# Classification of Alzheimer's Disease by Magnetic Resonance Imaging with an Ensemble Approach

**Gustavo S. Silva**[1] **, Omar A. C. Cortes**[2,1] **, Antonio F. L. Jacob Jr.**[1]

[1]Programa de Pós-graduação em Engenharia da Computação e Sistemas (PECS)
Universidade Estadual do Maranhão (UEMA)
Cidade Universitária Paulo VI – Caixa Postal 09 – São Luís, MA – Brazil

[2]Departamento de Computação (DComp) – Instituto Federal do Maranhão (IFMA)
Av. Getúlio Vargas, 04 – 65030-005 – São Luís, MA – Brazil

gustavosoares112@gmail.com, omar@ifma.edu.br

***Abstract.*** *Alzheimer's Disease (AD), affecting over 55 million people globally, demands reliable diagnostic tools. Single-model approaches using CNNs and traditional ML face critical limitations. This study proposes two frameworks: a stacking-CNN ensemble (VGG-16, ResNet-101, DenseNet-121) and two voting ML ensembles (Voting[all]: KNN, RF, SVC, LR, XGBoost; Voting[few]: KNN, RF, XGBoost). Evaluated on 6,400 MRIs, Voting[few] achieved the highest classification metrics (97.8% accuracy; 0.984 MCC; 93.8% F1macro), outperforming individual CNNs, validated through Friedman-Nemenyi tests. Results suggest, in this context, that simpler ML models might better capture the inherent characteristics of MRI data for AD diagnosis.*

## 1. Introduction

Alzheimer's Disease (AD) is the most prevalent form of dementia, characterized by a progression from episodic memory lapses to severe cognitive decline and functional disability [Alzheimer's & Dementia 2024, Planche et al. 2022]. Representing 60-70% of global dementia cases, AD affects more than 55 million people, with projections indicating a tripling by 2050 [World Health Organization 2023]. This scenario and the absence of curative therapies emphasize the need for enhanced diagnostic tools.

Structural magnetic resonance imaging (MRI) enables non-invasive detection of AD-related brain atrophy, particularly in the hippocampus and medial temporal lobe [Kumar et al. 2025]. Advances in Artificial Intelligence (AI), especially in Convolutional Neural Networks (CNNs), have optimized the classification of these images through automated feature extraction [Shastry 2024]. Architectures such as VGG-16, ResNet-101, and DenseNet-121 have shown prominence in identifying AD-specific pathological patterns [Silva et al. 2023]. Despite these advances, single-model CNN approaches face generalization limitations [Abimannan et al. 2023, Ganaie et al. 2022]. The variability in MRI data, such as differences in scanner protocols and image resolution, typically leads to inconsistent performance across datasets. Similarly, traditional Machine Learning (ML) models such as Logistic Regression (LR) and Support Vector Classifiers (SVC), while interpretable, fail to capture complex spatial patterns [Wadghiri et al. 2022, Ganaie et al. 2022].

This study addresses these challenges through two ensemble frameworks to overcome the generalization limitations of individual CNNs and the inability of

traditional ML models to capture complex spatial patterns. The approach combines: (1) a stacking-CNN ensemble combining VGG-16 [Simonyan and Zisserman 2015], ResNet-101 [He et al. 2015], and DenseNet-121 [Huang et al. 2018], exploring their complementary capabilities; and (2) a voting-based ML ensemble, combining K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), Support Vector Classifier (SVC), and eXtreme Gradient Boost (XGBoost), aiming for greater interpretability and robustness [Ganaie et al. 2022]. The remainder of this document details the related works im Section 2, materials and methods in Section 3, computational experiments, setup and results in Section 4, and the investigation conclusions and directions for future work in Section 5.

## 2. Related Works

The application of AI techniques in medicine has demonstrated significant advances in diagnosing and prognosis of various pathologies, particularly in MRI-based disease diagnosis. A meta-analysis conducted by [Zhang et al. 2023] on using DL in medical imaging revealed diagnostic accuracy comparable to or surpassing that of specialists across various imaging modalities, including X-ray, computed tomography, and MRI. Studies further demonstrate that models such as CNNs have been successfully employed in MR image classification for various applications, including brain tumor detection [Kaplan et al. 2023] and cardiac disease identification [Özaltın 2025, Ganaie et al. 2022]. Pre-trained models, such as DenseNet, MobileNet, ResNet, and ShuffleNet, have been utilized for feature extraction, proving effective due to their ability to adapt to new tasks with reduced training requirements [Özaltın 2025].

Ensemble methods, characterized by integrating multiple classifiers, have demonstrated efficacy in diagnostic medical image analysis. The ensemble combining DenseNet, VGG, and Xception showed superior performance in tasks such as COVID-19 detection in computed tomography, achieving accuracy, precision, and recall exceeding 97% [Silva et al. 2023]. Similarly, [Mahmud et al. 2025] combined VGG-19, ResNet-152, and EfficientNetB1 models, achieving 97.16% validation accuracy in AD classification through majority voting. In this context, [Chatterjee and Byun 2022] proposed a voting ensemble approach integrating Support Vector Machine (SVM), LR, Naïve-Bayes (NB), and KNN classifiers, demonstrating superior performance over individual models in dementia classification with 96.4% accuracy.

The issue of class imbalance in medical data mining arises when skewed distributions introduce bias into model learning, particularly affecting minority classes, which often signify critical disease states [Salmi et al. 2024, Lopes et al. 2024]. In the past decade, various strategies have been implemented to address imbalance in medical data, including pre-processing techniques, learning algorithms, and hybrid methods [Salmi et al. 2024]. The study by [Ahmed et al. 2022] demonstrated that the implementation of the Adaptive Synthetic Sampling (ADASYN) method, which generates synthetic instances for minority classes during the classification of Alzheimer's Disease (AD) stages, considerably enhances model performance. To assess models under these circumstances, the Matthews Correlation Coefficient (MCC) is identified as one of the nine primary classification metrics in the literature [Salmi et al. 2024]. For instance, the research by [Lopes et al. 2024] employs the MCC to evaluate outcomes when utilizing an approach for the detection of *leishmania* amastigotes through the Deep Metric Learning

method.

In summary, the literature demonstrates a growing interest in applying ensemble models for MRI classification, particularly in distinguishing dementia stages, such as Alzheimer's disease. However, gaps remain to be explored, such as the need for validating these models across diverse and heterogeneous clinical datasets to ensure their generalization and practical applicability.

## 3. Material and Methods

### 3.1. Convolutional Neural Networks

CNNs are a class of DL models that have gained prominence in image analysis due to their ability to learn hierarchical features from data automatically. In medical imaging applications, CNNs have demonstrated a remarkable capacity to automatically learn relevant features from complex visual data, making them particularly suitable for diagnostic support systems and disease classification tasks [Mahmud et al. 2025].

CNNs integrate three main components: convolutional layers, which apply learnable filters to extract local features (e.g., textures, edges) through spatial convolutions; activation functions (e.g., ReLU), which introduce non-linearity to increase model expressiveness; and pooling layers, which reduce spatial dimensionality while preserving essential features [Krichen 2023, Alzubaidi et al. 2021, Saleem et al. 2022]. High-level features are aggregated by fully connected layers for classification or regression, as demonstrated in architectures such as AlexNet [Krizhevsky 2014] and VGG [Simonyan and Zisserman 2015]. This hierarchical structure enables CNNs to learn discriminative representations directly from raw pixel data, validated across various image domains [Krichen 2023].

### 3.2. Ensemble Learning

Ensemble models, which combine predictions from multiple base learners to increase generalization and accuracy, have become a formidable framework within both traditional ML and DL paradigms. Based on fundamental theories such as bagging [Breiman 1996] and boosting [Freund and Schapire 1997], ensemble methods effectively reduce the biases and variance associated with individual models, particularly in intricate domains such as medical imaging. Hybrid approaches that integrate DL architectures, such as CNNs, with conventional ML algorithms like support vector SVM and XGBoost harness their complementary strengths: CNNs to extract hierarchical features, while ML models enhance interpretability and efficiency [Dietterich 2000, Sharmin et al. 2023].

The architecture of the ensemble models can vary significantly depending on the combination of algorithms employed. Common strategies include bagging, boosting, and stacking [Patel et al. 2011]. Bagging methods, such as Random Forests, involve training multiple instances of the same model on different data subsets and averaging their predictions to reduce variance [Breiman 1996, Dietterich 2000] and boosting techniques, such as XGBoost, train models sequentially, focusing on errors made by previous models to improve overall accuracy. Stacking, on the other hand, combines multiple models (including ML and DL models) and uses a meta-learner to make final predictions based on the output of the base models [Dietterich 2000, Patel et al. 2011].

### 3.3. Ensemble Pretrained CNNs and ML models

In this study, three pre-trained CNN architectures were selected due to their complementary properties: ResNet-101, which employs residual connections to mitigate the issue of gradient vanishing [He et al. 2015]; DenseNet-121, characterized by its dense connections that facilitate feature reuse [Huang et al. 2018]; and VGG-16, a more compact architecture utilizing 3×3 convolutions for structured feature extraction [Simonyan and Zisserman 2015]. These architectures are integrated through a stacking approach (Stacking-CNN), in which their outputs are combined into a single feature vector to serve as input for a single-layer neural network responsible for the final classification [Dietterich 2000]. The Stacking-CNN diagram is presented in Figure 1.

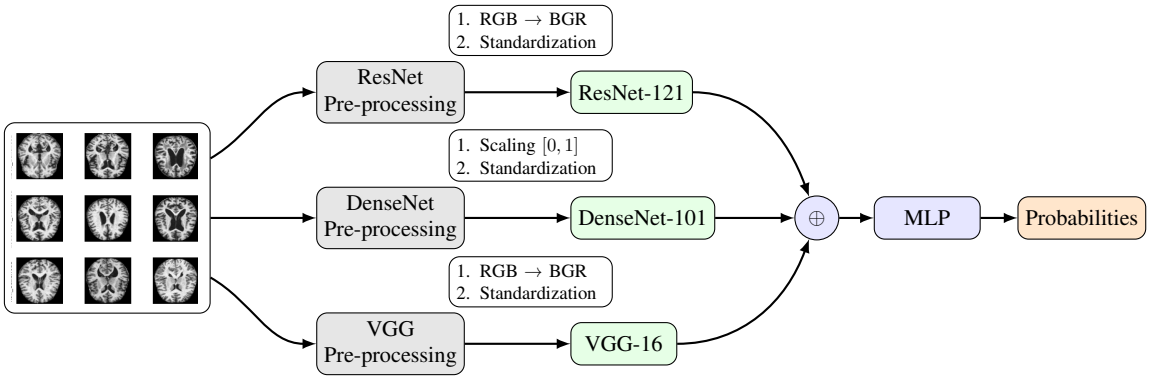



**Figure 1. Workflow of the STACKING-CNN Model**

In addition to the method involving CNN architectures, two voting-based ensemble approaches were incorporated into the framework, as shown in Figure 2.

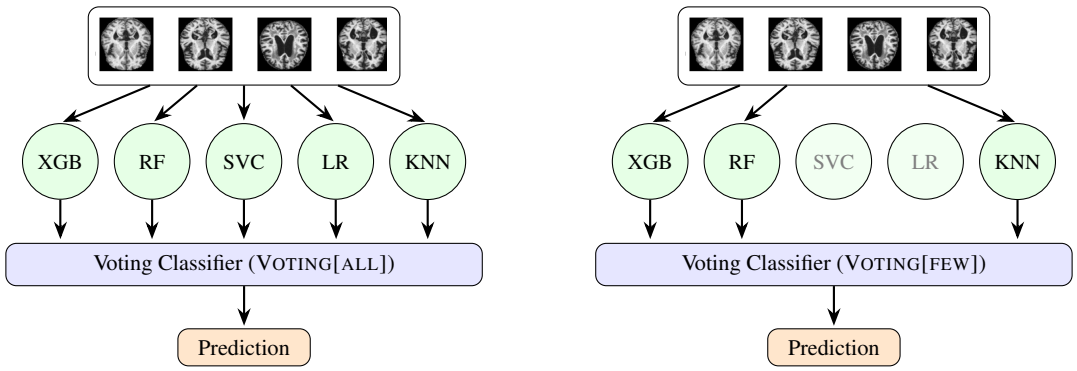



**Figure 2. Comparative Voting Ensemble Architectures**

The first ensemble (Voting[all]), based on the work of [Chatterjee and Byun 2022], integrates five classifiers: KNN, LR, RF, SVC, and XGBoost. The second ensemble (Voting[few]) presents a refined version, retaining only KNN, RF, and XGBoost. Despite the literature on techniques such as model pruning [Bhatnagar et al. 2014, Bhardwaj and Bhatnagar 2015, Alzubi et al. 2020], the model selection was based on the exclusion of classifiers with lower individual performance through trial and error.

## 4. Computational Experiments

### 4.1. Setup

The experiments[1] were conducted in the *Google Colab* environment with an A100 GPU. The DenseNet-121, ResNet-101, and VGG-16 architectures, using *ImageNet* [Deng et al. 2009] weights, were trained with Adam optimizer and Categorical Crossentropy loss function, selected for their suitability for multiclass classification tasks. For hyperparameters, the learning rate was set to $10^{-3}$ for 50 epochs, with a reduction strategy of 10% if reaching a *plateau* after 10 evaluation epochs.

For the Voting ensemble models, parameter selection followed the methodology of [Chatterjee and Byun 2022]. Tree-based models (XGBoost and RF) were instantiated with 120 estimators, and the number of neighbors for KNN was set to 5. The parameters of the remaining models were kept unchanged.

The training phase employed a batch size of 32 and implemented a stratified k-fold cross-validation methodology with five folds. This approach aimed to assess the architectures' performance across different dataset partitions, thereby reducing result variability and enhancing the reliability of the performance analysis. To address class imbalance within the dataset, the ADASYN technique was applied to the training data in each fold, adaptively oversampling the minority class to create a more balanced training distribution.

This investigation's results were evaluated using appropriate metrics for the classification task, employing accuracy, MCC, and *F1*-macro. Accuracy quantifies the overall rate of correct classifications, MCC offers a balanced evaluation regardless of class imbalance, while *F1*-macro weighs the relationship between precision and *recall* depending on the class distribution in the dataset.

We adopted the methodology described by [Demsar 2006] for statistical significance evaluation, which involves applying the Friedman chi-square test after validating normality and variance assumptions through Shapiro-Wilk and Levene tests, respectively. The Friedman test is conducted under the null hypothesis ($H_0$) that the architectures are equivalent and show no significant performance differences. If the null hypothesis is rejected, we proceed with Nemenyi's *post-hoc* test to verify which architecture pairs differ from each other. This statistical approach was selected for its suitability in comparing multiple classifiers ($> 2$) in independent sets.

### 4.2. Dataset

This investigation utilized the ALZHEIMER MRI dataset [Falah.G.Salieh 2023][2], accessed through the HuggingFace library. The dataset contains 6,400 magnetic resonance images, which are categorized into the classes classes Non-Demented (ND, 3,200 images), Very Mild Demented (VMD, 2,240 images), Mild Demented (MD, 896 images), and Moderated Demented (MOD, 64 images). A representative sample from each class in the dataset is shown in Figure 3.

[1]The code used for the experiments is publicly available at `https://github.com/gustaph/SBCAS-2025-AD-Classification-using-Ensemble-Methods`

[2]`https://huggingface.co/datasets/Falah/Alzheimer_MRI`

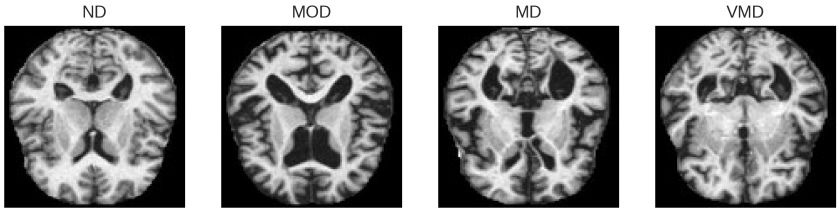



**Figure 3. Examples of dataset classes**

As the figure shows, this is a grayscale dataset with only one channel. In addition to being imbalanced, the images exhibit high variability in the shapes and structures of the brain in the MRIs.

### 4.3. Pre-Processing

Data preprocessing, an essential step to optimize model performance [Fan et al. 2021], was implemented following a specific transformation flow for image data and labeling.

As part of image pre-processing, we implemented the channel replication technique [Xie and Richmond 2019], where the single dimension of grayscale images is replicated three times. This approach was necessary to ensure compatibility with the pre-trained architectures used, which were initially designed to process three-channel RGB images.

Additionally, the model implementations suggest using specific preprocessing functions before feeding images into the models. The process consists of pixel and channel normalization for DenseNet. For VGG and ResNet architectures, the process involves RGB to BGR channel conversion and channel normalization.

Finally, the image classes were transformed into binary vectors to accommodate the chosen loss function for the experiment. No image preprocessing was performed for the voting models.

## 5. Results

This section outlines the results obtained from the experimental methodology. To clarify the learning process, Figure 4 illustrates the trends observed in training accuracy and loss metrics throughout the model training period.

The Stacking-CNN (S-CNN) achieved about 80% validation accuracy with converging curves, indicating good generalization. ResNet-101 (RN), DenseNet-121 (DN), and VGG-16 (VGG) showed 65% accuracy, with diverging curves and high validation loss, indicating overfitting. S-CNN's fluctuations between epochs 10 and 25, can be attributed to the variability introduced by ADASYN-generated synthetic data for the minority class. The subsequent stabilization of S-CNN indicates that the ensemble approach successfully generalizes with integrated augmented data, showing the effectiveness of the CNN ensemble approach, which, through model combination, achieved better predictive performance in classifying Alzheimer's patient dementia levels than individual models, while individual models struggled with overfitting on the combined original and synthetic data.

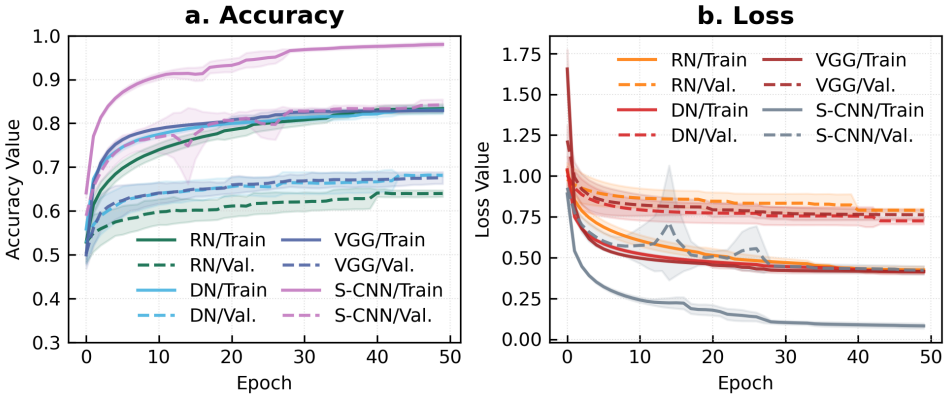



**Figure 4. Evolution of a) accuracy and b) loss during training**

Subsequent to the training phase conducted via 5-fold cross-validation, the quantitative performance of each model was assessed. These metrics are presented in Table 1.

**Table 1. Classification results for the models**

| Model | Accuracy | MCC | F1-Macro | Precision | Recall |
|---|---|---|---|---|---|
| **DenseNet-121** | 0.681 (± 0.003) | 0.488 (± 0.007) | 0.697 (± 0.022) | 0.692 (± 0.016) | 0.708 (± 0.028) |
| **ResNet-101** | 0.640 (± 0.007) | 0.399 (± 0.016) | 0.550 (± 0.036) | 0.558 (± 0.040) | 0.549 (± 0.034) |
| **VGG-16** | 0.675 (± 0.014) | 0.481 (± 0.013) | 0.667 (± 0.031) | 0.659 (± 0.025) | 0.688 (± 0.053) |
| **Stacking-CNN** | 0.842 (± 0.013) | 0.742 (± 0.019) | 0.862 (± 0.014) | 0.879 (± 0.017) | 0.855 (± 0.028) |
| **KNN** | 0.984 (± 0.003) | 0.973 (± 0.005) | 0.984 (± 0.004) | 0.985 (± 0.006) | 0.984 (± 0.009) |
| **LR** | 0.858 (± 0.011) | 0.767 (± 0.018) | 0.872 (± 0.026) | 0.874 (± 0.018) | 0.871 (± 0.037) |
| **RF** | 0.938 (± 0.006) | 0.898 (± 0.009) | 0.891 (± 0.050) | 0.965 (± 0.003) | 0.850 (± 0.062) |
| **XGBoost** | 0.978 (± 0.004) | 0.964 (± 0.006) | 0.961 (± 0.025) | 0.987 (± 0.003) | 0.942 (± 0.041) |
| **SVC** | 0.829 (± 0.017) | 0.716 (± 0.028) | 0.849 (± 0.028) | 0.883 (± 0.015) | 0.826 (± 0.039) |
| **Voting[few]** | **0.990 (± 0.003)** | **0.984 (± 0.004)** | **0.988 (± 0.009)** | **0.994 (± 0.001)** | **0.982 (± 0.016)** |
| **Voting[all]** | 0.968 (± 0.003) | 0.947 (± 0.005) | 0.971 (± 0.009) | 0.981 (± 0.002) | 0.963 (± 0.016) |
| **Mean (± std.)** | 0.853 (± 0.007) | 0.760 (± 0.012) | 0.845 (± 0.023) | 0.860 (± 0.013) | 0.838 (± 0.033) |

Among traditional ML models, KNN and XGBoost achieved high MCC values of 0.973 (± 0.005) and 0.964 (± 0.006), respectively, matching their accuracy ranks (98.4% and 97.8%). Their precision (98.5% and 98.7%) and recall (98.4% and 94.2%) indicate very low false-positive and false-negative rates, making them reliable for detecting both positive and negative cases. In DL models, however, MCC diverges; e.g., ResNet-101 has an accuracy of 64%≈, F1-macro of 70%≈, and MCC of 0.40≈, highlighting an imbalance between true positive and negative rates. DenseNet-121 and VGG-16 also underperformed, with ResNet-101 showing the worst MCC (0.399±0.016), indicating high misclassification.

Notable is the performance of the reduced ensemble Voting[few], which combines only the most effective models, achieving the best metrics. This behavior suggests that model selection for ensemble composition can be more effective than including all available classifiers. The models have an average accuracy of 85.3% and an MCC of 0.760, indicating good agreement with some variability in reliability. The average F1-macro score reflects moderate class balance, and mean precision slightly exceeds mean

recall, implying fewer false positives than false negatives on average.

The statistical analysis followed three sequential steps to determine the most appropriate comparison test. First, the Shapiro-Wilk test was applied to verify data normality. The results confirmed normal distribution for all models ($p > 0.05$). Next, variance homogeneity was evaluated through the Levene test, whose results are presented in Table 2.

**Table 2. Results of Levene's statistical test**

| | |
|---|---|
| $W$ | 1.364 |
| p-value | 0.228 |
| Reject H0 | ✗ |

The Levene's test affirmed the homogeneity of variance ($p = 0.228$) as presented in Table 2. This finding permitted the subsequent execution of the Friedman chi-square test, which was based on the F1 score in the validation set (see Table tab:fried).

**Table 3. Friedman $\chi^2$ statistical test**

| | |
|---|---|
| $\chi^2$ | 49.02 |
| p-value | $3.74 \times 10^{-7}$ |
| Reject H0 | ✓ |

The Friedman test revealed significant performance differences between models ($\chi^2 = 49.02$, $p = 3.74 \times 10^{-7} < \alpha = 0.05$), thus rejecting the null hypothesis that the models exhibit equivalent performance. To specifically identify which model pairs show statistically significant differences, a *post-hoc* analysis was conducted using the Nemenyi test. The results of this analysis are presented in Figure 5.

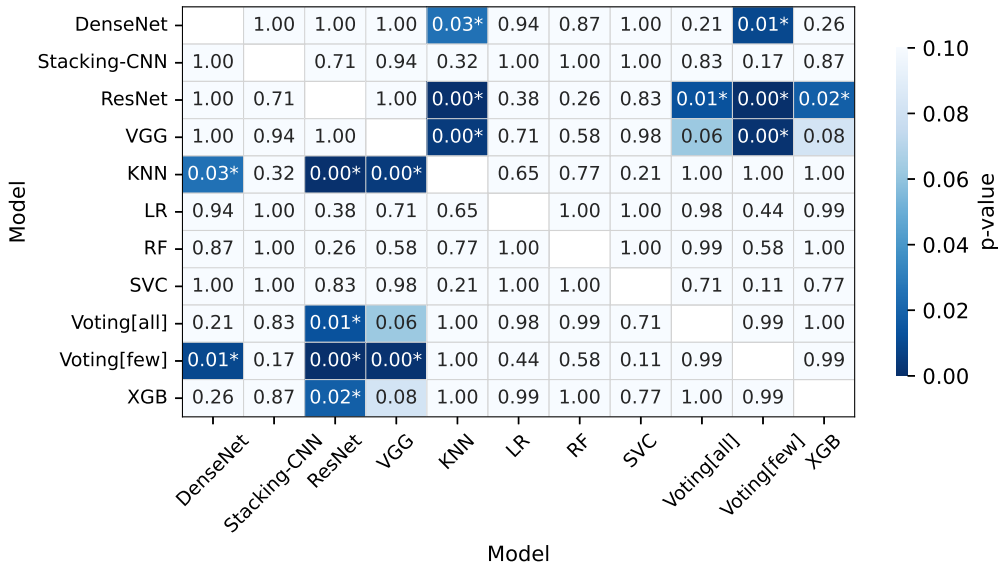


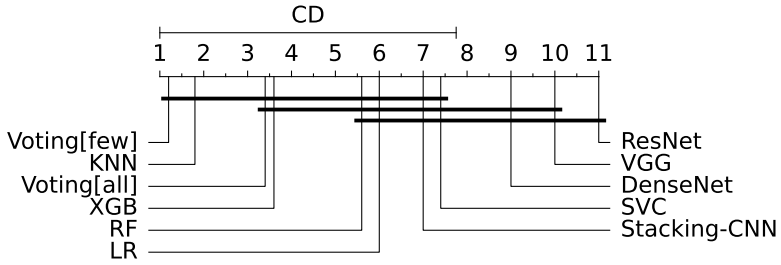

**Figure 5. Nemenyi *post-hoc* tests**

The Nemenyi *post-hoc* test shows no significant difference in average ranks between Voting[all], Voting[few], KNN, LR, RF, and XGB ($p > 0.05$). Between

the ensembles themselves, Voting[few] shows no significant difference from Voting[all] ($p = 0.99$), indicating that a smaller voter pool does not affect rank performance for this case. Furthermore, neither voting model demonstrates a statistically significant improvement over the DL-only stacking approach.

The Nemenyi test results are reinforced by Figure 6, which presents the Critical Differences (CD) diagram. This diagram provides a complementary visualization by grouping models into statistically equivalent performance clusters.

**Figure 6. Critical Differences between models**




The CD diagram groups the models into three performance clusters with statistically equivalent performance: superior (Voting[few], KNN, XGB, Voting[all]), intermediate (RF, LR, Stacking-CNN), and inferior (DenseNet, VGG, ResNet, and SVC). This analysis corroborates the previous accuracy, MCC, and F1-macro results, confirming the superiority of traditional models and ensembles over individual DL architectures, with emphasis the voting models, which consistently demonstrated significant differences when contrasted with CNN models.

## 6. Conclusions and Future Work

Alzheimer's Disease is a progressive neurodegenerative disorder affecting millions worldwide that requires accurate early diagnosis to enable timely intervention. This study investigated the effectiveness of ensemble approaches for AD classification using MRI images, comparing traditional machine learning ensembles with deep learning models.

The results demonstrated that the reduced voting ensemble (Voting[few]), which integrates KNN, Random Forest and XGBoost, achieved the highest classification metrics (97.8% accuracy; 0.984 MCC; 93.8% F1-macro), significantly surpassing individual CNN architectures. Friedman and Nemenyi analyses corroborate that its performance advantage over all DL models is statistically significant.

The superior performance of simpler, traditional ML models over complex DL architectures in this context suggests that these algorithms might better capture the inherent characteristics of MRI data for AD classification. This finding challenges the common assumption that DL models necessarily provide better results for medical image analysis tasks, highlighting the importance of considering multiple approaches when developing diagnostic support systems.

Future work should focus on validating these findings across different MRI datasets and exploring the potential of hybrid approaches that combine CNNs' feature extraction capabilities with the classification performance of traditional ML ensembles.

Further research could explore dynamic ensemble pruning techniques to optimize model selection and computational efficiency.

## References


Abimannan, S., El-Alfy, E.-S. M., Chang, Y.-S., Hussain, S., Shukla, S., and Satheesh, D. (2023). Ensemble multifeatured deep learning models and applications: A survey. *IEEE Access*, 11:107194–107217.

Ahmed, G., Er, M. J., Fareed, M. M. S., Zikria, S., Mahmood, S., He, J., Asad, M., Jilani, S. F., and Aslam, M. (2022). Dad-net: Classification of alzheimer's disease using adasyn oversampling technique and optimized neural network. *Molecules*, 27(20):7085.

Alzheimer's & Dementia (2024). 2024 Alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 20(5):3708–3821.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53.

Alzubi, O. A., Alzubi, J. A., Alweshah, M., Qiqieh, I., Al-Shami, S., and Ramachandran, M. (2020). An optimal pruning algorithm of classifier ensembles: Dynamic programming approach. *Neural Computing and Applications*, 32(20):16091–16107.

Bhardwaj, M. and Bhatnagar, V. (2015). Towards an optimally pruned classifier ensemble. *International Journal of Machine Learning and Cybernetics*, 6(5):699–718.

Bhatnagar, V., Bhardwaj, M., Sharma, S., and Haroon, S. (2014). Accuracy–diversity based pruning of classifier ensembles. *Progress in Artificial Intelligence*, 2(2-3):97–111.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Chatterjee, S. and Byun, Y.-C. (2022). Voting Ensemble Approach for Enhancing Alzheimer's Disease Classification. *Sensors*, 22(19):7661.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, and Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL. IEEE.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In Goos, G., Hartmanis, J., and Van Leeuwen, J., editors, *Multiple Classifier Systems*, volume 1857, pages 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg.

Falah.G.Salieh (2023). Alzheimer MRI dataset.

Fan, C., Chen, M., Wang, X., Wang, J., and Huang, B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research*, 9.

Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Ganaie, M., Hu, M., Malik, A., Tanveer, M., and Suganthan, P. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2018). Densely Connected Convolutional Networks.

Kaplan, E., Baygin, M., Barua, P. D., Dogan, S., Tuncer, T., Altunisik, E., Palmer, E. E., and Acharya, U. R. (2023). ExHiF: Alzheimer's disease detection using exemplar histogram-based features with CT and MR images. *Medical Engineering & Physics*, 115:103971.

Krichen, M. (2023). Convolutional neural networks: A survey. *Computers*, 12(8):151.

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks.

Kumar, A., Sidhu, J., Lui, F., and Tsao, J. W. (2025). Alzheimer Disease. In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

Lopes, C. E. F., Lisboa, E., Ribeiro, Y., and Queiroz, F. (2024). A patch-based microscopic image analysis for visceral leishmaniasis screening using a deep metric learning approach. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 166–177. SBC.

Mahmud, T., Aziz, M. T., Uddin, M. K., Barua, K., Rahman, T., Sharmen, N., Shamim Kaiser, M., Sazzad Hossain, Md., Hossain, M. S., and Andersson, K. (2025). Ensemble Learning Approaches for Alzheimer's Disease Classification in Brain Imaging Data. In Mahmud, M., Kaiser, M. S., Bandyopadhyay, A., Ray, K., and Al Mamun, S., editors, *Proceedings of Trends in Electronics and Health Informatics*, volume 1034, pages 133–147. Springer Nature Singapore, Singapore.

Özaltın, Ö. (2025). Early Detection of Alzheimer's Disease from MR Images Using Fine-Tuning Neighborhood Component Analysis and Convolutional Neural Networks. *Arabian Journal for Science and Engineering*.

Patel, H., Ganatra, A. P., Bhensdadia, C. K., and Kosta, Y. (2011). Experimental study and review of boosting algorithms. *Artificial Intelligent Systems and Machine Learning*, 3:31–41.

Planche, V., Manjon, J. V., Mansencal, B., Lanuza, E., Tourdias, T., Catheline, G., and Coupé, P. (2022). Structural progression of alzheimer's disease over decades: the mri staging scheme. *Brain Communications*, 4(3).

Saleem, M. A., Senan, N., Wahid, F., Aamir, M., Samad, A., and Khan, M. (2022). Comparative Analysis of Recent Architecture of Convolutional Neural Network. *Mathematical Problems in Engineering*, 2022:1–9.

Salmi, M., Atif, D., Oliva, D., Abraham, A., and Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, 57(10).

Sharmin, S., Ahammad, T., Talukder, M. A., and Ghose, P. (2023). A Hybrid Dependable Deep Feature Extraction and Ensemble-Based Machine Learning Approach for Breast Cancer Detection. *IEEE Access*, 11:87694–87708.

Shastry, K. A. (2024). Deep Learning-Based Classification of Alzheimer's Disease Using MRI Scans: A Customized Convolutional Neural Network Approach. *SN Computer Science*, 5(7):917.

Silva, L. F. D. J., Cortes, O. A. C., and Diniz, J. O. B. (2023). A novel ensemble CNN model for COVID-19 classification in computerized tomography scans. *Results in Control and Optimization*, 11:100215.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition.

Wadghiri, M., Idri, A., El Idrissi, T., and Hakkoum, H. (2022). Ensemble blood glucose prediction in diabetes mellitus: A review. *Computers in Biology and Medicine*, 147:105674.

World Health Organization (2023). Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia.

Xie, Y. and Richmond, D. (2019). Pre-training on grayscale imagenet improves medical image classification. In *Computer Vision – ECCV 2018 Workshops*, page 476–484. Springer International Publishing.

Zhang, J., Li, Z., Lin, H., Xue, M., Wang, H., Fang, Y., Liu, S., Huo, T., Zhou, H., Yang, J., Xie, Y., Xie, M., Lu, L., Liu, P., and Ye, Z. (2023). Deep learning assisted diagnosis system: improving the diagnostic accuracy of distal radius fractures. *Frontiers in Medicine*, 10.