scientific reports

Check for updates

OPEN

Applying ensemble machine learning techniques to MRI scans to predict Alzheimer's disease

Georgios Theocharidis[1]✉, Sotirios Bisdas[2,3], John Pazarzis[7], Stavros Zervoudakis[4], Dongnanzi Zheng[5], Georgios Angelidis[6] & Antonios Saravanos[4]✉

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia globally. Early prediction, prior to the onset of symptoms, is critical for enabling timely interventions. We present a machine learning framework that predicts future AD conversion in cognitively normal (CN) individuals using only structural magnetic resonance imaging (MRI) data. The approach leverages transfer learning with a pre-trained VGG16 model for feature extraction and processes five representative 2D slices per brain MRI scan to generate compact imaging descriptors. These features are classified using an ensemble composed of support vector machines (SVM), random forests (RF), and artificial neural networks (ANN), with outputs combined through soft voting. The model was evaluated using person-wise stratified cross-validation on 1,093 subjects from the OASIS-3 dataset, ensuring no data leakage and providing realistic performance estimates. Across 200 randomized runs, the ensemble achieved a median AUC-ROC of 0.951, accuracy of 0.872, recall of 0.923, and F1 score of 0.811. These results demonstrate that ensemble machine learning can detect preclinical AD signatures from structural MRI, offering a practical, relatively accessible, and cost-effective tool for early risk identification and intervention.

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that accounts for up to 80% of dementia cases worldwide[1]. As the global population ages, the prevalence of AD is expected to rise[2]. In the United States, it is the sixth leading cause of death overall, and the fifth among individuals over 65[3]. Detecting AD during its preclinical stage, before cognitive symptoms emerge, is critical to enabling timely interventions that may delay progression and improve patient outcomes. In this study, preclinical AD is defined as the stage at which individuals are cognitively normal (CN) at the time of imaging but later develop clinical AD dementia. These individuals are referred to as *converters*, while those who remain CN throughout the study are termed *non-converters*.

The brain performs complex processes to record, store, and retrieve information through memories; however, AD impairs these functions, resulting in gradual and progressive memory loss[4]. While both AD and normal aging are associated with brain shrinkage, the underlying mechanisms and rates of brain shrinkage differ between the two conditions. AD is characterized by specific patterns of brain atrophy, including hippocampal and entorhinal cortex atrophy, which are not observed in normal aging.[5] Current diagnostic practice relies on medical history, neuropsychological testing, and sometimes neuroimaging, but these approaches have limited sensitivity to early-stage structural brain changes[6,7]. Magnetic resonance imaging (MRI) studies have identified patterns of hippocampal and entorhinal cortex atrophy in AD, yet these features often overlap with normal aging, making early diagnosis challenging[5,8].

Machine learning (ML) techniques, particularly deep learning, have increasingly been applied to neuroimaging to uncover latent patterns indicative of disease. Convolutional neural networks (CNNs) are well-suited to this task due to their ability to learn spatial features directly from imaging data[9,10]. These models have shown promise in distinguishing between AD, mild cognitive impairment (MCI), and healthy aging, particularly when trained on three-dimensional MRI inputs. Ensemble learning strategies, which integrate multiple classifiers, have further improved diagnostic accuracy by leveraging model diversity[11,12].

Nevertheless, much of the existing literature suffers from several limitations. First, many studies focus on patients already exhibiting symptoms, targeting diagnosis rather than early prediction. This reduces their

[1]Independent Researcher, Athens, Greece. [2]Department of Translational Neuroscience and Stroke, Institute of Neurology, University College London, London, UK. [3]Department of Neuroradiology, Athens Medical Centre, Athens, Greece. [4]New York University, New York, NY 10003, USA. [5]Columbia University, New York, NY 10025, USA. [6]School of Medicine, University of Crete, Heraklion, Crete 71003, Greece. [7]Independent Researcher, New York, NY, USA. ✉email: theocharidisg@outlook.com; saravanos@nyu.edu

clinical utility for preventive screening. Second, numerous approaches rely on multimodal data, such as positron emission tomography (PET) scans, cognitive scores, or genetic information, which are often unavailable in routine clinical practice. In contrast, MRI is already widely used in diagnostic workflows, is non-invasive, and is substantially more accessible and cost-effective than PET or CSF-based assessments. A predictive model that relies solely on MRI is therefore more feasible for broad clinical deployment, especially in settings with limited access to specialized diagnostic tools. Third, many reported models suffer from data leakage—a methodological flaw in which multiple scans from the same individual are included in both training and testing sets, leading to artificially inflated performance estimates. Wen et al. reviewed 30 CNN-based studies and found that over half were affected by this issue, calling into question the generalizability of their results[13].

This study seeks to address these limitations by developing a machine learning framework that predicts AD conversion in CN individuals using only structural MRI scans. The model leverages transfer learning for feature extraction, employing the convolutional layers of the VGG16 architecture pre-trained on natural image data. These extracted features are then passed to a set of three classifiers—artificial neural networks (ANN), support vector machines (SVM), and random forests (RF)—whose predictions are combined in an ensemble. To ensure rigorous evaluation, a person-wise stratification strategy is adopted. All MRI scans from a single individual are assigned exclusively to one data partition (training, validation, or testing). This approach eliminates data leakage and yields more realistic performance estimates.

Rather than processing entire volumetric scans, this framework selects five representative 2D slices per subject to capture essential anatomical regions. These include three axial slices at varying superior–inferior levels, one mid-sagittal slice, and one coronal slice. Each slice is resized and passed through the convolutional layers of the VGG16 model to extract salient features. To reduce dimensionality while preserving discriminative information, global average pooling is applied to the resulting feature maps. This generates a compact feature representation for each slice, which is concatenated into a single feature vector and used as input to the classifiers. This design balances computational efficiency with anatomical coverage and avoids the overhead of full-volume 3D processing.

The classifiers operate independently and predict the probability of each scan belonging to the AD-positive class. These predictions are aggregated through soft voting. A decision threshold of 0.6, empirically determined through validation set optimization, is used to label each scan. This threshold strikes a balance between precision and recall, aiming to minimize false positives while maintaining sensitivity in the context of class imbalance.

A key challenge in this domain is the skewed class distribution inherent to longitudinal imaging datasets such as OASIS-3. The majority of participants remain cognitively stable, with a relatively small number transitioning to AD. This imbalance can bias classifiers toward the majority class, reducing sensitivity to converters. To address this, we evaluated several techniques including cost-sensitive learning[14] and synthetic oversampling via SMOTE[15], ultimately selecting a modest undersampling strategy that retained clinical realism while improving minority class performance.

While earlier ensemble approaches have demonstrated strong performance, they typically rely on data from symptomatic individuals and often combine imaging data with cognitive or genetic metrics[11,12]. In contrast, the present study focuses solely on structural MRI and includes only participants who were CN at the time of scanning. No cognitive scores, PET data, or genetic information were used in model training. This design ensures that any predictive signal identified by the model is derived exclusively from imaging data obtained before the onset of detectable clinical symptoms.

Furthermore, the issue of data leakage is directly addressed through strict person-level partitioning. By preventing overlap in subject identity across dataset splits, the model's generalization capacity is more accurately assessed. This is particularly important in longitudinal studies, where a single individual may contribute multiple scans over time. Without careful stratification, performance metrics can be misleadingly high, as the model may exploit subject-specific rather than disease-specific features[13].

In summary, this work presents a machine learning approach designed to predict the development of Alzheimer's disease in CN individuals using only structural MRI. The method addresses key limitations in the literature by (i) focusing on preclinical prediction rather than diagnosis, (ii) using single-modality imaging data, (iii) applying transfer learning for efficient feature extraction, (iv) combining multiple classifiers in a robust ensemble, and (v) adopting a person-wise stratification scheme to prevent data leakage. A central strength of the proposed approach lies in its exclusive reliance on structural MRI. An imaging modality that is non-invasive, widely available in clinical settings, and substantially more affordable than alternatives. Because MRI is already a routine part of many diagnostic workflows, our solution can be readily integrated into existing healthcare infrastructure without requiring additional equipment, invasive procedures, or multimodal data. These advantages make the framework particularly well-suited for large-scale, early-stage AD screening, including in resource-constrained environments where access to advanced diagnostic tools may be limited.

Results

In this section, we present the results of our Alzheimer's disease (AD) prediction pipeline, beginning with a detailed description of the dataset and preprocessing steps, followed by the feature extraction strategy, partitioning methodology, and techniques employed to address class imbalance. We then describe the ensemble classification framework and evaluate its performance across multiple randomized experiments. The results highlight both the stability and discriminative power of the proposed approach, demonstrating its effectiveness for early AD detection in clinical MRI data.

Data structure

We used the Open Access Series of Imaging Studies (OASIS-3) dataset[16], which provides magnetic resonance imaging (MRI) scans from 1,093 individuals. Fig. 1 illustrates representative MRI slices from the dataset. Of

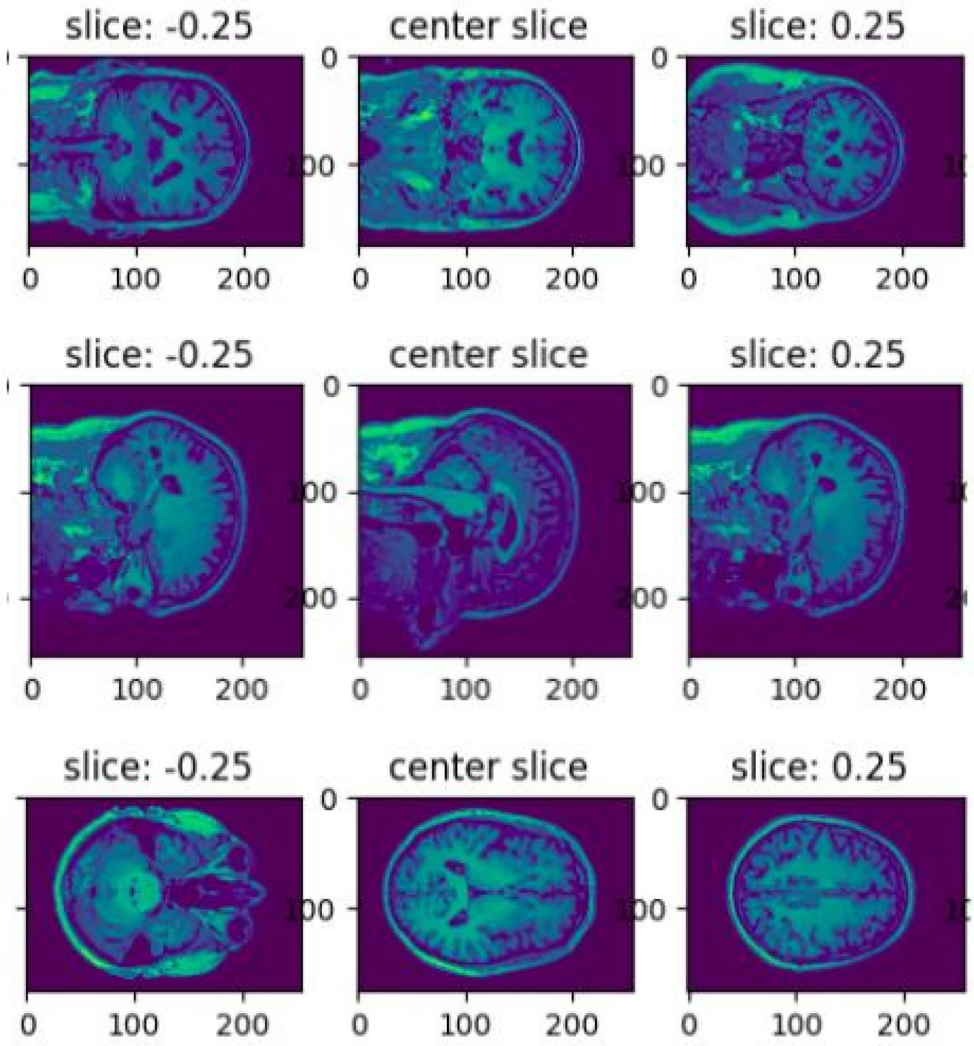


Fig. 1. Example images showing MRI slices from the OASIS-3 dataset.

these, 41 had corrupted MRI scans and were thus unusable, from the remainder 659 remained cognitively normal (CN) throughout the study period, whereas 393 exhibited varying degrees of cognitive decline. A subset of these individuals presented conflicting information across the clinical dementia rating (CDR), mini-mental state examination (MMSE), and clinical diagnosis. These "uncertain" cases, typically associated with mild cognitive impairment (MCI), represent patients with conflicting diagnostic parameters, making it difficult to definitively classify their cognitive status. To ensure clarity and consistency in model training, all uncertain cases, including those with MCI or other conflicting diagnostic information, were excluded from the entire model development process. This exclusion was necessary to minimize potential noise in the data and to ensure that the model was trained solely on participants with clear diagnostic outcomes, either classified as CN or suffering from AD.

Each MRI scan consists of a three-dimensional volume made up of sequential two-dimensional slices. To balance computational efficiency with predictive accuracy, we tested configurations using between one and nine representative slices per scan and selected the configuration that maximized AUC-ROC. Because the dataset includes participants, as illustrated in Fig. 2, with diverse cognitive trajectories, a consistent and tractable input format was necessary. The analysis identified five slices as optimal: three axial slices sampled along the superior–inferior axis to capture both cortical and subcortical regions, one mid-sagittal slice to represent midline anatomy, and one coronal slice to capture bilateral structures. For example, an axial slice at the medial temporal level includes the hippocampus and ventricular system, while the mid-sagittal slice highlights callosal and brainstem landmarks. This carefully selected set retained key anatomical information relevant to AD prediction while significantly reducing computational overhead. All selected slices were resized to 224×224 pixels using the OpenCV library[17] to ensure standardized input dimensions across the dataset. The original MRI data were stored in NIfTI format[18], a widely used standard in neuroimaging. To ensure compatibility with Python-based machine learning libraries, the NIfTI files were converted into NumPy arrays using the NiBabel library[19]. This transformation produced 2D numerical arrays, where each value represents the brightness of a voxel (i.e., a volumetric pixel), facilitating seamless integration with the machine learning pipeline.

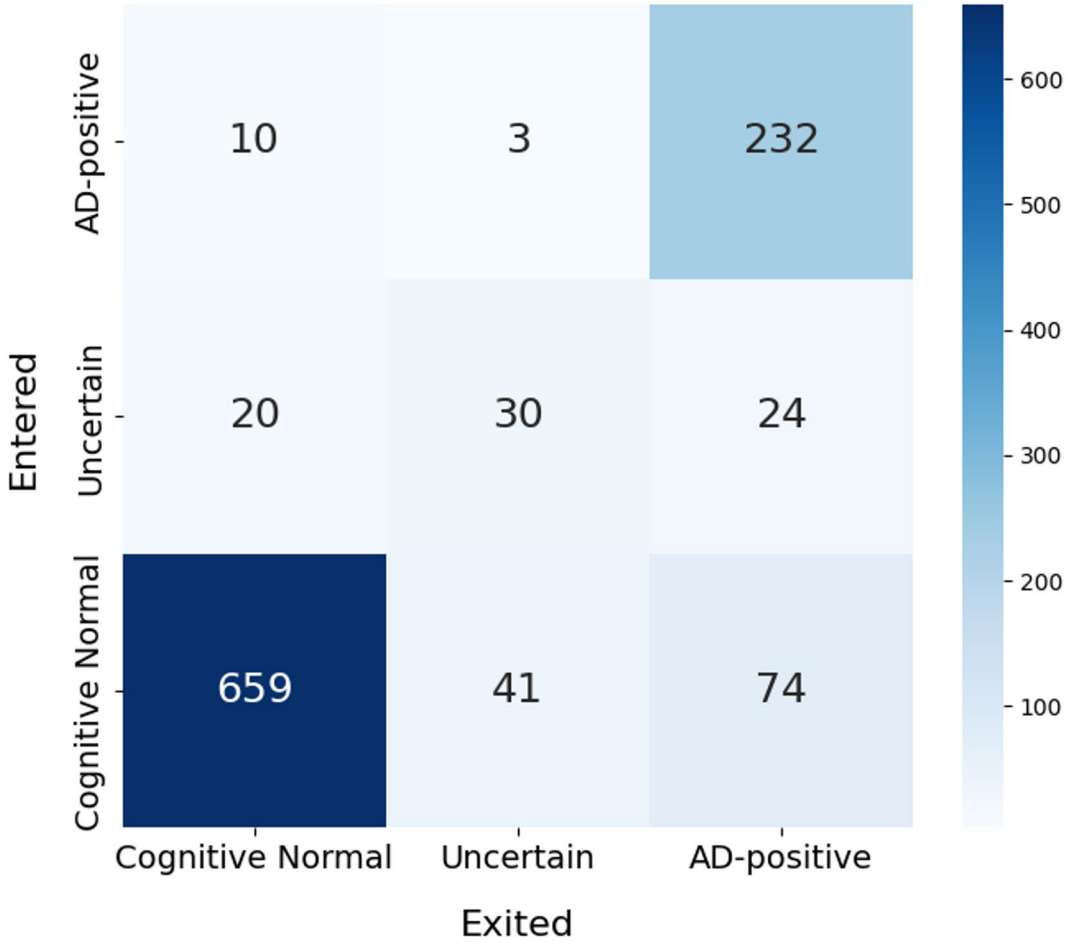


Fig. 2. Confusion matrix of cognitive state transitions among the 1,093 subjects in the OASIS-3 dataset.

To address the ambiguity arising from conflicting information in clinical diagnosis (dx1), Mini-Mental State Examination (MMSE), and Clinical Dementia Rating (CDR) scores, a composite metric combining all three parameters was developed to improve the consistency of cognitive state classification. Fig. 3 shows a scatter plot of MMSE versus CDR scores, with each data point color-coded by clinician diagnosis (i.e., CN, uncertain, or AD-positive), illustrating how the integration of these metrics helps distinguish between diagnostic categories. This composite approach classified each subject as CN, uncertain, or AD-positive, with the proportions of each group shown in Fig. 2. For model development, only participants with clear diagnostic outcomes were included, specifically those labeled as either CN or AD-positive. This approach ensured alignment with clinical practice and reduced the impact of label noise on model training. After preprocessing, all extracted MRI slices, along with their associated labels and metadata, were serialized using Python's pickle module. This process converted the data into a byte stream, preserving the internal structure of Python objects such as NumPy arrays and dictionaries[19,20]. Storing the data in this format allowed for efficient reuse and loading while avoiding the need to repeat time-consuming preprocessing steps in downstream tasks.

Feature extraction and representation using transfer learning

For feature extraction, we employed a pre-trained 2D convolutional neural network (CNN), VGG16, instead of deeper or fully 3D CNN architectures. Pre-trained CNNs, particularly those trained on large-scale datasets such as ImageNet, have been successfully applied to AD classification tasks, offering a strong foundation for learning transferable visual features. Compared to training a full 3D network from scratch, using VGG16 significantly reduces computational and memory requirements while also mitigating the need for large-scale labeled datasets. Training 3D models effectively would require substantially more data to capture the increased spatial complexity in volumetric inputs. This design decision aligns with recent research emphasizing the importance of computational efficiency in clinical MRI workflows. In our pipeline, VGG16 provided an effective compromise between leveraging transfer learning and maintaining manageable resource demands. The model was used to perform both dimensionality reduction and feature extraction from high-resolution MRI data.

Each 3D MRI scan was reduced to five anatomically representative 2D slices: three axial (sampled along the superior–inferior axis), one mid-sagittal, and one coronal. These five slices were selected based on their ability to capture key anatomical structures implicated in AD, including the hippocampus, entorhinal cortex, and corpus callosum. All slices were resized to 224×224 pixels using OpenCV [17] to match the VGG16 input requirements.

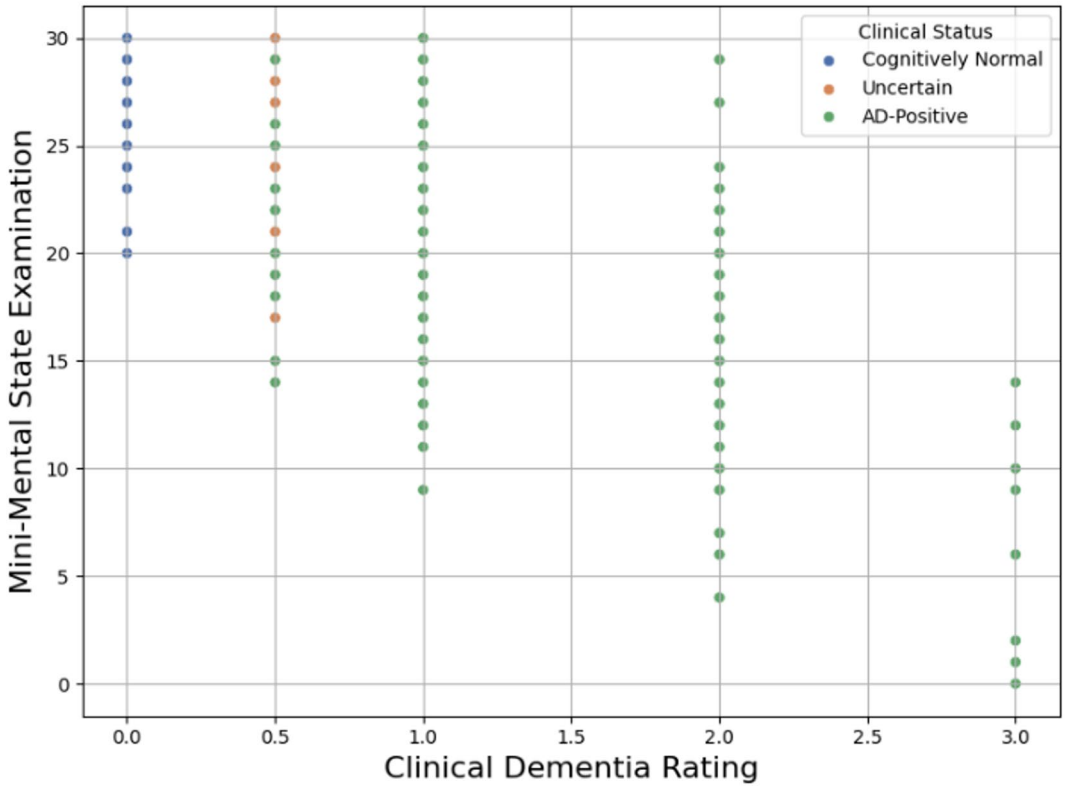


Fig. 3. Scatter plot showing MMSE vs. CDR scores colored by diagnosed cognitive state.

The original MRI volumes were in NIfTI format [18], loaded via NiBabel [19], and converted to NumPy arrays [20] for downstream processing.

Although the resized images retained important anatomical structures, a significant portion of the pixel space contained non-informative background. To address this, we applied the pre-trained VGG16 model to extract the most salient features from each slice [21]. Originally trained on over 14 million images across approximately 22,000 categories, including animals, vehicles, plants, and household objects, VGG16 was able to transfer its general pattern recognition capabilities to our medical imaging task.

Normally, the final stage of the VGG16 model involves using a fully connected layer for classification; however, in our approach, we bypassed this step. Instead, we utilized the output from the last convolutional layer, which provides a detailed analysis of the images, to create a compact set of 512 feature maps, each measuring 6×6 pixels, for each MRI slice. The 6×6 size of these feature maps is a result of both VGG16's architecture and the size of our input images. These feature maps preserve the essential features of the original images while eliminating non-informative pixels, thus reducing the dimensionality of the data. By extracting features at this stage, we preserved the critical details from the images while reducing the computational complexity and memory footprint of the subsequent model training process. This approach not only optimized the use of the available data but also allowed faster inference during model deployment. Furthermore, transfer learning allowed us to leverage the generalized knowledge embedded in the VGG16 model, transferring its learned representations to the specific medical imaging task in this study with no fine-tuning required (Fig. 4).

Global average pooling was then applied to optimize the efficiency of the feature maps by simplifying them while minimizing the loss of useful information. This operation transforms each 6×6 feature map into a single value representing the average of its pixels, thereby averaging the spatial information and making the model more robust to spatial translations of its input. The Keras function `GlobalAveragePooling2D` was used to reduce the feature map size, resulting in the 512 6×6 input feature maps being replaced by a one-dimensional array containing 512 values. Minimizing the complexity of the data helped reduce the risk of overfitting and improved the model's generalization ability on unseen data. Additionally, reducing the size of the feature map made the model faster and less computationally demanding, allowing for more comprehensive hyperparameter tuning later in the modeling process.

After condensing the MRI scan into five arrays of 512 essential features (one corresponding to each respective slice), these arrays were concatenated into a single, one-dimensional numerical array containing 2,560 values. This compact representation was then used as an input for the various classifier algorithms, i.e., the support vector machine (SVM), random forest (RF), and artificial neural network (ANN) models. From this point onward, the original scan slices were no longer utilized in the modeling workflow, as the classifiers relied solely on the derived global averages of the feature maps.

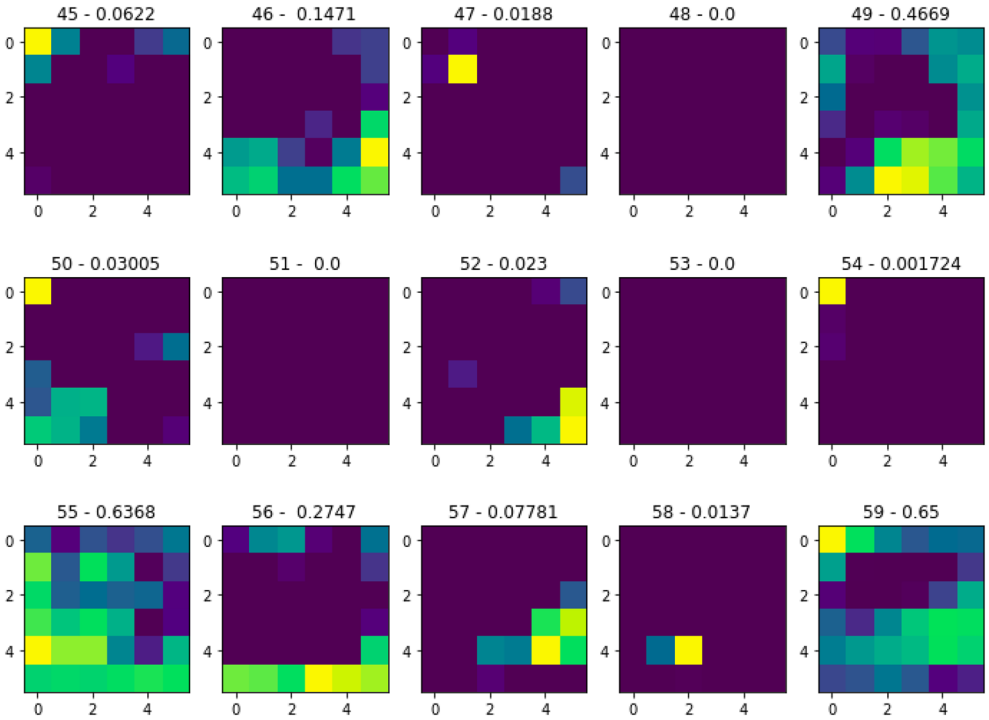


Fig. 4. Sample feature maps obtained from the VGG16 model. Each feature map highlights different spatial patterns learned from the data.

For feature extraction, we employed a pre-trained 2D convolutional neural network (CNN), VGG16, instead of deeper or fully 3D CNN architectures. Pre-trained CNNs, particularly those trained on large-scale datasets such as ImageNet, have been successfully applied to AD classification tasks, offering a strong foundation for learning transferable visual features. Compared to training a full 3D network from scratch, using VGG16 significantly reduces computational and memory requirements while also mitigating the need for large-scale labeled datasets. Training 3D models effectively would require substantially more data to capture the increased spatial complexity in volumetric inputs. This design decision aligns with recent research emphasizing the importance of computational efficiency in clinical MRI workflows. In our pipeline, VGG16 provided an effective compromise between leveraging transfer learning and maintaining manageable resource demands. The model was used to perform both dimensionality reduction and feature extraction from high-resolution MRI data. After preprocessing, each subject's scan was reduced to five selected slices, resized to 224×224 pixels. Although these images retained important anatomical structures, a significant portion of the pixel space contained non-informative background. To address this, we applied the pre-trained VGG16 model to extract the most salient features from each slice[21]. Originally trained on over 14 million images across approximately 22,000 categories, including animals, vehicles, plants, and household objects. VGG16 was able to transfer its general pattern recognition capabilities to our medical imaging task.

Data partitioning and stratification strategy

When developing a machine learning model to predict AD, having a sufficient amount of data for training is essential. However, the size of the available dataset was limited, as it contained multiple brain scans from the same individuals taken at different times. This presented a significant challenge in terms of training and evaluating the models.

The primary concern with this dataset was that including scans from the same individuals in both the training and validation sets could lead to the model learning specific details unique to those individuals' scans. Consequently, the model might appear to perform exceptionally well on the validation set; however, this behavior is not due to its ability to generalize extremely well to new, unseen data but instead because it has effectively "memorized" the scans it encountered during training. This phenomenon, known as data leakage, results in artificially inflated model performance, thus hindering the evaluation of the model's true predictive capabilities. This issue was described by Wen et al.[13], who, in their review, reported that over half of 30 studied papers suffered from the data leakage phenomenon, which artificially inflated these studies' results[13].

To mitigate this issue, we employ a technique called "person-wise stratification". This approach ensures that all scans from an individual are contained within one dataset partition, be it the training set, the validation set, or the test set. By enforcing this segregation, we ensured that the trained model was tested on entirely unseen data from different individuals during the validation and testing phases. This method prevented any overlap between the training and validation sets, thus eliminating the risk of data leakage and allowing the model's AD prediction performance to be more accurately assessed in terms of its generalization capabilities on unseen data.

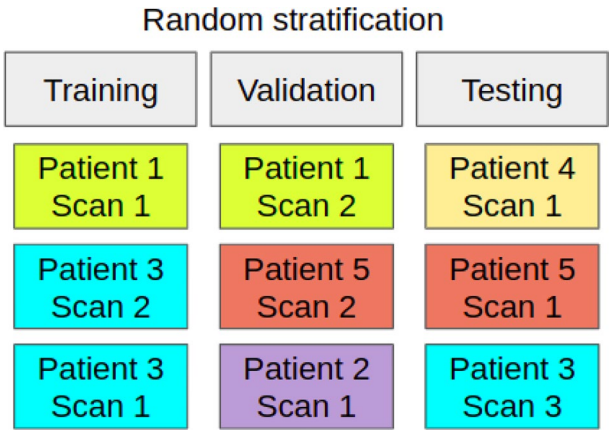


Fig. 5. Random stratification example.

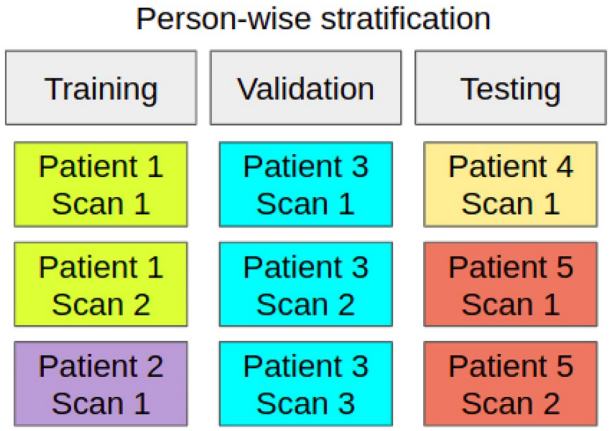


Fig. 6. Person-wise stratification example.

A comparison between traditional (i.e., random) stratification and person-wise stratification is shown in Figs. 5 and 6.

Addressing class imbalance

As shown in Fig. 7, the dataset was imbalanced, with cognitively normal (CN) participants comprising the majority of the analyzed cohort. Given the minority prevalence of converters, we evaluated several approaches to address class imbalance, including cost-sensitive learning [14] and SMOTE[15]. The final pipeline used modest undersampling of CN in the training set only, maintaining some imbalance to reflect clinical prevalence. Validation and test sets preserved natural class proportions and were never resampled.

Ensemble

The final stage of the pipeline involved feeding the concatenated one-dimensional feature vectors into an ensemble of three classifiers: support vector machine (SVM), random forest (RF), and artificial neural network (ANN), as illustrated in Fig. 8. Each model independently predicted the probability that a scan belonged to the AD-positive class. The ensemble combined these outputs using a soft voting strategy, computing the mean of the predicted probabilities and applying a decision threshold to determine the final label. We empirically selected a threshold of 0.6. If the average predicted probability across the three models exceeded 0.6, the scan was labeled as AD-positive; otherwise, it was classified as CN[22].

To evaluate the robustness of our ensemble binary classification model, we ran 200 randomized experiments in which the participants were randomly placed in one of the three groups (i.e., training, validation, and testing). Given that AUC-ROC, F1, F2, accuracy, precision, and recall are bounded between 0 and 1 and exhibited skew with significant departures from normality (Shapiro–Wilk $p \leq 0.05$), we summarize performance using the median and 95% confidence intervals for the median, which are more robust to outlying runs. The results summarized in Table 1 demonstrated consistently strong performance across evaluation metrics. We first report the area under the receiver operating characteristic curve (AUC-ROC), which measures how well a model can distinguish between positive and negative classes across all possible thresholds. The ROC curve plots the true positive rate (recall) against the false positive rate, generated by varying the decision threshold. AUC-ROC

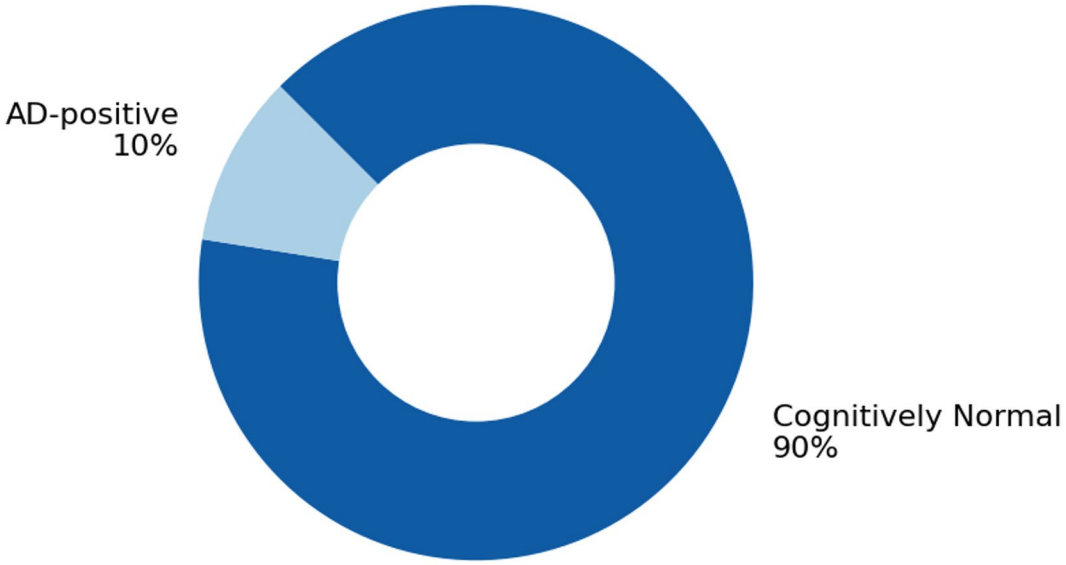


Fig. 7. Pie chart showing the proportion of each cognitive class in our dataset.

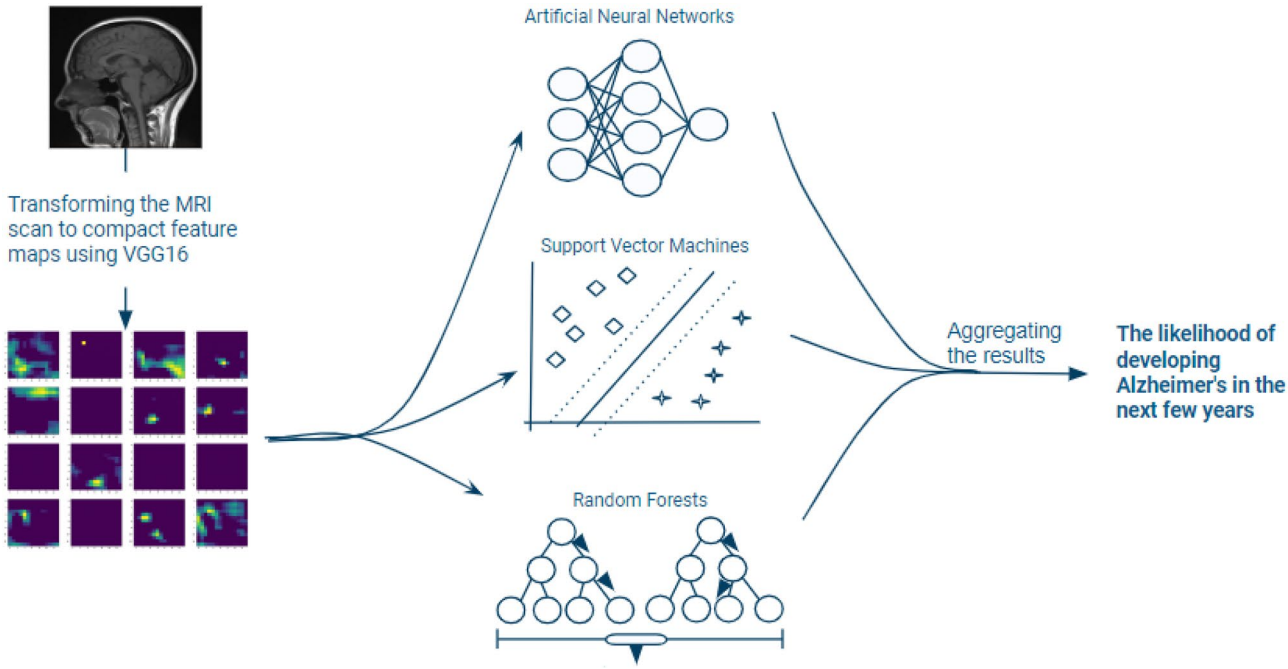


Fig. 8. Overview of the ensemble classification pipeline.

Metric	Median	95% CI (Median Low)	95% CI (Median High)	Best Model
F1	0.811	0.805	0.817	0.941
Accuracy	0.872	0.869	0.875	0.974
AUC	0.951	0.949	0.953	1.000
Precision	0.714	0.706	0.722	1.000
Recall	0.923	0.919	0.927	0.888

Table 1. Model performance metrics with medians and 95% confidence intervals.

ranges from 0 to 1, with higher values indicating better discriminative performance; a score of 0.5 suggests performance no better than random guessing. In our study, the median AUC-ROC was 0.951 (95% CI: 0.949–0.953), confirming excellent discriminative ability.

Alongside AUC-ROC, we report threshold-dependent metrics at the prespecified operating point. The F1 score[23] is the harmonic mean of precision (the percentage of true positive predictions among all positive predictions) and recall (the percentage of true positives among all actual positives). In imbalanced datasets, F1 provides a concise summary of performance at a chosen threshold[24], while the F2 score places greater weight on recall. In our experiments, the median F1 was 0.811 (95% CI: 0.805–0.817) and the median F2 was 0.750 (95% CI: 0.743–0.757), reflecting strong and stable performance across person-wise randomized splits.

Other metrics further contextualize model behavior: median accuracy was 0.872 (95% CI: 0.869–0.875), precision had a median of 0.714 (95% CI: 0.706–0.722), and recall reached a higher median of 0.923 (95% CI: 0.919–0.927). The higher recall relative to precision suggests that the model prioritizes sensitivity, successfully identifying most true positives while tolerating more false positives. This trade-off minimizes missed detections, which is often advantageous in applications where false negatives carry greater risk than false alarms. Fig. 9 shows the distributions for all metrics; AUC-ROC values are tightly concentrated near 0.95 across runs, consistent with the aggregate summary. The ROC curve demonstrates the strong discriminative power of the ensemble, consistent with the high AUC values observed across randomized experiments.

Finally, we found that the ensemble approach provided a balanced trade-off between sensitivity and specificity, outperforming single-model baselines in both stability and generalization. This highlights the benefit of combining diverse classifiers for robust early AD prediction[25,26]. These results indicate that the ensemble classifier generalizes well and maintains robust discriminative ability despite fluctuations in the proportion of CN to AD-positive subjects in the training data. This level of stability across varying prevalence scenarios supports the model's potential for real-world deployment in diverse clinical settings.

Discussion

Our MRI-based ensemble model demonstrated a strong ability to predict preclinical Alzheimer's disease (AD) in asymptomatic individuals. Across 200 randomized person-wise experiments, the median AUC-ROC was approximately 0.951; at the prespecified operating point the model prioritized sensitivity (recall 0.923) while maintaining moderate precision (0.714), yielding an F1 score of 0.811. Such performance is notable given the challenging nature of identifying AD before clinical symptoms emerge. Several design choices underpin this strong performance. First, our model relies exclusively on structural magnetic resonance imaging (MRI), a widely available and non-invasive modality. We chose this approach to maximize scalability and equity. MRI machines are widely available, the scanning procedure is non-invasive, and using MRI does not involve the special tracers required for PET scans. This makes our approach practical even in settings that lack access to PET imaging or cerebrospinal fluid (CSF) testing. Second, the use of an ensemble, combining support vector machines (SVM), random forests (RF), and artificial neural networks (ANN), allowed our ensemble to harness the strengths of diverse classifiers (see Table 2), improving stability and generalization. Third, we enforced strict person-wise separation between training and testing data, thereby avoiding the data leakage concerns that have plagued earlier studies[13]. Under these rigorous conditions, our model still achieved high discrimination and sensitivity, indicating that it is capturing generalizable features related to disease progression rather than idiosyncrasies of individual scans.

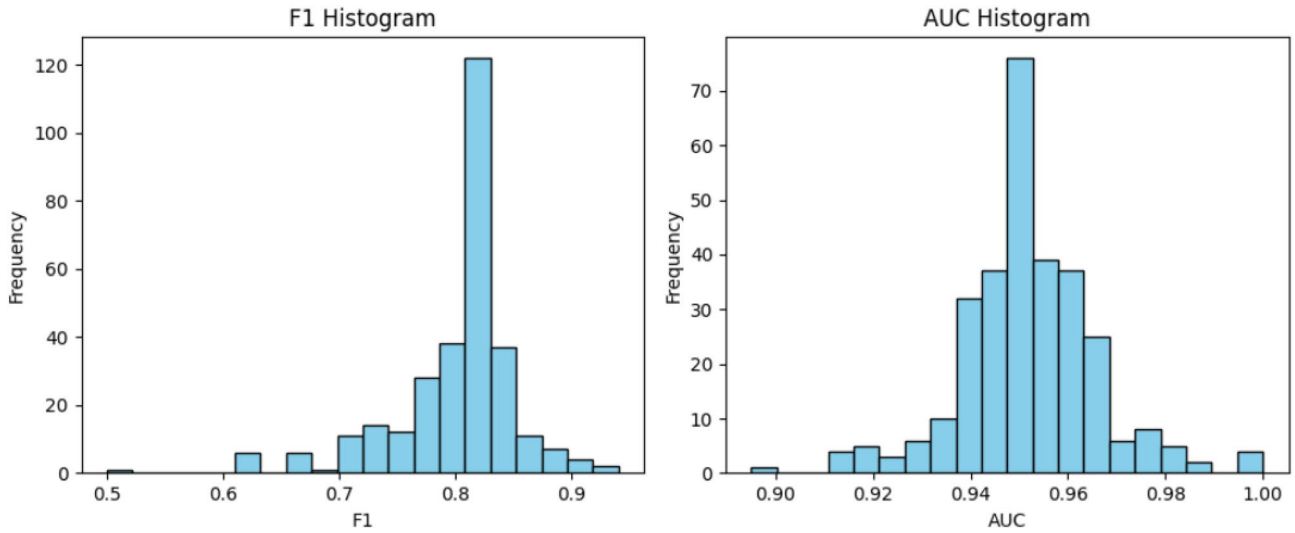


Fig. 9. Distribution of F1 scores (left) and AUC-ROC values (right) obtained across 200 randomized train/validation/test splits. The concentration of scores around their median values indicates stable performance and robust generalization of the ensemble classifier across repeated evaluations.

Model	Strengths	Weaknesses
ANN	• Captures complex nonlinear relationships • Learns hierarchical feature representations	• Requires large training datasets • Prone to overfitting without sufficient data • Difficult to interpret ("black-box")
SVM	• Effective in high-dimensional spaces • Robust with limited samples via kernel methods	• Sensitive to kernel and parameter choices • Not inherently multiclass (requires one-vs-rest schemes) • Computationally intensive on very large datasets
RF	• Ensemble of trees reduces overfitting • Provides feature importance measures • Handles noisy or missing data well	• Can be biased toward features with many levels • Depending on the data, RF may require many trees for high accuracy • Results remain complex to interpret

Table 2. Strengths and weaknesses of the ANN, SVM, and RF models.

Contextualized against recent work, our results compare favorably with models that either combine multiple modalities or target symptomatic endpoints. Castellano et al.[27] fused 3D T1-weighted MRI with amyloid PET on OASIS-3 and reported 95% accuracy for AD vs. CN with AUCs of 0.93 for fusion and 0.94 for MRI alone. Venugopalan et al.[28] demonstrated that multimodal deep learning (genetics and cognitive tests with MRI) improves accuracy, precision, recall, and F1 over unimodal baselines, though confined to symptomatic staging in ADNI. MRI-only symptomatic classifiers can report very high numbers: Şener et al.[29] achieved ∼98–99% accuracy on binary ADNI tasks using transfer-learned CNNs, and Zolfaghari et al.[30] reported 99.06% four-class accuracy with dual-CNN features and a classical-ML ensemble. These settings differ materially from our preclinical aim and often evaluate at the slice or symptomatic level, which is easier than forecasting conversion among currently asymptomatic individuals.

Several recent MRI-only pipelines illustrate this distinction. A deep pipeline combining improved DeepLabV3+ segmentation with an enhanced ResNeXt classifier reported 98.12% accuracy and AUC 0.97 for four-class symptomatic staging[31]. A vision-transformer model with inverted residual self-attention achieved 96.1% accuracy and 96.05% precision after augmentation[32]. A multiobjective evolutionary approach classified early MCI vs. CN at >93% accuracy on a small cohort[33]. By contrast, models emphasizing generalizability across cohorts tend to report more conservative values: Liu et al.[34] trained a 3D CNN on ADNI and validated on NACC, achieving AUC 0.851 for CN vs. (MCI+mild AD) and 0.624 for MCI detection. Under leakage-safe evaluation, our median AUC near 0.95 for preclinical forecasting suggests that structural MRI alone can support competitive early prediction when paired with compact representations and ensembling.

Multimodal and generated-data approaches can further raise headline metrics but at the cost of feasibility or interpretability. A hybrid AI system that fused structured clinical variables with an MRI module achieved 96.19% accuracy on ADNI MRI and 99.82% on NACC structured data[35]. A generative framework that simulates future MRIs with vision-transformer GANs and predicts CN to AD conversion via a 3D CNN reported accuracy 0.85 and F1 0.86 up to 10 years[36]. Graph-neural-network models that integrate structural MRI with comorbidity data from electronic health records achieved 0.98 accuracy for multi-class classification and 0.99/0.93/0.94 for AD vs. CN, AD vs. MCI, and CN vs. MCI, respectively[37]. While promising, these methods rely on additional modalities, synthetic data, or integrated EHRs that may not generalize to centers without such resources. Closer to our scope, Saoud and AlMarzouqi's ROI-wise vision-transformer ensemble on MRI alone reached 90%/0.91 (accuracy/AUC) for AD vs. CN, but only 84%/0.81 for CN vs. MCI and 74%/0.60 for AD vs. MCI[38]; Lee et al.[39] reported AUC 0.875 for MCI to AD conversion using gradient boosting with demographics, MRI features, and amyloid PET SUVr. Together these comparisons reinforce three points: symptomatic classification is substantially easier than preclinical forecasting; subject-wise and cross-cohort evaluation narrows headline metrics; and multimodal fusion can add signal but reduces practicality.

In the context of existing literature, these results compare favorably with work that either leverages multimodal inputs or addresses easier symptomatic endpoints. For example, Battineni et al.[40] reported an SVM (RBF) with F1 = 0.72 and sensitivity = 0.54 on MRI; our ensemble attains F1 = 0.811 and recall = 0.923 while tackling the harder task of predicting future conversion in asymptomatic individuals. Lebedev et al.[41] reached accuracy comparable to ours by combining MRI with genetic and demographic data, underscoring that structural MRI alone—when paired with appropriate representation learning and ensembling—can carry substantial predictive signal. Very high accuracies reported by recent ensemble deep-learning studies typically reflect symptomatic classification (e.g., AD vs. CN at diagnosis) rather than preclinical forecasting.

In particular, Fathi et al.[42] applied an ensemble of six CNN classifiers on ADNI data, reporting $\sim 98.6\% - \sim 99.8\%$ accuracy for binary tasks (e.g., NC vs. AD, LMCI vs. AD) and $\sim 94\%$ for four-class classification. Topsakal et al.[43] fine-tuned and stacked pretrained CNNs like ResNet50 and DenseNet201, achieving $\sim 95\%$ across precision, recall, F1, and accuracy. Mahanty et al.[44] used a snapshot ensemble of an enhanced Xception network and reached 99.14% accuracy in a four-class AD staging task. Similarly, Yue et al.[45] used ensemble feature-selection methods on a large longitudinal Chinese cohort, attaining 99.2% accuracy for AD vs. control. Detection in symptomatic indiviudals being a substantially easier task than forecasting conversion in asymptomatic individuals. Efforts closer to our objective, such as Qiu et al.[46], achieved approximately 80% accuracy up to four years pre-symptom onset but relied primarily on accuracy, which is sensitive to class imbalance, and provided limited detail on subject-wise separation. Under leakage-safe evaluation, our median AUC near 0.95 indicates that structural MRI alone can support competitive preclinical prediction.

Several design choices likely contributed to performance. First, relying exclusively on structural MRI emphasizes scalability and equity. MRI is non-invasive, widely available, and avoids the logistics of PET tracers or CSF collection. Second, transfer learning with VGG16 compresses five anatomically targeted slices into compact,

translation-tolerant descriptors via convolutional features and global average pooling, reducing overfitting risk while preserving salient neuroanatomy[21]. Third, combining heterogeneous learners (i.e., a margin-based SVM, a tree-based RF, and a neural ANN) stabilizes predictions under class imbalance and distributional variability, consistent with classic evidence for the benefits of diversity in ensembles[25,26]. Inference is efficient because feature extraction dominates compute and the classifier heads are lightweight, facilitating integration into routine workflows.

These findings have practical implications. They suggest that subtle structural changes present in routine MRI can indicate elevated future AD risk before clinical symptoms emerge, enabling imaging-led triage for confirmatory biomarkers, risk counseling, or trial referral. They also show that methodical controls (e.g., person-wise splitting, compact representations, and heterogeneous ensembling) can unlock robust performance from single-modality data. Beyond AD, the same pipeline could extend to other neurocognitive disorders where early MRI signatures are subtle; for example, Parkinson's disease remains diagnostically challenging and has been explored with MRI-based prediction[47].

Our model's high AUC, recall, and F1 score are therefore especially meaningful in context. Unlike many high-performing models in the literature, these metrics are achieved with MRI alone (i.e., without cognitive test scores, PET, or CSF data). Although studies such as Lebedev et al.[41] attained comparable accuracy by integrating additional biomarkers, our findings indicate that carefully extracted and aggregated structural features (e.g., those obtained via VGG16) can yield similar performance independently[21]. Taken together, these results suggest that routine MRI contains predictive signal relevant to future AD risk, that ensemble modeling is a robust strategy in data-limited or class-imbalanced settings, and that accessible imaging can feasibly support early-intervention pipelines.

Limitations and future research directions

While our findings are promising, several limitations should be acknowledged which concurrently offer promising avenues for future development of this work. Most notably, the model's performance is unlikely to improve significantly without the introduction of additional data, in particular scans of patients who transitioned from cognitively normal to AD-positive during the study. Given the consistency of performance across 200 randomized trials and the narrow variance in AUC-ROC, it appears the model has reached a performance plateau on this dataset. Accordingly, adding new data, especially from converters, is likely the most effective strategy for achieving further gains.

Another key challenge is label ambiguity. The dataset includes three partially overlapping variables: clinician diagnosis, Clinical Dementia Rating (CDR), and Mini-Mental State Examination (MMSE), that occasionally provide inconsistent assessments of cognitive state. While we constructed a composite metric to harmonize these labels and excluded uncertain cases, some residual misclassification may persist, introducing noise into both the training and evaluation sets. This label noise may partly limit model performance. If MRI scans were accompanied by positron emission tomography (PET) or cerebrospinal fluid (CSF) biomarkers, we could better evaluate whether the model identifies at-risk individuals more consistently than clinicians.

Our use of MRI-only data represents a trade-off between clinical feasibility and predictive richness. While this design improves accessibility and integration into standard workflows, it excludes potentially informative modalities such as PET, CSF, or genetic data. Similarly, reducing each scan to five representative 2D slices enhances efficiency but omits full volumetric and inter-slice context, which may carry subtle predictive signals. Each slice is processed independently, which further limits the model's ability to learn spatial patterns across orientations. Future research should explore full 3D CNNs, hybrid 2D–3D models, or sequence-aware architectures (e.g., temporal CNNs or RNNs) that can integrate cross-slice and longitudinal dynamics for richer feature extraction.

Our approach was developed and tested entirely within OASIS-3. Although multi-site in nature, this dataset may not fully capture the heterogeneity encountered in real-world clinical settings. Differences in scanner protocols, population demographics, and disease prevalence can affect model performance under domain shift. We did not conduct external validation or apply harmonization techniques across imaging sites, so generalizability remains an open question. While many prior studies report results solely within a single dataset, often without addressing data leakage or class imbalance, our use of strict person-wise stratification and repeated evaluation across 200 randomized splits offers a more conservative and realistic estimate of generalization. Nonetheless, as with most existing work, our model has not yet been externally validated on independent cohorts. Future studies should assess performance across datasets such as ADNI or NACC to further test robustness under domain shift.

A further challenge is that some patients may not exhibit signs of AD until after their last recorded scan. In such cases, model predictions labeled as false negatives may actually be correct, highlighting a limitation of fixed-timepoint evaluation and the need for temporally aware metrics. Longitudinal follow-up and time-to-event modeling could provide a more accurate assessment of predictive validity over time.

Additionally, the VGG16 model used for feature extraction was pretrained on natural images, which may limit its ability to capture subtle neuroanatomical features. Although transfer learning improves efficiency and reduces data requirements, domain-specific feature extractors, such as medical CNN backbones or self-supervised models trained directly on neuroimaging data, could yield better representations. Our ensemble used soft voting with equal weighting across classifiers; more sophisticated meta-learning or stacked ensembles might further enhance predictive accuracy.

Despite these limitations, our results underscore the feasibility of using structural MRI alone for early prediction of Alzheimer's disease, without the need for multimodal biomarkers or invasive procedures. By leveraging transfer learning and ensemble modeling, we demonstrate that subtle neuroanatomical patterns in cognitively normal individuals can yield predictive signal for future conversion to dementia. These findings offer a compelling foundation for developing accessible, non-invasive screening tools deployable in routine clinical workflows. Future efforts that integrate longitudinal modeling, richer imaging modalities, and external

validation will be essential to translate this potential into real-world impact. Nonetheless, this study provides a meaningful step toward scalable, early identification of individuals at risk for Alzheimer's disease, at a time when preventive interventions may be most effective.

Materials and methods

This section outlines the data sources, preprocessing steps, model architecture, training protocol, evaluation procedures, and statistical analyses used in this study. We begin by describing the dataset and participant selection criteria, followed by a detailed overview of the magnetic resonance imaging (MRI) preprocessing and feature extraction pipeline. We then present the classification models and ensemble strategy, along with the experimental design used to ensure rigorous and reproducible evaluation. Finally, we summarize the statistical metrics and validation techniques used to assess performance and robustness.

Computational requirements

Our workflow relies on standard Python libraries and hardware. In practice, one would need a graphical processing unit (GPU) for initial model training/feature extraction as well as for routine predictions. The memory requirements are modest (the final models and feature arrays fit in a few hundred MB). Clinically, our method would require integration of the MRI preprocessing pipeline (e.g., normalization, and slicing) into the imaging workflow, but no specialized hardware beyond a typical research server. In summary, aside from the usual image preprocessing steps, the computational demands of our ensemble are modest and well within the capabilities of modern clinical/research computing systems.

Data acquisition and preprocessing

We used the Open Access Series of Imaging Studies (OASIS-3) dataset[16]. In our local snapshot, we identified 1,093 unique participants and 7,588 MRI scans across 2,496 scan days. The dataset includes clinician diagnoses and cognitive screening test results (MMSE and CDR). MRI scans and associated metadata were systematically organized after download from the OASIS-3 repository, and a comprehensive data-cleaning pass was performed using NumPy (v2.4.2) to ensure consistency and quality. After quality control and feature extraction, 1,052 participants had at least two MRI scans that yielded valid features. For the preclinical prediction task, we restricted the analysis to participants who were cognitively normal (CN) at their first scan (*analysis cohort* $n = 733$): 659 remained CN throughout follow-up and 74 converted to AD-positive ("converters"). Participants already AD-positive at entry and cases with conflicting labels (e.g., MCI/uncertain based on diagnosis, CDR, and MMSE) were excluded from model development ($n = 360$).

All MRI volumes were in NIfTI format[18] and loaded using NiBabel (v5.3.3)[19]. To enhance computational efficiency while retaining anatomical coverage, each 3D scan was condensed from nine into five representative 2D slices: three axial (superior–inferior sampling), one mid-sagittal, and one coronal. Each selected slice was resized to 224×224 pixels using OpenCV (v4.13.0.92)[17]. We did not apply per-slice z-scoring or ImageNet mean/standard-deviation normalization; normalization was performed at the feature level (see below). This was done after testing slice combinations to optimize AUR-ROC.

Composite labeling and censoring

To address inconsistencies between clinician diagnosis, MMSE, and CDR, we constructed a composite metric to categorize subjects as CN, uncertain, or AD-positive. Only subjects with clear outcomes (CN or AD-positive) were retained for modeling. Converters were defined as participants who were CN at their first OASIS-3 scan and subsequently received an AD dementia diagnosis (i.e., converted to AD-positive). To reduce proximity-to-event leakage, we included only scans acquired while CN. Non-converters contributed CN scans through last follow-up.

Feature extraction via transfer learning

We used the convolutional backbone of VGG16[21] pre-trained on ImageNet for fixed feature extraction (`include_top=False`; weights frozen; no fine-tuning). Each single-channel MRI slice was resized to 200 $\times$ 200 pixels and converted to a three-channel pseudo-RGB image by duplicating the same intensity matrix across all three channels. Unlike pipelines that use the pooled output, we used the output of the final VGG16 convolutional backbone, which in our implementation yields a $6 \times 6 \times 512$ activation map per 200×200 slice. We then applied `GlobalAveragePooling2D` over the 6×6 grid to obtain an array of 512 real numbers per slice, and concatenated the five slice-level descriptors to form a single 2,560-dimensional vector (5×512) per scan. Before training, these vectors were scaled to $[0, 1]$ with scikit-learn's `MinMaxScaler` fit on the training data only and applied unchanged to validation and testing to avoid leakage. Only the downstream classifiers (ANN, SVM, RF) were trained; the convolutional features from VGG16 were fixed and not updated during model training. We did not apply Keras vgg16.preprocess_input; instead, normalization was performed at the feature-vector level after VGG16 extraction.

Base classifiers

Artificial neural network (ANN)

The artificial neural network (ANN) classifier was implemented in Keras with TensorFlow. Each input vector consisted of 2,560 features, obtained by concatenating the 512-dimensional global average pooled features from the five selected MRI slices. The ANN architecture comprised two fully connected hidden layers with rectified linear unit (ReLU) activation functions. The first hidden layer contained 5,120 units (twice the input dimensionality), followed by a dropout layer with a rate of 0.3. The second hidden layer contained 1,280 units

(half the input dimensionality), also followed by dropout with a rate of 0.3. The output layer consisted of a single neuron with a sigmoid activation function, producing a scalar probability for the AD-positive class.

Model hyperparameters were selected via grid search cross-validation over combinations of layer sizes, batch sizes $\{50, 100, 200\}$, and learning rates in the range $[10^{-5}, 10^{-3}]$. The final model was trained using the Adam optimizer with a fixed learning rate of 0.0001 and binary cross-entropy loss. Training was monitored using the area under the curve (AUC) on the validation set. Early stopping was applied with a patience of 10 epochs, based on validation AUC. Although a `ReduceLROnPlateau` callback (factor = 0.1, patience = 10) was defined during development, it was not used in the final configuration. Training was capped at 120 epochs unless early stopping was triggered. No batch normalization layers were used. Class imbalance during ANN training was handled through the training-set undersampling strategy described above, and validation and test sets were not resampled. Apart from the run-specific seed S_r used for data partitioning and undersampling, the ANN architecture and training configuration remained unchanged across runs.

Support vector machine (SVM)

The support vector machine (SVM) classifier was implemented using scikit-learn's SVC with a radial basis function (RBF) kernel. Hyperparameters were optimized via grid search cross-validation. We explored values of C from 1 to 1000 and γ from 10^{-4} to 0.1 (RBF kernel). The best performance was obtained at $C = 100$ and $\gamma = 0.01$. Probability estimates were enabled (probability=True) to support soft-voting in the ensemble. No SVM-specific class weighting was applied; class imbalance was addressed through the training-set undersampling strategy described above.

Random forest (RF)

The random forest (RF) classifier was implemented using scikit-learn. We tuned `n_estimators` (200, 500, 1000), max_depth {5, 10, None}, `min_samples_leaf` $\{3, 6, 12\}$, and `min_samples_split` $\{3, 6\}$, max_features {sqrt, log2, 0.5}, bootstrap {True, False}, and criterion {gini, entropy} via grid search cross-validation; the best configuration per run was selected by validation AUC.

Ensemble and operating point

We combined ANN, SVM, and RF via *equal-weight soft voting*: the three per-model AD-positive probabilities were averaged. A single decision threshold of 0.60, chosen during initial validation sweeps and then fixed a priori, was used across all runs (AD-positive if average probability > 0.60, otherwise CN).

Person-wise partitioning and primary evaluation protocol

To eliminate data leakage from longitudinal scans, all splits were performed at the *participant* level: all scans from a given individual were assigned to a single partition. The primary evaluation was comprised of 200 independent, subject-wise randomized runs with a 70/15/15 train/validation/test split. Within each run, hyperparameters were tuned using subject-wise 5-fold CV on the training set only. All fitted transforms (e.g., `MinMaxScaler`) and any resampling (below) were estimated on the training data and applied unchanged to validation and test sets.

Addressing class imbalance

Given the minority prevalence of converters, we evaluated cost-sensitive learning and SMOTE[14,15]. The final pipeline used *modest undersampling of CN* in the *training set only*, maintaining some imbalance to reflect clinical prevalence. Validation and test sets preserved natural class proportions and were never resampled.

Performance metrics and uncertainty quantification

Per run, we computed AUC-ROC, accuracy, precision, recall, and F1 using scikit-learn (`roc_auc_score`, `roc_curve`, `f1_score`, `accuracy_score`, `confusion_matrix`); plots were generated with Matplotlib (v3.5.3). Across the 200 runs, we report the *median* and *95% confidence intervals (CIs)* for the median via non-parametric bootstrap on the 200 run-level scores (1,000 resamples). Unless noted, headline figures (median AUC-ROC ≈ 0.951; accuracy ≈ 0.872; recall ≈ 0.923; F1 ≈ 0.811) refer to this repeated, subject-wise evaluation.

Sensitivity analysis: class-balance stability

To evaluate robustness to prevalence shifts, we performed an additional *class-balance stability* experiment in which a single held-out, subject-wise test set was fixed, while the CN:AD ratio in the training set was varied across 200 randomized runs (ratios sampled in $[0.5, 0.8]$). Evaluation on the fixed test set followed the same metrics; distributions and summary statistics are reported in the Results (see Fig. 9).

Randomness control and reproducibility

We fixed a master seed and derived per-run seeds to ensure reproducibility across the 200 runs. Let $S_0 = 42$; for run $r \in \{0, \ldots, 199\}$, $S_r = S_0 + r$. All randomized operations (dataset splitting, CV shuffling, undersampling, ANN training, and bootstrap resampling) used S_r. We set the Python hash seed and RNGs for `random`, NumPy, and TensorFlow/Keras, and enabled deterministic TensorFlow operations when available. Classical models/utilities used `random_state=`S_r (e.g., `train_test_split`, `StratifiedKFold`, `RandomUnderSampler`, `RandomForestClassifier`). The SVM with `probability=True` is deterministic given fixed data. Environment flags included `PYTHONHASHSEED`, `TF_DETERMINISTIC_OPS`, and `TF_CUDNN_DETERMINISTIC`.

Quality control, longitudinal handling, and harmonization

Automated QC excluded volumes with corrupted headers or failed feature extraction. Participants missing key cognitive fields (MMSE/CDR) or with conflicting composite labels were excluded; no imputation was performed. All scans from a participant were assigned to the same split; no participant contributed scans to more than one partition. We did not apply explicit inter-site or inter-scanner harmonization (e.g., ComBat) to the features, and we did not include age/sex covariates, to isolate the predictive value of structural MRI alone.

Software and computing

Analyses used Python v3.12.3 with NumPy (v2.4.2), SciPy (v1.17.1), NiBabel (v5.3.3), OpenCV (v4.13.0.92), scikit-learn (v1.8.0), TensorFlow (v2.20.0), Keras (v3.13.2), and Matplotlib (v3.10.8). Training and inference were conducted on a standard research workstation; memory requirements were modest because only $2{,}560$ -dimensional feature vectors per scan were used for classifier training.

Data availability

The MRI scan data and associated clinical and cognitive information that support the findings of this study are available from the OASIS-3 database. The dataset can be accessed via the following link: https://www.oasis-brains.org/

Code availability

The code used in this study is not publicly available at this time. All methodological details, fixed seeds, and determinism settings required to reproduce the analyses are described in the Materials and Methods.

Received: 11 October 2024; Accepted: 5 June 2026
Published online: 23 June 2026

References

1. DeTure, M. A. & Dickson, D. W. The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration* **14**, 32 (2019). doi:10.1186/s13024-019-0333-5
2. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the Global Burden of Disease Study 2019. *The Lancet Public Health* **7**(2), e105–e125 (2022). doi:10.1016/S2468-2667(21)00249-8
3. Centers for Disease Control and Prevention, National Center for Health Statistics. *National Vital Statistics System, Mortality* 2018–2021. CDC WONDER Online Database (2021). Available at: http://wonder.cdc.gov/ucd-icd10-expanded.html (Accessed: 11 October 2024).
4. Halliday, G. Pathology and hippocampal atrophy in Alzheimer's disease. *The Lancet Neurology* **16**(11), 862–864 (2017). doi:10.1016/S1474-4422(17)30343-5
5. Whitwell, J. L. Progression of atrophy in Alzheimer's disease and related disorders. *Neurotoxicity Research* **18**(3–4), 339–346 (2010). doi:10.1007/s12640-010-9175-1
6. Juganavar, A., Joshi, A. & Shegekar, T. Navigating early Alzheimer's diagnosis: A comprehensive review of diagnostic innovations. *Cureus* **15**(9), e44937 (2023). doi:10.7759/cureus.44937
7. Grossman, M. & Irwin, D. J. The mental status examination in patients with suspected dementia. *Continuum* **22**(2 Dementia), 385–403 (2016). doi:10.1212/CON.0000000000000298
8. Kothapalli, S. V. V. N. *et al.* Quantitative gradient echo MRI identifies dark matter as a new imaging biomarker of neurodegeneration that precedes tissue atrophy in early Alzheimer's disease. *Journal of Alzheimer's Disease* **85**(2), 905–924 (2022). doi:10.3233/JAD-210503
9. Basheera, S. & Sai Ram, M. S. Convolution neural network–based Alzheimer's disease classification using hybrid enhanced independent component analysis based segmented gray matter of T2-weighted magnetic resonance imaging with clinical valuation. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **5**, 974–986 (2019). doi:10.1016/j.trci.2019.10.001
10. Zhao, X. *et al.* A review of convolutional neural networks in computer vision. *Artificial Intelligence Review* **57**, 99 (2024). doi:10.1007/s10462-024-10721-6
11. Islam, J. & Zhang, Y. Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain Informatics* **5**, 2 (2018). https://doi.org/10.1186/s40708-018-0080-3
12. Nguyen, D., Nguyen, H., Ong, H., Le, H., Ha, H., Duc, N. T. & Ngo, H. T. Ensemble learning using traditional machine learning and deep neural network for diagnosis of Alzheimer's disease. *IBRO Neuroscience Reports* **13**, 255–263 (2022). doi:10.1016/j.ibneur.2022.08.010
13. Wen, J. *et al.* Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis* **63**, 101694 (2020). doi:10.1016/j.media.2020.101694
14. Pes, B. & Lai, G. Cost-sensitive learning strategies for high-dimensional and imbalanced data: A comparative study. *PeerJ Computer Science* **7**, e832 (2021). doi:10.7717/peerj-cs.832
15. Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H. & Ismail, M. SMOTE for handling imbalanced data problem: A review. In *2021 Sixth International Conference on Informatics and Computing (ICIC)*, 1–8 (2021). doi:10.1109/ICIC54025.2021.9632912
16. LaMontagne, P. *et al. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease*. medRxiv (2019). doi:10.1101/2019.12.13.19014902
17. Culjak, I., Abram, D., Pribanic, T., Dzapo, H. & Cifrek, M. A brief introduction to OpenCV. *2012 Proceedings of the 35th International Convention MIPRO* (2012). doi:10.1109/MIPRO.2012.6240859
18. Moore, C. *et al.* NIfTI (file format). *Radiopaedia* (2019). doi:10.53347/rID-72562
19. Halchenko, Y. O. et al. NiBabel: Neuroimaging in Python. *NiBabel documentation*. https://nipy.org/nibabel/ (accessed 11 October 2024).
20. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering* **13**(2), 22–30 (2011). doi:10.1109/MCSE.2011.37
21. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations* (2015). doi:10.48550/arXiv.1409.1556.
22. Mohammed, A. & Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University – Computer and Information Sciences* **35**(2), 757–774 (2023). doi:10.1016/j.jksuci.2023.01.014
23. Hicks, S. A. *et al.* On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports* **12**, 5979 (2022). doi:10.1038/s41598-022-09954-8

24. Ahsan, M. M., Luna, S. A. & Siddique, Z. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare* **10**(3), 541 (2022). doi:10.3390/healthcare10030541
25. Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*, Lecture Notes in Computer Science **1857**, 1–15. Springer, Berlin, Heidelberg (2000). doi:10.1007/3-540-45014-9_1
26. Opitz, D. & Maclin, R. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research* **11**, 169–198 (1999). doi:10.1613/jair.614
27. Castellano, G., Esposito, A., Lella, E., Montanaro, G. & Vessio, G. Automated detection of Alzheimer's disease: A multi-modal approach with 3D MRI and amyloid PET. *Scientific Reports* **14**, 5210 (2024). doi:10.1038/s41598-024-56001-9
28. Venugopalan, J., Tong, L., Hassanzadeh, H. R. & Wang, M. D. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports* **11**, 3254 (2021). doi:10.1038/s41598-020-74399-w
29. Şener, B., Acici, K. & Sümer, E. Categorization of Alzheimer's disease stages using deep learning approaches with McNemar's test. *PeerJ Computer Science* **10**, e1877 (2024). doi:10.7717/peerj-cs.1877
30. Zolfaghari, S., Joudaki, A. & Sarbaz, Y. A hybrid learning approach for MRI-based detection of Alzheimer's disease stages using dual CNNs and ensemble classifier. *Scientific Reports* **15**, 25342 (2025). doi:10.1038/s41598-025-11743-y
31. Vinukonda, E. R. & Jagadesh, B. N. An integrated deep learning model for early and multi-class diagnosis of Alzheimer's disease from MRI scans. *Scientific Reports* **15**, 17169 (2025). doi:10.1038/s41598-025-01845-y
32. Ibrar, W. *et al.* A novel interpreted deep network for Alzheimer's disease prediction based on inverted self attention and vision transformer. *Scientific Reports* **15**, 29974 (2025). doi:10.1038/s41598-025-15007-7
33. Zamani, J., Sadr, A. & Javadi, A.-H. Diagnosis of early mild cognitive impairment using a multiobjective optimization algorithm based on T1-MRI data. *Scientific Reports* **12**, 1020 (2022). doi:10.1038/s41598-022-04943-3
34. Liu, S. *et al.* Generalizable deep learning model for early Alzheimer's disease detection from structural MRIs. *Scientific Reports* **12**, 17106 (2022). doi:10.1038/s41598-022-20674-x
35. Mmadumbu, A. C., Saeed, F., Ghaleb, F. & Qasem, S. N. Early detection of Alzheimer's disease using deep learning methods. *Alzheimer's & Dementia* **21**(5), e70175 (2025). doi:10.1002/alz.70175
36. Aghaei, A. & Moghaddam, M. E. An integrated predictive model for Alzheimer's disease progression from cognitively normal subjects using generated MRI and interpretable AI. *Scientific Reports* **15**, 28340 (2025). doi:10.1038/s41598-025-13478-2
37. Abuhantash, F., Abu Hantash, M. K. & AlShehhi, A. Comorbidity-based framework for Alzheimer's disease classification using graph neural networks. *Scientific Reports* **14**, 21061 (2024). doi:10.1038/s41598-024-72321-2
38. Saoud, L. S. & AlMarzouqi, H. Explainable early detection of Alzheimer's disease using ROIs and an ensemble of 138 3D vision transformers. *Scientific Reports* **14**, 27756 (2024). doi:10.1038/s41598-024-76313-0
39. Lee, M.-W. *et al.* A multimodal machine learning model for predicting dementia conversion in Alzheimer's disease. *Scientific Reports* **14**, 12276 (2024). doi:10.1038/s41598-024-60134-2
40. Battineni, G., Chintalapudi, N. & Amenta, F. Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). *Informatics in Medicine Unlocked* **16**, 100200 (2019). doi:10.1016/j.imu.2019.100200
41. Lebedev, A. V. *et al.* Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical* **6**, 115–125 (2014). doi:10.1016/j.nicl.2014.08.023
42. Fathi, S., Ahmadi, A., Dehnad, A., Almasi-Dooghaee, M. & Sadegh, M. for the Alzheimer's Disease Neuroimaging Initiative. A deep learning-based ensemble method for early diagnosis of Alzheimer's disease using MRI images. *Neuroinformatics* **22**, 89–105 (2024). doi:10.1007/s12021-023-09646-2
43. Topsakal, O. & Lenkala, S. Enhancing Alzheimer's disease detection through ensemble learning of fine-tuned pre-trained neural networks. *Electronics* **13**(17), 3452 (2024). doi:10.3390/electronics13173452
44. Mahanty, C. *et al.* Effective Alzheimer's disease detection using enhanced Xception blending with snapshot ensemble. *Scientific Reports* **14**, 29263 (2024). doi:10.1038/s41598-024-80548-2
45. Yue, L., Chen, W.-G., Liu, S.-C., Chen, S.-B. & Xiao, S.-F. An explainable machine learning based prediction model for Alzheimer's disease in China Longitudinal Aging Study. *Frontiers in Aging Neuroscience* **15**, 1267020 (2023). doi:10.3389/fnagi.2023.1267020
46. Qiu, A., Xu, L., Liu, C. & Alzheimer's Disease Neuroimaging Initiative. Predicting diagnosis 4 years prior to Alzheimer's disease incident. NeuroImage: *Clinical* **34**, 102993 (2022). doi:10.1016/j.nicl.2022.102993
47. Heim, B., Krismer, F., De Marzi, R. & Seppi, K. Magnetic resonance imaging for the diagnosis of Parkinson's disease. *Journal of Neural Transmission* **124**, 915–964 (2017). doi:10.1007/s00702-017-1717-8

Author contributions

G.T., J.P., S.B.: Conceptualization, methodology, investigation. G.T.: Writing—original draft preparation. A.S., G.T.: Writing—review and editing. J.P.: Data curation, software, formal analysis, visualization, validation. D.Z., G.A., S.Z.: Writing—review and editing. A.S., S.B.: Supervision, resources. A.S.: Project administration. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study involved secondary analysis of de-identified data from the Open Access Series of Imaging Studies (OASIS-3) and complied with the dataset's data-use terms. The project was determined to be exempt from further review by the New York University Institutional Review Board (protocol IRB-FY2025-10272). All methods were carried out in accordance with relevant guidelines and regulations.

Additional information

Correspondence and requests for materials should be addressed to G.T. or A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026