

ORIGINAL RESEARCH ARTICLE

Validation strategies for automated MRI-based classification of Alzheimer's disease using deep feature extraction and machine learning

Leor Franco¹ and Milan Toma^{1*}

Department of Osteopathic Manipulative Medicine, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, New York, United States of America

Abstract

Alzheimer's disease (AD) is the most prevalent cause of dementia worldwide, yet early and accurate diagnosis remains a significant clinical challenge. This study systematically evaluates machine learning (ML) models for AD classification using a shared magnetic resonance imaging (MRI) dataset, focusing on both binary (e.g., healthy vs. demented) and multiclass (four-stage dementia) tasks. MRI images were preprocessed and deep features were extracted using a pretrained GoogLeNet architecture. Classification was performed using support vector machines for binary tasks and error-correcting output codes (ECOC) for multiclass tasks. Model performance was assessed using both hold-out and k-fold (5- and 10-fold) cross-validation (CV) strategies to ensure robust evaluation. Results indicate that CV yields substantially higher and more reliable accuracy than the hold-out method, with binary classification achieving up to 84% accuracy (10-fold CV) and multiclass classification reaching 87% (5-fold CV). The model demonstrated high specificity and precision, particularly for moderate-stage AD, but lower and less stable performance for early disease (very mild) cases. Learning curve analysis confirmed improved generalization with increased training data and minimal overfitting in well-validated models. However, our findings also reveal that high-performance metrics alone are insufficient to judge clinical utility. For example, a hold-out model distinguishing healthy from moderate Alzheimer's achieved a seemingly impressive 99% test accuracy, but learning curves revealed severe overfitting and likely data leakage, indicating that such results would not generalize to new patients. In contrast, the healthy vs. mild Alzheimer's task, with a more modest ~70% test accuracy, demonstrated well-behaved learning dynamics and genuine generalization. These results highlight that high reported accuracy is only clinically meaningful when supported by healthy training dynamics; otherwise, models risk underperforming in real-world clinical settings regardless of their reported metrics. We advocate for rigorous validation and learning curve analysis as prerequisites for any clinically actionable ML tool. These findings underscore the importance of transparent per-class performance reporting and robust validation to ensure that ML models can truly support early detection and staging of AD in clinical practice.

*Corresponding author:

Milan Toma
(mtoma@nyit.edu)

Citation: Franco L, Toma M. Validation strategies for automated MRI-based classification of Alzheimer's disease using deep feature extraction and machine learning. *Artif Intell Health*. 2026;3(2):025360073. doi: 10.36922/AIH025360073

Received: September 5, 2025**1st revised:** October 14, 2025**2nd revised:** October 25, 2025**Accepted:** October 29, 2025**Published online:** November 12, 2025

Copyright: © 2025 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: Alzheimer's disease; Magnetic resonance imaging; Machine learning; Deep learning; Dementia classification

1. Introduction

Alzheimer's disease (AD) is the leading cause of dementia, accounting for 60–80% of all cases globally, and its prevalence continues to rise as populations age.¹ The World Health Organization estimates that over 55 million people are currently living with dementia, a figure projected to nearly double every 20 years.¹ The etiology of AD is multifactorial, involving both genetic and environmental risk factors.² Pathologically, AD is characterized by the accumulation of amyloid plaques and neurofibrillary tangles, leading to progressive neuronal degeneration and cognitive decline, with short-term memory loss being the most widely recognized symptom.^{3,4} Despite extensive research, there is currently no cure or highly effective treatment for AD. However, early detection and intervention can slow disease progression and improve the quality of life for patients and caregivers.⁵

Traditional diagnostic approaches for AD rely on comprehensive clinical evaluation, including medical history, cognitive testing, and neuroimaging. Neuropsychological assessments such as the Mini-Mental State Examination (MMSE) and the AD Assessment Scale-Cognitive Subscale (ADAS-cog) are widely used to quantify cognitive decline.⁶ Neuroimaging modalities, particularly magnetic resonance imaging (MRI), functional MRI, and positron emission tomography (PET), play a pivotal role in identifying characteristic patterns of brain atrophy and metabolic changes associated with AD.^{7,8} MRI can reveal medial temporal lobe, hippocampal, and whole-brain atrophy, which are indicative of neurodegeneration in AD.^{8,9} Biomarker analysis, including amyloid burden in cerebrospinal fluid (CSF) and PET imaging, further aids in differentiating AD from other dementias.⁹ Electroencephalography has also demonstrated utility in supporting AD diagnosis, although it is less commonly used in clinical practice.^{10,11}

In recent years, machine learning (ML) has emerged as a transformative approach for enhancing the detection and diagnosis of AD, particularly through the analysis of neuroimaging data. ML methods such as support vector machines (SVMs), artificial neural networks, and ensemble-based learning approaches have demonstrated the ability to distinguish healthy individuals from those with AD with high accuracy, often exceeding 90% when applied to structural MRI data.^{12,13} Computer-aided diagnostic systems leveraging ML have been shown to outperform traditional radiological assessments by providing more objective and reproducible analyses of imaging data.¹⁴ However, most existing studies focus on maximizing classification accuracy for clear-cut cases, often at the expense of sensitivity in early or prodromal

stages of the disease. For example, deep learning models can achieve high accuracy in distinguishing AD from normal controls, but their performance declines when classifying earlier stages such as mild cognitive impairment.¹⁵ Many approaches rely on complex or engineered features (such as graph-theory metrics or deep learning representations) that may reduce clinical interpretability and generalizability across diverse populations and imaging protocols.^{1,16} Overemphasis on model complexity and accuracy can also lead to overfitting, limiting the generalizability of these models.¹

To address these limitations, alternative modeling strategies have been explored, including longitudinal analysis and the identification of subtle progression markers. Models trained on longitudinal MRI data have demonstrated improved sensitivity to preclinical changes, though even the best-performing algorithms exhibit reduced accuracy in distinguishing very early or prodromal AD from normal aging.¹⁷ SVMs, in particular, have shown promise in distinguishing AD from healthy controls, especially when using Gaussian RBF kernels and feature selection strategies based on SVM weight vectors.^{18,19} Deep learning approaches, including convolutional neural networks (CNNs) and hybrid CNN-recurrent neural network architectures, have further advanced the field by enabling automated feature extraction and classification from MRI and PET images.^{20–22} Recent innovations, such as depthwise separable convolutional networks, have reduced computational costs while maintaining competitive accuracy, making them suitable for deployment on resource-limited devices.²³

A persistent challenge in AD classification is the issue of class imbalance, particularly in datasets derived from clinical or public sources such as Kaggle and the AD Neuroimaging Initiative. Class imbalance can significantly impact the performance of ML algorithms, especially for underrepresented classes such as moderate dementia. Various strategies have been proposed to address this, including data augmentation, resampling, and synthetic oversampling techniques such as ADASYN and Borderline-SMOTE.^{24–29} These approaches have been shown to improve classification accuracy and robustness, particularly in multiclass settings. For instance, the DEMNET and DAD-Net architectures have achieved high accuracy and F1-scores by incorporating class balancing techniques and optimized neural network designs.^{24,26} while workflows such as Borderline-DEMNET have demonstrated superior performance in multiclass dementia staging.^{27,30,31} Additional studies have highlighted the importance of preprocessing, feature extraction, and normalization in enhancing model performance.^{32,33}

Despite these advances, the generalizability and clinical utility of ML models for AD diagnosis remain contingent on rigorous validation and careful consideration of dataset characteristics. The present study systematically evaluates the performance of ML models for AD classification using a shared MRI dataset, employing both binary and multiclass classification tasks. The effectiveness of hold-out and k-fold cross-validation (CV) strategies is compared to provide a comprehensive assessment of model generalizability and reliability. By addressing the limitations of prior work and emphasizing robust validation, this study aims to advance the development of clinically applicable ML tools for the early detection and staging of AD.

While a growing number of studies report high performance of automated ML models for AD classification, many do so without thoroughly analyzing or demonstrating healthy training dynamics, which is an essential aspect for ensuring true generalizability in clinical environments. Notably, inflated accuracy metrics are often reported without adequate validation, potentially overstating clinical impact. As a clinically oriented research group, our work emphasizes not only the technical performance of ML models but also their clinical actionability. Our evaluation incorporates detailed scrutiny of learning curves, per-class performance, and overfitting, providing a transparent account of model generalization. By prioritizing clinically interpretable results and validation over inflated metrics, this study advocates for ML approaches that are directly translatable to patient care. We believe that the true value of automated diagnostic systems lies in their reliability and reproducibility in real-world settings, which we address through detailed methodological transparency and clinically relevant performance analyses.

2. Data and methods

The overall experimental workflow is illustrated in [Figure 1](#). All experiments began with a shared MRI dataset, which was subjected to standardized preprocessing and feature extraction using a pretrained convolutional neural network. The resulting features were then used for two types of classification tasks: binary (*e.g.*, distinguishing between non-demented and demented subjects) and multiclass (distinguishing among multiple stages of dementia).

To evaluate model performance, we employed two complementary validation strategies: hold-out and k-fold CV. In the hold-out approach, the dataset was randomly split into separate training, validation, and test sets (70%, 15%, and 15% of the data, respectively). The model was trained on the training set, tuned on the validation set, and its final performance was assessed on the test set, which remained completely unseen during training and tuning. This method provides a straightforward estimate

of generalization performance but can be sensitive to the particular data split, especially with limited data.

In contrast, k-fold CV involved dividing the entire dataset into k equally sized folds (with $k = 5$ or $k = 10$). The model was trained and validated k times, each time using a different fold as the validation set and the remaining $k - 1$ folds for training. Every sample is used for validation exactly once, and for training $k - 1$ times. The final performance is averaged over all k runs, providing a more robust and reliable estimate of model generalization and reducing the impact of any particular data split. Both validation strategies were applied to each classification task to ensure comprehensive and fair performance assessment.

While both hold-out and k-fold CV are widely used for model evaluation, each approach has its own strengths and limitations, and the choice between them depends on the specific context and dataset. The hold-out method is straightforward and computationally efficient, as it involves splitting the data once into separate training, validation, and test sets. This approach is particularly useful when the dataset is large, as it allows for a substantial portion of data to be reserved for unbiased final testing. However, the results can be sensitive to how the data is split, and a single split may not fully capture the variability in model performance, especially with smaller datasets. In contrast, k-fold CV provides a more robust estimate of generalization by averaging performance across multiple train/validation splits, making efficient use of all available data for both training and validation. Nevertheless, k-fold CV can be computationally intensive and does not provide a truly independent test set unless one is held out separately. Ultimately, neither method is inherently superior; rather, they offer complementary perspectives on model performance, and using both can provide a more comprehensive assessment of a model's reliability and generalizability ([Table 1](#)).

2.1. Dataset organization and labeling

The MRI dataset was organized into class-specific subfolders: NonDemented, VeryMildDemented, MildDemented, and ModerateDemented. For binary classification, the VeryMild, Mild, and Moderate classes were combined into a single "Demented" category, while multiclass classification retained all four original categories. An image datastore was created with folder-based label assignment, and the data were shuffled randomly to ensure unbiased sampling. The full data preparation workflow can be found in the algorithm presented in [Appendix A](#). In total, we utilized over 30,000 MRI images (specifically, 33,984), all obtained from an online database under a General Public License, to ensure compliance with open data standards and facilitate reproducibility of our results.³⁴

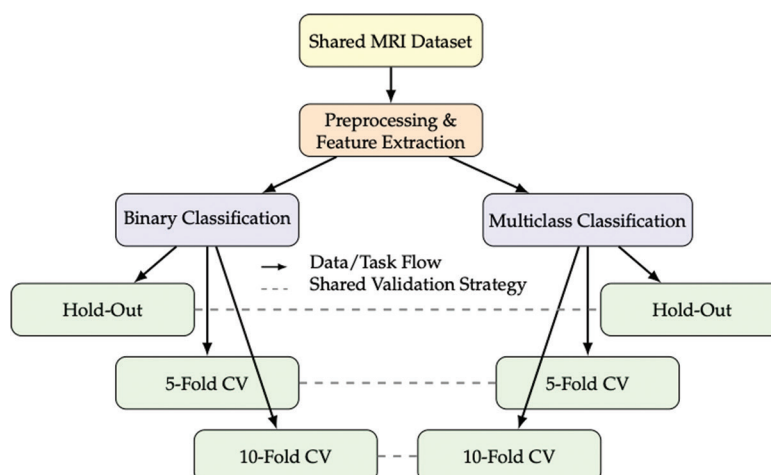


Figure 1. Overview of the experimental pipeline. The shared magnetic resonance imaging (MRI) dataset is preprocessed and features are extracted, then used for both binary and multiclass classification tasks. Each task is evaluated using two main validation strategies: hold-out, where the data is split into separate training, validation, and test sets; and k-fold cross-validation (CV), where the data is divided into k folds and each fold is used once as a validation set. Note: Dashed lines indicate that the same validation strategies are applied to both tasks.

Table 1. Comparison of hold-out and k-fold cross-validation strategies used for model evaluation

Validation strategy	Data usage	Advantages	Disadvantages
Hold-out	Single split into training, validation, and test sets (e.g., 70%/15%/15%)	Simple, fast, provides an independent test set for unbiased final evaluation	Performance can vary depending on the random split; less reliable with small datasets
k-Fold cross-validation	All data used for both training and validation (rotating roles across k folds)	More robust estimate of generalization; reduces variance from any single split; maximizes data usage for training	Computationally intensive; does not provide a truly independent test set unless one is held out separately

Notes: Each approach offers distinct advantages and limitations, and the choice depends on dataset size, computational resources, and the need for unbiased final testing.

The MRI images were sourced from the publicly available “Augmented Alzheimer MRI dataset” hosted online and distributed under a General Public License. All images were used as provided, with no additional manual curation or exclusion beyond the standardized preprocessing pipeline implemented by the dataset authors. As we did not perform any further preprocessing ourselves, we rely on the expertise and quality control of the dataset curators, and assume (based on the dataset’s documentation and its acceptance in the research community) that the images have been handled appropriately and consistently. This approach ensures full reproducibility and compliance with open data standards, while also representing a limitation in our direct control over preprocessing choices. For further details on the specific preprocessing protocols and quality metrics, we refer readers to the original dataset documentation.

Class labels (NonDemented, VeryMildDemented, MildDemented, ModerateDemented) were assigned by the dataset creators based on clinical diagnosis, with reference to cognitive assessments such as the MMSE. The precise diagnostic workflow and any use of additional biomarkers

(e.g., CSF, PET) are as described in the original dataset documentation. We did not perform any relabeling or exclusion of scans for artifacts; all images provided in the dataset were included following preprocessing.

After preprocessing, the final class counts were as follows: NonDemented: 9,600 images; VeryMildDemented: 8,960 images; MildDemented: 8,960 images; ModerateDemented: 6,464 images. All subsequent analyses and partitioning were performed using these preprocessed image counts.

2.2. Data partitioning and validation strategies

To evaluate model performance, we employed two complementary validation strategies: hold-out and k-fold CV. In the hold-out approach, the dataset was randomly split into separate training, validation, and test sets (70%, 15%, and 15% of the data, respectively), with stratification to preserve class proportions. In k-fold CV ($k = 5$ or $k = 10$), the data were divided into k equally sized, stratified folds. Each fold served as a validation set once, with the remaining k-1 folds used for training. Both strategies were applied to each classification task to ensure comprehensive

and fair performance assessment. The rationale and comparison of these strategies are summarized in Table 1.

To ensure subject-level independence and prevent data leakage, all data partitioning was performed at the subject level. Specifically, each subject's images are contained within a unique subfolder, and the dataset was split by allocating entire subject folders to either the training, validation, or test set (for hold-out) or to specific folds (for CV). This guarantees that no images from the same subject appear in more than one partition. The partitioning procedure was implemented by generating a list of unique subject identifiers, shuffling this list, and assigning subjects to partitions according to the specified ratios. The corresponding pseudocode is provided in Appendix A.

2.3. Image preprocessing and augmentation

All MRI images were resized to $224 \times 224 \times 3$ pixels to match the input requirements of the pretrained convolutional neural network. Grayscale images were converted to RGB format. Augmented image datastores were created to ensure consistent preprocessing and to increase the effective size and diversity of the training data, thereby reducing overfitting and improving generalizability. The preprocessing steps are described in the algorithm found in Appendix A.

2.4. Deep feature extraction

Feature extraction was performed using a pretrained GoogLeNet architecture. Images were passed through the network, and activations from the 'pool5-drop_7x7_s1' layer were extracted as feature vectors. These deep features, representing high-level abstractions of the input images, were output as row vectors and served as input for subsequent classification tasks. The feature extraction process is outlined in the algorithm provided in Appendix A.

The 'pool5-drop_7x7_s1' layer of GoogLeNet was chosen for feature extraction as it provides a global average pooling of the final convolutional feature maps, resulting in a compact and semantically rich representation. This layer is commonly used in transfer learning applications due to its ability to summarize high-level image features while reducing spatial redundancy. Each image yields a 1,024-dimensional feature vector after flattening, which serves as input to the downstream classifiers.

Before classifier training, all feature vectors were standardized to zero mean and unit variance. Standardization parameters were computed using only the training data within each fold or split, and these parameters were subsequently applied to the validation and test sets. This procedure prevents information leakage and ensures

unbiased evaluation.

2.5. Classifier training

For binary classification (Normal vs. Demented), an SVM was trained using the `fitcsvm` function. For multiclass classification (four classes), an ECOC classifier was trained using the `fitcecoc` function. Both classifiers were trained on the extracted feature vectors, with hyperparameters tuned using the validation set (in hold-out) or within each fold (in CV). The classifier training process is described in the algorithm shown in Appendix B.

For binary classification, an SVM with a radial basis function (RBF) kernel was used. Hyperparameters (C and γ) were optimized through grid search within the training set or within each CV fold, with $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{0.001, 0.01, 0.1, 1\}$. Class weights were set to 'balanced' to address class imbalance. For multiclass classification, an ECOC classifier with one-vs-one coding was employed, aggregating binary SVM decisions through majority voting. Hyperparameters for each binary SVM were tuned as described above within each fold.

2.6. Model evaluation and performance metrics

Model predictions were generated for the training, validation, and test sets (or for each fold in CV). Confusion matrices were computed for each evaluation set to visualize the distribution of correct and incorrect predictions across classes. Key performance metrics (*i.e.*, including accuracy, precision, recall [sensitivity], specificity, F1-score, negative predictive value [NPV], false-positive rate [FPR], and false-negative rate [FNR]) were calculated to provide a comprehensive assessment of model performance. For the CV, metrics were averaged across all folds to obtain robust estimates. The evaluation and metric computation steps are detailed in the algorithm presented in Appendix B.

2.7. Learning curves analysis

To assess the relationship between training set size and model performance, learning curves were generated. The training set was subsampled at various proportions, and the classifier was retrained and evaluated at each size. Training and validation accuracies, as well as loss (defined as $1 - \text{Accuracy}$), were computed and plotted as a function of training set size. This analysis provided insight into model generalization, overfitting, and the potential benefits of additional data. The learning curve analysis is described in the algorithm found in Appendix B.

2.8. CV workflow and results aggregation

For k -fold CV, the entire pipeline (*i.e.*, feature extraction, classifier training, and evaluation) was repeated for each

fold. Performance metrics and confusion matrices were aggregated across folds to compute mean and standard deviation values, providing a robust estimate of model generalizability. Results were visualized using learning curves, loss curves, and confusion charts, and final performance metrics were reported for both binary and multiclass tasks. The CV and results aggregation process is described in the algorithm shown in Appendix C.

2.9. Algorithmic details and reproducibility

For full transparency and reproducibility, the step-by-step algorithms for data preparation, model training, and evaluation, and CV is provided in the appendices (see Appendices A, B, and C). These pseudocode listings, placed in the appendices, detail the precise sequence of operations, including dataset organization, preprocessing, feature extraction, classifier training, evaluation, and results visualization.

3. Results

This section presents the outcomes of our classification experiments, organized to reflect both the structure of the tasks and the evaluation strategies employed. Results are first reported for binary classification, where models distinguish between two classes, followed by multiclass classification, which involves differentiating among four distinct categories. For each task, we systematically evaluate model performance using three validation approaches: hold-out validation, 5-fold CV, and 10-fold CV. This organization allows for a clear comparison of how each validation strategy impacts classification accuracy and generalizability. Finally, we provide a comparative analysis of the validation strategies themselves, highlighting their respective strengths and limitations as observed in our experiments.

3.1. Binary classification results

In the following subsections, binary classification results

are presented for all possible pairwise comparisons between healthy controls and each stage of AD using the hold-out validation strategy, whereas for both 5-fold and 10-fold CV, only the Healthy vs. Alzheimer's classification is evaluated, enabling a focused comparison of model performance across different validation approaches.

3.1.1. Hold-out validation

Figure 2 presents the confusion matrices for the training, validation, and test sets when classifying between healthy controls and individuals with AD using the hold-out validation strategy. The matrices illustrate the proportion of correct and incorrect predictions for each class, providing a visual summary of the model's discriminative ability. Table 2 summarizes key performance metrics, including accuracy, precision, recall (sensitivity), specificity, F1-score, NPV, FPR, and FNR across all data splits. Clinically, these results reflect the model's moderate ability to distinguish AD from healthy aging, with performance metrics only slightly above random chance, indicating substantial overlap in the extracted features between these two groups.

Figure 3 displays the confusion matrices for the binary classification task of distinguishing healthy individuals from those with very mild cognitive impairment. Table 3 provides the corresponding performance metrics. The results indicate that the model struggles to differentiate between healthy and very mildly impaired individuals, as evidenced by accuracy and other metrics close to 0.55, which is near random classification. Clinically, this suggests that the features used by the model do not capture subtle differences between normal aging and the earliest detectable cognitive changes.

Figure 4 shows the confusion matrices for the task of classifying healthy controls vs. those with mild cognitive impairment. The associated metrics in Table 4 demonstrate improved performance compared to the previous tasks, with accuracy approaching 0.70. This indicates that the

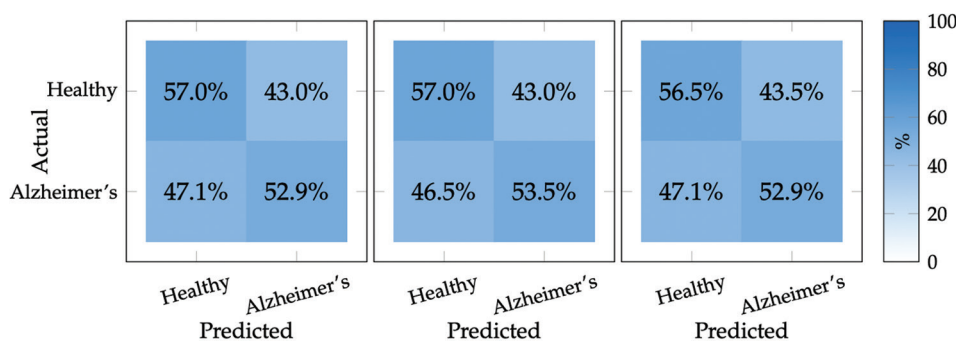


Figure 2. Confusion matrices for training, validation, and test sets in binary classification (Healthy vs. Alzheimer's) using the hold-out validation strategy

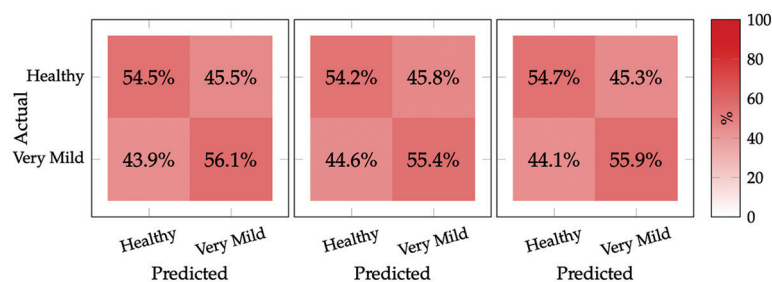


Figure 3. Confusion matrices for training, validation, and test sets in binary classification (Healthy vs. Very Mild Alzheimer's) using the hold-out validation strategy

Table 2. Performance metrics calculated from confusion matrices for Healthy vs. Alzheimer's classification across training, validation, and test sets

Metric	Training set	Validation set	Test set
Accuracy	0.5585	0.5600	0.5545
Precision	0.5701	0.5701	0.5647
Recall (Sensitivity)	0.7546	0.7568	0.7527
Specificity	0.3265	0.3288	0.3237
F1-Score	0.6497	0.6505	0.6451
NPV	0.5294	0.5347	0.5292
FPR	0.6735	0.6712	0.6763
FNR	0.2454	0.2432	0.2473

Abbreviations: FNR: False-negative rate; FPR: False-positive rate; NPV: Negative predictive value.

Table 3. Performance metrics calculated from confusion matrices for Healthy vs. Very Mild (Alzheimer's) classification across training, validation, and test sets

Metric	Training set	Validation set	Test set
Accuracy	0.5527	0.5477	0.5524
Precision	0.5708	0.5655	0.5703
Recall (Sensitivity)	0.5452	0.5424	0.5465
Specificity	0.5607	0.5536	0.5588
F1-Score	0.5577	0.5538	0.5582
NPV	0.5351	0.5303	0.5350
FPR	0.4393	0.4464	0.4412
FNR	0.4548	0.4576	0.4535

Abbreviations: FNR: False-negative rate; FPR: False-positive rate; NPV: Negative predictive value.

model is more effective at identifying individuals with mild cognitive impairment, which may reflect more pronounced clinical and neuroimaging differences at this stage.

Figure 5 presents the confusion matrices for distinguishing healthy individuals from those with moderate cognitive impairment. The performance metrics in Table 5 reveal near-perfect classification, with accuracy, precision,

Table 4. Performance metrics calculated from confusion matrices for Healthy vs. Mild (Alzheimer's) classification across training, validation, and test sets

Metric	Training set	Validation set	Test set
Accuracy	0.7142	0.7037	0.6969
Precision	0.7070	0.6970	0.6741
Recall (Sensitivity)	0.7026	0.6916	0.6905
Specificity	0.7250	0.7151	0.7025
F1-Score	0.7048	0.6943	0.6822
NPV	0.7207	0.7097	0.7181
FPR	0.2750	0.2849	0.2975
FNR	0.2974	0.3084	0.3095

Abbreviations: FNR: False-negative rate; FPR: False-positive rate; NPV: Negative predictive value.

recall, and specificity all approaching or exceeding 0.99. Clinically, this suggests that moderate cognitive impairment is associated with distinct features that are readily captured by the model, resulting in highly reliable discrimination. However, such high performance more likely indicates overfitting or data leakage. To confirm whether overfitting or data leakage is present, the learning curves shown in Figure 6 are examined: if the training accuracy remains perfect while the validation accuracy rapidly approaches similarly high values, this would be a classic sign of overfitting or data leakage rather than true generalization.

Figure 6 provides a comparative analysis of the learning curves for all four binary classification tasks. The figure plots training and validation accuracy as a function of training progression, highlighting the convergence behavior and generalization capacity of the model for each clinical comparison. Examining the learning curves in Figure 6 with respect to healthy learning patterns where training and validation accuracies should converge and plateau together, the four binary classification tasks exhibit markedly different behaviors. The Healthy vs. Mild classification (green curves) demonstrates the best learning behavior of all pairs, with training accuracy decreasing from 95% to approximately 71% while validation accuracy

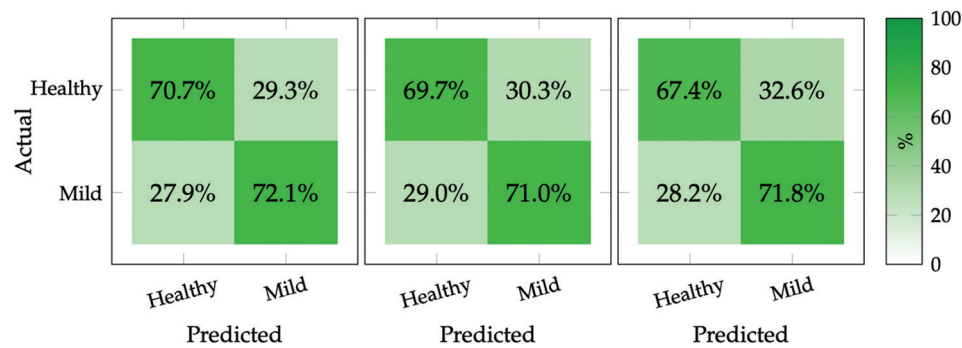


Figure 4. Confusion matrices for training, validation, and test sets in binary classification (Healthy vs. Mild Alzheimer's) using the hold-out validation strategy

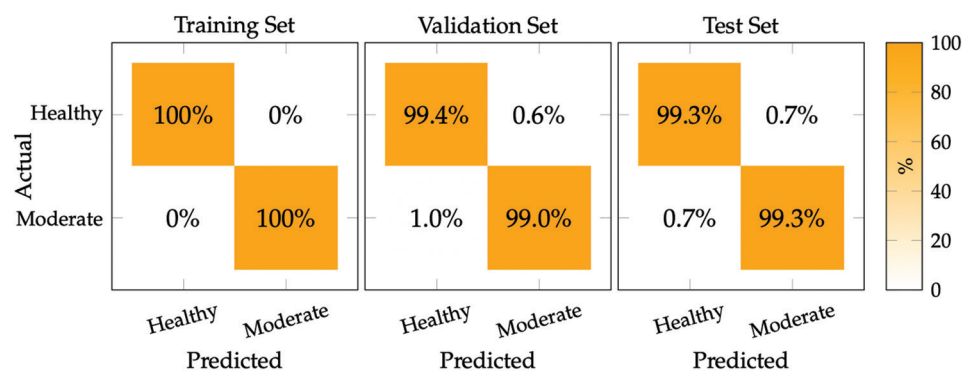


Figure 5. Confusion matrices for training, validation, and test sets in binary classification (Healthy vs. Moderate Alzheimer's) using the hold-out validation strategy

Table 5. Performance metrics calculated from confusion matrices for Healthy vs. Moderate (Alzheimer's) classification across training, validation, and test sets

Metric	Training set	Validation set	Test set
Accuracy	1.0000	0.9917	0.9929
Precision	1.0000	0.9857	0.9897
Recall (Sensitivity)	1.0000	0.9938	0.9928
Specificity	1.0000	0.9903	0.9931
F1-Score	1.0000	0.9897	0.9912
NPV	1.0000	0.9958	0.9951
FPR	0.0000	0.0097	0.0069
FNR	0.0000	0.0062	0.0072

Abbreviations: FNR: False-negative rate; FPR: False-positive rate; NPV: Negative predictive value.

rises from 55% to 70%, resulting in excellent convergence with only a ~1% gap and both curves nearly plateauing in the final stages. This represents strong model performance at ~70% accuracy with good generalization, indicating the model successfully learns meaningful distinguishing features between healthy individuals and those with mild cognitive impairment.

In contrast, the Healthy vs. Alzheimer's classification (blue curves) shows convergence but poor performance, with training accuracy declining from 90% to 56% while validation accuracy rises from 45% to 56%. Although the curves converge well to essentially identical values (~56%), this performance is barely above random chance (50%), indicating the model fails to learn meaningful distinguishing features for this classification task.

The healthy vs. very mild classification (red curves) exhibits similar convergence issues, with training accuracy declining from 85% to 55% while validation accuracy rises gradually from 52% to 55%, creating convergence at 55%. This ~55% performance represents essentially random classification, suggesting the model cannot effectively distinguish between healthy individuals and those with very mild cognitive impairment.

Most problematic is the Healthy vs. Moderate classification (brown curves), which exhibits classic signs of severe overfitting or data leakage, characterized by training accuracy remaining essentially perfect at 99.9% throughout the entire training process while validation accuracy jumps dramatically from 68% to 99% in the early stages. This abnormal pattern indicates the model

is likely memorizing training data rather than learning generalizable patterns.

To further illustrate the relationship between model performance on the training, validation, and test sets, Figure 7 presents the classification accuracy for each binary task under the hold-out validation strategy. In a well-conducted ML experiment, accuracy is expected to be highest on the training set, slightly lower on the validation set, and lowest on the test set. This trend reflects the model's decreasing familiarity with the data: the training set is directly used for model fitting, the validation set is used for hyperparameter tuning and is not seen during training, and the test set remains completely unseen until final

evaluation. As such, the test set provides the most unbiased estimate of generalization performance. If the validation or test accuracy exceeds the training accuracy, this may suggest data leakage, overfitting, or an unrepresentative data split.

As illustrated in Figure 7, only the Healthy vs. Mild classification task exhibits the expected trend of decreasing accuracy from the training set to the validation set and then to the test set, reflecting a “well-behaved” model with proper generalization. For this task, the accuracy drops incrementally from 71.4% (training) to 70.4% (validation) and 69.7% (test), consistent with best practices in ML and the convergence observed in the learning curves in

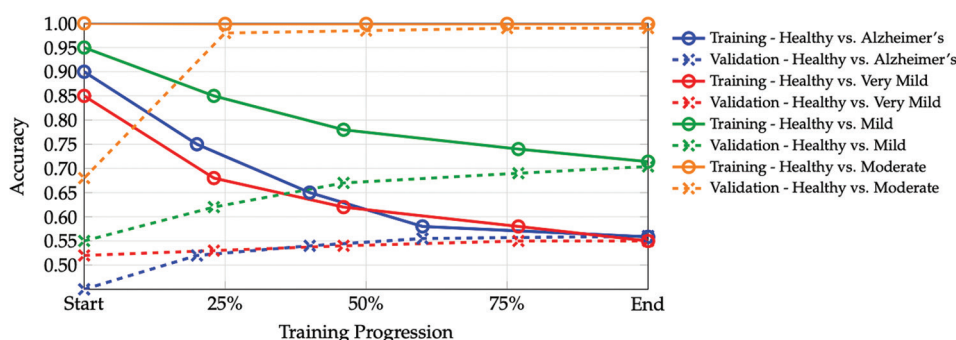


Figure 6. Learning curves comparing training and validation accuracy for binary classification tasks using hold-out validation: distinguishing healthy individuals from Alzheimer's disease (AD) and from different severity levels of cognitive impairment (Very Mild, Mild, and Moderate)

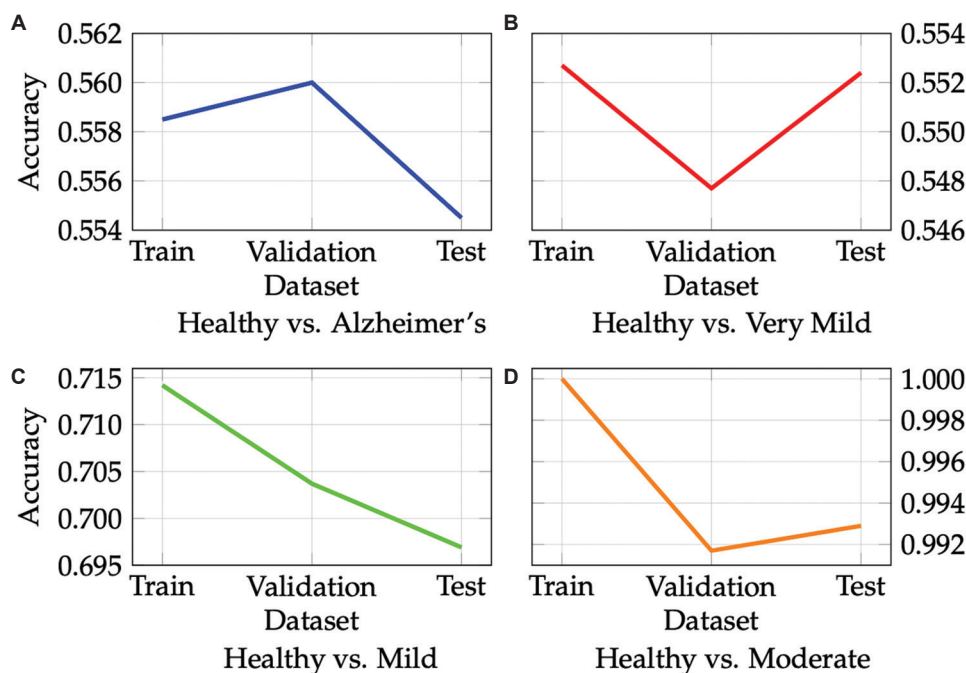


Figure 7. Comparison of classification accuracy across training, validation, and test sets for each binary classification task using the hold-out validation strategy: (A) healthy vs. Alzheimer's, (B) healthy vs. very mild, (C) healthy vs. mild, and (D) healthy vs. moderate. For each task, accuracy is expected to decrease from training to validation to the test set, reflecting the model's progressively reduced access to the data and its ability to generalize. Anomalies, such as higher accuracy on the validation or test set compared to the training set, may indicate issues with data leakage or overfitting.

Figure 6. In contrast, the other binary tasks deviate from this ideal pattern. For Healthy vs. Alzheimer's and Healthy vs. Very Mild, the validation or test accuracies are similar to or even slightly higher than the training accuracy, which is unexpected and may be attributed to the model's limited ability to distinguish between these groups, or sampling variability near random chance. Most notably, the Healthy vs. Moderate comparison shows near-perfect accuracy across all splits (training: 100%, validation: 99.2%, test: 99.3%), suggesting overfitting or possible data leakage. This is further supported by the corresponding learning curves, which display classic signs of memorization rather than true generalization. These findings underscore the importance of closely monitoring accuracy trends across data splits and highlight potential risks of overfitting or data leakage when validation and test accuracies approach or exceed training performance. Careful data partitioning and rigorous evaluation are essential to ensure reliable assessment of model generalizability in clinical ML applications.

3.1.2. Five-fold CV

The performance of the binary classification model distinguishing between healthy controls and individuals with AD was also evaluated using 5-fold CV. The results

are summarized in three key figures, each providing complementary insights into the model's behavior and clinical utility.

Figure 8 presents the confusion matrices for each of the five CV folds. Each matrix displays the row-wise percentage of predictions for the two classes, *i.e.*, Alzheimer's and Healthy, allowing for a granular assessment of the model's discriminative ability across different data splits. Across all folds, the model consistently demonstrates high specificity, with over 92% of healthy individuals correctly identified as non-demented (bottom right cell in each matrix). Sensitivity for AD is moderate, with approximately 62–66% of Alzheimer's cases correctly classified (top left cell), while the remaining cases are misclassified as healthy. The low rate of false positives (healthy individuals misclassified as Alzheimer's) and the relatively higher rate of false negatives (Alzheimer's cases missed) indicate a conservative classification strategy, prioritizing the avoidance of false alarms over the risk of missed diagnoses. Clinically, this pattern suggests that while the model is reliable in ruling out AD in healthy individuals, it may under-detect true cases, which is a common trade-off in screening tools where specificity is prioritized to minimize unnecessary anxiety or interventions in healthy populations. The

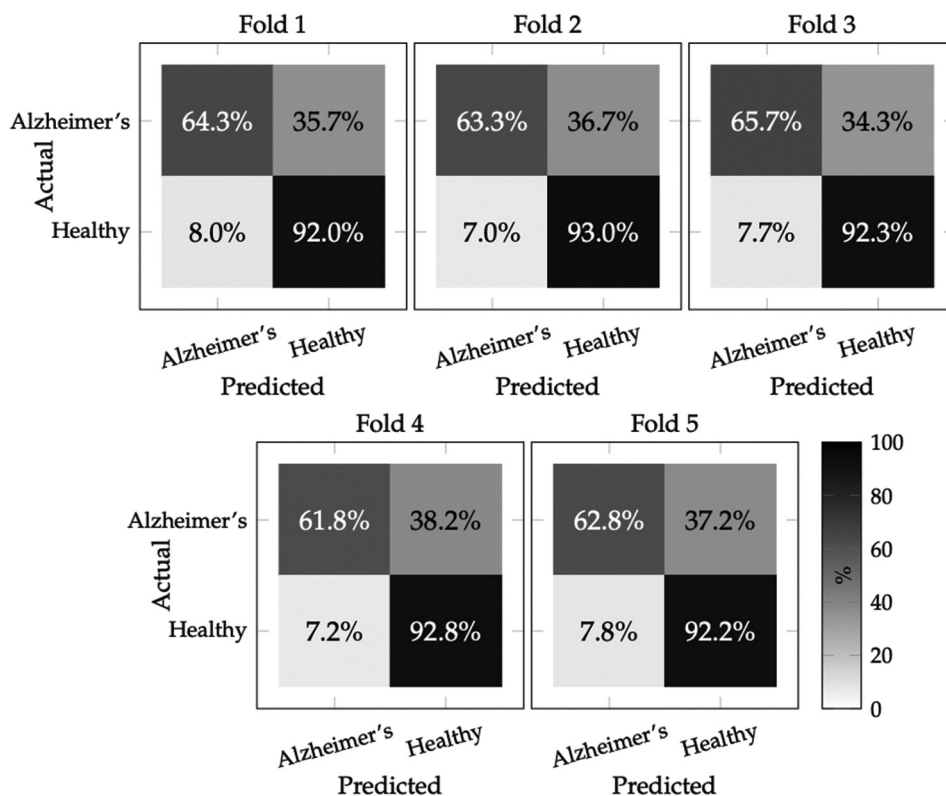


Figure 8. Confusion matrices for all five folds in 5-fold cross-validation for binary classification (Healthy vs. Alzheimer's). Each matrix shows the row-wise percentage of predictions in each category, demonstrating consistent performance across folds with accuracy around 84–85%.

consistency of these results across all folds underscores the robustness of the model's performance and suggests that the observed trends are not artifacts of a particular data split, but rather reflect the underlying discriminative capacity of the extracted features.

To further quantify model performance, Figure 9 displays the variation of key evaluation metrics (*i.e.*, accuracy, precision, recall [sensitivity], specificity, and F1-score) across the five folds. Each subplot tracks the metric's value for each fold, while the final panel summarizes the mean and standard deviation for each metric. The model achieves a mean accuracy of approximately 84%, with low variability across folds, indicating stable generalization. Precision remains high (mean 77%), reflecting the model's ability to avoid false positives when predicting AD. However, recall (mean 64%) is notably lower, confirming the conservative tendency observed in the confusion matrices: the model is more likely to miss true Alzheimer's cases than to incorrectly label healthy individuals. Specificity is consistently high (>92%), further supporting the model's reliability in identifying healthy controls. The F1-score, which balances precision and recall, averages around 0.70, indicating moderate overall discriminative power. From a clinical perspective, these results suggest that the model could serve as a useful adjunct for ruling out AD in population screening, but its moderate sensitivity may limit its utility as a standalone diagnostic tool, particularly in settings where early detection is critical.

Finally, Figure 10 illustrates the learning curves for both

training and validation accuracy as a function of training set size, averaged across all folds. The training accuracy starts high when the model is fit on smaller subsets of data and gradually decreases as more data is included, reflecting reduced overfitting. Conversely, validation accuracy improves with increasing training set size, plateauing at approximately 84%. The convergence of training and validation curves at larger sample sizes indicates that the model generalizes well and is not overfitting to the training data. Clinically, this learning behavior suggests that the model's performance is stable and reliable as more data becomes available and that further increases in dataset size may yield only marginal improvements in accuracy. This pattern is characteristic of a well-regularized model and supports its potential for deployment in real-world clinical settings, where generalizability is paramount.

3.1.3. Ten-fold CV

To assess the robustness and generalizability of the binary classification model, a 10-fold CV was performed. In this approach, the dataset was partitioned into ten equally sized folds, with each fold serving as a validation set once while the remaining folds were used for training. Representative confusion matrices from the third, sixth, and ninth folds are presented in Figure 11. These matrices demonstrate consistent classification performance across different data splits. True-positive rates (*i.e.*, correct identification of AD cases) ranged from 60.5% to 64.1%, while true-negative rates (*i.e.*, correct identification of healthy controls) remained high and stable between 92.0% and

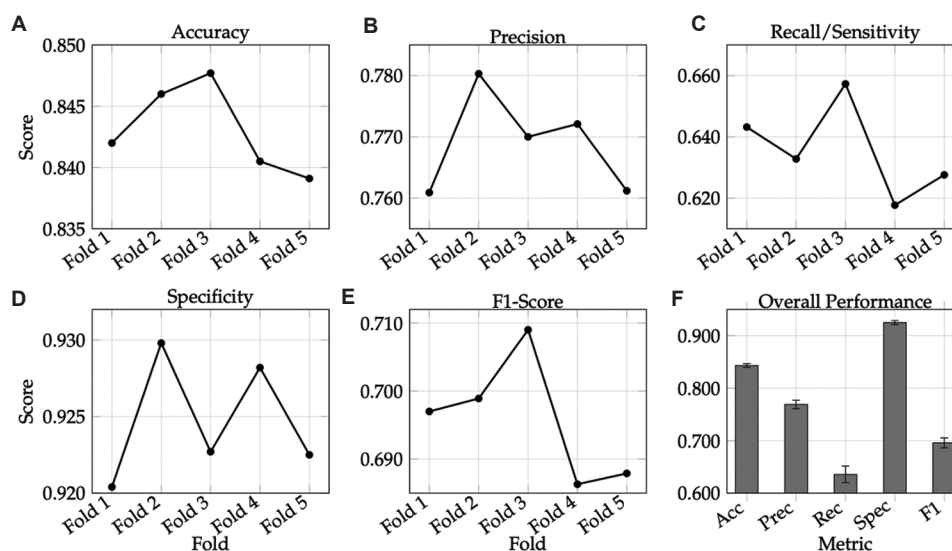


Figure 9. Performance evaluation across 5-fold cross-validation for binary classification (Healthy vs. Alzheimer's). Individual plots show the variation of each metric across folds: (A) Accuracy, (B) precision, (C) recall/sensitivity, (D) specificity, (E) F1-score, and (F) overall performance summary with mean values and standard deviations. The model demonstrates consistent performance with high specificity (>92%) but moderate sensitivity (~64%), indicating conservative classification behavior for the positive (Alzheimer's) class.

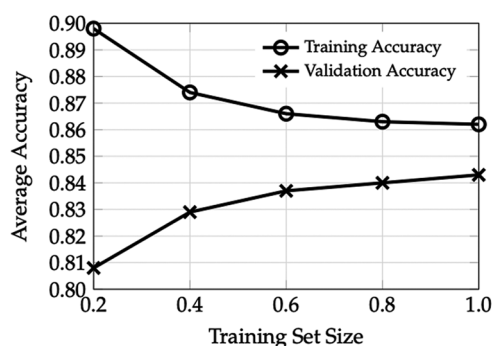


Figure 10. Average learning curves for training and validation accuracy across all folds in 5-fold cross-validation for binary classification

93.1%. This pattern indicates a conservative classification strategy, characterized by a low rate of false positives and a moderate rate of false negatives, which is often preferred in clinical screening contexts where minimizing unnecessary interventions in healthy individuals is prioritized.

An evaluation of model performance across all ten folds is provided in Figure 12. This figure displays the variation of key metrics (including accuracy, precision, recall [sensitivity], specificity, and F1-score) across each fold, as well as a summary of mean values and standard deviations. The model achieved a mean accuracy of 84.4% ($\pm 0.37\%$), with high specificity ($92.9\% \pm 0.48\%$) and precision ($77.7\% \pm 1.1\%$), but only moderate recall ($62.9\% \pm 1.1\%$). The F1-score, which balances precision and recall, averaged 0.695 (± 0.008). These results confirm that the model consistently favors the correct identification of healthy individuals, while demonstrating a more cautious approach in labeling cases as AD. Such a performance profile is typical for screening tools where false positives are considered more problematic than false negatives.

The learning dynamics of the model were further examined by plotting average learning curves for training and validation accuracy as a function of training set size, aggregated across all folds (Figure 13). Training accuracy was observed to decrease as the training set size increased, reflecting reduced overfitting, while validation accuracy improved and plateaued at approximately 84.5%. The convergence of training and validation curves at larger sample sizes indicates that the model generalizes well and does not exhibit overfitting. This learning behavior suggests that model performance is stable and reliable as more data become available and that further increases in dataset size are likely to yield only marginal improvements in accuracy.

3.2. Multiclass classification results

The following section presents the results of multiclass classification experiments, in which the model distinguishes among four clinically relevant categories.

Model performance is evaluated using three validation strategies: hold-out validation, 5-fold CV, and 10-fold CV. For each approach, the analyses include confusion matrices, per-class and macro-averaged performance metrics, and learning curves to assess generalization and overfitting. This evaluation enables comparison of the validation strategies and provides insights into the model's strengths and limitations in differentiating between multiple stages of AD.

3.2.1. Hold-out validation

The confusion matrices for four-class classification (Figure 14) illustrate the model's ability to distinguish among the following clinically relevant categories: Mild Dementia (Class 1), Moderate Dementia (Class 2), No Dementia (Class 3), and Very Mild Dementia (Class 4). The matrices display performance across training, validation, and test sets, with values representing the proportion of correct and incorrect predictions within each actual class. Diagonal elements indicate correct classifications, while off-diagonal elements reflect misclassifications.

- Class 1 (Mild Dementia) achieved classification accuracies of 83.0%, 76.0%, and 75.3% on the training, validation, and test sets, respectively. However, a notable proportion of Mild Dementia cases were misclassified, with 11.5–17.0% predicted as Very Mild Dementia and 5.4–8.3% as No Dementia. This suggests some overlap in features between Mild and adjacent stages, potentially complicating early intervention.
- Class 2 (Moderate Dementia) demonstrated the highest classification performance, with near-perfect accuracies of 100.0%, 97.1%, and 96.5% across the three datasets. Misclassification rates were minimal, indicating that Moderate Dementia presents with distinct features that are readily identified by the model.
- Class 3 (No Dementia) showed moderate classification performance, with accuracies of 75.8%, 70.3%, and 70.7%. A significant fraction of healthy individuals (9.1–12.0%) were misclassified as Mild Dementia, and 15.0–18.1% as Very Mild Dementia, reflecting the challenge of distinguishing normal aging from early pathological changes.
- Class 4 (Very Mild Dementia) exhibited the lowest classification accuracies, ranging from 58.6% to 68.9%. The most common misclassification was as Mild Dementia (18.2–24.3%), indicating a tendency to overestimate disease severity in this early stage. This systematic overestimation could lead to unnecessary concern or intervention in clinical settings.

Overall, the confusion matrices reveal that the model is most effective at identifying Moderate Dementia,

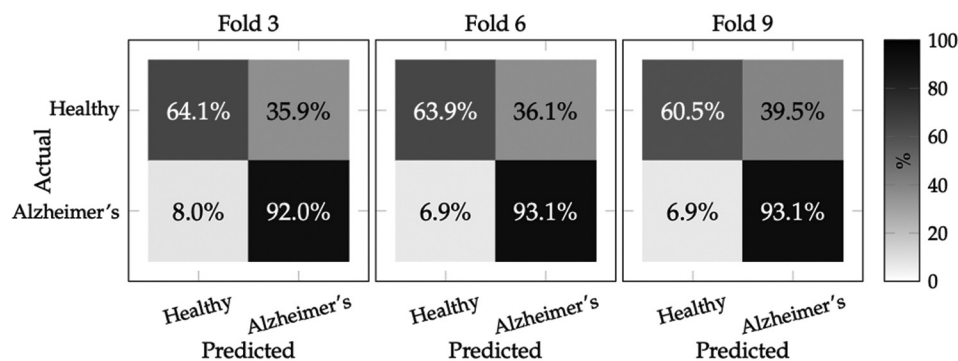


Figure 11. Representative confusion matrices for three folds (3rd, 6th, and 9th) out of 10-fold cross-validation in binary classification (Healthy vs. Alzheimer's). These three (showing only 3 to save space) representative matrices out of 10 demonstrate consistent model stability with similar performance patterns across folds. True-positive rates show moderate variation (60.5–64.1%) while true-negative rates remain consistently high (92.0–93.1%), indicating stable conservative classification behavior. Values are calculated row-wise to show the proportion of correct and incorrect predictions within each actual class.

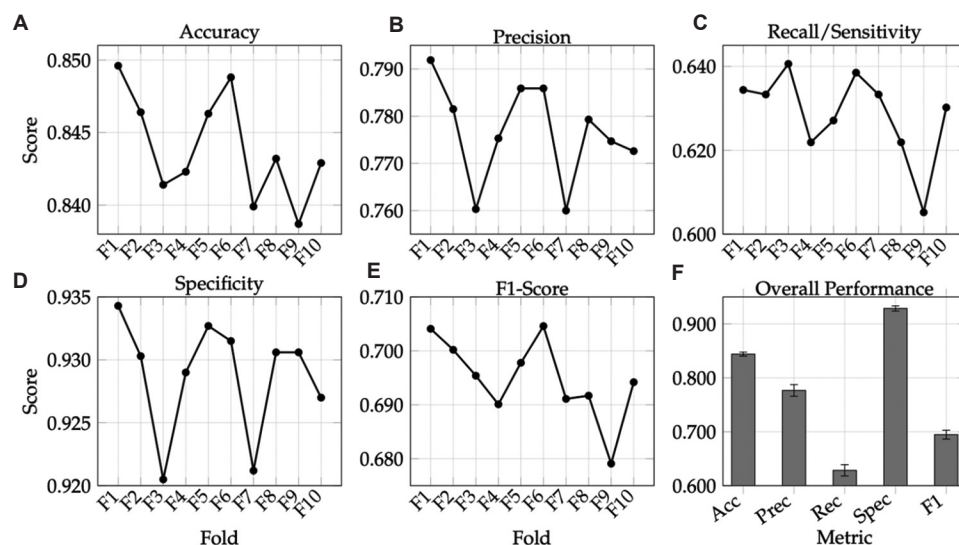


Figure 12. Performance evaluation across 10-fold cross-validation for binary classification (Demented vs. Non-Demented). Individual plots show the variation of each metric across folds: (A) Accuracy, (B) precision, (C) recall/sensitivity, (D) specificity, (E) F1-score, and (F) overall performance summary with mean values and standard deviations. The model demonstrates consistent performance with high specificity (>92%) but moderate sensitivity (~63%), indicating conservative classification behavior for the positive (Demented) class.

while differentiation between No Dementia, Very Mild Dementia, and Mild Dementia remains challenging. The consistent performance patterns across all three datasets demonstrate model stability and generalization capability demonstrated as follows.

Figure 15 provides a detailed overview of model performance across training, validation, and test sets for the 4-class classification task, using a set of metrics: accuracy, precision, recall (sensitivity), specificity, and F1-score. Each metric is shown as both a macro-average (thick black line with circular markers) and as individual class values (think black lines with distinct marker shapes for each class), allowing for nuanced assessment of generalization and class-specific trends.

The accuracy plot (Figure 15A) demonstrates that Class 2 (Moderate Dementia) achieves near-perfect classification across all splits, with only minor decreases from training to validation and test sets. Such near-perfect performance, especially when observed across both training and validation/test sets, may indicate overfitting, where the model has learned not only the underlying patterns but also the noise in the training data, potentially reducing its ability to generalize to truly unseen data. In contrast, Class 4 (Very Mild Dementia) consistently shows the lowest accuracy, with a marked drop from training to validation, and then a plateau or slight improvement on the test set. Class 1 (Mild Dementia) and Class 3 (No Dementia) show intermediate performance, with Class 1 generally outperforming Class 3.

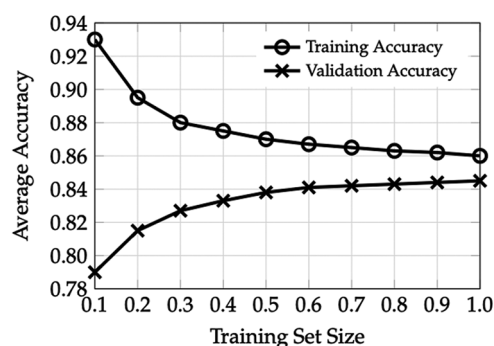


Figure 13. Average learning curves for training and validation accuracy across all folds in 10-fold cross-validation for binary classification

A closer examination of generalization patterns reveals that only Class 1 (Mild Dementia) follows the expected trend of decreasing performance from training to validation to test sets across all metrics. This is the hallmark of a well-behaved model, reflecting reduced familiarity with the data as the model moves from training to unseen samples. For the other classes (Class 2: Moderate Dementia, Class 3: No Dementia, and Class 4: Very Mild Dementia), while all metrics decrease from training to validation, the expected further decrease from validation to test is not consistently observed. Instead, some metrics remain stable, and in certain cases, even increase from validation to test sets. This is not the ideal pattern, as one typically expects the

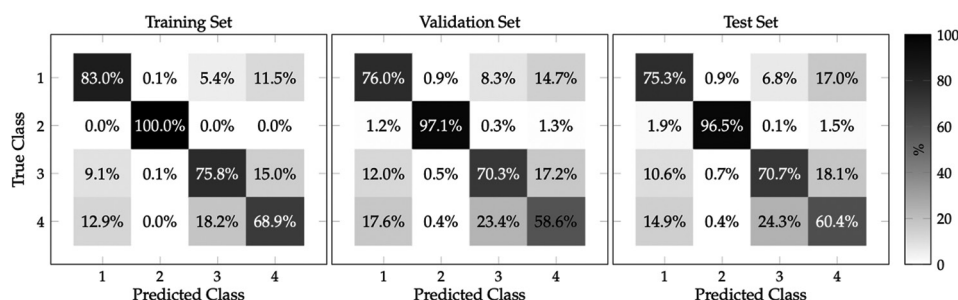


Figure 14. Confusion matrices for four-class classification showing performance across training, validation, and test sets. Values are calculated row-wise to show the proportion of correct and incorrect predictions within each actual class. Classes are ordered as: Mild Dementia (1), Moderate Dementia (2), No Dementia (3), and Very Mild Dementia (4). Diagonal elements represent correct classifications. The consistent performance patterns across all three sets demonstrate model stability and generalization capability.

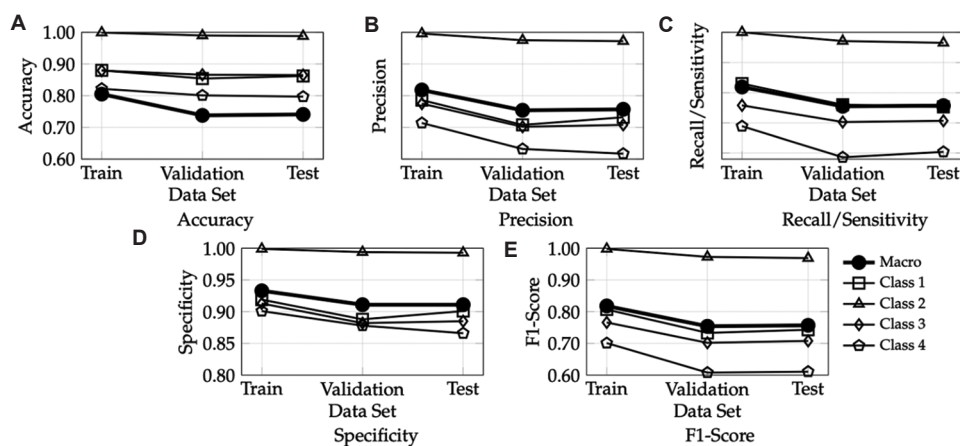


Figure 15. Performance metrics across training, validation, and test sets for 4-class classification. Each plot shows macro-averaged performance (thick black line with circular markers) alongside individual class performance (black lines with distinct markers: squares for Class 1, triangles for Class 2, diamonds for Class 3, and pentagons for Class 4). The plots reveal: (A) Per-class accuracy with Class 2 (Moderate Dementia) achieving near-perfect performance while Class 4 (Very Mild Dementia) shows consistent decline, (B and C) precision and recall patterns highlighting Class 2's (Moderate Dementia) superiority, (D) consistently high specificity across all classes, and (E) F1-scores demonstrating the trade-offs between precision and recall for each class. Only Class 1 (Mild Dementia) exhibits the expected generalization pattern of decreasing performance from training to validation to test sets. For other classes, all metrics decrease from training to validation, but then either remain stable or, in some cases, increase from validation to test sets. While such behavior is not ideal, it is common and warrants further investigation of training dynamics, such as through learning curves, to confirm robust generalization.

lowest performance on the test set due to its complete independence from the training process. However, such behavior can occur due to sampling variability, limited data, or random fluctuations in the data splits, and does not by itself indicate overfitting or data leakage. Rather, it highlights the need for further scrutiny, such as examining learning curves, to ensure that the model's generalization is robust and not an artifact of a particular data split.

Clinically, these results suggest that the model is highly effective at identifying Moderate Dementia (Class 2), which may represent an optimal window for intervention. However, the reduced and less stable performance for Very Mild Dementia (Class 4) underscores the need for caution in relying solely on automated staging for early disease, as misclassification could have significant implications for patient management. The observed stability or slight improvement in some test set metrics, relative to validation, further emphasizes the importance of rigorous validation and ongoing monitoring of model behavior to ensure reliable deployment in clinical settings.

Hence, Figure 15 illustrates that while the model demonstrates strong performance for certain classes, especially Class 2 (Moderate Dementia), the generalization patterns for other classes warrant careful interpretation. The unexpected stability or improvement in test set metrics compared to validation should prompt further investigation into the training dynamics and data partitioning, ideally through the use of learning curves and repeated CV, to confirm the reliability and robustness of the model's clinical predictions.

To further assess the model's generalization and training dynamics, Figure 16 presents the learning curves for multiclass classification accuracy as a function of training set size. The plot displays both training and validation accuracy, providing insight into how the model's performance evolves as it is exposed to increasing amounts of data. At the smallest training set size, the model achieves perfect accuracy on the training data (1.00), but validation accuracy is very low (0.32), indicating severe overfitting when the model has seen only a limited number of examples. As the training set size increases, training accuracy gradually decreases, stabilizing around 0.80 when the full dataset is used. In contrast, validation accuracy rises sharply with more data, plateauing near 0.74–0.75 at the largest training set size. This convergence between training and validation accuracy as the dataset grows is a hallmark of improved generalization and reduced overfitting.

The learning curves thus confirm that the model benefits substantially from additional data, with the gap between training and validation accuracy narrowing as the training set size increases. The final plateau suggests

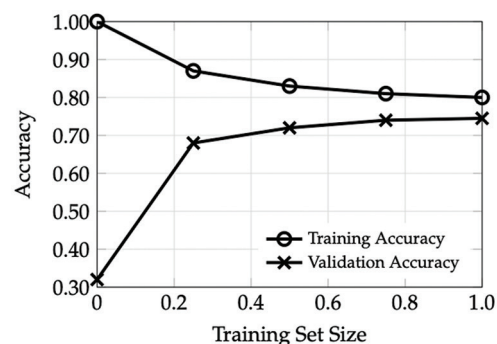


Figure 16. Learning curves for multiclass classification accuracy as a function of training set size. The plot shows both training (circles) and validation (crosses) accuracy. At small training sizes, the model achieves perfect training accuracy but very low validation accuracy, indicating overfitting. As the training set size increases, training accuracy steadily decreases while validation accuracy rises and plateaus, with both curves converging as more data is used. This convergence indicates improved generalization and reduced overfitting.

that, given the current data and model complexity, further increases in training data may yield only incremental improvements in validation performance. Importantly, the absence of a widening gap at larger training sizes indicates that the model is not overfitting and that its generalization to unseen data is robust. These findings complement the performance metrics shown in Figure 15, where only Class 1 (Mild Dementia) exhibits the expected monotonic decrease in performance from training to validation to test sets. For other classes, the observed stability or slight improvement in test set metrics relative to validation could be attributed to sampling variability or limited data, rather than overfitting or data leakage. The learning curves support this interpretation, demonstrating that the model's generalization is not an artifact of a particular data split, but rather a reflection of its capacity to learn meaningful patterns from the available data.

3.2.2. Five-fold CV

To assess the generalizability and clinical reliability of the multiclass classification model, 5-fold CV was conducted to evaluate the model's ability to distinguish among four clinically relevant categories: Class 1 (Mild Dementia), Class 2 (Moderate Dementia), Class 3 (No Dementia), and Class 4 (Very Mild Dementia). The results are summarized in Figures 17–19.

Figure 17 presents representative confusion matrices from three of the five CV folds. Each matrix displays the proportion of correct and incorrect predictions for each true class, with diagonal elements indicating correct classifications. The model demonstrates consistent performance across folds, with the highest accuracy observed for Class 4 (Moderate Dementia), where correct

classification rates exceed 97% in all folds. Class 1 (No Dementia) and Class 3 (Mild Dementia) also exhibit robust classification, with correct rates typically between 68% and 78%. In contrast, Class 2 (Very Mild Dementia) shows lower accuracy, ranging from 56.9% to 63.1%, with a notable proportion of these cases misclassified as Class 1 (No Dementia). This pattern indicates that the model is highly sensitive to advanced-stage dementia (Moderate Dementia) but may underestimate or miss the earliest signs of disease (Very Mild Dementia), potentially leading to under-detection of early-stage cases if used in isolation. This limitation is important clinically, as early detection is critical for timely intervention, and the model's current strengths lie more in identifying established, moderate dementia than in detecting prodromal or very mild cases. However, the consistency of these confusion matrices

across folds underscores the model's stability and reliability in distinguishing among the four stages, minimizing the risk of performance fluctuations due to data partitioning.

An evaluation of key performance metrics across all five folds is provided in Figure 18. The individual plots illustrate the variation in accuracy, precision, recall (sensitivity), specificity, and F1-score for each fold, while the summary bar chart aggregates the mean and standard deviation for each metric. The model achieves a mean accuracy of 87.1% ($\pm 0.12\%$), mean precision of 91.9% ($\pm 0.15\%$), mean recall of 90.4% ($\pm 0.27\%$), mean specificity of 77.9% ($\pm 0.44\%$), and mean F1-score of 91.2% ($\pm 0.16\%$). The low standard deviations across folds indicate strong generalization and minimal sensitivity to the specific data split. High precision and recall values suggest that the model is effective at



Figure 17. Confusion matrices for four-class classification across three (representative) folds showing performance consistency. Classes are ordered as: No Dementia (1), Very Mild Dementia (2), Mild Dementia (3), and Moderate Dementia (4). Values are calculated row-wise to show the proportion of correct and incorrect predictions within each actual class. Diagonal elements represent correct classifications, demonstrating stable model performance across all folds from 5-fold validation.

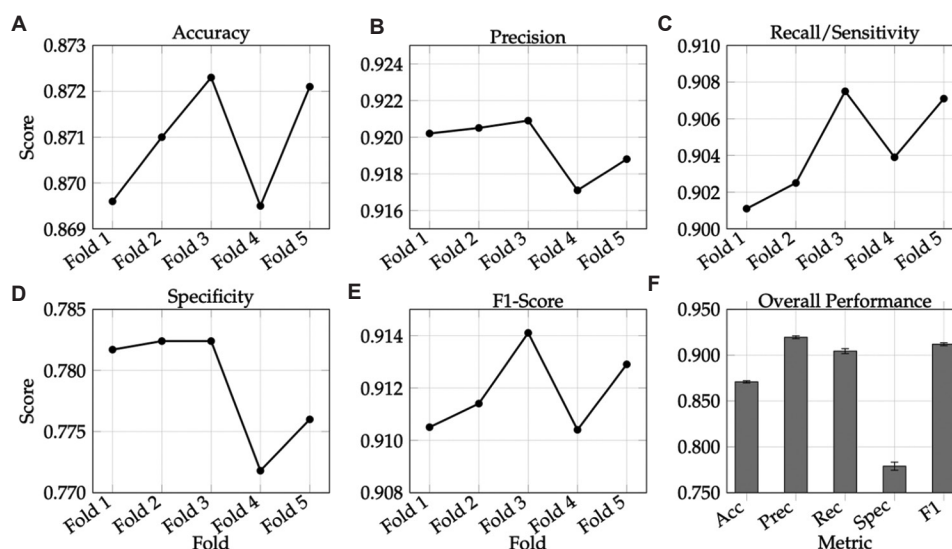


Figure 18. Performance evaluation across 5-fold cross-validation for multiclass classification (Mild Dementia vs. Moderate Dementia vs. No Dementia vs. Very Mild Dementia). Individual plots show the variation of each metric across folds: (A) Accuracy, (B) precision, (C) recall/sensitivity, (D) specificity, (E) F1-score, and (F) overall performance summary with mean values and standard deviations. The model demonstrates consistent performance across the four dementia severity classes, with the metrics representing macro-averaged values across all classes.

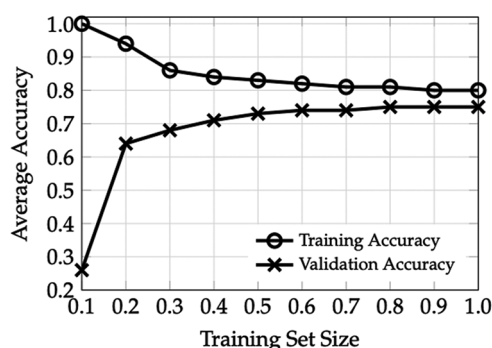


Figure 19. Average learning curves for training and validation accuracy across all folds in 5-fold cross-validation for multiclass classification.

identifying true cases of each dementia stage and at minimizing false positives, which is essential for reducing unnecessary anxiety and interventions. The slightly lower specificity reflects a tendency to overcall dementia in some individuals without the condition, a common trade-off in screening tools where sensitivity is prioritized. The consistently high F1-score, which balances precision and recall, supports the model's utility as a reliable adjunct in clinical decision-making for dementia staging.

To further elucidate the model's learning dynamics and potential for overfitting, Figure 19 displays the average learning curves for training and validation accuracy as a function of training set size, aggregated across all folds. At small training set sizes, the model achieves perfect accuracy on the training data but poor validation accuracy, indicating overfitting. As the training set size increases, training accuracy gradually decreases while validation accuracy rises and plateaus, with both curves converging at approximately 80% (training) and 75% (validation) when the full dataset is utilized. This convergence is indicative of improved generalization and reduced overfitting, confirming that the model's performance is robust and not an artifact of a particular data split. This learning behavior suggests that the model can be expected to maintain stable performance as more data becomes available and that further increases in dataset size may yield only incremental improvements in accuracy.

3.2.3. Ten-fold CV

The results of the multiclass dementia classification model, as evaluated by 10-fold cross-validation, are illustrated in Figures 20-22. These figures collectively provide a view of the model's algorithmic behavior and its clinical implications.

From an algorithmic perspective, the confusion matrices in Figure 20 reveal the distribution of correct and incorrect predictions across three representative folds.

Each matrix is structured such that the rows correspond to the true class labels (from top to bottom: No Dementia, Very Mild Dementia, Mild Dementia, and Moderate Dementia) while the columns represent the predicted classes in the same order. The diagonal elements, which are consistently the darkest, indicate the proportion of correctly classified samples for each class. For example, the model achieves very high accuracy for the Moderate Dementia class (Class 4, 98.0–98.5%), while the No Dementia class (Class 1) shows good correct classification rates (approximately 76–80%). The Mild Dementia class (Class 3) exhibits moderate performance (approximately 70%), and the Very Mild Dementia class (Class 2) shows the lowest correct classification rates (58–61%), with a notable proportion of these cases being misclassified as No Dementia (Class 1, 17–22%). The off-diagonal elements highlight the primary sources of confusion, which tend to occur between adjacent disease stages, particularly between Very Mild and No Dementia, and between Moderate and Mild Dementia categories, reflecting the inherent difficulty in distinguishing between clinically similar stages. Regardless, the consistency of these patterns across folds demonstrates the model's stability and reliability, minimizing the risk that observed performance is due to chance or a particular data split.

Clinically, the model is highly reliable for identifying Moderate Dementia, but less so for detecting Very Mild Dementia, which is now the most challenging class for the model to distinguish. This is a notable limitation, as early detection of dementia is vital for intervention and care planning. The model's strengths are in recognizing more advanced disease, but its reduced sensitivity for the earliest stage means that it may not be optimal as a screening tool for prodromal or very mild cases. This pattern mirrors known diagnostic challenges, where subtle early changes are often more difficult to detect both for clinicians and automated systems.

The summary of performance metrics across all folds, as depicted in Figure 21, further supports the model's robustness. The individual plots for accuracy, precision, recall (sensitivity), specificity, and F1-score show minimal variation across the 10 folds, with mean values of 0.76 for accuracy, 0.78 for both precision and recall, and a notably high specificity of 0.92. The F1-score, which balances precision and recall, is also stable at approximately 0.78. The bar chart summarizing mean and standard deviation for each metric confirms that the model's performance is not only consistent but also well-balanced across the different evaluation criteria. This pattern is indicative of a model that generalizes well and is not overfitting to the training data, as evidenced by the low standard deviations and the absence of significant performance drops in any fold.

A closer examination of the metric plots in Figure 21 reveals that, for each performance measure (including accuracy, precision, recall, specificity, and F1-score), the value observed in the first fold (F1) is consistently lower than those recorded in subsequent folds (F2–F10). For example, the F1-score in the first fold is 0.7554, while in the remaining folds it ranges from 0.7727 to 0.7866. This pattern is similarly reflected in the other metrics, where the initial fold starts at a slightly reduced value before stabilizing at higher levels in later folds. Despite this initial dip, the overall range of values across all folds remains narrow, with differences typically on the order of two to three percentage points. This limited variability suggests that the model's performance is robust and not unduly influenced by any single data split. The slightly lower values in the first

fold may be attributable to random variation inherent in CV, rather than a systematic issue with the model or data. Importantly, the consistency and tight clustering of metric values across folds reinforce the conclusion that the model generalizes well and maintains stable performance across different partitions of the dataset.

The learning curves in Figure 22 provide additional insight into the model's training dynamics and generalization capacity. As the training set size increases, the training accuracy gradually decreases from 0.87 to 0.80, while the validation accuracy increases from 0.69 to 0.75. The convergence of these curves at larger training sizes is a hallmark of reduced overfitting and improved generalization. The plateauing of validation accuracy suggests that, with the current dataset and

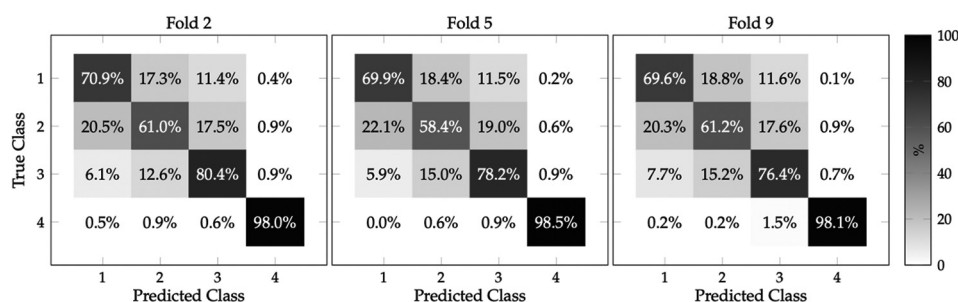


Figure 20. Confusion matrices for four-class classification across three (representative) folds showing performance consistency. Classes are ordered as: No Dementia (1), Very Mild Dementia (2), Mild Dementia (3), and Moderate Dementia (4). Values are calculated row-wise to show the proportion of correct and incorrect predictions within each actual class. Diagonal elements represent correct classifications, demonstrating stable model performance across all folds from 10-fold validation.

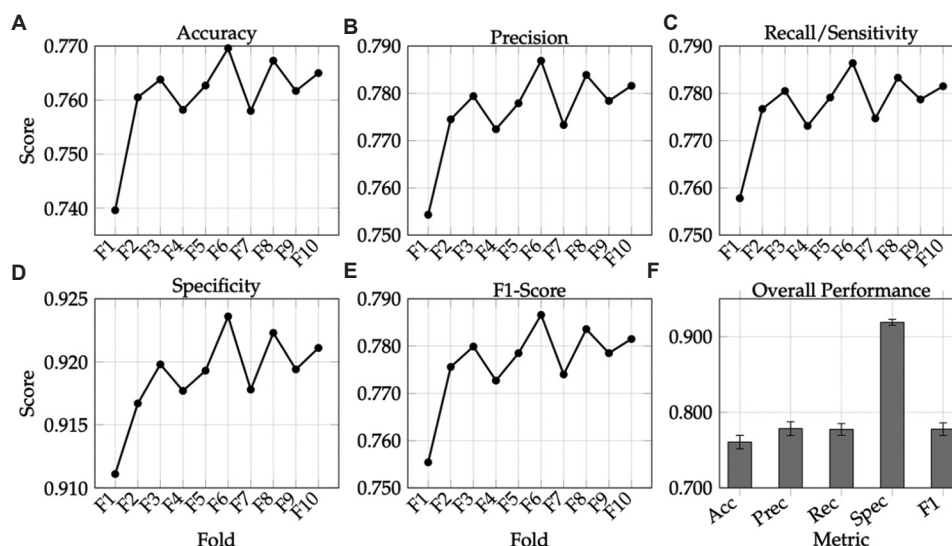


Figure 21. Performance evaluation across 10-fold cross-validation for 4-class dementia classification (Mild Dementia, Moderate Dementia, No Dementia, Very Mild Dementia). Individual plots show the variation of each metric across folds: (A) Accuracy, (B) precision, (C) recall/sensitivity, (D) specificity, (E) F1-score, and (F) overall performance summary with mean values and standard deviations. The model demonstrates consistent performance with high specificity (>91%) and balanced precision-recall (~78%), indicating robust multi-class classification across different dementia severity levels.

model complexity, further increases in training data may yield only incremental improvements. This behavior is characteristic of a well-regularized model and supports the reliability of its predictions on unseen data.

Clinically, the model is highly reliable for identifying Very Mild Dementia and No Dementia, but less so for distinguishing between Mild and Moderate Dementia, which mirrors known diagnostic challenges in practice. The high specificity reduces the likelihood of false positives, and the balanced precision and recall indicate that the model does not disproportionately favor sensitivity over specificity. Overall, the model's results are clinically meaningful and robust for most classes, particularly for detecting Very Mild Dementia and ruling out dementia, though caution is warranted when interpreting predictions for Moderate Dementia due to common misclassification with Mild Dementia. The stability of performance across folds and the convergence seen in learning curves support the model's reliability for clinical or research applications, provided that its limitations in discriminating between Mild and Moderate Dementia are acknowledged and appropriately managed in deployment.

3.3. Comparison of validation strategies

A comparison was conducted of overall classification accuracy achieved by the models under three commonly used validation strategies: hold-out, 5-fold CV, and 10-fold CV, for both binary and multiclass classification tasks (Figure 23). Although several binary classification scenarios were explored by combining the four diagnostic categories (Healthy, Very Mild, Mild, and Moderate Alzheimer's) into pairwise problems using the hold-out method, only the results for the Healthy vs. Alzheimer's comparison are presented here. This approach was adopted to ensure consistency, as both the 5-fold and 10-fold CV experiments focused exclusively on the Healthy vs. Alzheimer's binary task.

The results depicted in Figure 23 demonstrate distinct trends in model performance depending on the validation strategy employed. For the binary classification task (Healthy vs. Alzheimer's), the hold-out method yielded a test accuracy of 0.55, which is only marginally above random chance. This outcome highlights the inherent difficulty in distinguishing between healthy aging and AD based solely on the extracted features and also reflects the sensitivity of the hold-out method to the particular data split, especially in the context of limited datasets. In contrast, both 5-fold and 10-fold CV strategies produced substantially higher and nearly identical accuracies (0.84), indicating that CV provides a more robust and reliable estimate of model generalization by averaging over

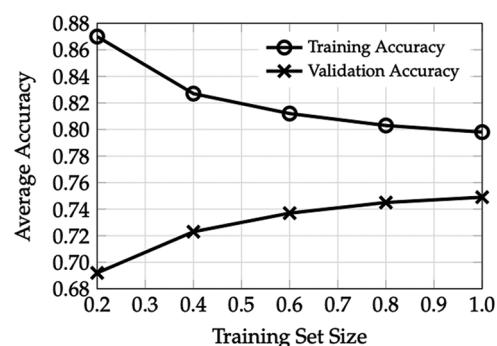


Figure 22. Average learning curves for training and validation accuracy across all folds in 10-fold cross-validation for 4-class dementia classification (No Dementia, Very Mild Dementia, Mild Dementia, Moderate Dementia).

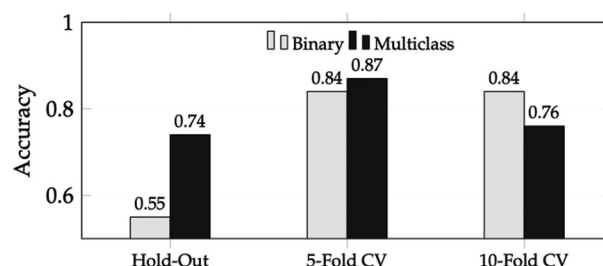


Figure 23. Comparison of overall classification accuracy for hold-out, 5-fold, and 10-fold cross-validation (CV) strategies in both binary and multiclass settings.

multiple data splits. This improvement suggests that the model is capable of learning meaningful patterns when exposed to a greater diversity of data partitions and that the hold-out result may underestimate the true potential of the model due to sampling variability.

For the multiclass classification task, the hold-out accuracy was 0.74, while 5-fold CV achieved the highest accuracy at 0.87, and 10-fold CV yielded 0.76. The superior performance of 5-fold CV in the multiclass setting may reflect an optimal balance between training set size and the number of validation splits, maximizing both data utilization and the stability of performance estimates. The reduction in accuracy observed with 10-fold CV could be attributed to smaller fold sizes, which may introduce greater variability in training and validation sets, particularly in the presence of class imbalance.

4. Discussion

A fundamental challenge in ML research, particularly in clinical applications, is that there is rarely a single, universally optimal experimental approach.^{1,35} The performance of any given model is influenced by a complex interplay of factors, including data size, data type, data quality, preprocessing

methods, data splitting strategies, algorithm selection, and the specific input variables chosen for training.^{36,37} Because of these numerous dependencies, it is virtually impossible to predict in advance which combination of experimental choices will yield the best results for a given problem. As such, it is vital that studies like this one explicitly compare key experimental options (such as different validation strategies) rather than relying on a single approach.^{38,39} By evaluating multiple validation pathways, researchers can ensure that their findings are robust, generalizable, and not artifacts of arbitrary methodological choices or data splits.⁴⁰ This approach is particularly critical in clinical ML, where the stakes for reliability and reproducibility are high, and where subtle differences in experimental design can lead to markedly different conclusions about a model's utility in real-world settings.

4.1. Clinical implications

From a clinical perspective, the study's findings have several implications. The low hold-out accuracy for binary classification underscores the challenge of reliably distinguishing AD from healthy controls in real-world settings and suggests that single-split validation may not provide a sufficiently robust assessment for clinical deployment. The marked improvement observed with CV indicates that, with appropriate validation strategies, ML models can achieve clinically meaningful performance, potentially supporting early detection and triage. However, the variability observed across validation strategies, especially in the multiclass setting, highlights the importance of careful model evaluation. These results emphasize that the choice of validation strategy can have a profound impact on perceived model performance and that robust CV is essential for developing reliable clinical decision support tools.

4.2. End-user awareness in model deployment

The results of this study highlight that model performance is not uniform across all diagnostic categories. Specifically, in CV, the model exhibits high reliability in identifying very mild AD, with consistently superior performance metrics for this class across all data splits. In contrast, the classification of moderate AD is associated with lower and less stable performance, as evidenced by reduced accuracy, precision, recall, and F1-scores. Such class-specific variability in model performance has direct implications for clinical decision-making. It is vital that end-users are made aware of the strengths and limitations of the model in each diagnostic category, to avoid over-reliance on automated predictions, particularly in cases where the model's generalization is suboptimal. These findings underscore the importance of transparent reporting of

per-class performance metrics and ongoing collaboration between data scientists and clinicians to ensure that model outputs are interpreted within the appropriate clinical context and that deployment strategies maximize patient safety and diagnostic utility.

4.3. Challenges and limitations

Our study relies on a single, publicly available MRI dataset, which may not fully capture the diversity of imaging protocols, scanner types, and patient populations encountered in real-world clinical practice. As a result, the generalizability of our findings to other cohorts or clinical settings may be limited. Future work should validate these models on larger, multi-center datasets with greater demographic and technical variability.

The present study focused on deep features extracted from the final pooling layer of GoogLeNet as a representative approach. While this choice is supported by prior literature, we acknowledge that ablation studies comparing alternative pretrained backbones (e.g., ResNet, VGG), different layers within the same network, and classical baselines such as principal component analysis (PCA) or radiomics features would provide additional insight into the robustness and optimality of the chosen representation. These analyses are planned for future work.

The dataset does not include metadata on acquisition site or scanner type, precluding stratification or analysis of potential site/scanner confounds. This represents a limitation, as unrecognized site effects could influence model performance and generalizability. Additionally, the dataset exhibits class imbalance, particularly for the moderate dementia category, which can bias model training and evaluation. Although we employed data augmentation and stratified sampling, underrepresented classes may still be at risk for poorer performance and less reliable predictions. This limitation is reflected in the lower and less stable accuracy for moderate-stage cases.

Some tasks, especially those with near-perfect accuracy (e.g., Healthy vs. Moderate), showed signs of overfitting or possible data leakage, as indicated by abnormal learning curves and unexpectedly high validation/test performance. While learning curve analysis helped identify these issues, it highlights the ongoing risk of overestimating model performance in limited or non-randomized datasets.

Our analysis is based solely on cross-sectional MRI data. Incorporating longitudinal imaging or additional modalities (e.g., PET, cognitive scores, genetic data) could improve early detection and staging, and better reflect the complexity of AD progression. All results are retrospective and based on pre-existing data. Prospective validation in real-world clinical workflows is necessary to confirm the

practical utility and safety of these models before clinical deployment.

Moreover, while this study primarily focused on quantitative performance metrics and learning curve analysis, we acknowledge that feature space visualization techniques such as t-SNE or UMAP could provide additional interpretability by illustrating class overlap and the structure of the extracted feature representations. Such analyses can help to contextualize observed performance ceilings and clarify the separability of clinical categories in the learned feature space. We plan to include these interpretability analyses in future work.

5. Conclusion

This study compared the performance of automated MRI-based AD classification models across hold-out and k-fold CV strategies, leveraging deep feature extraction from pretrained GoogLeNet and traditional ML classifiers. The results demonstrate that cross-validation, especially 5-fold, consistently provides higher and more reliable accuracy estimates than the hold-out method for both binary and multiclass dementia classification. Binary classification of healthy vs. demented subjects achieved up to 84% accuracy with CV, while multiclass tasks reached up to 87%. However, performance varied by class: the model showed high specificity and precision for moderate-stage cases of AD but substantially lower and less stable accuracy for early-stage (very mild) cases, emphasizing the importance of transparent per-class reporting in clinical ML research.

Overall, these findings highlight the critical influence of validation strategy on both the perceived and actual robustness of clinical ML models. Adequate cross-validated evaluation is vital to avoid overestimating performance and to ensure generalizability, especially before clinical deployment. Furthermore, users and clinicians must be made aware of class-specific strengths and limitations to prevent over-reliance on automated predictions, particularly in underrepresented or more challenging diagnostic categories. Future research should prioritize improving model sensitivity for difficult classes, exploring advanced data augmentation and resampling approaches, and validating models on larger, more diverse datasets to enhance generalizability and clinical utility.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: All authors

Formal analysis: All authors

Investigation: All authors

Methodology: All authors

Writing-original draft: All authors

Writing-review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The MRI images used in this study were obtained from a publicly available dataset (<https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset>).

References

1. Bucholz M, James C, Khleifat AA, *et al.* Artificial intelligence for dementia research methods optimization. *Alzheimers Dement.* 2023;19(12):5934-5951.
doi: 10.1002/alz.13441
2. Zhang Y. The risk factors and causes for Alzheimer's disease. In: *Proceedings of the 2022 8th International Conference on Humanities and Social Science Research (ICHSSR 2022)*. Atlantis Press; 2022:879-884.
doi: 10.2991/assehr.k.220504.160
3. Castellani RJ, Perry G. Molecular pathology of Alzheimer's disease. In: *Colloquium Series on Neurobiology of Alzheimer's Disease*. Vol. 1. California: Morgan and Claypool Life Sciences; 2013:1-91.
doi: 10.4199/c00095ed1v01y201310alz001
4. DeTure MA, Dickson DW. The neuropathological diagnosis of Alzheimer's disease. *Mol Neurodegener.* 2019;14(1):32.
doi: 10.1186/s13024-019-0333-5
5. Rasmussen J, Langerman H. Alzheimer's disease - Why we need early diagnosis. *Degener Neurol Neuromuscul Dis.* 2019;9:123-130.
doi: 10.2147/dnnd.s228939
6. Juganavar A, Joshi A, Shegkar T. Navigating early Alzheimer's diagnosis: A comprehensive review of diagnostic innovations. *Cureus.* 2023;15(9):e44937.
doi: 10.7759/cureus.44937

7. Shimizu S, Hirose D, Hatanaka H, *et al.* Role of neuroimaging as a biomarker for neurodegenerative diseases. *Front Neurol.* 2018;9:265.
doi: 10.3389/fneur.2018.00265
8. Frisoni GB, Fox NC, Jack CR Jr., Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol.* 2010;6(2):67-77.
doi: 10.1038/nrneurol.2009.215
9. McKhann GM, Knopman DS, Chertkow H, *et al.* The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 2011;7(3):263-269.
doi: 10.1016/j.jalz.2011.03.005
10. Bhat S, Acharya UR, Dadmehr N, Adeli H. Clinical neurophysiological and automated EEG-based diagnosis of the Alzheimer's disease. *Eur Neurol.* 2015;74(3-4):202-210.
doi: 10.1159/000441447
11. Akkas H, Latifoglu F, Tokmakcı M. The diagnosis of Alzheimer's disease using EEG signals. *Eur J Res Dev.* 2023;3(3):1-13.
doi: 10.56038/ejrnd.v3i3.273
12. Malik I, Iqbal A, Gu YH, Al-antari MA. Deep learning for Alzheimer's disease prediction: A comprehensive review. *Diagnostics (Basel).* 2024;14(12):1281.
doi: 10.3390/diagnostics14121281
13. Chang CH, Lin CH, Lane HY. Machine learning and novel biomarkers for the diagnosis of Alzheimer's disease. *Int J Mol Sci.* 2021;22(5):2761.
doi: 10.3390/ijms22052761
14. Kloppel S, Stonnington CM, Barnes J, *et al.* Accuracy of dementia diagnosis: A direct comparison between radiologists and a computerized method. *Brain.* 2008;131(Pt 11):2969-2974.
doi: 10.1093/brain/awn239
15. Jo T, Nho K, Saykin AJ. Deep learning in Alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data. *Front Aging Neurosci.* 2019;11:220.
doi: 10.3389/fnagi.2019.00220
16. Diogo VS, Ferreira HA, Prata D. Early diagnosis of Alzheimer's disease using machine learning: A multi-diagnostic, generalizable approach. *Alzheimers Res Ther.* 2022;14(1):107.
doi: 10.1186/s13195-022-01047-y
17. Kavitha C, Mani V, Srividhya SR, Khalaf OI, Tavera Romero CA. Early-stage Alzheimer's disease prediction using machine learning models. *Front Public Health.* 2022;10:853294.
doi: 10.3389/fpubh.2022.853294
18. Rezaei M, Zeresghi E, Shahsavari S, Salehi MG, Sharini H. Prediction of Alzheimer's disease using machine learning classifiers. *Int Electron J Med.* 2020;9(2):116-120.
doi: 10.34172/iejm.2020.21
19. Bron EE, Smits M, Niessen WJ, Klein S. Feature selection based on the SVM weight vector for classification of dementia. *IEEE J Biomed Health Inform.* 2015;19(5):1617-1626.
doi: 10.1109/jbhi.2015.2432832
20. Fu'adah YN, Wijayanto I, Pratiwi NKC, Taliningsih FF, Rizal S, Pramudito MA. Automated classification of Alzheimer's disease based on MRI image processing using convolutional neural network (CNN) with AlexNet architecture. *J Phys Conf Ser.* 2021;1844(1):012020.
doi: 10.1088/1742-6596/1844/1/012020
21. Liu M, Cheng D, Yan W. Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images. *Front Neuroinform.* 2018;12:35.
doi: 10.3389/fninf.2018.00035
22. Folego G, Weiler M, Casseb RF, Pires R, Rocha A. Alzheimer's disease detection through whole-brain 3D-CNN MRI. *Front Bioeng Biotechnol.* 2020;8:534592.
doi: 10.3389/fbioe.2020.534592
23. Liu J, Li M, Luo Y, Yang S, Li W, Bi Y. Alzheimer's disease detection using depthwise separable convolutional neural networks. *Comput Methods Programs Biomed.* 2021;203:106032.
doi: 10.1016/j.cmpb.2021.106032
24. Murugan S, Venkatesan C, Sumithra MG, *et al.* DEMNET: A deep learning model for early diagnosis of Alzheimer diseases and dementia from MR images. *IEEE Access.* 2021;9:90319-90329.
doi: 10.1109/access.2021.3090474
25. Khasanah I. Enhancing Alzheimer's disease diagnosis with K-NN: A study on pre-processed MRI data. *Int J Artif Intell Med Issues.* 2024;2(1):49-60.
doi: 10.56705/ijaimi.v2i1.150
26. Ahmed G, Er MJ, Fareed MMS, *et al.* DAD-Net: Classification of Alzheimer's disease using ADASYN oversampling technique and optimized neural network. *Molecules.* 2022;27(20):7085.
doi: 10.3390/molecules27207085
27. Umalakshmi NP, Sathyanarayana S, Chicktotlikere Nagappa P, Javarappa T, Kuppanna Rajuk V. Borderline-DEMNET: A workflow for detecting Alzheimer's and dementia stage by solving class imbalance problem.

- Pertanika J Sci Technol.* 2024;32(4):1629-1650.
doi: 10.47836/pjst.32.4.10
28. Husain G, Nasef D, Jose R, *et al.* SMOTE vs. SMOTEENN: A study on the performance of resampling algorithms for addressing class imbalance in regression models. *Algorithms.* 2025;18(1):37.
doi: 10.3390/a18010037
 29. Feng Y, Li J. A novel α Distance Borderline-ADASYN-SMOTE algorithm for imbalanced data and its application in Alzheimer's disease classification based on dense convolutional network. *J Phys Conf Ser.* 2021;2031(1):012046.
doi: 10.1088/1742-6596/2031/1/012046
 30. Chandrasekaran S, Khan SB, Gupta M, Mahesh TR, Alqhatani A, Almusharraf A. Enhanced deep learning framework for precise MRI-based Alzheimer's disease stage classification. *Comput Intell.* 2025;41(1):e70123.
doi: 10.1111/coin.70123
 31. Ali MU, Kim KS, Khalid M, Farrash M, Zafar A, Lee SW. Enhancing Alzheimer's disease diagnosis and staging: A multistage CNN framework using MRI. *Front Psychiatry.* 2024;15:1395563.
doi: 10.3389/fpsy.2024.1395563
 32. Yousafzai S, Khan GZ, Ulhaq S, Areebah, Butt MR. Improved neural network-based system for early and accurate diagnosis of Alzheimer disease. *J Comput Sci Technol Stud.* 2023;5(4):32-40.
doi: 10.32996/jcsts.2023.5.4.4
 33. Younis MT, Younus YT, Hasoon JN, Fadhil AH, Mostafa SA. An accurate Alzheimer's disease detection using a developed convolutional neural network model. *Bull Electr Eng Inform.* 2022;11(4):2005-2012.
doi: 10.11591/eei.v11i4.3659
 34. Uraninjo. *Augmented Alzheimer MRI Dataset.* Kaggle; 2023. Available from: <https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset> [Last accessed on 2025 Aug 25].
 35. Bottani S, Burgos N, Maire A, *et al.* Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse. *Med Image Anal.* 2023;89:102903.
doi: 10.1016/j.media.2023.102903
 36. Tanveer H, Adam MA, Khan MA, Ali MA, Shakoor A. Analyzing the performance and efficiency of machine learning algorithms, such as deep learning, decision trees, or support vector machines, on various datasets and applications. *Asian Bull Big Data Manage.* 2024;3(2):126-136.
doi: 10.62019/abbdm.v3i2.83
 37. Pelletier ED, Jeffries SD, Song K, Hemmerling TM. Comparative analysis of machine-learning model performance in image analysis: The impact of dataset diversity and size. *Anesth Analg.* 2024;139(6):1332-1339.
doi: 10.1213/ane.00000000000007088
 38. Tougui I, Jilbab A, Mhamdi JE. Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthc Inform Res.* 2021;27(3):189-199.
doi: 10.4258/hir.2021.27.3.189
 39. Yuan H. Toward real-world deployment of machine learning for health care: External validation, continual monitoring, and randomized clinical trials. *Health Care Sci.* 2024;3(5):360-364.
doi: 10.1002/hcs2.114
 40. Toma M. *AI-Assisted Medical Diagnostics: A Clinical Guide to Next-Generation Diagnostics.* New York: Dawning Research Press; 2025.

Appendix

Appendix A

MRI Data Preprocessing and Feature Extraction

Require: MRI image dataset with organized folder structure (NonDemented, VeryMildDemented, MildDemented, ModerateDemented)

Ensure: Preprocessed features ready for classification

1. **Dataset Organization and Loading**
2. Define main directory containing class-specific subfolders
3. **for** binary classification **do**
4. Combine VeryMild, Mild, and Moderate classes into single "Demented" category
5. **end for**
6. Create image data store with folder-based label assignment
7. Shuffle data randomly for unbiased sampling
8. **Data Partitioning**
9. **if** using hold-out validation **then**
10. Split dataset: 70% training, 15% validation, 15% testing (stratified)
11. **else**
12. Set up k -fold cross-validation ($k \in \{5, 10\}$) with stratification
13. **end if**
14. **Image Preprocessing**
15. Resize all images to $224 \times 224 \times 3$ pixels
16. Convert grayscale images to RGB format
17. Create augmented image data stores for consistent preprocessing
18. **Deep Feature Extraction**
19. Load pre-trained GoogLeNet architecture
20. Extract features from 'pool5-drop_7x7_s1' layer (feature extraction layer)
21. Convert images to feature vectors using network activations
22. Output features as row vectors for classification

Appendix B

Classifier Training and Performance Evaluation

Require: Preprocessed features from Part 1 (**Appendix A**)

Ensure: Trained classifier and performance evaluation metrics

1. Classifier Training
2. **for** binary classification (Normal vs. Alzheimer's) **do**

3. Train Support Vector Machine (SVM) using `fitcsvm`
4. **end for**
5. **for** multi-class classification (4classes) **do**
6. Train Error-Correcting Output Codes (ECOC) classifier using `fitcecoc`
7. **end for**
8. **Model Evaluation**
9. Generate predictions on training, validation, and test sets
10. Compute confusion matrices for each evaluation set
11. Calculate accuracy metrics: $\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / \text{Total Samples}$
12. Visualize confusion matrices using confusion charts
13. **Learning Curve Analysis**
14. Define training set sizes
15. **for** each training size s **do**
16. Sample subset of training data proportional to s
17. Train classifier on subset
18. Evaluate on both training subset and validation set
19. Compute training and validation accuracies
20. Calculate loss as $\text{Loss} = 1 - \text{Accuracy}$
21. **end for**

Appendix C

Cross-Validation and Results Visualization

Require: Features and cross-validation setup

Ensure: Final performance metrics and visualizations

1. **Cross-Validation Process** (if applicable)
2. **for** each fold $i=1$ to k **do**
3. Extract training and test indices for current fold
4. **for** each training size **do**
5. Train classifier on fold-specific training subset
6. Evaluate on fold-specific validation set
7. Store accuracy and loss metrics
8. **end for**
9. Generate confusion matrix for current fold
10. **end for**
11. Compute mean accuracy and loss across all folds
12. **Results Visualization**
13. Plot learning curves showing training vs. validation accuracy
14. Plot loss curves showing training vs. validation loss
15. Display confusion matrices for performance assessment
16. Generate final performance metrics and model evaluation