Function | -/ [Sigmoid, Tanh]/-/- | -/-/Tanh | - | 3 × 384 |
| 5 | GRU_3 (GRU) | 416–576/-/-/0.001–0.1 | -/320/- | -/0.004 | 3 × 384 |
| 6 | Activation Function | -/ [Sigmoid, Tanh]/-/- | -/-/Tanh | -/- | 3 × 384 |
| 7 | Flattening Layer | -/-/-/- | -/-/- | -/- | 1152 + 17 |
| 8 | Dense_1 (Dense) | 256–512/-/-/- | -/512/- | - | 512 |
| 9 | Dropout_1 | -/-/-/0.1–0.5 | -/-/- | 0.4 | 512 |
| 10 | Dense _2 (Dense) | -/-/-/- | -/64/- | - | 64 |
| 11 | Dense_3 (Dense) | -/-/-/- | Softmax | - | 2 |

treatment. Additionally, this would reduce the time and costs associated with conducting clinical trials. This study included data from three classes including cognitively normal subjects, progressed to AD, and AD subjects. Note that all progressed to AD subjects are converted by the month 48. The defining criteria of each group are defined as follows. (1) Cognitively normal subjects: MMSE scores between 24 and 30 (inclusive), a CDR of 0, non-depressed, non-MCI, and nondemented; (2) Converted subjects: MMSE scores between 24 and 30 (inclusive), a memory complaint, have objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia; (3) AD: MMSE scores between 20 and 26 (inclusive), CDR of 0.5 or 1.0, and meets NINCDS/ADRDA criteria for probable AD. In this study, we employed 1692 (564 × 3) MRI volumes collected at three time points: baseline (BL), month 6 (M06), and month 12 (M12). Our model predicts the changes in patients' status after three years, based on the final assessment visit, which took place at month 48 (M48). The dataset utilized in this study comprised 3 T T1-weighted anatomical sequences that utilized the volumetric 3D MPRAGE protocol, with a voxel size of 1 × 1 × 1 mm. In contrast to popular datasets, such as the Open Access Series of Imaging Studies (OASIS) [59], and Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) [60], the ADNI program collected patient data at regular interval of six months. A standard preprocessing pipeline was applied to all 564 longitudinal MRI volumes in the dataset. In this study, 282 participants were cognitively normal across all three time points. A total of 182 individuals were diagnosed with AD throughout the assessment period, whereas 100 subjects were cognitively normal during the initial visits, but later converted to AD three years after their last visit in M12 (at M48). Similar to the existing studies [24,22], we integrated the 100 converted subjects with the subjects diagnosed with AD, resulting in a total of 282 subjects, with 182 subjects being diagnosed with AD from baseline until M48, and 100 subjects diagnosed with AD after conversion from cognitively normal at M48. In addition to exploring the MRI data, this study also considered the role of other critical modalities selected by our medical domain experts, such as cognitive scores and biomarkers, in disease identification process. Scores such as the ADAS13, FAQ, MMSE, and RAVLT, as well as well-known biomarkers such as APOE4 and hippocampal volume, which have also been widely used in literature studies [31,61] and in real-world medical practice, were analyzed. A detailed description of these features is summarized in Table 5.

**Table 5**
Descriptive features of the selected patients from the ADNI dataset.

| Scores | Cognitively Normal (CN) | Converted Patients | Alzheimer's Disease (AD) |
|---|---|---|---|
| | n = 282 | n = 100 | n = 182 |
| Gender (M/F) | 191/228 | 86/54 | 187/152 |
| Age (years) | 73.84 ± 05.78 | 73.89 ± 06.84 | 75.01 ± 07.81 |
| Education | 16.43 ± 02.70 | 16.13 ± 02.71 | 15.13 ± 02.98 |
| ADAS-13 | 05.34 ± 03.16 | 07.15 ± 03.14 | 20.64 ± 7.58 |
| FDG | 0.30 ± 00.50 | 0.50 ± 0.50 | 0.88 ± 0.72 |
| TAU | 937.27 ± 663.5 | 488.15 ± 535.1 | 532.0 ± 401 |
| PTAU | 225.61 ± 248.9 | 128.96 ± 122.64 | 300 ± 209.16 |
| CDRSB | 41.36 ± 139.11 | 12.53 ± 12.48 | 35.72 ± 49.0 |
| MMSE | 27.13 ± 04.87 | 25.63 ± 04.03 | 23.62 ± 03.99 |
| RAVLT Immediate | 28.01 ± 08.20 | 29.42 ± 00.81 | 22.78 ± 03.03 |
| RAVLT Learning | 43.36 ± 14.07 | 42.59 ± 07.64 | 21.37 ± 08.83 |
| RAVLT Forgetting | 08.65 ± 10.22 | 05.57 ± 02.12 | 02.57 ± 04.46 |
| RAVLT Percentage Forgetting | 06.02 ± 10.48 | 03.52 ± 02.65 | 07.80 ± 17.20 |
| FAQ | 69.26 ± 40.48 | 103.34 ± 51.58 | 159.74 ± 107.62 |
| MOCA | 07.62 ± 31.41 | 0.37 ± 00.99 | 15.77 ± 24.64 |
| Hippocampus | 11.70 ± 29.18 | 93.82 ± 61.61 | 31.29 ± 92.92 |
| APOE4 | 01.10 ± 00.75 | 0.30 ± 00.68 | 0.27 ± 00.76 |

### 4.2. Implementation of standard AE

We compared the quality of the feature embeddings extracted from the proposed AE with that of other standard architecture-based CAEs, such as VGG-CAE and UNET-CAE. We implemented VGG-CAE from scratch, as described in this study [62]. We also implemented a U-NET architecture for creating feature embeddings owing to its encoder-decoder architectural design. The reason for choosing these architectures to obtain the latent feature embedding is two-fold: (1) The encoder-decoder modules are symmetrical in shape; (2) to the best of our knowledge, there has not been any official implementation available for autoencoders using standard methods such as ResNet, Inception, or DenseNet-based CAE. Therefore, we implemented VGG-CAE from scratch, whereas for U-NET, we adopted the Pytorch implementation of the following code [63]. The implementation details of the comparative CAEs are presented in the subsections below.

**VGG-16-based AE:** VGG-16 is an effective image classification model because of its remarkable feature extraction capability and unique architectural design that produces sparsity in deeper layers, making it simpler than other deep learning models (e.g., ResNet and InceptionNet). It has been used in various image-processing tasks, such as classification, detection, segmentation, and image enhancement. In contrast to handcrafted features, deep models can capture inherent features in the input data, making them an ideal choice for our encoder-

decoder network to provide a latent feature space for the input image [64].

In this study, we implemented a VGG-16-based convolutional autoencoder (VGG-CAE) developed by XU et al. [62] to acquire the abstract representation of an input image and its compact feature set. To convert the VGG model into a CAE, we initially removed its fully connected layers and appended a decoder module by incorporating convolutional layers in a symmetrical fashion with the encoder module. The input images, initially scaled to $128 \times 128 \times 1$ grayscale, were inputted into the encoder part of the network to extract the latent feature vector. Simultaneously, the decoder module reconstructed the input image by minimizing the binary cross-entropy loss between the reconstructed image and the input image. Hyperparameter tuning was carried out using an Adam optimizer with a learning rate of $1 \times 10^{-4}$. The model was trained for 120 epochs, with the inclusion of an early stopping callback mechanism to prevent overfitting. Fig. 3 shows an example of the input and reconstructed image using VGG-CAE.

To obtain an abstract representation of the input image in the latent feature space, we first removed the decoder module of the trained model and then collected the feature embeddings for each 2D slice at each time step. We took the output feature maps from the encoder module ($8 \times 8 \times 512$), fed them to the PCA, and produced a compact feature vector for the input image. This process was applied to obtain a compact set of features for all patients at each time step (i.e., BL, M06, and M12). The collected feature vector was then used to optimize a pool of time-series models using the Bayesian optimization process. We collected feature embeddings using VGG-CAE in the same manner as we did for the proposed CAE.

**UNET-based CAE:** UNET is a well-known DNN model originally designed for semantic segmentation tasks using biomedical data [65]. The UNET model comprises two subnetworks, that is, an encoder and a decoder. The encoding module comprises a sequence of convolution, max-pooling, and batch normalization layers. The size of the kernels in each layer was $3 \times 3$ throughout the network, with ReLU as a non-linear activation function. In the encoder module, the spatial dimensions of the input image are consecutively decreased, whereas the depth dimensions increase. The number of convolution kernels in the encoder module was as follows: 64, 128, 256, and 512. The layers were arranged in groups, and each group was applied twice to the input image. The group comprised convolution, batch normalization, and max-pooling layers. The output of the encoder module is 1024 dimensions feature map that reduces the original image by a ratio of 1/16th. The decoder module of the UNET model maintains the same arrangement of kernel settings, but in the reverse order, that is, 512, 256, 128, and 64. The max-pooling layer in the decoder module was replaced with a transposed convolution layer. The output image after every second group was upsampled by two, which increased the spatial dimensions and decreased the depth dimensions. The 64-channel output at the final layer was mapped to the original image size, and the binary cross-entropy was calculated using the segmentation mask for the image.

In this study, we leveraged the UNET model's capability to generate a latent feature representation of the input image. Rather than using a segmentation mask to compute the segmentation loss, we employed the same input image as a label map in the training process. During the backpropagation step, binary cross-entropy was utilized to facilitate the learning of image reconstruction instead of segmentation masks. The model was trained for 120 epochs, with an input image shape of $128 \times 128 \times 1$ and a learning rate of $1 \times 10^{-3}$. Hyperparameter tuning was conducted using an Adam optimizer, and an early stopping callback mechanism was employed to prevent overfitting. Finally, to achieve a compact representation of the feature embeddings using PCA, we followed the same procedures for creating feature embeddings for UNET-CAE as we discussed for VGG and the proposed CAE in earlier sections. Fig. 3 shows an example of the input and reconstructed image using UNET-CAE. In addition to visual inspection, quantitative metrics were also employed to evaluate the performance of the autoencoders. Specifically, metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) were utilized to assess reconstruction accuracy and feature representation quality.

Table 6 shows the comparison of performance between the Proposed CAE, VGG-CAE, and U-NET-CAE, where the results indicated notable differences. The Proposed CAE demonstrated superior performance in terms of both reconstruction accuracy (visual inspection: Fig. 3, qualitative assessment: Table 5) and feature representation quality (results Section 5) compared to VGG-CAE and U-NET-CAE. This superiority was evident through lower MSE values, higher PSNR scores, and greater SSIM values, indicating better image fidelity and feature preservation in the reconstructed images. Additionally, qualitative analysis complemented these findings, further reinforcing the efficacy of the Proposed CAE in generating compact and high-quality feature representations of input images.

### 4.3. Evaluation metrics

We evaluated the proposed framework using several evaluation metrics to test how well the model generalizes to the training data. These metrics were accuracy, precision, recall, F1 score, and AUC. The mathematical representation of each metric is given by Eqs. 8–11. *Accuracy* of the model refers to the correctly detected samples (CN/AD) in the predicted data. *Precision* of a model refers to the ratio of correctly classified (AD) patients to predicted positive (AD) samples. The *recall* of the model refers to the ratio of correctly classified (AD) patients to the total number of AD patients in the dataset. The F1 scores of the model refer to the weighted averages of the precision and recall. The AUC score represents the evaluation results of a model at different classification thresholds, that is, it specifies the relationship between the true positive rate and the false positive rate.

$$Accuracy = \frac{TN + TP}{TS} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{11}$$

TP refers to true positives, FP refers to false positives, and TN and TS refer to true negatives and total samples, respectively.

## 5. Results and discussion

We collected embeddings from the VGG-CAE, UNET-CAE, and the proposed CAE. The collected embeddings for each time step (i.e., BL, M06, and M12) were further compressed through PCA to obtain more compact representational features of the input data, as discussed in Section 2. Sequences for each time step were used to optimize the time series model (i.e., 1D CNN, LSTM, and GRU). Fig. 4 shows the experimental route map used in our study. We designed the experiments as follows: (1) In Experiment 1, we used MRI embeddings to optimize and evaluate each time-series model. (2) In Experiment 2, we extended Experiment 1 by fusing a set of medically important cognitive scores

**Table 6**
Qualitative analysis of the reconstructed image of the comparative CAEs.

| CAE | MSE | PSNR | SSIM (%) |
|---|---|---|---|
| VGG-CAE | 780.06 | 18.21 | 63.71 |
| UNET-CAE | 554.18 | 20.26 | 83.36 |
| Proposed CAE | 298.68 | 24.83 | 88.61 |

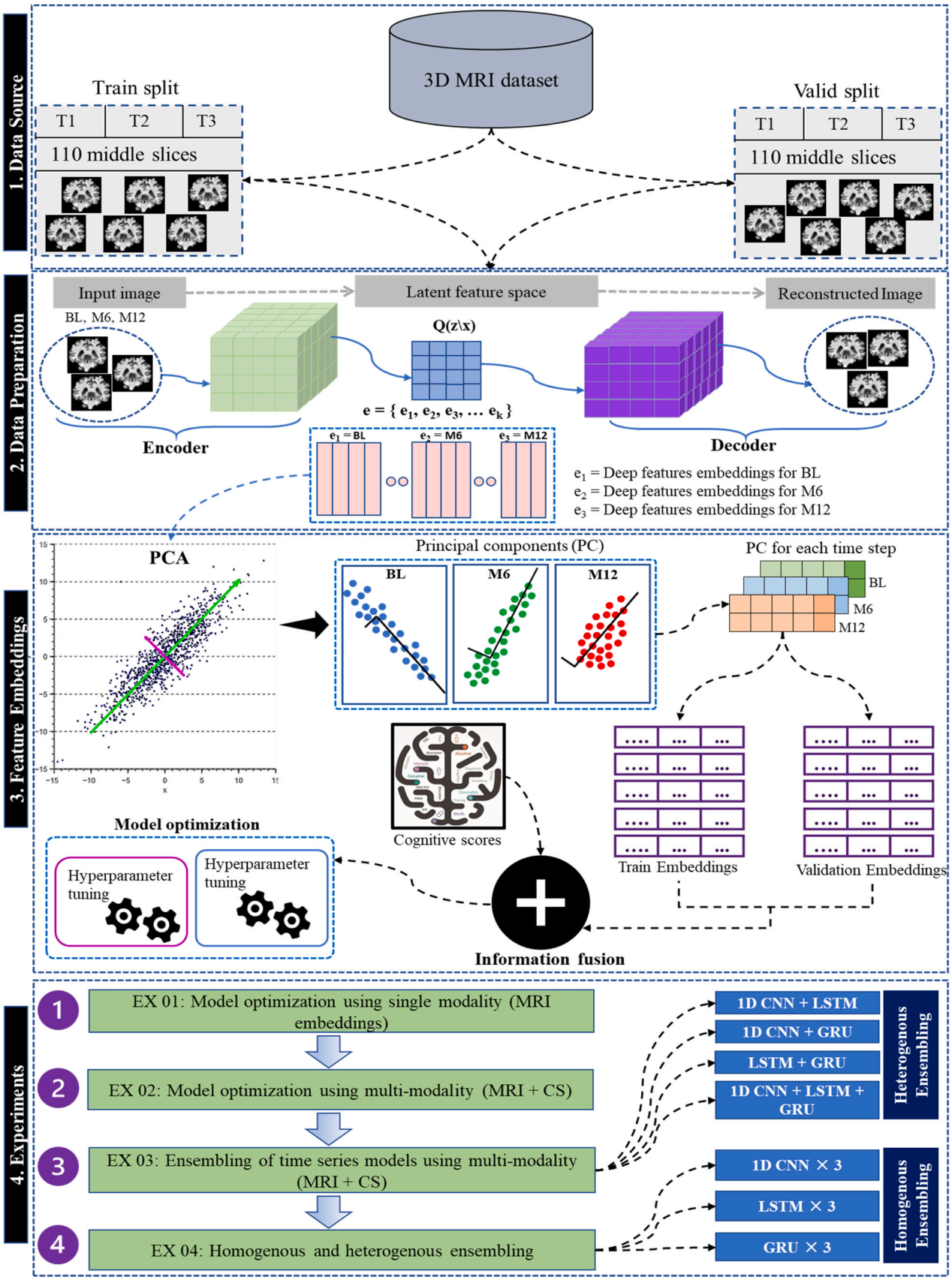


**Fig. 4.** The experimental route map with data representation and fusion over the stages of the proposed framework.

with the extracted MRI embeddings and evaluating each model. (3) In Experiment 3, we extended Experiment 2 by building ensemble models of different combinations of the optimized time-series models in Experiment 2 and tuned these models to investigate the performance of AD progression detection. (4) In Experiment 4, we investigated the effectiveness of homogenous and heterogeneous ensemble networks in AD progression detection. The decision made by the EL model is based not only on a single MRI slice, but also on 110 middle slices from the MRI volume in a longitudinal manner. Fig. 5 shows the decision-making process of the EL model for a single patient. Multimodal data of longitudinal MRI embeddings fused with cognitive scores were passed through the EL network, and the probabilities were averaged to determine AD progression. All results were reported as percentages.

### 5.1. Experiment 1: network optimization using single time series modality (MRI embeddings only)

In Experiment 1, we performed network optimization of various time-series models using a single modality of data (MRI embeddings only). We used the MRI embeddings collected from the PCA algorithm, as discussed in Section 2. We optimized each time-series model using the embeddings collected from the proposed CAE. We also performed the same optimization process for embeddings collected from other comparative CAEs (i.e., VGG-CAE and UNET-CAE). During the experimentation process, each time-series model was optimized with a single time step of data (BL), two-time steps (BL-M06), and three-time steps (BL-M12). The Bayesian optimizer returns the most critical hyperparameter set for each time series model. Table 7 reports the five-evaluation metrics achieved by each time-series model during Bayesian optimization. The reported evaluation metrics were the means of precision, recall, F1 score, AUC, and accuracy. The bold text represents the superior results for a specified category of time-series models for a particular feature embedding.

For instance, by optimizing the 1D CNN with VGG-CAE-based embeddings, the reported accuracies at the BL~M06 time step outperformed the accuracies achieved at BL and BL~M12. At BL~M06, 1D CNN achieved a precision of 73.28 ± 3.12, a recall of 75.43 ± 3.01, an F1 score of 74.72 ± 4.24, an AUC of 76.05 ± 2.68, and an accuracy of 75.91 ± 1.57. In the case of UNET and the proposed CAE-based feature embeddings, the 1D CNN achieved an improvement in the upward pattern. Such behavior of a model refers to the improvement and stability of the diagnostic model, as it is exposed to the subsequent time steps of the patient's longitudinal data. With the UNET feature embeddings, the achieved accuracies were as follows: precision: 77.72 ± 2.19, recall: 76.59 ± 2.11, F1 score: 76.76 ± 3.39, AUC: 77.25 ± 2.01, and accuracy: 75.13 ± 3.16 at BL~M12. With the proposed CAE-based feature embeddings, the reported accuracies were as follows: precision: 81.37 ± 2.50, recall: 81.22 ± 1.41, F1 score: 81.84 ± 2.09, AUC: 80.55 ± 1.52, and accuracy: 80.43 ± 1.63. These results suggest a significant improvement with the increase in longitudinal time steps of data. The same optimization process was repeated for the LSTM and GRU models. By evaluating the LSTM model with different CAE-based feature embeddings, the model reported the best accuracies with UNET and the proposed CAE-based feature embeddings. With UNET, the LSTM model achieved the following values for the evaluation metrics: precision: 84.14 ± 3.42, recall: 84.92 ± 4.23, F1 score: 84.80 ± 4.37, AUC: 83.96 ± 2.24, and accuracy: 80.20 ± 2.09. Furthermore, with the proposed CAE, the LSTM model achieved the following values for the evaluation metrics: precision: 78.14 ± 1.42, recall: 77.92 ± 2.03, F1 score: 78.20 ± 1.37, AUC: 79.96 ± 2.24, and accuracy: 77.20 ± 1.09. All accuracies reported by the GRU model showed an upward trend, indicating that the GRU model can accurately detect the progressive pattern of AD from all types of feature embeddings. For each type of feature embedding, the GRU model achieved improvements in all the evaluation metrics by increasing the longitudinal time steps. For instance, with VGG-CAE feature embedding, the GRU model achieved a precision of 82.17 ± 3.44, a recall of 83.71 ± 3.61, an F1 score of 82.67 ± 2.69, an AUC of 83.73 ± 2.32, and an accuracy of 79.78 ± 3.11. Furthermore, the reported accuracies with UNET based feature embeddings were as follows: precision: 80.17 ± 1.44, recall: 82.71 ± 2.61, F1 score: 83.67 ± 1.69, AUC: 81.73 ± 1.32, and accuracy: 79.78 ± 2.11. In contrast, the same model achieved a precision of 81.43 ± 1.32, a recall of 81.33 ± 2.02, an F1 score of 80.73 ± 1.25, an AUC of 79.53 ± 2.01, and an accuracy of 77.59 ± 1.31 with the proposed CAE based feature embeddings. Notably, in the results reported by each time series model using the proposed CAE-based feature embeddings, all accuracy metrics achieved improvement by increasing the longitudinal time steps of data. However, this was not true for VGG- and UNET-CAE-based feature embedding. The improvement in model performance with the increase in longitudinal training data also makes sense medically because doctors get more insights into the progressive patterns of a disease as they monitor patients in follow-up visits. Furthermore, the statistics shown in Table 7 indicate that the variance of the optimized models with the UNET and VGG-CAE feature embeddings was very high (i.e., > 2), indicating the negative effect of the noise available in the training data over the model; however, in the case of the proposed CAE feature embeddings, the variance remained very low (i.e., 1–2).

Fig. 6 shows a performance comparison of the three-time series

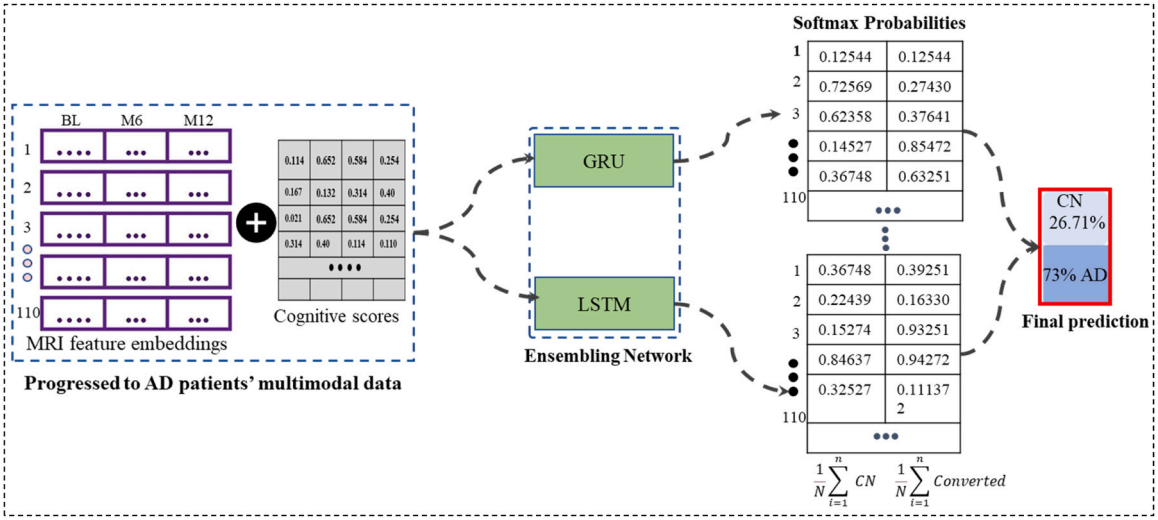


**Fig. 5.** Data processing of single patients' multimodal data an ensemble network.

**Table 7**
Performance results of the optimized time series model with longitudinal feature embeddings obtained from different CAEs.

| Timeseries Model | Convolution Autoencoder | Times Steps | Precision (%) | Recall (%) | F1 Score (%) | AUC (%) | Accuracy (%) |
|---|---|---|---|---|---|---|---|
| 1D CNN | VGG-CAE | BL | 70.43 ± 3.10 | 69.91 ± 3.09 | 70.83 ± 2.38 | 72.33 ± 1.10 | 69.91 ± 3.01 |
| | | **BL~M06*** | **73.28 ± 3.12** | **75.43 ± 3.01** | **74.72 ± 3.24** | **76.05 ± 2.68** | **75.91 ± 2.57** |
| | | BL~M12 | 72.11 ± 2.62 | 73.21 ± 1.53 | 73.07 ± 1.51 | 71.03 ± 3.65 | 70.96 ± 2.11 |
| | UNET-CAE | BL | 73.75 ± 2.87 | 72.62 ± 2.10 | 73.77 ± 3.66 | 73.19 ± 1.30 | 70.83 ± 3.04 |
| | | BL~M06 | 76.23 ± 3.04 | 78.82 ± 3.96 | 76.71 ± 3.74 | 78.14 ± 1.79 | 75.22 ± 1.41 |
| | | **BL~M12*** | **77.72 ± 2.19** | **76.59 ± 2.11** | **76.76 ± 3.39** | **77.25 ± 2.01** | **75.63 ± 3.16** |
| | Proposed-CAE | BL | 79.40 ± 1.51 | 78.59 ± 2.92 | 78.11 ± 1.31 | 77.48 ± 1.51 | 78.91 ± 2.86 |
| | | BL~M06 | 80.38 ± 2.01 | 80.81 ± 2.78 | 80.52 ± 1.28 | 79.48 ± 2.68 | 78.61 ± 1.49 |
| | | **BL~M12*** | **81.37 ± 2.50** | **81.22 ± 1.41** | **81.84 ± 2.09** | **80.55 ± 1.52** | **80.43 ± 1.63** |
| LSTM | VGG-CAE | **BL*** | **76.43 ± 4.10** | **76.56 ± 3.83** | **77.82 ± 2.11** | **77.53 ± 2.67** | **76.79 ± 4.24** |
| | | BL~M06 | 72.20 ± 4.34 | 71.89 ± 2.17 | 72.71 ± 2.39 | 70.80 ± 3.50 | 71.19 ± 4.72 |
| | | BL~M12 | 74.89 ± 2.33 | 75.62 ± 2.04 | 75.92 ± 3.41 | 74.50 ± 2.23 | 73.87 ± 3.95 |
| | UNET-CAE | BL | 74.03 ± 3.48 | 75.71 ± 3.06 | 75.57 ± 4.04 | 74.54 ± 1.54 | 73.71 ± 3.06 |
| | | BL~M06 | 81.88 ± 2.53 | 80.75 ± 2.62 | 81.75 ± 3.01 | 79.39 ± 2.32 | 79.82 ± 3.16 |
| | | **BL~M12**** | **84.14 ± 3.42** | **84.92 ± 4.23** | **84.80 ± 4.37** | **83.96 ± 2.24** | **80.20 ± 2.09** |
| | Proposed-CAE | BL | 75.18 ± 2.69 | 75.73 ± 2.01 | 75.84 ± 2.04 | 75.81 ± 2.54 | 75.24 ± 2.76 |
| | | BL~M06 | 76.18 ± 1.69 | 76.73 ± 2.01 | 77.84 ± 1.04 | 77.81 ± 1.64 | 75.24 ± 2.56 |
| | | **BL~M12*** | **78.14 ± 1.42** | **77.92 ± 2.03** | **78.20 ± 1.37** | **79.96 ± 2.24** | **77.20 ± 1.09** |
| GRU | VGG-CAE | BL | 74.31 ± 3.75 | 74.24 ± 2.91 | 73.41 ± 2.47 | 74.08 ± 3.61 | 75.12 ± 2.11 |
| | | BL~M06 | 82.11 ± 3.56 | 81.44 ± 2.87 | 82.22 ± 3.43 | 82.65 ± 3.99 | 80.37 ± 3.91 |
| | | **BL~M12*** | **82.17 ± 3.44** | **83.71 ± 3.61** | **82.67 ± 2.69** | **83.73 ± 2.32** | **81.08 ± 3.11** |
| | UNET-CAE | BL | 76.30 ± 3.61 | 76.13 ± 3.67 | 76.01 ± 2.05 | 77.87 ± 3.51 | 75.51 ± 2.26 |
| | | BL~M06 | 75.71 ± 2.91 | 75.99 ± 2.74 | 74.99 ± 3.02 | 75.25 ± 1.69 | 74.82 ± 3.19 |
| | | **BL~M12*** | **80.17 ± 1.44** | **82.71 ± 2.61** | **83.67 ± 1.69** | **81.73 ± 1.32** | **79.78 ± 2.11** |
| | Proposed-CAE | BL | 76.77 ± 2.86 | 76.63 ± 2.25 | 76.29 ± 2.31 | 77.65 ± 2.15 | 75.06 ± 2.32 |
| | | BL~M06 | 77.21 ± 1.86 | 77.54 ± 2.08 | 77.06 ± 1.92 | 76.83 ± 2.95 | 76.17 ± 2.56 |
| | | **BL~M12*** | **81.43 ± 1.32** | **81.33 ± 2.02** | **80.73 ± 1.25** | **79.53 ± 2.01** | **79.89 ± 1.31** |

*: best accuracy at a particular time step.
**: best accuracy outperforming all other accuracies.

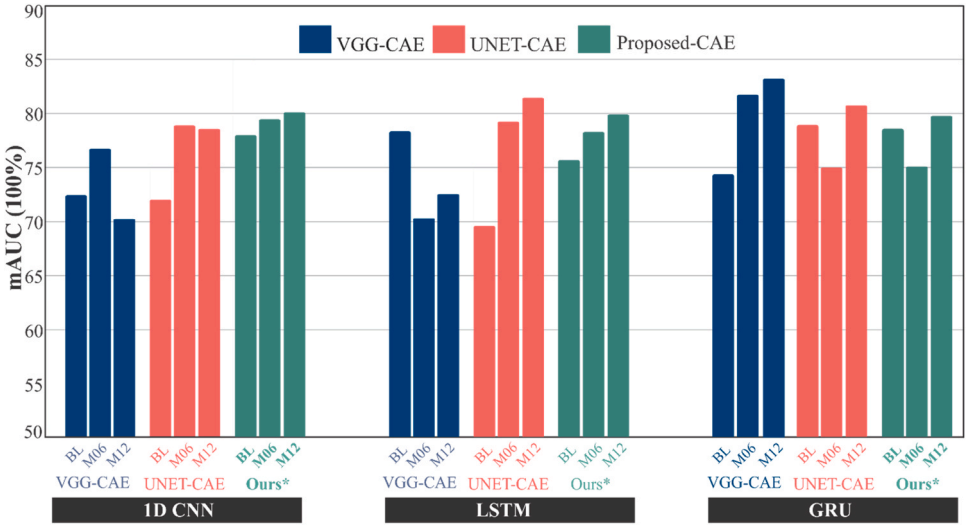


**Fig. 6.** Comparison of time series models based on embeddings collected from different CAEs at different timesteps.

models optimized with the feature embeddings collected from the three CAEs when only using the MRI modality with mAUC as the metric for comparison. The evaluation metric used in the performance comparison was mAUC. It is noteworthy that all evaluation metrics are consistent for all models; therefore, using mAUC is a representative indicator for the rest of the evaluation metrics. The performance of each time series model was investigated based on longitudinal input data. As shown in Fig. 6, the mAUC achieved by the 1D CNN at BL was 72, 78, and 78 with the feature embeddings collected from VGG-CAE, UNET-CAE, and proposed-CAE, respectively. With the MRI embeddings at two-time steps, i.e., BL~M6 visits, 1D CNN improved the mAUC by achieving 4%, 5%, and 2% improvements with VGG, UNET, and proposed-CAE feature embeddings, respectively. Adding more timesteps to the training set, i.e., BL~M12, 1D CNN improved further with the proposed-CAE based feature embeddings only, and the mAUC reached 80%. The LSTM model achieved an mAUC of 77 at BL with VGG-CAE, which is the highest compared with those of BL~06 and BL~M12. With the UNET and proposed CAE-based embeddings, the highest mAUC achieved was for BL~M12, i.e., 83% and 79%, respectively. In the case of the GRU model, the highest mAUC score was achieved for BL~M12 for all three types of feature embeddings. The achieved accuracies with VGG-CAE feature embeddings were 83%, that with UNET-CAE feature embeddings was 81%, and that with the proposed-CAE feature embeddings was 79%.

We noticed that the achieved mAUC scores with respect to the increase in the longitudinal time steps were not always upward trending. Unstable behavior with increasing time steps was observed for all three time series models with UNET and VGG-CAE feature embeddings. With

the proposed CAE-based feature embeddings, the 1D CNN and LSTM progressively tracked the disease diagnosis; however, the GRU could not show the same pattern. We observed from the literature that [5,66,31] disease diagnosis using multimodality improves the overall diagnostic process compared with using a single data modality. In Experiment 2, we explored the role of multimodality.

### 5.2. Experiment 2: network evaluation using multimodality (MRI embeddings + cognitive scores)

The main aim of Experiment 2 was to optimize time series models using multimodal time series data (i.e., MRI embeddings + cognitive scores). This study aimed to investigate the progression of AD by adding knowledge from a patient's cognitive abilities. The cognitive scores implemented in this study were gathered at the baseline visit of a patient and proved to be significant in the disease identification process. It includes genetic and physical biomarkers, along with behavioral tests on patients' health. A detailed description of these scores is presented in Table 5. In this experiment, we optimized the list of time-series models discussed in Experiment 1. A Bayesian optimizer was used to obtain the best set of hyperparameters for each time series model. The optimized hyperparameter sets returned by the Bayesian optimizer are listed in Table 2, Table 3, and Table 4 for the 1D CNN, LSTM, and GRU models, respectively. *By fusing cognitive scores with the longitudinal MRI embeddings, each time series model improved the overall results and stability*. This medically makes sense. From the machine learning perspective, this means that multimodal data added extra knowledge to the resulting feature set, which helped the models improve their decision boundaries. These results are consistent with those of previous studies that confirmed the positive role of multimodal data in improving model performance [19,14,16]. Using multimodal data, the proposed model learns the essential patterns related to disease progression from each modality during the training process.

Table 8 shows a performance comparison of a list of time-series models evaluated with different feature embeddings. For instance, the 1D CNN achieved the best accuracies at BL~M06 compared with BL and BL~M12 using VGG-CAE feature embeddings. At BL~M06, 1D CNN achieved a precision of 76.38 ± 2.90, a recall of 76.53 ± 2.90, an F1 score of 77.83 ± 3.89, an AUC of 77.15 ± 2.58, and an accuracy of 75.02 ± 1.47. Using UNET and the proposed CAE-based feature embeddings, 1D CNN outperformed at BL~M12 by achieving a precision of 80.83 ± 2.08, a recall of 80.71 ± 1.01, an F1 score of 81.86 ± 3.29, an AUC of 79.35 ± 2.94, and an accuracy of 79.24 ± 3.06 and a precision of 86.51 ± 1.41, a recall of 87.69 ± 2.82, an F1 score of 86.22 ± 1.22, an AUC of 87.58 ± 1.41, and an accuracy of 85.01 ± 2.75.

Notably, using multimodal training data, all accuracy metrics were improved compared with using only a single modality in disease diagnosis. Furthermore, comparing the accuracies with different sets of feature embeddings (i.e., 1D CNN, LSTM, and GRU) showed that 1D CNN outperformed BL~M12 using the proposed CAE-based feature embeddings. The LSTM model with VGG-CAE-based feature embeddings did not report consistent results, and the highest accuracies were shown at BL time steps instead of BL~M12. This behavior indicates that LSTM could not capture progressive patterns for the longitudinal data. Instead, the LSTM model reported very stable accuracies with UNET and the proposed CAE-based feature embeddings. The reported accuracies showed an upward trend with an increase in the longitudinal time steps of the training data, which is the desired output from the proposed framework. Finally, the GRU model, compared to both the 1D CNN and LSTM models, reported very consistent accuracies for all three sets of feature embeddings. The baseline accuracies consistently improved with an increase in the longitudinal time steps of the data. At BL~M12, GRU model achieved a precision of 84.53 ± 3.61, a recall of 85.43 ± 4.21, an F1 score of 85.83 ± 3.65, an AUC of 85.64 ± 2.31, and an accuracy of 83.69 ± 1.41 using VGG-CAE, and a precision of 86.28 ± 3.34, a recall of 85.82 ± 3.59, a F1 score of 84.77 ± 1.59, an AUC of 84.84 ± 1.22, and an accuracy of 84.88 ± 3.35 using UNET. Finally, using the proposed CAE-based feature embeddings, the GRU model outperformed all other models and achieved the following accuracies: precision: 89.42 ± 2.16, recall: 88.54 ± 3.27, F1 score: 88.02 ± 1.24, AUC: 90.86 ± 2.17, and accuracy: 87.99 ± 3.07. Combining MRI features with cognitive scores significantly improved all-time series models compared with using only the input data from the MRI modality. In particular, the overall accuracy of the GRU model reached an AUC of 90.86, indicating that the model utilized each modality very well during the training phase for recognizing AD progression.

**Table 8**
Performance results of the optimized time series model with longitudinal feature embeddings obtained from different CAEs.

| Timeseries Model | Convolution Autoencoder | Times Steps | Precision (%) | Recall (%) | F1 Score (%) | AUC (%) | Accuracy (%) |
|---|---|---|---|---|---|---|---|
| 1D-CNN | VGG-CAE | BL | 75.53 ± 3.00 | 75.01 ± 2.99 | 74.93 ± 1.28 | 75.44 ± 0.99 | 74.52 ± 2.91 |
| | | **BL~M06*** | **76.38 ± 2.90** | **76.53 ± 2.90** | **77.83 ± 3.89** | **77.15 ± 2.58** | **75.82 ± 1.47** |
| | | BL~M12 | 73.21 ± 2.51 | 72.31 ± 1.43 | 72.08 ± 1.41 | 71.13 ± 3.54 | 70.86 ± 2.22 |
| | UNET-CAE | BL | 75.85 ± 2.77 | 74.72 ± 3.11 | 75.88 ± 3.55 | 74.29 ± 2.19 | 72.83 ± 2.94 |
| | | BL~M06 | 78.33 ± 2.93 | 77.93 ± 3.86 | 77.82 ± 3.64 | 77.21 ± 2.68 | 75.62 ± 2.31 |
| | | **BL~M12*** | **80.83 ± 2.08** | **80.71 ± 1.01** | **81.86 ± 3.29** | **79.35 ± 2.94** | **79.84 ± 3.06** |
| | Proposed-AE | BL | 78.47 ± 2.40 | 79.33 ± 2.29 | 78.94 ± 1.99 | 79.65 ± 1.41 | 77.53 ± 2.53 |
| | | BL~M06 | 83.91 ± 1.91 | 84.91 ± 2.68 | 84.62 ± 1.17 | 83.58 ± 2.58 | 81.71 ± 1.38 |
| | | **BL~M12*** | **86.51 ± 1.41** | **87.69 ± 2.82** | **86.22 ± 1.22** | **87.58 ± 1.41** | **85.01 ± 2.75** |
| LSTM | VGG-CAE | **BL*** | **77.53 ± 4.69** | **78.66 ± 3.72** | **78.93 ± 2.14** | **78.63 ± 1.57** | **76.89 ± 3.89** |
| | | BL~M06 | 74.32 ± 4.23 | 72.14 ± 2.07 | 73.81 ± 2.29 | 74.91 ± 3.41 | 72.29 ± 4.62 |
| | | BL~M12 | 77.99 ± 2.23 | 77.73 ± 1.93 | 76.03 ± 3.11 | 76.61 ± 2.13 | 76.98 ± 3.44 |
| | UNET-CAE | BL | 75.98 ± 2.42 | 74.86 ± 2.52 | 75.85 ± 2.92 | 74.49 ± 2.22 | 73.43 ± 3.06 |
| | | BL~M06 | 75.14 ± 3.37 | 75.81 ± 2.96 | 76.67 ± 3.94 | 76.64 ± 1.44 | 75.81 ± 2.95 |
| | | **BL~M12*** | **77.13 ± 3.93** | **78.48 ± 2.28** | **77.06 ± 2.73** | **78.24 ± 2.81** | **76.55 ± 3.75** |
| | Proposed-AE | BL | 82.24 ± 3.32 | 81.02 ± 4.13 | 82.94 ± 4.26 | 82.06 ± 2.13 | 80.37 ± 1.99 |
| | | BL~M06 | 82.28 ± 1.59 | 83.81 ± 2.93 | 84.95 ± 1.94 | 84.92 ± 1.44 | 81.54 ± 3.65 |
| | | **BL~M12*** | **86.12 ± 2.33** | **86.59 ± 2.05** | **86.81 ± 1.97** | **87.52 ± 3.56** | **85.98 ± 2.27** |
| GRU | VGG-CAE | BL | 78.42 ± 4.64 | 78.35 ± 3.81 | 77.51 ± 3.37 | 79.19 ± 3.51 | 78.22 ± 4.45 |
| | | BL~M06 | 83.21 ± 3.45 | 84.55 ± 1.77 | 83.32 ± 2.33 | 82.75 ± 3.89 | 81.87 ± 3.84 |
| | | **BL~M12** | **84.53 ± 3.61** | **85.43 ± 4.21** | **85.83 ± 3.65** | **85.64 ± 2.31** | **83.69 ± 1.41** |
| | UNET-CAE | BL | 77.41 ± 2.51 | 79.23 ± 3.56 | 78.11 ± 1.94 | 78.98 ± 3.41 | 76.82 ± 2.16 |
| | | BL~M06 | 82.81 ± 2.81 | 82.11 ± 2.64 | 83.10 ± 2.91 | 82.35 ± 1.59 | 80.92 ± 3.08 |
| | | **BL~M12*** | **86.28 ± 3.34** | **85.82 ± 3.59** | **84.77 ± 1.59** | **85.84 ± 1.22** | **84.88 ± 3.35** |
| | Proposed-CAE | BL | 82.87 ± 2.76 | 82.74 ± 2.15 | 83.39 ± 2.19 | 84.69 ± 4.22 | 81.87 ± 2.22 |
| | | BL~M06 | 84.11 ± 3.56 | 85.64 ± 2.98 | 85.16 ± 4.62 | 86.94 ± 2.85 | 83.27 ± 1.96 |
| | | **BL~M12**** | **89.42 ± 2.16** | **88.54 ± 3.27** | **88.02 ± 1.24** | **90.86 ± 2.17** | **87.99 ± 3.07** |

Fig. 7 depicts the mAUC comparison of the Bayesian optimized time-series models with multimodal data. Each model was evaluated using patients' MRI fused with cognitive scores. Overall, after fusion, the trend of mAUC values was moving upward and was stable, showing that most models improved in mAUC as they were exposed to an increase in longitudinal time steps of data for the same patient.

As shown in Fig. 7, the mAUC achieved by the 1D CNN based on UNET and proposed CAE feature embeddings showed an upward trend, that is, a model became stable and accurate in disease prediction as more longitudinal data were added to the training set. Furthermore, by comparing the performance for single modality (i.e., MRI) with that of multimodal data (MRI + CS), 1D CNN achieved 3% improvement at BL∼M12 with UNET-based feature embeddings and 7% improvement with proposed CAE-based feature embeddings, reaching an mAUC from 77% to 79% and 80–87%, respectively. However, with VGG-CAE feature embeddings, the 1D CNN exhibited unstable and fluctuating results. In the case of the LSTM model, the improvement in the mACU was reported with VGG and proposed CAE-based featured embeddings only i.e., mAUC reached from 74% to 76% and 79–87%, respectively. No improvement in the mAUC metric was reported for UNET-based feature embeddings. In the case of the GRU model, all-time series models reported significant improvement i.e., the mAUC achieved using MRI features + CS improved from 83-85%, 81–85%, and 79–90% for VGG, UNET, and the proposed CAE feature embeddings, respectively.

We noticed from the experiments that combining cognitive scores with MRI features improved overall stability and model performance in disease diagnosis. In particular, each time-series model with the proposed CAE-based feature embeddings became more stable and accurate as they were exposed to subsequent time steps in the longitudinal data. Moreover, in Experiment 3, we further investigated how combining the decision-making process of multiple optimized networks improved the disease diagnostic process compared to a single model.

### 5.3. *Experiment 3: ensemble of the optimized networks using multimodal data (MRI + cognitive scores)*

After optimizing each time series model with the Bayesian optimizer, we conducted various experiments with a single modality (MRI) and multiple modalities (MRI + cognitive scores) to investigate the effect of each modality on disease diagnosis. In Experiment 2, we found that each time series model became more stable and accurate in the disease diagnosis as it was exposed to longitudinal data from the subsequent time steps. In this section, we present combinations of optimized time-series models and accumulate the decisions of multiple networks to make the final decision. We tested various combinations of heterogeneous EL models, including combinations of two network ensembles (i.e., 1D CNN + LSTM, 1D CNN + GRU, and LSTM + GRU) and three network ensembles (i.e., 1D CNN + LSTM + GRU). Table 10 lists various combinations of heterogenous EL networks. We tested different combinations of EL networks to determine the best combination of base classifiers. Each combination was evaluated using one-, two-, and three-time steps of longitudinal data fused with cognitive scores at baseline. Furthermore, we tested each combination of base classifiers with UNET, VGG, and the proposed CAE-based feature embeddings.

*1D CNN + LSTM*:1D CNN + LSTM: Overall, the EL model of 1D CNN + LSTM showed improved accuracy metrics as more time steps of the longitudinal data were added to the training set. At BL, 1D CNN + LSTM EL model achieved a precision of 77.41 ± 2.20, recall of 76.13 ± 3.10, F1 score of 77.14 ± 3.12, AUC of 77.13 ± 3.10, and accuracy of 76.01 ± 2.05 with the proposed CAE based feature embeddings and outperformed the accuracies achieved with VGG and UNET features embeddings. However, with two-time steps of longitudinal data, the 1D CNN + LSTM EL model achieved the best accuracies with UNET-CAE-based feature embeddings (i.e., precision: 84.38 ± 4.20, recall: 84.23 ± 2.10, F1 score: 83.34 ± 3.04, AUC: 84.15 ± 4.32, and accuracy: 83.54 ± 3.47). With three-time steps of data, the model achieved the best accuracy with the proposed CAE-based feature embeddings, outperforming the accuracies achieved with the other feature embeddings.

*1D CNN + GRU:* This combination of EL models also reported stable but slightly lower results compared to other combinations of EL networks based on two-network-based classifiers. No significant improvement at BL∼M12 was found that can make 1D CNN + GRU better than the others. However, a performance improvement with respect to longitudinal time steps in the training data was found for all sets of feature embeddings. Compared with VGG- and UNET-CAE-based feature embeddings, 1D CNN + GRU reported the highest accuracies at BL∼M12 with the proposed CAE-based feature embeddings, with a precision of 89.12 ± 2.33, recall of 90.59 ± 3.05, F1-score of 89.81 ± 1.97, AUC of 89.52 ± 3.56, and accuracy of 87.98 ± 2.27.

*LSTM + GRU*: The combination of the LSTM + GRU model showed

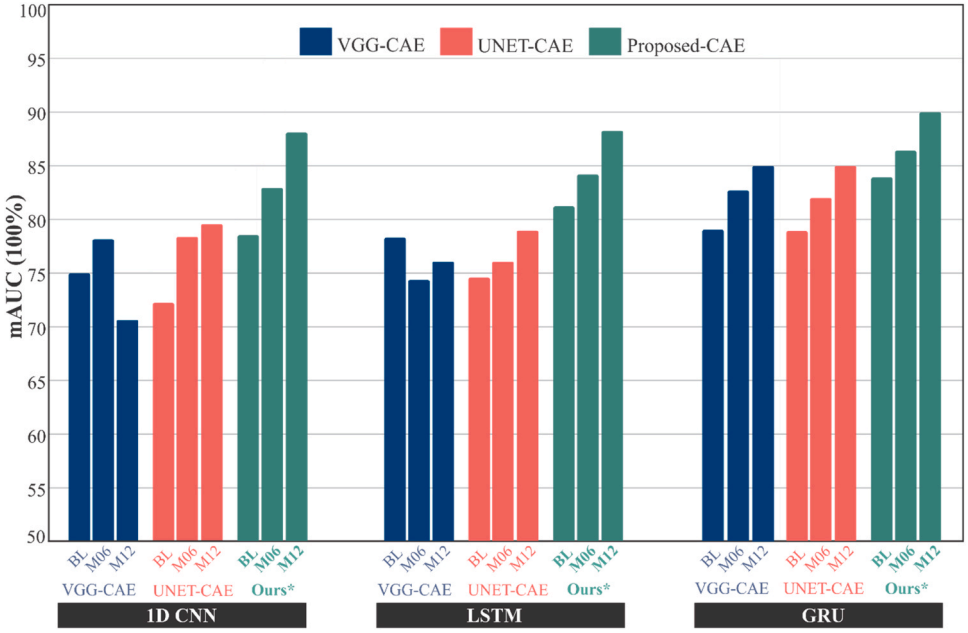


**Fig. 7.** Performance comparison of 1D CNN, LSTM, and GRU using two modalities (Feature embeddings + Cognitive scores).

**Table 9**
Homogenous ensemble of different combinations of time series models with multimodal training data (Features embedding + cognitive scores).

| Timeseries Model | Convolution Autoencoder | Times Steps | Precision (%) | Recall (%) | F1-Score (%) | AUC (%) | Accuracy (%) |
|---|---|---|---|---|---|---|---|
| 1DCNN + LSTM | VGG-CAE | BL | 74.38 ± 4.20 | 75.63 ± 4.10 | 75.15 ± 3.04 | 74.15 ±4.38 | 73.51 ± 1.07 |
| | | BL~M06 | 80.38 ± 4.20 | 81.23 ± 4.10 | 81.15 ± 3.04 | 81.15 ±4.38 | 79.61 ± 1.07 |
| | | **BL~M12*** | **87.31 ± 2.51** | **87.31 ± 1.13** | **87.08 ± 2.41** | **86.41 ±3.54** | **85.76 ± 2.32** |
| | UNET-CAE | BL | 76.38 ± 4.20 | 75.13 ± 2.30 | 76.14 ± 3.02 | 76.12 ±3.30 | 74.01 ± 3.02 |
| | | BL~M06 | 84.38 ± 4.20 | 84.23 ± 2.10 | 83.34 ± 3.04 | 84.15 ±4.32 | 83.54 ± 3.47 |
| | | **BL~M12*** | **91.53 ± 2.28** | **92.21 ± 1.01** | **91.86 ± 3.68** | **91.35 ±1.24** | **89.84 ± 3.06** |
| | Proposed-CAE | BL | 77.41 ± 2.20 | 76.13 ± 3.10 | 77.14 ± 3.12 | 77.13 ±3.10 | 76.01 ± 2.05 |
| | | BL~M06 | 82.38 ± 2.20 | 82.21 ± 3.20 | 82.18 ± 1.44 | 83.13 ±2.35 | 82.54 ± 2.22 |
| | | **BL~M12*** | **92.51 ± 1.41** | **91.69 ± 2.82** | **92.22 ± 1.21** | **90.54 ±1.41** | **90.01 ± 2.75** |
| 1D CNN + GRU | VGG-CAE | BL | 78.10 ± 3.20 | 78.15 ± 3.50 | 78.13 ± 3.02 | 79.12 ±3.30 | 77.01 ± 3.22 |
| | | BL~M06 | 83.23 ± 2.10 | 82.13 ± 1.90 | 83.15 ± 3.20 | 82.15 ± 4132 | 81.54 ± 3.42 |
| | | **BL~M12*** | **85.43 ± 2.23** | **85.73 ± 3.14** | **86.03 ± 2.41** | **84.41 ± 2.30** | **84.77 ± 2.01** |
| | UNET-CAE | BL | 85.13 ± 3.22 | 85.63 ± 3.26 | 84.33 ± 2.24 | 84.51 ± 2.30 | 84.75 ± 3.01 |
| | | BL~M06 | 79.21 ± 3.20 | 79.33 ± 3.40 | 80.65 ± 2.60 | 79.14 ± 4132 | 77.74 ± 3.52 |
| | | **BL~M12*** | **89.13 ± 3.23** | **88.41 ± 1.22** | **89.06 ± 1.23** | **88.24 ± 2.81** | **87.55 ± 3.75** |
| | Proposed- CAE | BL | 83.15 ± 3.12 | 82.43 ± 3.56 | 83.53 ± 3.44 | 83.51 ± 2.30 | 81.75 ± 3.01 |
| | | BL~M06 | 84.12 ± 3.33 | 83.91 ± 2.62 | 83.62 ± 2.83 | 85.21 ± 2.11 | 83.75 ± 3.75 |
| | | **BL~M12*** | **89.12 ± 2.33** | **90.59 ± 2.05** | **89.81 ± 1.97** | **89.52 ± 2.56** | **87.98 ± 2.27** |
| LSTM + GRU | VGG-CAE | BL | 80.12 ± 2.12 | 80.43 ± 3.43 | 79.43 ± 3.11 | 80.54 ± 2.64 | 78.65 ± 3.01 |
| | | BL~M06 | 80.12 ± 3.33 | 81.51 ± 2.62 | 82.52 ± 3.23 | 81.21 ± 2.31 | 80.25 ± 3.55 |
| | | **BL~M12*** | **86.53 ± 2.31** | **87.33 ± 4.21** | **86.83 ± 3.05** | **86.64 ± 2.21** | **84.69 ± 3.41** |
| | UNET-CAE | BL | 76.12 ± 3.36 | 75.51 ± 4.62 | 76.52 ± 4.53 | 75.21 ± 2.31 | 74.55 ± 3.19 |
| | | BL~M06 | 78.14 ± 3.13 | 79.51 ± 4.62 | 78.52 ± 3.26 | 77.32 ± 2.23 | 76.55 ± 3.25 |
| | | **BL~M12*** | **82.28 ± 3.34** | **83.32 ± 3.59** | **81.70 ± 3.59** | **82.84 ± 2.22** | **80.88 ± 2.25** |
| | Proposed- CAE | BL | 79.88 ± 3.44 | 78.22 ± 2.14 | 78.71 ± 3.32 | 79.34 ± 3.12 | 78.00 ± 2.75 |
| | | BL~M06 | 89.38 ± 3.44 | 90.02 ± 2.91 | 91.31 ± 2.72 | 89.32 ± 2.12 | 87.70 ± 2.75 |
| | | **BL~M12*** | **91.42 ± 2.16** | **92.04 ± 1.21** | **91.02 ± 2.34** | **91.29 ± 2.14** | **89.59 ± 2.07** |
| 1DCNN+LSTM + GRU | VGG-CAE | BL | 88.38 ± 3.41 | 87.42 ± 3.21 | 88.31 ± 2.72 | 89.32 ± 2.12 | 86.70 ± 2.75 |
| | | BL~M06 | 89.38 ± 2.41 | 89.12 ± 2.14 | 88.72 ± 2.12 | 89.42 ± 3.82 | 87.80 ± 2.05 |
| | | **BL~M12*** | **94.11 ± 2.10** | **94.24 ± 2.11** | **93.02 ± 3.31** | **93.46 ± 2.10** | **92.29 ± 3.07** |
| | UNET-CAE | BL | 88.32 ± 3.01 | 89.82 ± 3.54 | 88.41 ± 2.44 | 89.35 ± 3.82 | 87.70 ± 2.75 |
| | | BL~M06 | 90.38 ± 3.41 | 90.12 ± 2.14 | 90.72 ± 2.12 | 91.42 ± 3.17 | 89.70 ± 2.55 |
| | | **BL~M12*** | **95.11 ± 2.48** | **94.24 ± 2.18** | **94.02 ± 2.31** | **95.84 ± 2.10** | **95.61 ± 2.02** |
| | Proposed- CAE | BL | 89.13 ± 2.23 | 89.41 ± 2.22 | 88.06 ± 1.23 | 88.24 ± 2.81 | 87.65 ± 2.75 |
| | | BL~M06 | 93.41 ± 2.31 | 93.65 ± 2.71 | 94.42 ± 2.21 | 93.54 ± 1.41 | 92.71 ± 2.75 |
| | | **BL~M12**** | **96.51 ± 1.14** | **96.44 ± 1.38** | **96.22 ± 1.30** | **97.04 ± 2.12** | **96.83 ± 2.32** |

*: best accuracy at a particular time step
**: best accuracy outperforming all other accuracies.

very stable results for all three types of feature embeddings. The LSTM + GRU EL model reported improvements in all achieved accuracies with an increase in the longitudinal time steps of the training data. This behavior indicates that the model captures the progressive pattern of AD with respect to time. Furthermore, the model achieved the highest accuracies at BL~M12 with the proposed CAE feature embeddings (i.e., precision: 91.42 ± 2.16, recall: 92.04 ± 1.21, F1 score: 91.02 ± 2.34, AUC: 91.29 ± 2.14, and accuracy: 89.59 ± 2.07), outperforming the other types of feature embeddings.

*1D CNN + LSTM + GRU:* By combining all three time-series models as base classifiers for a single EL model, a significant improvement in disease identification was achieved. For instance, by comparing the accuracies of the two- and three-network EL models at each time step, the three network EL models outperformed all other combinations. For example, the best-achieved accuracies at the BL timestep were reported by 1D CNN + GRU with UNET-CAE based feature embeddings, i.e., precision: 85.13 ± 3.22, recall: 85.63 ± 3.26, F1-score: 84.33 ± 2.24, AUC: 84.51 ± 2.30, and accuracy: 85.75 ± 3.01. Compared with the two network EL model, the three network EL model achieved a precision of 89.13 ± 3.23, a recall of 89.41 ± 2.22, an F1-score of 88.06 ± 1.23, an AUC of 88.24 ± 2.81, and an accuracy of 87.25 ± 3.75 at BL using the proposed CAE feature embeddings. At BL ~ M06, the best-achieved accuracies by the two network EL models were reported by LSTM + GRU using the proposed CAE feature embeddings, that is, a precision of 89.38 ± 3.44, recall of 90.02 ± 2.91, F1-score of 91.31 ± 2.72, AUC of 89.32 ± 2.12, and accuracy of 86.70 ± 2.75. In the case of the three-network EL model at BL ~ M06, each accuracy metric got improved by 3 ~ 7%. Finally, at BL ~ M12, the 1DCNN + LSTM + GRU model achieved the highest accuracy with all three sets of feature embeddings; in particular, the proposed CAE-based feature embeddings had a precision of 96.11 ± 1.14, recall of 96.24 ± 2.38, F1-score of 96.22 ± 1.30, AUC of 97.04 ± 2.12, and accuracy of 95.83 ± 2.32, outperforming all existing combinations of the two EL network models.

Fig. 8 shows the mAUC comparison of different combinations of heterogeneous EL models based on the mAUC metric only. The evaluation of each combination of the EL models was performed using three different sets of convolutional feature embeddings fused with cognitive scores. The mAUC achieved by each combination of EL networks showed a very accurate and stable value with increased longitudinal time steps of training data. Moreover, it is noteworthy that compared to the two-network EL model, the three-network EL model showed significant improvement at each time step and outperformed all combinations of the two-network EL models. Especially, three three-network EL model with the proposed CAE-based feature embedding achieved the highest mAUC of 97%.

### 5.4. Experiment 4: evaluating homogeneous and heterogeneous EL networks using multimodal data

We also conducted experiments with a homogenous network ensemble and compared them with the accuracies achieved using a heterogeneous network ensemble. We tested an ensemble network of each base classifier in three networks of homogenous ensemble format to investigate the effect of each type of feature embedding in disease diagnosis. We followed an experimental setup similar to that used in Experiment 3. Table 9 lists the accuracies achieved for each combination of homogenous EL networks. Using the LSTM x 3 EL model, the highest accuracies with all three types of feature embeddings were achieved at

**Table 10**
Homogenous and Heterogeneous network ensemble model with multimodal training data (MRI feature embeddings + Cognitive sores).

| Time-series Models | | Convolution Autoencoder | Times Steps | Precision (%) | Recall (%) | F1 Score (%) | AUC (%) | Accuracy (%) |
|---|---|---|---|---|---|---|---|---|
| Homogenous EL | 1D CNN + 1D CNN + 1D CNN | VGG-CAE | BL | 75.13 ± 3.10 | 74.71 ± 3.09 | 75.81 ± 1.38 | 74.31 ± 1.50 | 74.61 ± 3.01 |
| | | | BL~M06 | 74.23 ± 3.12 | 73.63 ± 3.01 | 73.82 ± 4.24 | 72.35 ± 2.38 | 72.51 ± 1.57 |
| | | | **BL~M12*** | **76.17 ± 2.62** | **75.31 ± 1.53** | **75.24 ± 1.51** | **75.43 ± 4.05** | **75.64 ± 2.31** |
| | | UNET-CAE | BL | 71.75 ± 2.37 | 72.42 ± 2.40 | 73.07 ± 3.76 | 73.14 ± 1.30 | 72.52 ± 3.24 |
| | | | BL~M06 | 75.44 ± 3.43 | 74.34 ± 3.15 | 74.71 ± 3.74 | 74.14 ± 3.77 | 73.55 ± 1.23 |
| | | | **BL~M12** | **74.74 ± 2.45** | **76.04 ± 4.34** | **75.63 ± 3.44** | **76.24 ± 2.45** | **75.33 ± 4.36** |
| | | Proposed- CAE | BL | 73.74 ± 3.45 | 72.04 ± 4.34 | 73.43 ± 2.24 | 73.24 ± 3.45 | 72.33 ± 3.36 |
| | | | BL~M06 | 74.79 ± 2.01 | 72.81 ± 3.78 | 73.52 ± 2.28 | 73.48 ± 3.68 | 72.61 ± 2.44 |
| | | | **BL~M12*** | **76.44 ± 2.52** | **75.59 ± 2.92** | **75.16 ± 2.32** | **75.56 ± 1.54** | **74.94 ± 2.46** |
| | LSTM+ LSTM + LSTM | VGG-CAE | BL | 74.12 ± 2.12 | 73.43 ± 3.43 | 74.43 ± 3.11 | 73.64 ± 2.22 | 72.64 ± 3.51 |
| | | | **BL~M06*** | **76.32 ± 3.33** | **77.51 ± 4.62** | **76.43 ± 2.43** | **76.23 ± 3.32** | **75.55 ± 3.55** |
| | | | BL~M12 | 75.52 ± 3.35 | 76.33 ± 2.21 | 76.49 ± 2.35 | 75.34 ± 2.24 | 75.74 ± 2.34 |
| | | UNET-CAE | BL | 76.46 ± 3.45 | 75.55 ± 3.64 | 76.54 ± 4.25 | 76.41 ± 4.33 | 75.55 ± 3.39 |
| | | | **BL~M06*** | **78.14 ± 3.13** | **78.51 ± 4.62** | **79.52 ± 3.26** | **78.32 ± 2.23** | **77.23 ± 3.25** |
| | | | BL~M12 | 79.56 ± 2.15 | 78.82 ± 3.32 | 78.77 ± 3.62 | 79.88 ± 2.32 | 77.75 ± 2.65 |
| | | Proposed- CAE | BL | 72.88 ± 3.44 | 73.33 ± 3.34 | 73.77 ± 3.15 | 72.34 ± 3.22 | 71.70 ± 3.25 |
| | | | BL~M06 | 75.74 ± 3.64 | 74.02 ± 3.42 | 74.47 ± 2.35 | 75.33 ± 2.32 | 74.45 ± 2.45 |
| | | | **BL~M12*** | **77.45 ± 2.33** | **78.26 ± 2.44** | **77.34 ± 1.43** | **78.39 ± 2.24** | **75.99 ± 2.47** |
| | GRU + GRU + GRU+ | VGG-CAE | **BL*** | **75.32 ±3.47** | **74.54 ± 3.34** | **75.44 ± 2.73** | **75.12 ± 2.52** | **74.75 ± 3.55** |
| | | | BL~M06 | 74.38 ± 3.41 | 75.12 ± 2.31 | 75.72 ± 2.32 | 74.36 ± 3.42 | 73.66 ± 2.25 |
| | | | BL~M12 | 73.12 ± 2.40 | 72.34 ± 2.42 | 73.32 ± 2.33 | 73.16 ± 2.16 | 72.39 ± 3.53 |
| | | UNET-CAE | BL | 74.34 ± 3.43 | 75.24 ± 3.34 | 74.43 ± 3.14 | 75.62 ± 3.42 | 74.70 ± 2.25 |
| | | | BL~M06 | 76.67 ± 3.23 | 76.31 ± 2.44 | 75.73 ± 3.22 | 77.42 ± 3.65 | 75.43 ± 3.55 |
| | | | **BL~M12*** | **79.11 ± 2.28** | **80.44 ± 2.28** | **79.33 ± 3.23** | **80.74 ± 2.35** | **79.33 ± 2.53** |
| | | Proposed- CAE | BL | 74.33 ± 3.33 | 74.44 ± 3.32 | 75.36 ± 2.33 | 75.34 ± 2.82 | 73.55 ± 3.72 |
| | | | BL~M06 | 76.42 ± 3.32 | 76.54 ± 2.72 | 75.12 ± 3.23 | 77.24 ± 2.43 | 74.71 ± 2.32 |
| | | | **BL~M12*** | **77.12 ± 2.21** | **78.41 ± 1.37** | **78.43 ± 2.35** | **78.34 ± 2.83** | **77.43 ± 2.52** |
| Heterogeneous EL | 1D CNN + LSTM + GRU | VGG-CAE | BL | 88.38 ± 3.41 | 87.42 ± 3.21 | 88.31 ± 2.72 | 89.32 ± 2.12 | 87.20 ± 2.75 |
| | | | BL~M06 | 89.38 ± 2.41 | 89.12 ± 2.14 | 88.72 ± 2.12 | 89.42 ± 3.82 | 87.60 ± 2.05 |
| | | | **BL~M12*** | **94.11 ± 2.10** | **94.24 ± 2.11** | **93.02 ± 3.31** | **93.46 ± 2.10** | **93.29 ± 3.07** |
| | | UNET-CAE | BL | 88.32 ± 3.01 | 89.82 ± 3.54 | 88.41 ± 2.44 | 89.35 ± 3.82 | 87.80 ± 2.75 |
| | | | BL~M06 | 90.38 ± 3.41 | 90.12 ± 2.14 | 90.72 ± 2.12 | 91.42 ± 3.17 | 89.50 ± 2.55 |
| | | | **BL~M12** | **95.11 ± 2.48** | **95.24 ± 2.18** | **94.02 ± 2.31** | **95.84 ± 2.10** | **94.51 ± 2.02** |
| | | Proposed- CAE | BL | 89.13 ± 2.23 | 89.41 ± 2.22 | 88.06 ± 1.23 | 88.24 ± 2.81 | 87.75 ± 2.75 |
| | | | BL~M06 | 93.41 ± 2.31 | 93.65 ± 2.71 | 94.42 ± 2.21 | 93.54 ± 1.41 | 92.51 ± 2.75 |
| | | | **BL~M12**** | **96.11 ± 1.14** | **96.24 ± 1.38** | **96.22 ± 1.30** | **97.04 ± 2.12** | **95.83 ± 2.32** |

*: best accuracy at a particular time step
**: best accuracy outperforming all other accuracies.

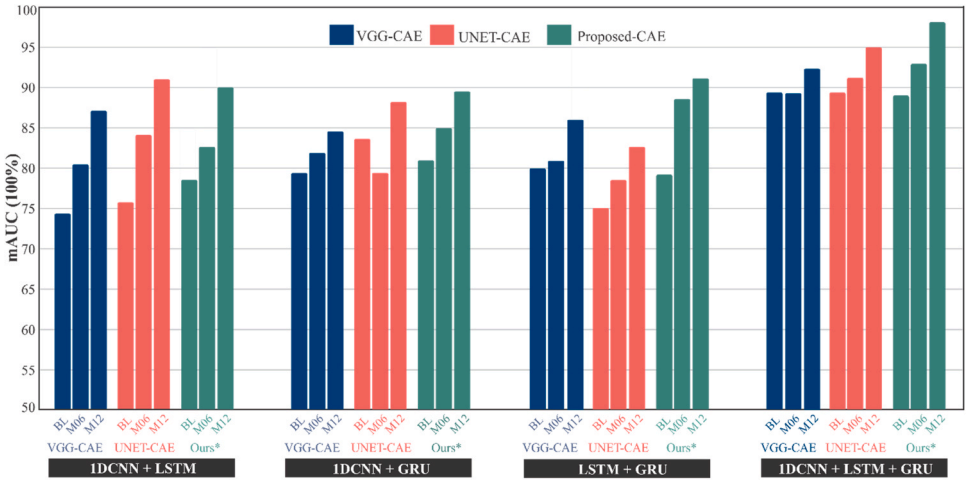


**Fig. 8.** Comparison of different combinations of time series models trained with MRI embeddings collected from different CAEs at different timesteps.

BL~M06 instead of BL~M12. Such behavior of the model showed that the model did not perform well in disease identification with increasing time steps of longitudinal data. With the 1D CNN, the model achieved the highest accuracies at BL~M12 with all three types of feature embeddings. No significant difference was observed between the accuracies achieved with each set of feature embedding. The GRU x 3 reported accuracies in a stable and upward performance, specifying that the model improved with increasing time steps of longitudinal data. We also tested more combinations of the base classifiers in the homogeneous EL network, that is, five and seven. We did not observe any significant improvement by repeating the same network more than three times. By comparing the performance results of the homogenous ensemble with those of the heterogeneous ensemble, we observed that the heterogeneous ensemble network performed well with the highest accuracy. In contrast, all combinations of homogenous EL showed accuracies below 80.

Fig. 9 shows the mAUC comparison of homogenous and heterogeneous ensemble networks. The heterogeneous EL network with three base classifiers outperformed all combinations of homogenous ensemble networks. In addition, in heterogeneous EL networks, no proper patterns in the accuracy with respect to increasing longitudinal time steps were found. In addition, none of the homogenous combinations of the EL network achieved more than 80% mAUC score. In the case of heterogenous EL networks, we noticed a significant improvement with the combination of three network ensembles. Furthermore, the model became more stable for each combination of feature embeddings with an increased number of longitudinal time steps.

## 6. Performance comparison of the state-of-the-art methods

This section presents a comparative analysis of the literature on deep ensemble modeling for AD progression detection using the proposed ensemble model. Notably, several evaluation metrics are reported in Table 11 for a fair comparison with existing studies. In addition, we noticed that many studies in Table 11 did not publish their training data, which made it impossible to reproduce their results. We relied on their published results and compared them with our results for an illustrative comparison. In the proposed EL framework, we utilized a soft voting technique that combined the probabilities of all the base classifiers and averaged them for the final output. We noticed that the proposed EL technique performed the best for several reasons. *First*, each base classifier was optimized using multimodal longitudinal data. The training data were composed of longitudinal data (MRI feature embeddings) and cross-sectional biomarkers (cognitive scores). *Second*, for each patient, a large portion of the data (110 MRI slices at each time step) was analyzed using each base classifier for the disease diagnostic process. Consequently, the model's applicability can be extended to numerous medical scenarios, and it is especially well suited for use in developing nations.

Table 11 summarizes the key points of EL-based studies published in the AD domain. We noticed that most studies concentrated on either a

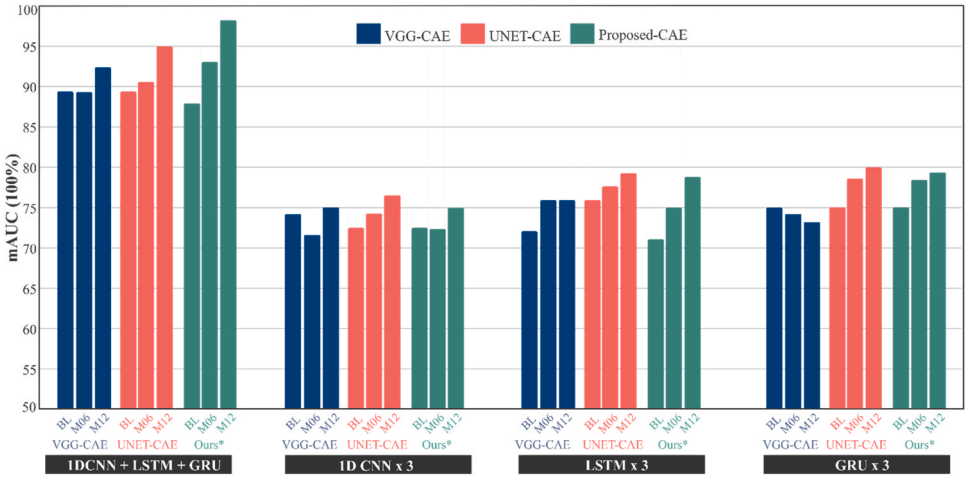


**Fig. 9.** Comparison of homogenous and heterogeneous EL models trained with MRI embeddings collected from different CAEs at different timesteps.

**Table 11**
Comparison of EL networks from the literature with the proposed EL framework.

| Ref., Year | Dataset | Modality | Time Series | Performance | ML Method | Task |
|---|---|---|---|---|---|---|
| [69], 2014 | OASIS: 85 | MRI | NO | 93.75/-/87.5/-/- | SVM, RF, and MLP based voting | Diagnosis (CN vs AD) |
| [74], 2015 | ADNI: 750 | NSB, FDG PET | NO | 96.3/96.3/96.3/-/- | Features selection and classification | Diagnosis (CN vs AD) |
| [32], 2016 | ADNI: 509 | MRI | NO | 92.30/-/-/-/- | Weighted sum rule-based voting and SVM | Diagnosis (CN vs AD) |
| [67], 2017 | ADRC: 41 | FMRI | NO | 95.00/-/-/-/- | SVM | Diagnosis (CN vs AD) |
| [27], 2017 | ADNI: 1737 | MRI, CSs, and CSF | YES (-) | 73.00 | RF | CN vs MCI vs AD |
| [70], 2017 | ADNI: 45 | DTI | NO | 83.3/80.0/82.5/-/- | LR, RF, SVM, KNN | Diagnosis (CN vs AD) |
| [75], 2018 | NACC: 386 | MRI, MMSE, logical memory (LM) | NO | 90.90/92.6/96.3/ 94.4/- | Voting of CNN VGG-11 for MRI and MLP for MMSE and LM | Diagnosis (CN vs MCI) |
| [61], 2020 | Figshare: 333 | CSF protein biomarkers, demographics, amyloid proteins, tau, ptau, Aβ42 | NO | 93.8/-/91.3/-/95.2 | SVM | Diagnosis (CN vs MCI) |
| [71], 2020 | ADNI: 815 | MRI | NO | 93.84/-/-/-/- | Majority voting (GG16, GoogleNet, and AlexNet) | Diagnosis (CN vs AD) |
| [72], 2020 | ADNI: 509 | MRI | NO | 84.00/-/-/-/92.00 | 3D SENet classifier | AD progression |
| [66], 2020 | NACC: 23,165 | MH, HIS, CVD, UPDRS, NPIQ, GDS, FAQ | NO | 85.00/86.00/84.30/ 84.6/-/- | Sparse autoencoder followed by stacking CNN | Diagnosis (CN vs AD) |
| [68], 2021 | ADNI: 480 | MRI | NO | 83.33/-/-/-/- | 3D densely connected CNN | AD progression |
| [31], 2021 | ADNI: 1737 | MRI, PET, CSF, cognitive tests, and demographics | NO | 84.00/-/-/-/- | Dynamic ensemble | Diagnosis (CN vs MCI vs AD) |
| [73], 2021 | ADNI:116 | FMRI | NO | 83.47/-/90.90/-/- | Clustering-evolutionary weighted SVMs | AD progression |
| [7], 2021 | Dataset from USC (Spain): 203 | MEG, MRI | YES (-) | 87.00/-/-/-/- | AlexNet with voting + {SVM + LDA} | AD progression |
| [6], 2021 | ADNI: 2228 | MRI, Demographic, MMSE | NO | 83.50/-/79.4/-/- | Graph CNN with voting | AD progression |
| [52], 2021 | ADNI: 813 | MRI | NO | 85.27/-/87.32/-/- | Ensemble of three DNN followed by voting | Diagnosis (CN vs AD) |
| [40], 2022 | ADNI: 3925 | MRI | NO | 87.11/-/-/-/- | Multiresolutional PartialNet | Diagnosis (CN vs AD) |
| [5], 2022 | ADNI: 1371 | CSs, NSB, Static | YES(4) | 99.56/99.56/99.56/ 99.51/- | Stacking with XGB framework | AD progression |
| [30], 2022 | ADNI: 612 | FDG-PET | NO | 89.05/-/-/-/- | Deep EL model for 2D and 3D CNN models | Diagnosis (CN vs AD) |
| **Ours*, 2023** | **ADNI: 1692** | **MRI, CS** | **YES(3)** | **96.11/96.24/96.22/ 97.06/95.43** | **CAE-AE followed by time series EL framework** | **AD progression** |

Dataset (-/-): (Name of dataset/Number of subjects), Performance (-/-/-/-/-): (Accuracy/Precision/Recall/F1-Score/AUC),

single modality of data (i.e., MRI) or a multimodality of data (MRI + other critical biomarkers).

In addition, many studies deal with AD diagnostic problems as binary classification tasks and only predict patients' status as CN or AD, which does not have a significant impact on the real environment. For instance, studies such as [67,68,69,32,70,71,72,73,52,40,30] utilized only a single modality of neuroimaging data i.e., MRI, PET, or FMRI. No data fusion of multimodalities and no time-series aspect of the training data were utilized in these studies. Instead, each framework acts as a simple classification model (i.e., diseased or nondiseased). Furthermore, most of these studies were based on an ensemble network for binary classification tasks such as CN versus AD [69,32,70,71,52,30]. For example, Armananzas et al. [67] proposed an ML-based EL framework with fMRI neuroimaging data to distinguish patients with AD from non-AD patients. They first preprocessed the fMRI volume using the statistical parametric mapping toolbox to select activated brain regions in the demented and non-demented groups. Feature selection of the activated brain regions was performed to extract the most relevant voxels, and then, an ensemble classifier was trained with these voxels to distinguish between CN and AD patients. Ruiz et al. [68] proposed a probability-based fusion of multiple CNN models to diagnose AD severity using brain MRI. A probability-based fusion ensemble approach was implemented by combining the probabilistic results of the final classification layers of different individual networks. This study asserts that an EL network is better than a single classifier. Choi et al. [71] proposed an ensemble of deep CNN models for each projection of brain MRI (i.e., cross-section planes) and trained it using a newly designed loss function based on sequential quadratic programming. They reported 93% accuracy in the classification of CN vs. AD.

Other studies such as [75,27,74,61,66,31,6,5] used multimodal data that included neuroimaging data fused with cognitive scores, patient demographics, and medication details. For instance, Sivapriya et al. [74] performed particle swarm optimization on multimodal data to select the best features to train an EL classifier. They proposed an EL classifier based on Naïve Bayes, random forest, SVM, and C4.5, based on the selected features set from multiple biomarkers, and reported 96% classification accuracy. Qui et al. [75] proposed a predictive model for detecting cognitively impaired people by analyzing brain anatomical structures and other cognitive abilities (i.e., MRI, MMSE, and LM tests). Three ML classifiers were initially trained, each with an individual modality, and later, their predictive decision was combined using soft voting to exemplify multimodal data fusion. They reported a classification accuracy of 90.9%. An et al. [66] proposed a three-layered deep EL framework to predict AD severity accurately. The multisource training data were compressed using a sparse autoencoder and then used to train multiple base classifiers. Subsequently, a stacking layer was designed, and the decision of all base classifiers was ranked using a nonlinear feature-weighted method, where the highest-ranked output

specifies the severity of AD. They claimed to perform 4% better than other well-known EL approaches. Liu et al. [6] proposed a multi-atlas multi-view graph convolution network to extract multiview features of the same brain regions in cognitively impaired individuals. A view-pooling

strategy was used to remove the redundant feature set before training the two-step EL model. They reported an AUC of 90.8%. Giovannetti et al. [7] proposed a deep-Meg model using longitudinal MEG data fused with brain MRI. They trained an AlexNet model with multimodal data for extracting functionally connected (FC) indices and then used this information to build predictive models using an SVM.

Table 11 provides a summary of the five key features from each of the previously discussed approaches: the dataset name and number of participants in the study, the type of data utilized, whether time-series data were employed, the number of time intervals, evaluation results, and the architecture applied. Many EL methods disregard the temporal aspect of AD data, thereby ignoring its longitudinal aspect, which inevitably hinders 'system performance. As seen in Table 11, the proposed network outperformed most EL-based models, except [5], where multiple critical modalities with large-scale datasets and more time steps of the longitudinal data were used. Based on a uniquely designed framework and multimodal longitudinal data analysis for a single patient, the proposed framework represents an excellent starting point for building a clinical decision support system for detection of AD progression.

## 7. Explainable deep surrogate model

The resulting ensemble model is robust, but this is not enough to get the trust of medical experts. As a result of the high architectural complexity of our proposed model, it is difficult to interpret the decision-making process within the framework. To overcome this issue, we developed a simplified version of the proposed framework that acts as a surrogate model for explaining the decision-making process. Initially, we removed the decoder, PCA, GRU, and 1D CNN subnetworks from the framework shown in Fig. 2. Then, the encoder part is directly attached to the LSTM model. Since the architectural design of LSTM and GRU are analogous, we chose LSTM as the time series model. The simplified version of the proposed model was a hybrid CNN-LSTM model with shared backbone trained weights for each time step (i.e., BL, M6, and M12). This model took 2D corresponding MRI slices at each time step and processed them for extracting deep features using the backbone CNN model. The extracted deep features were then used to train an LSTM network that captured the temporal features. This way, the LSTM network identified the progression of AD from the longitudinal MRI. We interpreted the output of this model by extracting the activation maps of the final CNN layer for each time step and visualized them on top of the original input, as shown in Fig. 10. Note that the surrogate model gives explanations for the decision if and only if its decision is the same as the main proposed model.

We employed the MedCam library [76] to visualize the salient features that make significant contributions in determining the final output class. A random sample from each class (i.e., CN, Converted, and AD) was fed into the trained MedCam model. To generate the regions of interest in the saliency maps of the last convolutional layer, we leveraged the Guided Grad-CAM technique. Fig. 10 displays the activated brain regions superimposed on the original 2D slice alongside the correspondingly highlighted areas. We applied this approach for each subject and time point in a longitudinal fashion, enabling us to reveal the evolutionary patterns of AD across all categories over time. Additionally, statistically significant salient regions were provided to enable users to better comprehend the reasoning behind the model's decisions. These regions included indicators for CN and AD cases, as well as subjects who transitioned from normal to AD within three years. Finally, numerous 2D slices from each 3D volume were presented to showcase the active voxels in each specific area.

*Patterns of brain atrophy in CN patients*: In the first row of Fig. 10, we see the activated brain regions for CN individuals at different time points during the disease diagnostic process. These regions serve as important differentiators between CN individuals and those diagnosed with AD. Specifically, the hippocampus, Amygdala, and Parahippocampal Gyrus are prominent brain regions that aid in distinguishing CN from AD patients [77]. Further subregions within the Amygdala such as the medial, basolateral, lateral, entorhinal cortices, and caudal areas are also marked as key contributors. Notably, these regions remain consistently active in all three timepoints (BL, M06, and M12), indicating that the patient maintains a steady state without any signs of cognitive decline or brain region deterioration throughout the diagnostic journey.

*Patterns of brain atrophy in converted patients:* The second row of Fig. 10 displays the progression of a converted patient, with brain tissue analysis revealing increasing atrophy from a healthy state to AD. Compared to healthy brains, converted cases experience rapid changes from normal state to AD. For converted patients who were CN at BL, the highlighted regions are the same as those identified in CN cases over three-time steps, but subsequent MRIs show brain shrinkage (i.e., BL∼M06 and onwards). The results are consistent with medical diagnostic procedures, including enlarged ventricles in the temporal horns and smaller hippocampi [78]. Temporal lobe regions beyond the hippocampus, including the amygdala, middle and inferior temporal gyrus, and fusiform gyrus, are affected by the disease process [79]. The amygdala's involvement in AD is well-known from previous research, as it is connected and adjacent to the hippocampus. Our model highlighted other temporal lobe structures, such as the fusiform gyrus, parahippocampal gyri, and middle and inferior temporal gyri, in converted patients at BL∼M06. In addition, recent imaging investigations have reported decreased volumes of the thalamus in AD, which aligns with past autopsy observations showing reductions in both the putamen and thalamus [80,79]. Consistent with these findings, we detected infected thalamic regions in the converted patient compared to CN at Month 12 (M12).

Additionally, we identified prominently affected regions within the mesial temporal lobe, including the hippocampus, amygdala, and parahippocampal gyri, as well as the temporal horn of the lateral ventricle and the posterior temporal lobe [81]. Beyond the temporal lobe, significant atrophy was observed in the posterior cortices, particularly the parietal and occipital lobes [82]. Such patterns conform to established descriptions of AD, characterized by a posterior-to-anterior gradient in cerebral atrophy. At BL∼M12, the patient had fully transitioned to AD state, resulting in extensive damage across nearly all brain tissues, including all the abovementioned regions from prior timepoints.

*Patterns of brain atrophy in AD patients:* The third row in Fig. 10 displays the progressive deterioration of an AD case over time. As seen in the first column (AD, BL timestep), the patient's initial state exhibits widespread activation, implying substantial neural dysfunction. The most notable affected regions encompass the medial temporal lobe, including the hippocampus, amygdala, and basal ganglia including the caudate nucleus, and putamen, indicative of severe disease burden already present at baseline (BL) timestep [77]. The medial temporal lobe has long been acknowledged for its importance in early AD diagnosis due to the consistent activations in this region. However, apart from the frequently examined medial temporal lobe, the network pinpointed other areas, namely the temporal lobe, insular cortex, and orbitofrontal cortex, demonstrating broader involvement of the temporal lobe in AD [83]. In the next stage (BL∼M06), peak activations extended to the inferior and superior temporal gyri, along with the fusiform gyrus, which are involved in pattern recognition and previously implicated in AD. Moving on to the frontal lobe, heightened activity centered around the middle and inferior frontal gyrus, crucial for decision making and problem solving, reportedly highly damaged in AD patients [79]. Moreover, the network emphasizes the detrimental role of the precuneus in AD-related functional breakdown. Lastly, as shown in the final column (BL∼M12), significant changes occurred in the parietal lobe, especially the precuneus, while cerebellar atrophy remains a key factor

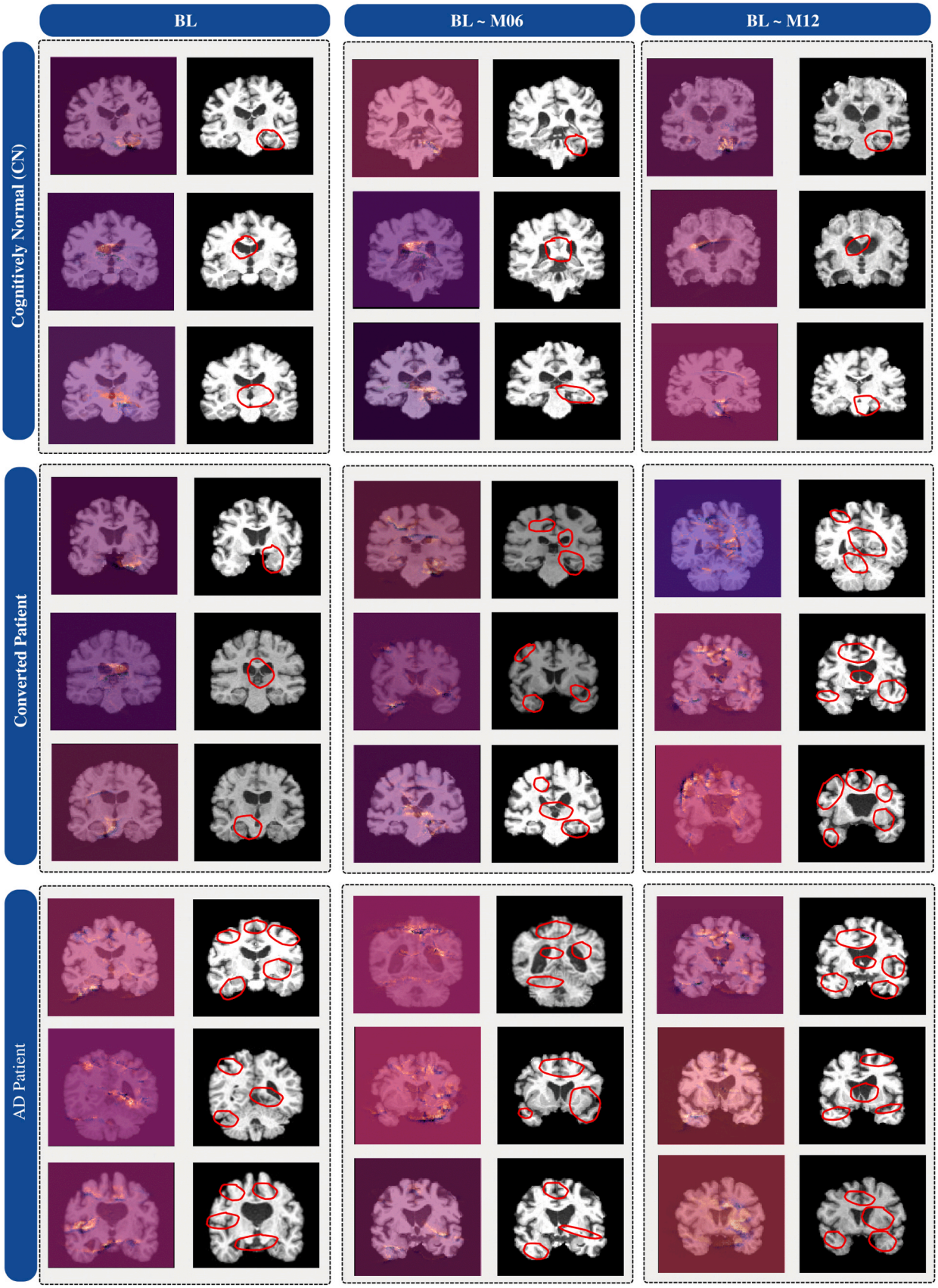


**Fig. 10.** Activated brain regions for CN, Converted, and AD classes across longitudinal timesteps.

in cognition and behavioral deterioration [82].

## 8. Model's generalizability evaluation

Our model has been optimized using the real-world ADNI dataset. To check the model stability and generalizability, we performed an external validation for the proposed model using a different dataset. This dataset is called the NACC [84] which was collected from patients from the European Union countries. Using this dataset, the final set of experiments conducted in this section evaluate the reliability and generalizability of the proposed model to an independently collected cohort. Table 12 displays the evaluation results of the proposed framework's robustness in terms of training on one distribution and testing it on another. This approach enables us to assess the robustness of the proposed framework in the disease identification process. In this experiment, the proposed Bayesian optimized heterogeneous ensembling framework (i.e., 1D CNN + LSTM + GRU) was tested on test subjects from the NACC dataset. The available NACC subjects for this experiment consisted of 31 CN subjects, 29 CN subjects that progressed to AD, and 27 AD subjects. To model the progression of AD within NACC, the data preprocessing steps applied to the NACC cohorts were identical to the ones applied to ADNI, as shown in Fig. 2. After processing, feature embeddings were extracted using each variant of CAE and PCA designed in the previous experiments. The obtained feature embeddings were then utilized to predict the health status in the NACC patients. Table 12 compares the modeling performance of the ADNI-trained model applied to both the ADNI and NACC test sets. As can be observed from the results, the previously selected configuration for training ADNI data is also a good choice when applied to NACC test data. Additionally, we observed improvement in disease identification with the longitudinal timestep from different variants of AEs.

For instance, at the BL timestep, the highest reported accuracy with the NACC test set is precision: 81.23 ± 2.51, recall: 85.63 ± 3.01, F1 score: 83.37 ± 2.27, AUC: 82.84 ± 2.59, and accuracy: 84.22 ± 1.61, which is achieved with UNET-CAE based feature embeddings. On the other hand, with the ADNI test set, the highest reported accuracy at the BL timestep was achieved with the proposed CAE-based feature embeddings and is as follows: precision: 89.13 ± 2.23, recall: 89.41 ± 2.22, F1-score: 88.06 ± 1.23, AUC: 88.24 ± 2.81, and accuracy: 87.75 ± 2.75. When data from the subsequent longitudinal time step is added to the disease diagnosis process, additional improvement in accuracy was reported for NACC patients. With two steps of data (i.e., at BL∼M06), the highest reported accuracy was achieved with proposed CAE-based feature embeddings, i.e., precision: 87.14 ± 3.53, recall: 85.24 ± 3.21, F1-score: 86.17 ± 2.28, AUC: 84.64 ± 3.15, and accuracy: 85.65 ± 2.25. However, we did not observe significant improvements using VGG-based feature embeddings, with accuracy values remaining in the range of 75–80%. Notably, the accuracy of our model continued to improve as longitudinal time steps increased. At BLM12, the proposed CAE-based feature embeddings showed significant improvement, with reported accuracies as follows: precision: 90.14 ± 3.04, recall: 91.24 ± 3.53, F1-score: 89.97 ± 2.43, AUC: 88.44 ± 3.34, and accuracy: 88.85 ± 2.15. Additionally, the reported accuracies of all five metrics for the ADNI test set were in the range of 95–97%. Our results suggest that the achieved accuracies with the NACC test set were lower than those achieved with ADNI subjects, likely due to the challenging nature of domain adaptation of multiple sources and cohorts.

*AUC comparison and model's robustness across cohorts:* Fig. 11 shows the mAUC comparison of a model trained on the ADNI dataset and tested on the NACC test set. We compared the achieved AUC at longitudinal time steps, including BL, BL∼M06, and BL∼M12, across different cohorts. Furthermore, we investigated the model's performance degradation over longitudinal time steps, which refers to the model's stability tested on data from different distributions. Though the performance achieved with NACC cohorts were not equal as ADNI subjects due to the different data distribution. However, the model was still able to capture the temporal features from the longitudinal timesteps of test data. The difference at BL time steps were significant, however the stability of the model gets improved with the additional subsequent time steps (i.e., BL∼M06, and on wards). The gap between the achieved accuracies reduces as shown in the case of VGG, and proposed CAE based features embeddings. In particular, the proposed light weight CAE-based feature embeddings exhibit better modeling performance than the VGG and UNET-based CAE models when applied to the NACC test set using the ADNI-trained model, highlighting the generalizability of the proposed approach across cohorts. It's worth noting that the NACC test set has a smaller number of subjects than ADNI, and both cohorts have imbalanced data. Specifically, the ADNI dataset comprises a total of 564 subjects across three classes (CN, Converted, and AD), whereas NACC has total of 87 subjects. The ADNI data covers a wide range of patients with varying health statuses, comprising many subjects. In contrast, the NACC test subjects are limited in number and are prone to noise. These factors may explain why the within-cohort ADNI performs better than the within-cohort NACC. Fig. 12 provides a visual illustration of this point, showing examples of 2D slices from both groups. The problem of domain adaptation in ML algorithms and healthcare, involving multiple

**Table 12**
Evaluating model robustness with test data.

| | | | Training on ADNI and test on NACC subjects | | | | | Train on ADNI and test on ADNI subjects | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | CAE | T-step | Pre. | Rec. | F1. | AUC | Acc. | Pre. | Rec. | F1. | AUC | Acc. |
| 1DCNN + LSTM + GRU | VGG-CAE | BL | 74.13 ± 2.01 | 72.21 ± 2.11 | 73.12 ± 2.48 | 73.14 ± 1.39 | 71.62 ± 2.11 | 88.38 ± 3.41 | 87.42 ± 3.21 | 88.31 ± 2.72 | 89.32 ± 2.12 | 87.20 ± 2.75 |
| | | BL∼M06 | 77.33 ± 3.04 | 75.28 ± 2.21 | 76.29 ± 2.11 | 74.44 ± 1.79 | 75.12 ± 3.51 | 89.38 ± 2.41 | 89.12 ± 2.14 | 88.72 ± 2.12 | 89.42 ± 3.82 | 87.60 ± 2.05 |
| | | **BL∼M12** | **87.13 ±2.25** | **85.68 ± 3.23** | **86.39 ± 2.70** | **86.34 ± 3.59** | **85.83 ± 3.11** | **94.11 ± 2.10** | **94.24 ± 2.11** | **93.02 ± 3.31** | **93.46 ± 2.10** | **93.29 ± 3.07** |
| | UNET-CAE | BL | 81.23 ± 2.51 | 85.63 ± 3.01 | 83.37 ± 2.27 | 82.84 ± 2.59 | 84.22 ± 1.61 | 88.32 ± 3.01 | 89.82 ± 3.54 | 88.41 ± 2.44 | 89.35 ± 3.82 | 87.80 ± 2.75 |
| | | BL∼M06 | 83.53 ± 3.32 | 84.63 ± 4.71 | 84.07 ± 2.35 | 84.22 ± 3.28 | 83.92 ± 3.65 | 90.38 ± 3.41 | 90.12 ± 2.14 | 90.72 ± 2.12 | 91.42 ± 3.17 | 89.50 ± 2.55 |
| | | **BL∼M12** | **88.54 ± 3.18** | **85.23 ± 3.74** | **86.85 ± 2.55** | **84.82 ± 3.73** | **84.72 ± 3.11** | **95.11 ± 2.48** | **95.24 ± 2.18** | **94.02 ± 2.31** | **95.84 ± 2.10** | **94.51 ± 2.02** |
| | Ours*-CAE | BL | 80.33 ± 3.13 | 83.44 ± 2.31 | 82.56 ± 2.43 | 81.24 ± 3.23 | 84.15 ± 3.35 | 89.13 ± 2.23 | 89.41 ± 2.22 | 88.06 ± 1.23 | 88.24 ± 2.81 | 87.75 ± 2.75 |
| | | BL∼M06 | 87.14 ± 3.53 | 85.24 ± 3.21 | 86.17 ± 2.28 | 84.64 ± 3.15 | 85.65 ± 2.25 | 93.41 ± 2.31 | 93.65 ± 2.71 | 94.42 ± 2.21 | 93.54 ± 1.41 | 92.51 ± 2.75 |
| | | **BL∼M12** | **90.14 ± 3.04** | **91.24 ± 3.53** | **89.97 ± 2.43** | **88.44 ± 3.34** | **88.85 ± 2.15** | **96.11 ± 1.14** | **96.24 ± 1.38** | **96.22 ± 1.30** | **97.04 ± 2.12** | **95.83 ± 2.32** |

CAE: convolutional auto encoder, T-step: longitudinal time step, Pre:precision, Rec:recall, F1:F1-score, Acc:accuracy

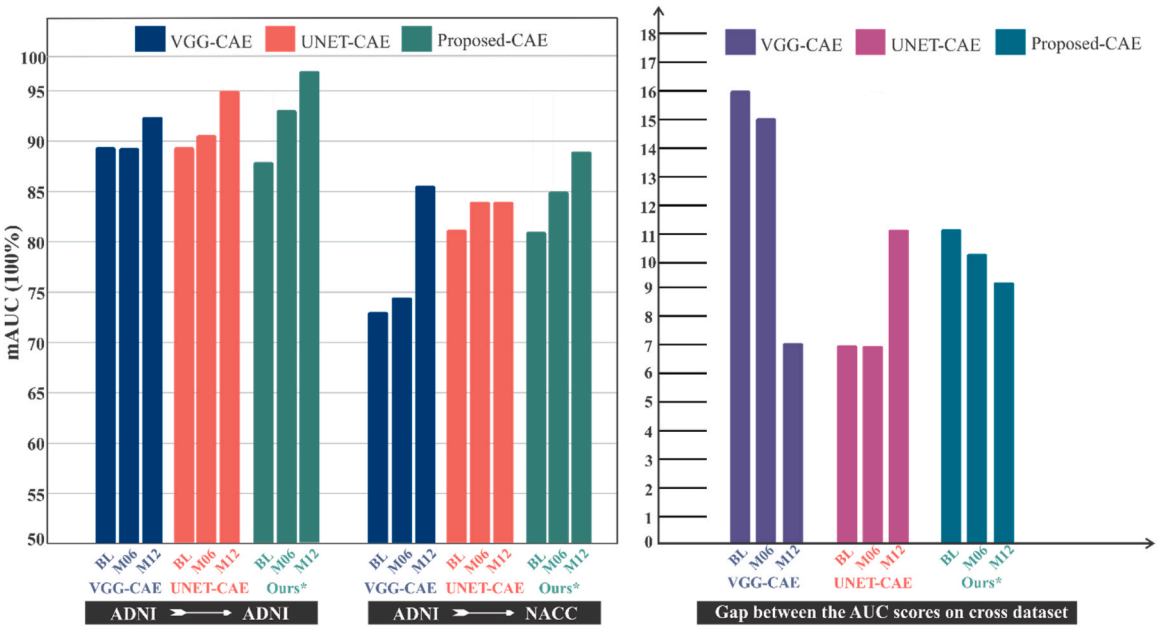


**Fig. 11.** mAUC comparison and accuracy gap between two different data distributions.

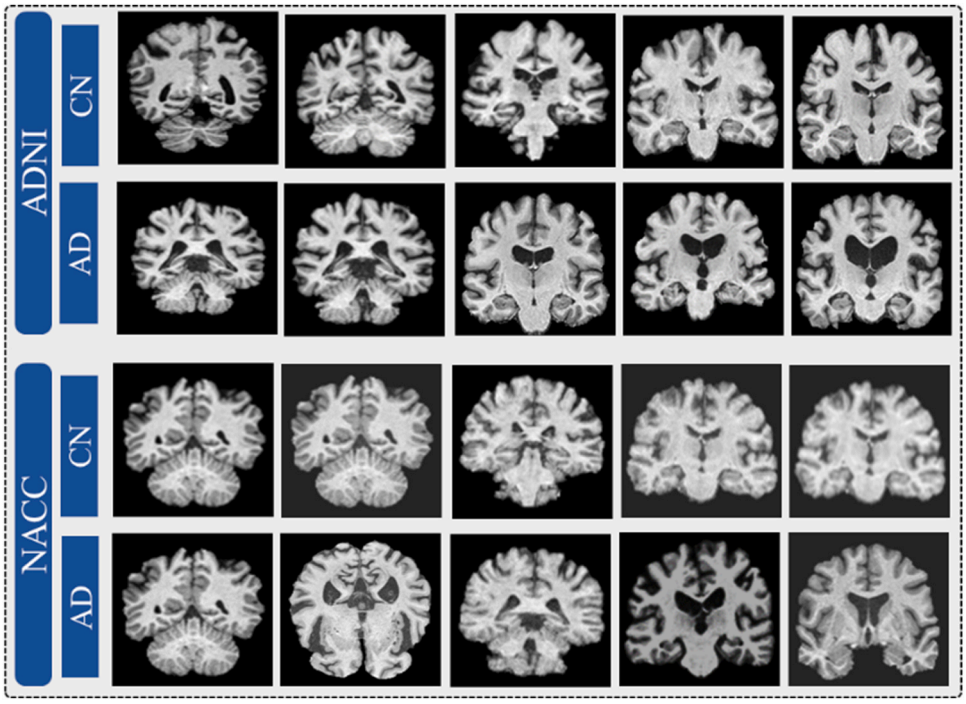


**Fig. 12.** Example of 2D middle of the brain slices from ADNI and NACC for CN and AD subjects.

sources and cohorts, presents significant challenges. Although Section 3.1 employs a preprocessing step to adjust/standardize the utilized each dataset, it's expected that applying a model trained on cleaned data (ADNI) to noisy (NACC) data with fewer biomarkers and measurements acquired from different sources/tools would result in a less accurate performance than applying it to data from the same source. In general, it can be deduced from Fig. 9 that the proposed heterogeneous ensembling model targets the same patterns in disease identification process across different distribution of data. That is, comparing the performance within ADNI and NACC reveals better results for ADNI, whereas a decrease in performance is observed across different cohorts. It can be inferred from the performance drop across cohorts that are likely the reason for this is cohort properties rather than model properties.

In this study, we conducted several types of validation to evaluate the performance and robustness of the optimized model. *First*, the effect of adding extra knowledge to the data in the form of new time steps increased accuracy and model stability. This was evident from the evaluation results, which showed that the selected configuration for training is optimal. *Second*, we performed **internal validation** of the model based on the well-known cross validation technique where we used the 5-fold cross validation. The robustness of the model was evaluated based on the stability of the results. This stability has been measured by the variance of the results among the different testing folds. There were no indications for overfitting or underfitting in the training

process based on the n-1 training folds (for n= 5). *Finally*, to further examine the generalization performance of the resulting model, we performed **external validation** based on a different AD benchmark dataset (i.e., NACC). The models appeared to generalize well to this unseen data from NACC dataset, indicating that it was appropriately trained and did not suffer from overfitting.

## 9. Limitations and future research

This study presents a pipeline for constructing a deep EL framework for detecting AD progression using a combination of longitudinal neuroimaging and cognitive scores. The key contributions of this study include the proposal of a lightweight CAE that produces latent feature representations for 2D MRI slices at each time point (i.e., BL, M06, and M12). The obtained deep features were further compressed into the finest set of feature vectors by applying PCA. Using a Bayesian optimizer, the obtained feature vectors were further utilized to tune the 1D CNN, LSTM, and GRU. The optimized model was further tested in homogenous and heterogeneous ensemble formats to improve AD diagnostics in healthcare systems. The best results were reported using heterogeneous ensemble networks, which outperformed all existing studies and other DL models. Although the proposed model surpasses other cutting-edge DL models in AD management, further refinement is necessary prior to its use in actual patient diagnosis. Future work will address these limitations; in real medical settings, experts require not only accurate decisions but also explanations that are based on context and user centered. For instance, the current study focuses mainly on performance, but there is a need for more interpretable and explainable results in the medical context. Therefore, our future plan involves the examination of XAI methods to demonstrate the role of each data modality in the decision-making process of ensemble models. In addition, the analyzed multimodal information consisted of MRI scans taken over a period of time and biomarker measurements taken only at the initial visit. It was not possible to examine the impact of cognitive scores at different points in time because this information was not available in our dataset. By combining MRI data with other neuroimaging modalities such as PET and DTI, it is anticipated that the accuracy of the model will increase and have greater medical significance. Finally, the model proposed in this study was created from scratch using ADNI data only. We could not evaluate our model with other datasets, such as OASIS [59], AIBL [85], and MIRIAD [60], because of compatibility issues, as discussed in Section 4.1. We will address these limitations in future studies.

## 10. Conclusion

AD is the most severe type of dementia at present, and there is no medically approved cure for it. The diagnostic methods available in the literature rely mostly on one-time data gathered during a patient's initial visit and do not consider the long-term nature of the patient's clinical information. This study presents a new method for detecting AD progression using longitudinal 3D MRI scans. We proposed a lightweight CAE to obtain high-level representational features of the 3D MRI volume in three steps (i.e., BL, M06, and M12). The output deep features were further compressed using PCA to preserve only the compact set of feature vectors from the convolutional feature maps. The obtained principal components in combination with patients' cognitive scores were further used by a Bayesian optimizer to tune a group of time-series models such as 1D CNN, LSTM, and GRU. The best-achieved accuracies were reported using multimodal data by the GRU model at BL∼M12, with a precision of 89.42 ± 2.16, recall of 88.54 ± 3.27, F1-score of 88.02 ± 1.24, AUC of 90.86 ± 2.17, and accuracy of 87.99 ± 3.07. We further tested various combinations of homogenous and heterogeneous EL networks by considering a large portion of the MRI volume (110 2D MRI slices) at each time step. Experimental results suggest that a heterogeneous ensemble of three-time series models (i.e., 1D CNN, LSTM, and GRU) outperformed each individual model and also other variants of EL models by achieving 96%, 96%, 96%, 97%, and 95% precision, recall, F1-score, AUC, and accuracy, respectively, with the proposed CAE based feature embeddings. These results are expected to be further enhanced in future studies by addressing the limitations discussed in the previous section.

## CRediT authorship contribution statement

**Shaker El-Sappagh:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. **Nasir Rahim:** Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Omar Amin El-serafy:** Visualization, Validation, Methodology, Investigation, Formal analysis. **Haytham Rizk:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis. **Tamer ABUHMED:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis.

## 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.

## Data availability

The authors do not have permission to share data.

## Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1011198), (Institute for Information & communications Technology Planning & Evaluation) (IITP) grant funded by the Korea government (MSIT) under the ICT Creative Consilience Program (IITP-2021-2020-0-01821), and AI Platform to Fully Adapt and Reflect Privacy-Policy Changes (No. 2022-0-00688).

## References

[1] Y. Liu, et al., Diffusion tensor imaging and tract-based spatial statistics in Alzheimer's disease and mild cognitive impairment, Neurobiol. Aging vol. 32 (9) (2011) 1558–1571, https://doi.org/10.1016/j.neurobiolaging.2009.10.006.
[2] J. Albright, Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm, Alzheimer'S. Dement. Transl. Res. Clin. Interv. vol. 5 (2019) 483–491, https://doi.org/10.1016/j.trci.2019.07.001.
[3] A. Abrol, M. Bhattarai, A. Fedorov, Y. Du, S. Plis, V. Calhoun, Deep residual learning for neuroimaging: an application to predict progression to Alzheimer's disease, J. Neurosci. Methods vol. 339 (2020) 108701, https://doi.org/10.1016/j.jneumeth.2020.108701.
[4] Z. Liu, T.S. Johnson, W. Shao, M. Zhang, J. Zhang, K. Huang, Optimal transport- and kernel-based early detection of mild cognitive impairment patients based on magnetic resonance and positron emission tomography images, Alzheimer'S. Res. Ther. vol. 14 (1) (2022) 1–12, https://doi.org/10.1186/s13195-021-00915-3.
[5] S. El-Sappagh, F. Ali, T. Abuhmed, J. Singh, J.M. Alonso, Automatic detection of Alzheimer's disease progression: an efficient information fusion approach with heterogeneous ensemble classifiers, Neurocomputing vol. 512 (2022) 203–224, https://doi.org/10.1016/j.neucom.2022.09.009.
[6] J. Liu, D. Zeng, R. Guo, M. Lu, F.X. Wu, J. Wang, MMHGE: detecting mild cognitive impairment based on multi-atlas multi-view hybrid graph convolutional networks and ensemble learning, Clust. Comput. vol. 24 (1) (2021) 103–113, https://doi.org/10.1007/s10586-020-03199-8.
[7] A. Giovannetti, et al., "Deep-MEG: spatiotemporal CNN features and multiband ensemble classification for predicting the early signs of Alzheimer's disease with magnetoencephalography, Neural Comput. Appl. vol. 33 (21) (2021) 14651–14667, https://doi.org/10.1007/S00521-021-06105-4/TABLES/4.
[8] Y. Fan, N. Batmanghelich, C.M. Clark, C. Davatzikos, Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline, Neuroimage vol. 39 (4) (2008) 1731–1743, https://doi.org/10.1016/j.neuroimage.2007.10.031.
[9] M. Tanveer, et al., Machine learning techniques for the diagnosis of alzheimer's disease: a review, ACM Trans. Multimed. Comput. Commun. Appl. vol. 16 (1s) (2020) 30, https://doi.org/10.1145/3344998.

[10] J. Jiang, L. Kang, J. Huang, T. Zhang, Deep learning based mild cognitive impairment diagnosis using structure MR images, Neurosci. Lett. vol. 730 (2020) 134971, https://doi.org/10.1016/j.neulet.2020.134971.

[11] N.T. Duc *et al.*, "3D-Deep Learning Based Automatic Diagnosis of Alzheimer's Disease with Joint MMSE Prediction Using Resting-State fMRI," 2021, doi: 10.1007/s12021-019-09419-w.

[12] Y. Zhang, et al., Multivariate approach for Alzheimer's disease detection using stationary wavelet entropy and predator-prey particle swarm optimization, J. Alzheimer'S. Dis. vol. 65 (3) (2018) 855–869, https://doi.org/10.3233/JAD-170069.

[13] D. Lu, K. Popuri, G.W. Ding, R. Balachandar, M.F. Beg, Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease, Med. Image Anal. vol. 46 (2018) 26–34, https://doi.org/10.1016/j.media.2018.02.002.

[14] L. Xu, X. Wu, K. Chen, L. Yao, Multi-modality sparse representation-based classification for Alzheimer's disease and mild cognitive impairment, Comput. Methods Prog. Biomed. vol. 122 (2) (Nov. 2015) 182–190, https://doi.org/10.1016/j.cmpb.2015.08.004.

[15] G. Muhammad, F. Alshehri, F. Karray, A.El Saddik, M. Alsulaiman, T.H. Falk, A Comprehensive Survey on Multimodal Medical Signals Fusion for Smart Healthcare Systems, in: Information Fusion, vol. 76, Elsevier, Dec. 2021, pp. 355–375, https://doi.org/10.1016/j.inffus.2021.06.007.

[16] S. El-Sappagh, T. Abuhmed, S.M. Riazul Islam, K.S. Kwak, Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data, Neurocomputing vol. 412 (2020) 197–215, https://doi.org/10.1016/j.neucom.2020.05.087.

[17] T. Abuhmed, S. El-Sappagh, J.M. Alonso, Robust hybrid deep learning models for Alzheimer's progression detection, Knowl. Based Syst. vol. 213 (2021) 106688, https://doi.org/10.1016/j.knosys.2020.106688.

[18] X.A. Bi, X. Hu, H. Wu, Y. Wang, Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest, IEEE J. Biomed. Heal. Inform. vol. 24 (10) (2020) 2973–2983, https://doi.org/10.1109/JBHI.2020.2973324.

[19] S. Huang, et al., "Identifying Alzheimer's disease-related brain regions from multi-modality neuroimaging data using sparse composite linear discrimination analysis, Adv. Neural Inf. Process. Syst. vol. 24 (2011).

[20] K.R. Gray, P. Aljabar, R.A. Heckemann, A. Hammers, D. Rueckert, Random forest-based similarity measures for multi-modal classification of Alzheimer's disease, Neuroimage vol. 65 (2013) 167–175, https://doi.org/10.1016/j.neuroimage.2012.09.065.

[21] Y. Zhang, S. Wang, K. Xia, Y. Jiang, P. Qian, Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion, Inf. Fusion vol. 66 (2021) 170–183, https://doi.org/10.1016/j.inffus.2020.09.002.

[22] S. El-Sappagh, et al., Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data, Futur. Gener. Comput. Syst. vol. 115 (2021) 680–699, https://doi.org/10.1016/j.future.2020.10.005.

[23] S. El-Sappagh, T. Abuhmed, K.S. Kwak, Alzheimer disease prediction model based on decision fusion of CNN-BiLSTM deep neural networks, Adv. Intell. Syst. Comput. (2021) 482–492, https://doi.org/10.1007/978-3-030-55190-2_36.

[24] N. Rahim, S. El-Sappagh, S. Ali, K. Muhammad, J. Del Ser, T. Abuhmed, Prediction of Alzheimer's progression based on multimodal DEep-learning-based Fusion and Visual Explainability of Time-series Data, Inf. Fusion vol. 92 (2023) 363–388, https://doi.org/10.1016/j.inffus.2022.11.028.

[25] A. Chincarini, et al., Integrating longitudinal information in hippocampal volume measurements for the early detection of Alzheimer's disease, Neuroimage vol. 125 (2016) 834–847, https://doi.org/10.1016/j.neuroimage.2015.10.065.

[26] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, Neuroimage vol. 104 (Jan. 2015) 398–412, https://doi.org/10.1016/j.neuroimage.2014.10.002.

[27] P.J. Moore, T.J. Lyons, J. Gallacher, Random forest prediction of Alzheimer's disease using pairwise selection from time series data, PLoS One vol. 14 (2) (2019), https://doi.org/10.1371/journal.pone.0211558.

[28] A. Holzinger, et al., Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion vol. 79 (2022) 263–278, https://doi.org/10.1016/j.inffus.2021.10.007.

[29] S. El-Sappagh, F. Hager Saleh, E.Amer Ali, Tamer Abuhmed, "Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time, Neural Comput. Appl. (2022) 1–23, https://doi.org/10.1007/s00521-022-07263-9.

[30] A. Yiğit, Y. Baştanlar, Z. Işık, Dementia diagnosis by ensemble deep neural networks using FDG-PET scans, Signal, Image Video Process vol. 16 (2022) 2203–2210, https://doi.org/10.1007/s11760-022-02185-4.

[31] M.N. Muhammed, P. Thiyagarajan, Alzheimer's classification using dynamic ensemble of classifiers selection algorithms: a performance analysis, Biomed. Signal Process. Control vol. 68 (Jul. 2021) 102729, https://doi.org/10.1016/J.BSPC.2021.102729.

[32] L. Nanni, C. Salvatore, A. Cerasa, I. Castiglioni, Combining multiple approaches for the early diagnosis of Alzheimer's Disease, Pattern Recognit. Lett. vol. 84 (2016) 259–266, https://doi.org/10.1016/j.patrec.2016.10.010.

[33] X. Tao, J.D. Velásquez, Multi-source information fusion for smart health with artificial intelligence, Inf. Fusion vol. 83–84 (2022) 93–95, https://doi.org/10.1016/J.INFFUS.2022.03.010.

[34] W. Wu, J. Venugopalan, M.D. Wang11C-PIB PET image analysis for Alzheimer's diagnosis using weighted voting ensembles Institute of Electrical and Electronics Engineers Inc , " in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS , Sep. 2017, , 3914–3917, 10.1109/EMBC.2017.8037712.

[35] A. Loddo, S. Buttau, C. Di Ruberto, Deep learning based pipelines for Alzheimer's disease diagnosis: a comparative study and a novel deep-ensemble method, Comput. Biol. Med. vol. 141 (2022), https://doi.org/10.1016/j.compbiomed.2021.105032.

[36] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM vol. 60 (6) (2017) 84–90, https://doi.org/10.1145/3065386.

[37] Z. Wu, C. Shen, A. van den Hengel, Wider or deeper: revisiting the resnet model for visual recognition, Pattern Recognit. vol. 90 (2019) 119–133, https://doi.org/10.1016/j.patcog.2019.01.006.

[38] M. Längkvist, L. Karlsson, A. Loutfi, Inception-v4, Inception-ResNet and the impact of residual connections on learning, Pattern Recognit. Lett. vol. 42 (1) (2014) 11–24. ⟨http://arxiv.org/abs/1512.00567⟩ ([Online]. Available).

[39] S.U. Sadat, H.H. Shomee, A. Awwal, S.N. Amin, M.T. Reza, M.Z. Parvez, Alzheimer's disease detection and classification using transfer learning technique and ensemble on convolutional neural networks, Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern. (2021) 1478–1481, https://doi.org/10.1109/SMC52423.2021.9659179.

[40] I. Razzak *et al.*, "Mutliresolutional ensemble PartialNet for Alzheimer detection using magnetic resonance imaging data," 2022, doi: 10.1002/int.22856.

[41] I.T. Jollife, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A: Math., Phys. Eng. Sci. vol. 374 (2065) (2016), https://doi.org/10.1098/rsta.2015.0202.

[42] S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, Int. J. Uncertain., Fuzziness Knowlege-Based Syst. vol. 6 (2) (Nov. 1998) 107–116, https://doi.org/10.1142/S0218488598000094.

[43] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," pp. 1–9, 2014, [Online]. Available: ⟨http://arxiv.org/abs/1412.3555⟩.

[44] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst. (2012) 2951–2959.

[45] M.F. Aslan, A. Durdu, A. Yusefi, K. Sabanci, C. Sungur, A tutorial: mobile robotics, SLAM, Bayesian filter, keyframe bundle adjustment and ROS applications, Stud. Comput. Intell. vol. 962 (2021) 227–269, https://doi.org/10.1007/978-3-030-75472-3_7/COVER.

[46] F. Guo, M. Ng, G. Kuling, G. Wright, Cardiac MRI segmentation with sparse annotations: ensembling deep learning uncertainty and shape priors, Med. Image Anal. vol. 81 (2022) 102532, https://doi.org/10.1016/j.media.2022.102532.

[47] T.G. DietterichlOnline]. Available MIT Press , vol. 40 The handbook of brain theory and neural networks-ensemble learning2002, ⟨https://courses.cs.washington.edu/courses/cse446/12wi/tgd-ensembles.pdf⟩.

[48] E. Kondrateva, P. Druzhinina, A. Kurmukov, K. Net, Do we really need all these preprocessing steps in brain MRI segmentation, Med. Imaging Deep Learn (2022) 2022.

[49] W. Kang, L. Lin, B. Zhang, X. Shen, S. Wu, Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis, Comput. Biol. Med. vol. 136 (2021) 104678, https://doi.org/10.1016/j.compbiomed.2021.104678.

[50] N. Rahim, T. Abuhmed, S. Mirjalili, S. El-Sappagh, K. Muhammad, Time-series visual explainability for Alzheimer's disease progression detection for smart healthcare, Alex. Eng. J. vol. 82 (no. August) (2023) 484–502, https://doi.org/10.1016/j.aej.2023.09.050.

[51] T. Che, et al., AMNet: adaptive multi-level network for deformable registration of 3D brain MR images, Med. Image Anal. vol. 85 (November 2022) (2023) 102740, https://doi.org/10.1016/j.media.2023.102740.

[52] M. Tanveer, A.H. Rashid, M.A. Ganaie, M. Reza, I. Razzak, K.L. Hua, Classification of Alzheimer's disease using ensemble of deep neural networks trained through transfer learning, IEEE J. Biomed. Heal. Inform. vol. 26 (4) (2022) 1453–1463, https://doi.org/10.1109/JBHI.2021.3083274.

[53] R. Gao *et al.*, "Technical Report: Quality Assessment Tool for Machine Learning with Clinical CT," pp. 1–18, 2021, [Online]. Available: ⟨http://arxiv.org/abs/2107.12842⟩.

[54] B.B. Avants, N.J. Tustison, and H.J. Johnson, "ANTs by stnava." [Online]. Available: ⟨http://stnava.github.io/ANTs/⟩.

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

[56] S.M. Muddamsetty, M.N.S. Jahromi, A.E. Ciontos, L.M. Fenoy, T.B. Moeslund, Visual explanation of black-box model: similarity difference and uniqueness (SIDU) method, Pattern Recognit. vol. 127 (2022) 108604, https://doi.org/10.1016/j.patcog.2022.108604.

[57] H.Il Suk, S.W. Lee, D. Shen, Deep ensemble learning of sparse regression models for brain disease diagnosis, Med. Image Anal. vol. 37 (2017) 101–113, https://doi.org/10.1016/j.media.2017.01.008.

[58] R.C. Petersen, et al., Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization, Neurology vol. 74 (3) (2010) 201–209, https://doi.org/10.1212/WNL.0b013e3181cb3e25.

[59] D.S. Marcus, A.F. Fotenos, J.G. Csernansky, J.C. Morris, R.L. Buckner, Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults, J. Cogn. Neurosci. vol. 22 (12) (2010) 2677–2684, https://doi.org/10.1162/jocn.2009.21407.

[60] I.B. Malone, et al., MIRIAD-Public release of a multiple time point Alzheimer's MR imaging dataset, Neuroimage vol. 70 (2013) 33–36, https://doi.org/10.1016/j.neuroimage.2012.12.044.

[61] A.H. Syed, T. Khan, A. Hassan, N.A. Alromema, M. Binsawad, A.O. Alsayed, An Ensemble-learning based application to predict the earlier stages of Alzheimer's disease (AD), IEEE Access vol. 8 (2020) 222126–222143, https://doi.org/10.1109/ACCESS.2020.3043715.

[62] Z. Xu, Q. Zhang, F. Hao, Z. Ren, Y. Kang, J. Cheng, VGG-CAE: unsupervised visual place recognition using VGG16-based convolutional autoencoder. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Science and Business Media Deutschland GmbH, 2021, pp. 91–102, https://doi.org/10.1007/978-3-030-88007-1_8.

[63] "GitHub - milesial/Pytorch-UNet: PyTorch implementation of the U-Net for image semantic segmentation with high quality images." Accessed: Nov. 13, 2022. [Online]. Available: ⟨https://github.com/milesial/Pytorch-UNet⟩.

[64] N. Rahim, J. Ahmad, K. Muhammad, A.K. Sangaiah, S.W. Baik, Privacy-preserving image retrieval for mobile devices with deep features on the cloud, Comput. Commun. vol. 127 (May) (2018) 75–85, https://doi.org/10.1016/j.comcom.2018.06.001.

[65] W. Weng, X. Zhu, INet: convolutional networks for biomedical image segmentation, IEEE Access vol. 9 (2021) 16591–16603, https://doi.org/10.1109/ACCESS.2021.3053408.

[66] N. An, H. Ding, J. Yang, R. Au, T.F.A. Ang, Deep ensemble learning for Alzheimer's disease classification, J. Biomed. Inform. vol. 105 (2020) 103411, https://doi.org/10.1016/j.jbi.2020.103411.

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

[68] J. Ruiz, M. Mahmud, M. Modasshir, M. Shamim Kaiser, 3D DenseNet ensemble in 4-way classification of Alzheimer's disease. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Science and Business Media Deutschland GmbH, 2020, pp. 85–96, https://doi.org/10.1007/978-3-030-59277-6_8.

[69] S. Farhan, M.A. Fahiem, H. Tauseef, An ensemble-of-classifiers based approach for early diagnosis of alzheimer's disease: classification using structural features of brain images, Comput. Math. Methods Med. vol. 2014 (2014), https://doi.org/10.1155/2014/862307.

[70] A. Ebadi, et al., Ensemble classification of Alzheimer's disease and mild cognitive impairment based on complex graph measures from diffusion tensor images, Front. Neurosci. vol. 11 (FEB) (Feb. 2017) 56, https://doi.org/10.3389/FNINS.2017.00056.

[71] J.Y. Choi, B. Lee, Combining of multiple deep networks via ensemble generalization loss, based on MRI images, for Alzheimer's disease classification, IEEE Signal Process. Lett. vol. 27 (2020) 206–210, https://doi.org/10.1109/LSP.2020.2964161.

[72] D. Pan, A. Zeng, L. Jia, Y. Huang, T. Frizzell, X. Song, Early detection of Alzheimer's disease using magnetic resonance imaging: a novel approach combining convolutional neural networks and ensemble learning, Front. Neurosci. vol. 14 (May) (2020) 1–19, https://doi.org/10.3389/fnins.2020.00259.

[73] X. an Bi, Y. Xie, H. Wu, L. Xu, Identification of differential brain regions in MCI progression via clustering-evolutionary weighted SVM ensemble algorithm, Front. Comput. Sci. vol. 15 (6) (Jan. 2021) 1–9, https://doi.org/10.1007/s11704-020-9520-3.

[74] T.R. Sivapriya, A.R.N.B. Kamal, P.R.J. Thangaiah, Ensemble merit merge feature selection for enhanced multinomial classification in Alzheimer's dementia, Comput. Math. Methods Med. vol. 2015 (2015), https://doi.org/10.1155/2015/676129.

[75] S. Qiu, G.H. Chang, M. Panagia, D.M. Gopal, R. Au, V.B. Kolachalama, Fusion of deep learning models of MRI scans, mini–mental state examination, and logical memory test enhances diagnosis of mild cognitive impairment, Alzheimer's Dement. Diagn. Assess. Dis. Monit. vol. 10 (2018) 737–749, https://doi.org/10.1016/j.dadm.2018.08.013.

[76] K. Gotkowski, C. Gonzalez, A. Bucher, A. Mukhopadhyay, M3d-CAM: A PyTorch Library to Generate 3D Attention Maps for Medical Deep Learning," in. Informatik aktuell, Springer Vieweg, Wiesbaden, 2021, pp. 217–222, https://doi.org/10.1007/978-3-658-33198-6_52.

[77] D.P. Devanand, et al., Hippocampal and entorhinal atrophy in mild cognitive impairment: Prediction of Alzheimer disease, Neurology vol. 68 (11) (2007) 828–836, https://doi.org/10.1212/01.wnl.0000256697.20968.d7.

[78] R.A. Heckemann, et al., Automatic morphometry in Alzheimer's disease and mild cognitive impairment, Neuroimage vol. 56 (4) (2011) 2024–2037, https://doi.org/10.1016/J.NEUROIMAGE.2011.03.014.

[79] G.W. Van Hoesen, J.C. Augustinack, J. Dierking, S.J. Redman, R. Thangavel, The parahippocampal gyrus in Alzheimer's disease. Clinical and preclinical neuroanatomical correlates. in Annals of the New York Academy of Sciences, New York Academy of Sciences, 2000, pp. 254–274, https://doi.org/10.1111/j.1749-6632.2000.tb06731.x.

[80] A. Cherubini, et al., Combined volumetry and DTI in subcortical structures of mild cognitive impairment and Alzheimer's disease patients, J. Alzheimer'S. Dis. vol. 19 (4) (2010) 1273–1282, https://doi.org/10.3233/JAD-2010-091186.

[81] M. Likeman, et al., Visual assessment of atrophy on magnetic resonance imaging in the diagnosis of pathologically confirmed young-onset dementias, Arch. Neurol. vol. 62 (9) (2005) 1410–1415, https://doi.org/10.1001/archneur.62.9.1410.

[82] J.D. Schmahmann, Cerebellum in Alzheimer's disease and frontotemporal dementia: not a silent bystander, Brain, pp. 1314–1318, May 01, Brain vol. 139 (5) (2016), https://doi.org/10.1093/brain/aww064.

[83] P.T. Nelson, et al., The amygdala as a locus of pathologic misfolding in neurodegenerative diseases, J. Neuropathol. Exp. Neurol. vol. 77 (1) (2018) 2–20, https://doi.org/10.1093/jnen/nlx099.

[84] D.L. Beekly, et al., The National Alzheimer's Coordinating Center (NACC) database: the uniform data set, Alzheimer Dis. Assoc. Disord. vol. 21 (3) (2007) 249–258, https://doi.org/10.1097/WAD.0b013e318142774e.

[85] K.A. Ellis, et al., The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease, Int. Psychogeriatr. vol. 21 (4) (2009) 672–687, https://doi.org/10.1017/S1041610209009405.