Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 14 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i14.57471

PAPER

Enhanced Alzheimer's Diagnosis Using Multimodal Data: A Comparative Study of CNN Architectures

Lailil Muflikhah (✉), Galih Restu Baihaqi, Shafatyra Redhita Shalsadilla, Achmad Ridok, Sri Soenarti

Brawijaya University, Malang, Indonesia

lailil@ub.ac.id

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that leads to severe cognitive decline, making early and accurate diagnosis essential for effective patient management. Traditional diagnostic approaches often rely on unimodal imaging data such as magnetic resonance imaging (MRI) or positron emission tomography (PET) scans, yet these methods are insufficient to capture the heterogeneous and multimodal characteristics of AD biomarkers. Similarly, conventional convolutional neural network (CNN) models demonstrate strong image recognition capabilities but remain limited in integrating diverse sources of clinical evidence. To address this gap, this study proposes an enhanced multimodal diagnostic framework that combines MRI, PET, and clinical metadata to provide a more holistic representation of disease progression. The framework is evaluated through a comparative analysis of state-of-the-art CNN architectures, including densely connected convolutional networks (DenseNet), residual network (ResNet), InceptionNet, VGG, MobileNet, and EfficientNet. Experiments were conducted on the ADNI dataset, which includes 372 subjects and a total of 63,777 image slices. The results clearly demonstrate that multimodal models outperform unimodal counterparts in classification performance. EfficientNetB1 emerged as the best-performing model, achieving 98.6% accuracy, precision, recall, and F1-score, highlighting the significant contribution of clinical metadata integration. However, this superior accuracy comes with higher computational requirements, as EfficientNetB1 required 22.16 seconds for prediction with a memory load of 31.6 GB. In contrast, lightweight models such as MobileNet offered faster inference speeds but sacrificed accuracy, reaching only about 76%. These findings emphasize the critical trade-off between computational efficiency and diagnostic performance in real-world clinical scenarios. Overall, the study provides strong evidence that multimodal CNN-based architectures offer robust and accurate tools for AD detection, while also underscoring the need to balance model complexity with resource constraints in practical healthcare implementation.

KEYWORDS

Alzheimer's disease (AD), classification, convolutional neural network (CNN) architecture, multimodal

Muflikhah, L., Baihaqi, G. R., Shalsadilla, S. R., Ridok, A., Soenarti, S. (2025). Enhanced Alzheimer's Diagnosis Using Multimodal Data: A Comparative Study of CNN Architectures. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(14), pp. 182–198. https://doi.org/10.3991/ijoe.v21i14.57471

Article submitted 2025-07-01. Revision uploaded 2025-10-02. Final acceptance 2025-10-02.

1 INTRODUCTION

Alzheimer's disease (AD) is a debilitating neurodegenerative disorder and the leading cause of dementia worldwide, affecting the majority of approximately 50 million people living with dementia today [1]. The disease progressively impairs cognitive function, memory, and daily activities, eventually leading to total dependence and death [2]. With a rapidly aging global population, the burden of AD is expected to increase significantly in the coming decades, posing substantial challenges to healthcare systems [3], [4]. This growing prevalence underscores the urgent need for early and accurate diagnostic tools that can enable timely intervention, slow disease progression, and improve patient outcomes.

In this study, we emphasize a novel contribution compared to previous works. While prior studies have combined imaging with demographic or clinical data, our work provides a systematic benchmarking of six widely used convolutional neural network (CNN) architectures (densely connected convolutional networks (DenseNet), residual network (ResNet), InceptionNet, VGG, MobileNet, and EfficientNet) in a multimodal framework integrating magnetic resonance imaging (MRI), positron emission tomography (PET), and clinical metadata. In addition, unlike earlier works that mainly focused on accuracy, we analyze the trade-off between diagnostic accuracy and computational efficiency (training time, inference latency, and memory load). Finally, we highlight how even simple clinical metadata such as age and gender significantly enhances CNN-based classification, underscoring the importance of lightweight but clinically relevant features for early Alzheimer's diagnosis.

The pathological processes underlying AD often begin years, or even decades, before clinical symptoms become apparent. This preclinical stage provides an important window for therapeutic intervention, as current treatments are more effective in slowing progression if administered early [5]–[7]. However, diagnosing AD at an early stage remains a significant challenge. Conventional diagnostic approaches rely on a combination of clinical assessment, neuropsychological tests, and imaging techniques such as MRI and PET. While these modalities provide valuable insights, they often require specialized expertise, are resource-intensive, and have inter-observer variability [8]–[10]. Furthermore, single modality analysis often fails to capture the multifaceted nature of AD, limiting its diagnostic accuracy [11], [12].

In recent years, advances in artificial intelligence (AI), particularly deep learning (DL), have changed the landscape of medical image analysis and disease diagnosis [13]–[17]. DL architectures such as DenseNet, ResNet, InceptionNet, VGG, MobileNet, and EfficientNet have achieved state-of-the-art performance across a wide range of tasks, including image classification. Their ability to automatically learn hierarchical feature representations from raw data makes them particularly suitable for analyzing complex data sets, such as medical images [18]–[22]. Beyond single modality analysis, DL has also shown tremendous potential in integrating multimodal data, combining insights from multiple sources to improve prediction accuracy. For AD, this could mean utilizing neuroimaging data alongside tabular clinical metadata (e.g., age, gender) to create a more holistic diagnostic framework [23]–[26].

Multimodal data integration is particularly important in the context of AD, as the disease manifests through structural, functional, and metabolic changes in the brain that are influenced by patient-specific factors [27], [28]. For example, MRI can reveal atrophy in areas such as the hippocampus, while PET imaging can highlight abnormal amyloid-beta deposition. Simultaneously, clinical data such as age and gender provide additional context that can refine diagnostic decisions [29]–[32].

However, integrating these disparate data sources poses significant challenges. Heterogeneity in data formats, missing information, and the complexity of cross-modal relationships necessitate the use of sophisticated computational models capable of learning from and harmonizing diverse data.

Despite these challenges, few studies have systematically explored the comparative performance of DL architectures for multimodal AD diagnosis [33], [34]. While models such as DenseNet, ResNet, InceptionNet, VGG, MobileNet and EfficientNet have shown superior performance in individual image-based tasks, their relative effectiveness in a multimodal setting remains poorly explored. Furthermore, considerations such as overfitting, interpretability, and computational efficiency have not been adequately addressed in previous research, which limits the clinical applicability of these approaches.

To address this gap, this study conducts a comprehensive evaluation of four leading DL architectures—DenseNet, ResNet, InceptionNet, VGG, MobileNet, and EfficientNet—for AD classification using multimodal data. Our approach integrates MRI neuroimaging modalities with clinical metadata to assess the impact of multimodal fusion on diagnostic performance. Specifically, we evaluate the performance of each model based on several metrics, including accuracy, precision, recall, F1 score, and inference time, and memory load while addressing the challenge of model interpretability.

2 RELATED WORK

DenseNet is a DL architecture that maximizes feature utilization by connecting each layer to all previous layers in the network. This unique design allows DenseNet to improve parameter efficiency, reduce overfitting, and strengthen gradient propagation. In the context of medical image analysis, DenseNet has shown excellent performance in detecting various conditions, including Alzheimer's diagnosis through MRI analysis. Wang et al. reported in their study that DenseNet achieved 97% accuracy in processing 3D Alzheimer's data, indicating that DenseNet performs quite well in processing medical images [35]. The study also confirmed that the DenseNet architecture is able to capture subtle structural changes in the brain, such as hippocampus atrophy, which is a key indicator in Alzheimer's diagnosis. This makes DenseNet a promising approach in the development of DL-based diagnosis systems. ResNet is one of the most influential DL architectures in recent years and has made significant contributions in AD classification. ResNet uses residual blocks with shortcut connections to overcome the vanishing gradient problem that often arises in very deep networks, allowing for more stable and accurate model training. Research conducted by Li and Yang (2021) compared the performance of ResNet with support vector machine (SVM) in AD classification [36]. The results showed that ResNet achieved a higher accuracy of 95% compared to SVM which only achieved 90% accuracy, indicating the superiority of ResNet in analyzing complex data such as medical images. Another DL model, InceptionNet, which was originally introduced in GoogleNet, adopts the Inception module design that allows the network to capture features from multiple scales. In the task of Alzheimer's diagnosis, InceptionNet's ability to incorporate multi-scale information is crucial in identifying structural and metabolic changes of the brain as a whole. Research conducted by Shamrat (2023) showed that InceptionNetV3 performed better than MobileNet and VGG16, with an accuracy of 96% using data from ADNI [37]. This improvement in accuracy was

achieved through the application of data preprocessing using the contrast limited adaptive histogram equalization (CLAHE) method, which improves image quality for deeper analysis.

Visual geometry group network (VGG) is known for the simplicity of its architecture that uses stacked convolutional layers with small kernels. VGG has been used extensively in medical image analysis, including Alzheimer's diagnosis. This simple yet effective network structure makes it excellent at recognizing atrophic patterns in the brain. However, one of the major drawbacks of VGG is its high memory consumption, which becomes a challenge when used on high-resolution data such as MRI. Research conducted by Song showed that VGG16 was able to achieve 96.4% accuracy in Alzheimer's classification using 3D data, confirming the potential of this architecture in medical image analysis despite its limitations on computational efficiency [38]. Another model, MobileNet, is designed as an efficient architecture with limited resources and has the advantage of fast inference and computational efficiency. MobileNet also performs well in AD classification. In a study conducted by Alwakid (2024), MobileNetV2 achieved an accuracy of 80.31% before preprocessing the data [39]. After applying preprocessing using the CLAHE method, MobileNet's accuracy increased significantly to 92.34%, demonstrating the effectiveness of preprocessing in improving the performance of this model for medical image analysis. Another model, EfficientNet, works by integrating a compound scaling strategy to optimize the size, depth, and resolution of the network simultaneously. This approach has proven to be state-of-the-art in various image classification tasks, including Alzheimer's diagnosis. EfficientNet's ability to balance computational efficiency with high performance makes it particularly suitable for MRI data analysis. Research conducted by Rao (2023) showed that EfficientNet was able to achieve up to 99% accuracy in AD segmentation and classification [40]. However, despite its high accuracy, EfficientNet has the disadvantage of a relatively long runtime, which can be an obstacle in clinical applications that require high speed. Beyond the conventional CNN-based approaches, recent studies have also introduced innovative deep learning frameworks for medical diagnosis. For instance, Al-Nawashi et al. (2024) proposed a machine learning-based technique for breast cancer detection that demonstrated the potential of integrating clinical and imaging features for improved performance [41]. Similarly, Gharaibeh et al. (2023) developed a Swin Transformer-based segmentation and feature fusion method for AD, highlighting the effectiveness of transformer models and multi-scale feature integration for advancing AD diagnosis [42]. These studies reinforce the importance of multimodal integration and advanced architectures, aligning with our work that systematically compares CNN-based multimodal frameworks for Alzheimer's disease.

Based on previous studies, various DL architectures such as DenseNet, ResNet, InceptionNet, VGG, MobileNet, and EfficientNet have shown great potential in Alzheimer's diagnosis. Each model has unique advantages, ranging from parameter efficiency in DenseNet, learning stability in ResNet, multi-scale capability in InceptionNet, architectural simplicity in VGG, and computational efficiency in MobileNet, to state-of-the-art performance in EfficientNet. In the context of multimodal data, the potential of these architectures is increasingly relevant as they enable the integration of different types of data, such as MRI and clinical metadata, to improve the accuracy and sensitivity of diagnosis. However, challenges such as data format compatibility, preprocessing requirements, and computational efficiency remain major concerns. Therefore, an in-depth evaluation of the performance of these models in a multimodal environment is required to identify the best approach that can support Alzheimer's diagnosis more accurately and efficiently.

3 METHODOLOGY

3.1 Dataset

The dataset used in this study originates from the Image & Data Archive (IDA), managed by the Laboratory of Neuro Imaging (LONI) at the University of Southern California. IDA serves as the custodian of the Alzheimer's Disease Neuroimaging Initiative (ADNI), a large-scale effort to collect and provide neuroimaging data for AD research. This multimodal dataset comprises 3D images such as MRI and PET scans, as well as non-image data, including sex and age. The dataset includes a total of 372 individuals, whose data have been carefully collected and processed.

To facilitate analysis, the 3D images of everyone were sliced into 2D sections, resulting in a total of 63,777 images (see Figure 1). This slicing simplifies data representation and optimizes compatibility with image-based machine learning models, which are generally more effective with 2D data. The dataset is categorized into two classes based on medical diagnoses: mild cognitive impairment (MCI) and cognitively normal (CN). The MCI class comprises individuals with mild cognitive impairment, a potential precursor to AD, while the CN class includes individuals with normal cognitive function. This categorization supports comparative analysis between cognitively healthy individuals and those with early signs of impairment.

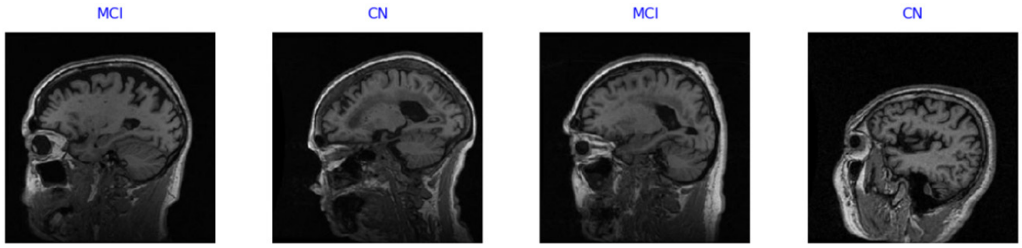

Fig. 1. Sample image dataset

Additional non-image data, such as gender and age, enriches the analysis by allowing the identification of biologically and demographically relevant patterns. Uniform processing parameters were applied during 3D image cropping to ensure data quality and consistency to minimize bias.

3.2 Model architecture

The architecture shown in Figure 2 is designed to process multimodal inputs, consisting of a combination of image data (sized 224 × 224 pixels) and tabular data (containing gender and age information), to produce a classification output. The first pathway is the image pathway, where images are processed using DL-based backbones. The deep learning models employed as backbones include DenseNet, ResNet, InceptionNet, VGG, MobileNet, and EfficientNet, which are responsible for extracting critical visual features from the images. Once features are extracted, the feature tensor is flattened using a flattened layer and subsequently passed through a dense layer for further processing. To enhance robustness and reduce the risk of overfitting, a dropout layer is employed in this pathway before the results are integrated with those of the other pathway.

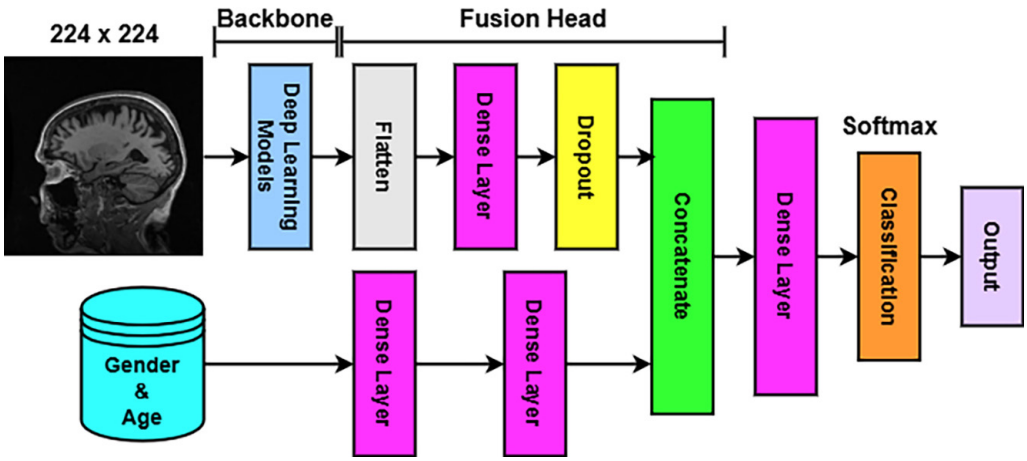

Fig. 2. Model architecture

The second pathway processes the tabular data, specifically gender and age. This data is passed through two dense layers that act as simple feature extractors, transforming the tabular data into meaningful vector representations. The dense layers in this pathway capture the non-linear relationships between gender and age to prepare the tabular data for integration with visual features. There is no explicit backbone for the gender and age pathway; the dense layers serve as a simple preprocessing mechanism for the tabular data.

These two pathways converge in the Fusion Head, where features from the image backbone and dense layers for gender and age are combined using a concatenation operation. After concatenation, the multimodal features are processed by additional dense layers to refine the combined feature representation. This stage ensures the alignment and optimization of the relationships between visual and tabular features.

Subsequently, the output of the Fusion Head is passed to the Classification Head, where additional dense layers process the combined features before producing the final output. The last layer is a classification layer, utilizing the Softmax activation function to generate class probabilities. This model is designed to leverage the strengths of multimodal data by combining visual and non-visual information to improve prediction accuracy. In this architecture, each pathway (including the backbones for image data) operates independently before integration, allowing for the comparison and evaluation of results from different types of backbones used.

3.3 Evaluation metrics

In this study, the evaluation of model performance is conducted using several key metrics, namely accuracy, precision, recall, and F1-Score, which are essential indicators for assessing the effectiveness of a model. Additionally, we assess resource efficiency by analyzing the memory load utilized during the model's training process, the time required to train the model to completion (training time), and the time needed to generate predictions (prediction time). The computation of accuracy, precision, recall, and F1-Score is based on the confusion matrix approach, which provides a detailed overview of the model's correct and incorrect predictions across each category. The formulas used to calculate accuracy, precision, recall, and F1-Score are presented in Equation (1) through Equation (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Remark:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

4 RESULTS AND DISCUSSION

The dataset used in this study consists of two classes: CN, which accounts for 52.7%, and MCI, which makes up the remaining 47.3% (see Figure 3). To ensure robust evaluation, the dataset was divided into three subsets: training (51,021 images), validation (6,378 images), and test (6,378 images). Prior to model training, all images were preprocessed by resizing them to a uniform dimension of 224 × 224 pixels and grouped into a batch size of 64. Model training was carried out for 70 epochs using the Adam optimizer with a learning rate of 0.001, executed on an NVIDIA A100 GPU equipped with 80 GB of RAM to handle the computational demands. A comprehensive set of CNN architectures was evaluated, including DenseNet121, DenseNet169, DenseNet201; EfficientNetB0 through B7; InceptionResNetV2; InceptionNetV3; MobileNetV1, V2, and V3; ResNet50, ResNet101, and ResNet152; as well as VGG16 and VGG19. This broad selection of models enabled systematic benchmarking of both lightweight and complex deep learning architectures for AD classification.

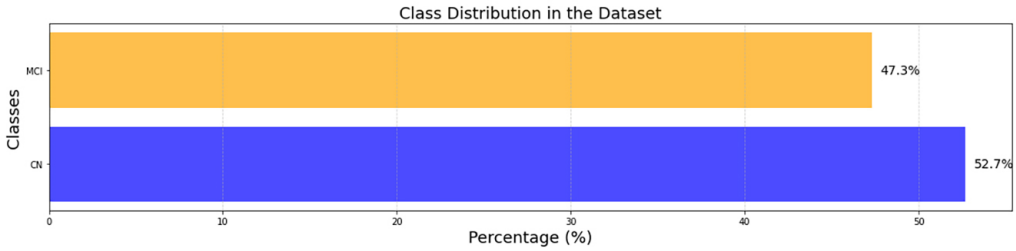

Fig. 3. Dataset distribution

4.1 Training time results

The training time analysis, as illustrated in Figure 4, shows a considerable variation across the evaluated models. Among all architectures, InceptionNetV3 demonstrated the fastest training process, completing in 4,231.81 seconds, whereas

EfficientNetB7 required the longest training time, reaching 72,335.28 seconds. This highlights how simpler models such as InceptionNet are computationally more efficient to train compared to more complex and deeper models such as EfficientNet, which, despite their accuracy, demand significantly greater computational resources.

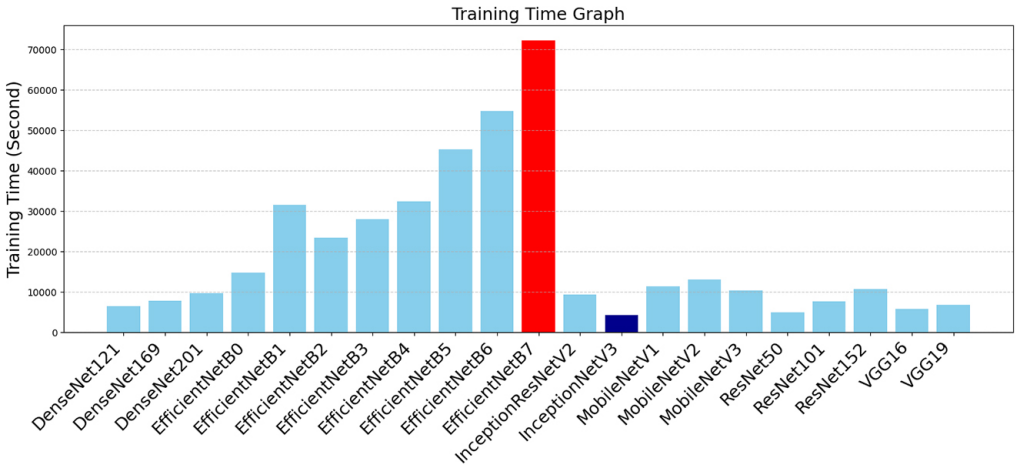

Fig. 4. Training time results

4.2 Testing time results

Figure 5 shows the model's prediction time. The results show that EfficientNetB1 has the longest prediction time of 22.16 seconds, while the fastest model is MobileNetV1 with a prediction time of 2.71 seconds. From these results, it can be concluded that MobileNetV1 offers much better prediction time efficiency than other models, making it an ideal choice for applications with fast response requirements, such as mobile-based systems or real-time applications. In contrast, the longer prediction time of EfficientNetB1 reflects the higher architectural complexity of the model, which is likely designed to improve accuracy or capture more detailed features of the data.

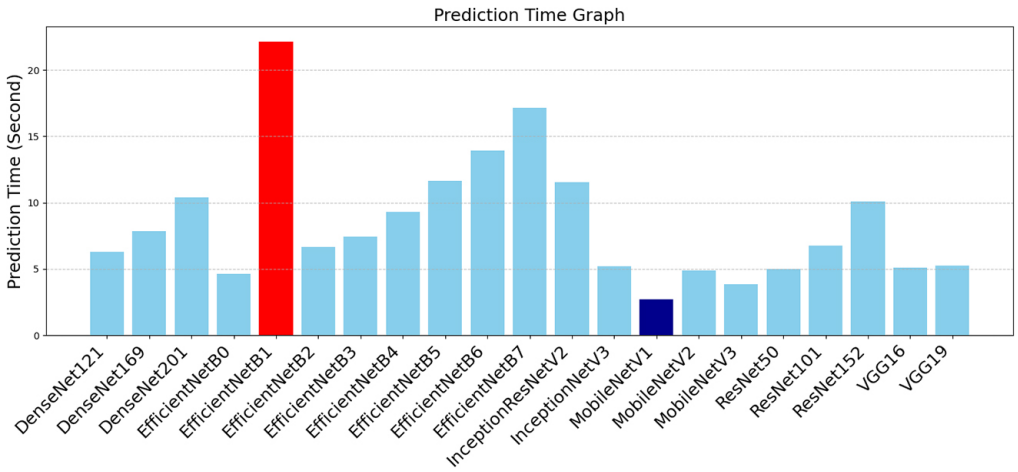

Fig. 5. Prediction time results

4.3 Memory load results

The test results in measuring memory load show that the EfficientNetB0 model has the largest memory load with a total load of 31761.68 MB. In contrast, the model with the lightest memory load during the process is VGG19, with a total load of 26349.94 MB (see Figure 6). This difference indicates that models with more complex architectures, such as EfficientNet, require more memory resources to perform their operations. This can be attributed to the larger number of parameters and more intensive computational processes, which are designed to improve accuracy and capture more details from the data used. In contrast, models such as VGG19, with a simpler architecture, tend to have a lower memory load, making them more efficient for applications that have limited hardware resources.

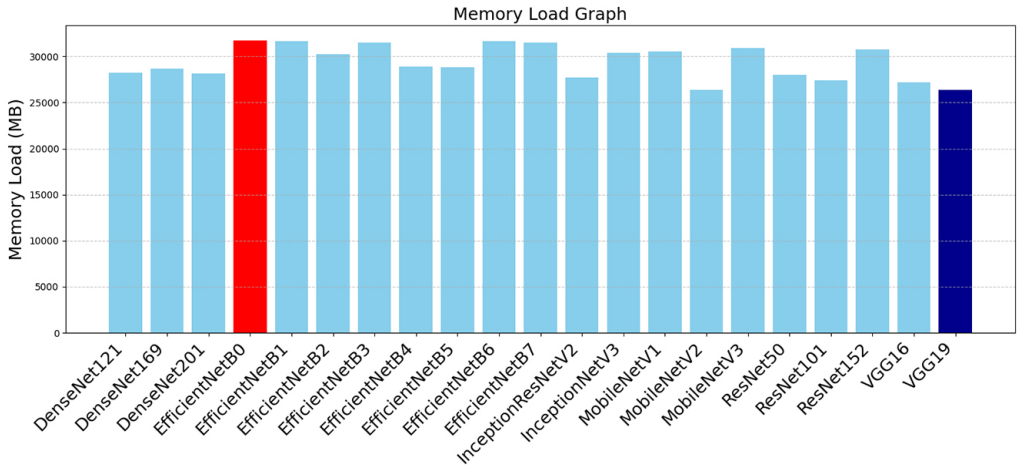

Fig. 6. Multimodal memory load results

4.4 Overall results

Table 1 shows the performance evaluation results of all tested models. All models from the DenseNet family, namely DenseNet121, DenseNet169, and DenseNet201, show excellent performance, with accuracy, precision, recall, and F1-score values consistently at 98%. This reflects the reliability of the DenseNet models in capturing the important features of the data used, while providing stable evaluation results on various metrics.

On the other hand, models from the EfficientNet family showed mixed results. The EfficientNetB0, EfficientNetB1, and EfficientNetB2 models managed to maintain equivalent performance to DenseNet, with overall accuracy, precision, recall, and F1-score values of 98%. However, the EfficientNetB3 to EfficientNetB7 models experienced a significant drop in performance. The accuracy value obtained for these models is 76%, with a precision of 80%, recall of 77%, and F1-score of 75%, except for EfficientNetB4 which is slightly better with an F1-score of 76%. This decrease is likely due to the increased complexity of the model which is not matched by the increased ability to generalize the data effectively. In addition, the EfficientNet family generally shows a greater burden in terms of training time, prediction time, and memory consumption, as shown in Figures 4–6.

The InceptionResNetV2 model also showed excellent performance, with accuracy, precision, recall, and F1-score of 98%, close to the results obtained by DenseNet

and EfficientNetB0-B2. In contrast, InceptionNetV3 experienced a decline in performance, with accuracy, precision, recall, and F1-score values of 97% each. This performance is still high but slightly lower than the other best models.

The MobileNet family (MobileNetV1-V3) showed lower performance, with accuracy, precision, recall, and F1-score values of 76%, 80%, 77%, and 75%, respectively. This model was followed by ResNet50 and ResNet101, which had similar evaluation results to MobileNet. However, ResNet152 showed significant improvement, with accuracy, precision, recall, and F1-score values of 83%. This indicates that the more complex the ResNet architecture is, the more the performance of the model can improve, albeit with the consequence of greater training time and memory load.

Finally, models from the VGG family, namely VGG16 and VGG19, have similar evaluation results to the MobileNet family, with accuracy, precision, recall, and F1-score values of 76%, 80%, 77%, and 75%, respectively. This suggests that simpler model architectures such as VGG tend to have lower performance than more complex models such as DenseNet or InceptionResNetV2.

Table 1. Overall results

Model	BE	TT (s)	PT (s)	ML (MB)	ACC (%)	PRE (%)	REC (%)	F1 (%)
DenseNet121	63	6422.88	6.27	28262.69	98	98	98	98
DenseNet169	67	7846.5	7.84	28657.21	98	98	98	98
DenseNet201	60	9650	10.41	28173.67	98	98	98	98
EfficientNetB0	66	14754.18	4.63	31761.68	98	98	98	98
EfficientNetB1	66	31481.07	22.16	31665.43	98	98	98	98
EfficientNetB2	69	23333.9	6.63	30265.93	98	98	98	98
EfficientNetB3	29	28025.67	7.41	31480.7	76	80	77	75
EfficientNetB4	28	32449.06	9.29	28895.77	76	80	77	76
EfficientNetB5	15	45272.21	11.63	28806.44	76	80	77	75
EfficientNetB6	16	54807.66	13.94	31694.73	76	80	77	75
EfficientNetB7	10	72335.28	17.17	31487.48	76	80	77	75
InceptionResNetV2	39	9284.20	11.55	27722.12	98	98	98	98
InceptionNetV3	69	4231.81	5.2	30384.01	97	97	97	97
MobileNetV1	20	11362.64	2.71	30547.16	76	80	77	75
MobileNetV2	19	13033.83	4.91	26352.16	76	80	77	75
MobileNetV3	18	10361.12	3.84	30904.53	76	80	77	75
ResNet50	24	4895.59	5	28029.37	76	80	77	75
ResNet101	39	7595.14	6.74	27428.51	76	80	77	75
ResNet152	8	10636.63	10.1	30744.37	83	83	83	83
VGG16	15	5774.18	5.1	27226.89	76	80	77	75
VGG19	29	6754.1	5.27	26349.94	76	80	77	75

Notes: BE = Best Epoch, TT, PT = Training Time, Prediction Time, ML = Memory Load, ACC, PRE, REC, F1 = Accuracy, Precision, Recall, F1-Score.

Overall, these results illustrate that model selection should consider the trade-off between model complexity and application-specific requirements. More complex models such as DenseNet, EfficientNet (B0-B2), and InceptionResNetV2 provide excellent evaluation results, but at a higher computational cost. In contrast, simpler models such as MobileNet and VGG can be an alternative for applications that require efficient use of computing resources, albeit with a compromise on performance.

Furthermore, the best models based on accuracy, precision, recall, and F1-score are models with more complex architectures, such as DenseNet121, DenseNet169, DenseNet201, EfficientNetB0, EfficientNetB1, EfficientNetB2, and InceptionResNetV2. The EfficientNetB1 model shows the best performance based on accuracy validation with an achievement of 98.6%, which is indicated by the red round symbol in Figure 7. The validation graph in the figure also shows a consistent upward trend, reflecting the model's ability to generalize well.

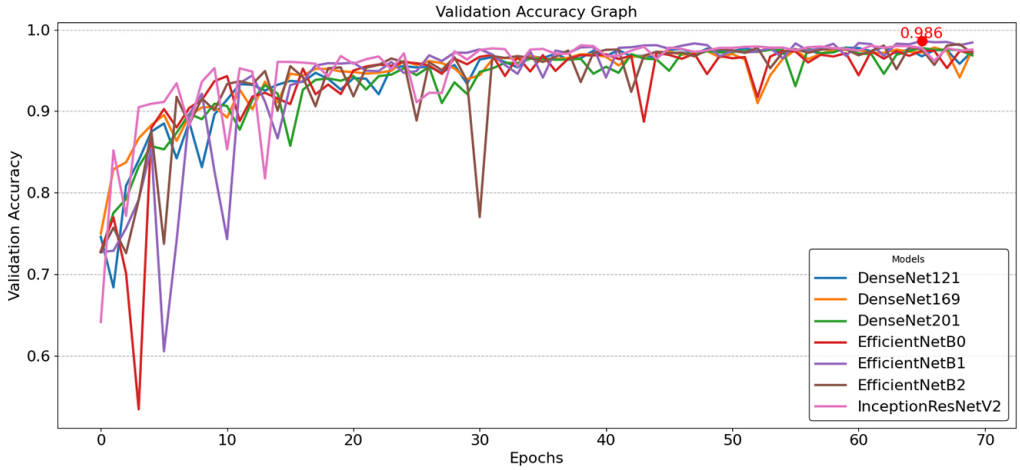

Fig. 7. Validation accuracy

However, the models in the EfficientNet family have high complexity on average, characterized by longer training and prediction times than other models. EfficientNetB1 recorded the highest prediction time of 22.16 seconds, which is one of the trade-offs of its superior performance. Nonetheless, based on the overall evaluation, EfficientNetB1 can be considered as the best model among all the tested models, as it is able to provide excellent accuracy and evaluation performance despite the greater computational burden.

4.5 Unimodal vs. Multimodal comparison results

The comparison between multimodal and unimodal models (without clinical data) clearly demonstrates that the removal of clinical metadata negatively impacts model performance. As shown in Figure 8 and Table 2, the best-performing multimodal model, EfficientNetB1, achieved an accuracy of 98.6%, but when evaluated in the unimodal setting its accuracy dropped to 98.1%. This indicates that clinical features, even simple ones such as age and gender, provide valuable information that enhances classification reliability.

Although unimodal models appear more efficient in terms of training time, prediction time, and memory load—with memory consumption reduced by nearly 50%—this computational efficiency comes at the expense of diagnostic accuracy.

Such efficiency gains are insufficient to outweigh the loss of clinically relevant information, confirming that multimodal integration plays a critical role in producing more reliable outcomes.

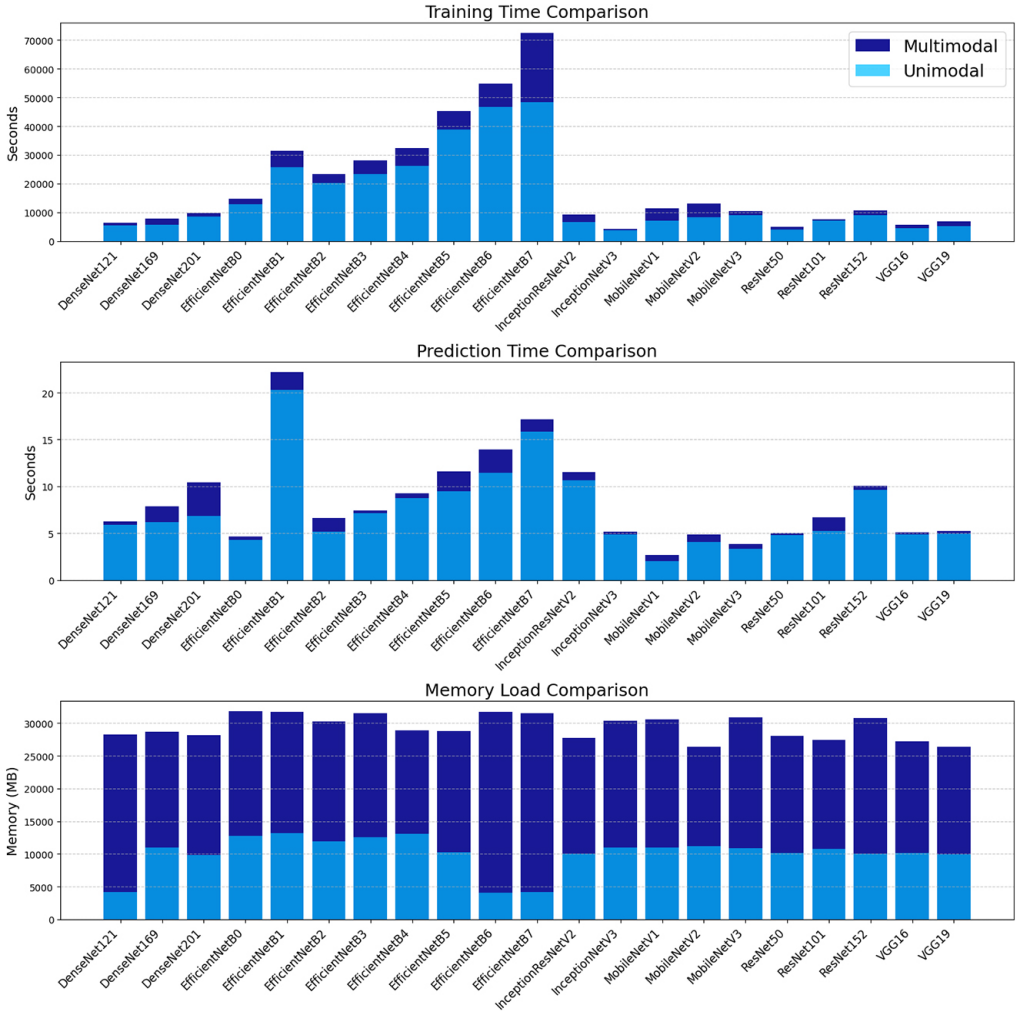

Fig. 8. Unimodal vs. multimodal performance results

Despite its superior accuracy, EfficientNetB1 presented practical challenges for clinical use due to its relatively long inference time of 22.16 seconds. To address this limitation, future work should explore model compression strategies such as pruning, quantization, and knowledge distillation to reduce latency while preserving diagnostic accuracy. These optimizations are essential to enable deployment in real-time hospital environments and mobile health applications.

Furthermore, the complexity of multimodal data can be viewed as analogous to nonlinear dynamical systems, where principles of adaptive and robust control may inspire improvements in CNN robustness. For example, adaptive fuzzy control provides stability under uncertainty, while backstepping allows adaptation to dynamic complexity. Within this analogy, EfficientNetB1 can be considered an "optimal controller"—delivering high performance but at higher computational cost—whereas MobileNet functions more like a "robust controller", offering efficiency with reduced accuracy. Embedding these control-inspired principles into CNN design could help strike a balance between accuracy, efficiency, and robustness in AD diagnosis.

Table 2. Comparison of performance evaluation

Model	Accuracy		Precision		Recall		F1-Score	
	Multi-Modal	Uni-Modal	Multi-Modal	Uni-Modal	Multi-Modal	Uni-Modal	Multi-Modal	Uni-Modal
DenseNet121	0.98	0.97	0.98	0.97	0.98	0.97	0.98	0.97
DenseNet169	0.98	0.97	0.98	0.97	0.98	0.97	0.98	0.97
DenseNet201	0.98	0.97	0.98	0.97	0.98	0.97	0.98	0.97
EfficientNetB0	0.98	0.53	0.98	0.28	0.98	0.53	0.98	0.36
EfficientNetB1	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
EfficientNetB2	0.98	0.53	0.98	0.28	0.98	0.53	0.98	0.36
EfficientNetB3	0.76	0.53	0.80	0.28	0.77	0.53	0.75	0.36
EfficientNetB4	0.76	0.98	0.80	0.98	0.77	0.98	0.76	0.98
EfficientNetB5	0.76	0.53	0.80	0.28	0.77	0.53	0.75	0.36
EfficientNetB6	0.76	0.97	0.80	0.97	0.77	0.97	0.75	0.97
EfficientNetB7	0.76	0.53	0.80	0.28	0.77	0.53	0.75	0.36
InceptionResNetV2	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
InceptionNetV3	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
MobileNetV1	0.76	0.53	0.80	0.28	0.77	0.53	0.75	0.36
MobileNetV2	0.76	0.53	0.80	0.28	0.77	0.53	0.75	0.36
MobileNetV3	0.76	0.53	0.80	0.28	0.77	0.53	0.75	0.36
ResNet50	0.76	0.53	0.80	0.28	0.77	0.53	0.75	0.36
ResNet101	0.76	0.84	0.80	0.84	0.77	0.84	0.75	0.84
ResNet152	0.83	0.97	0.83	0.97	0.83	0.97	0.83	0.97
VGG16	0.76	0.53	0.80	0.28	0.77	0.53	0.75	0.36
VGG19	0.76	0.53	0.80	0.28	0.77	0.53	0.75	0.36

5 CONCLUSION

This study evaluated the performance of various deep learning models for Alzheimer's diagnosis using multimodal data. Models with complex architectures, such as DenseNet121, DenseNet169, DenseNet201, EfficientNetB0, EfficientNetB1, EfficientNetB2, and InceptionResNetV2, demonstrated superior performance, consistently achieving high accuracy, precision, recall, and F1-score values close to 98%. Among them, EfficientNetB1 emerged as the best-performing model, reaching a validation accuracy of 98.6% with a steady upward trend, thereby proving its robustness in multimodal classification. However, this enhanced performance came at the expense of computational efficiency, with EfficientNetB1 recording the longest prediction time (22.16 seconds) and higher memory consumption, highlighting the trade-off between model complexity and resource requirements.

The comparative analysis between multimodal and unimodal models further reinforced the importance of clinical metadata (e.g., age and gender) in enhancing

diagnostic reliability. For instance, EfficientNetB1, when trained as a multimodal model, achieved an accuracy of 98.6%, while its unimodal counterpart (without clinical metadata) experienced a slight drop to 98.1%. Although unimodal models were computationally more efficient, requiring less training time, prediction time, and memory load, the results demonstrated that multimodal integration provides significant gains in classification accuracy by leveraging clinically relevant features. Even though unimodal models reduced memory consumption by nearly 50%, this efficiency did not outweigh the decline in diagnostic performance.

Conversely, lighter architectures such as MobileNet and VGG yielded lower evaluation metrics, with F1-scores around 75%, but offered practical advantages in terms of faster training and lower memory usage. These characteristics make them suitable for resource-constrained environments or real-time healthcare applications, where efficiency is prioritized over peak accuracy.

This revised study clarifies the novelty of benchmarking six CNN architectures in a multimodal setting, expands the related work to include RNNs, GNNs, and transformers, and discusses potential lightweight optimization strategies (e.g., pruning, quantization, and knowledge distillation) to improve clinical usability. In addition, drawing parallels with adaptive and robust control theory enriches the discussion on model robustness and efficiency. These insights position CNNs not only as strong baselines but also as foundational models upon which more advanced multimodal pipelines—such as transformer-based or hybrid fusion models—can be developed to further advance AD diagnosis.

6 ACKNOWLEDGEMENTS

This study is financially supported by the Faculty of Computer Science on Doctoral Research Grant under contract number 02245/UN10.F1501/B/PT.01.05.1/12024 was dated 12 July, 2024.

7 REFERENCES

[1] L. Langnickel *et al.*, "Information extraction from German clinical care documents in context of Alzheimer's disease," *Applied Sciences*, vol. 11, no. 22, p. 10717, 2021. https://doi.org/10.3390/app112210717

[2] S. De Marchi, F. Lot, F. Marchetti, and D. Poggiali, "Variably scaled persistence Kernels (VSPKs) for persistent homology applications," *Journal of Computational Mathematics and Data Science*, vol. 4, p. 100050, 2022. https://doi.org/10.1016/j.jcmds.2022.100050

[3] K. Yang, X. Yang, P. Yin, M. Zhou, and Y. Tang, "Temporal trend and attributable risk factors of Alzheimer's disease and other dementias burden in China: Findings from the global burden of disease study 2021," *Alzheimer's and Dementia*, vol. 20, no. 11, pp. 7871–7884, 2024. https://doi.org/10.1002/alz.14254

[4] L. Xia, F. Xiaojin, S. Xiaodong, N. Hou, H. Fang, and L. Yongping, "Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2019," *Front Aging Neurosci.*, vol. 14, 2022. https://doi.org/10.3389/fnagi.2022.937486

[5] N. S. Raghavan *et al.*, "Association between common variants in RBFOX1, an RNA-binding protein, and brain amyloidosis in early and preclinical Alzheimer disease," *JAMA Neurol.*, vol. 77, no. 10, pp. 12880–1298, 2020. https://doi.org/10.1001/jamaneurol.2020.1760

[6] G. Bonifazi *et al.*, "The nonlinear meccano of hyperactivity in Alzheimer," *bioRxiv*, 2023. https://doi.org/10.1101/2023.10.09.561541

[7] R. Petersen, "Detecting Alzheimer disease clinically," *Neurology*, vol. 98, pp. 607–608, 2022. https://doi.org/10.1212/WNL.0000000000200172

[8] Y.-E. Quek, Y. L. Fung, P. Bourgeat, S. J. Vogrin, S. J. Collins, and S. C. Bowden, "Detecting early Alzheimer's disease using neuropsychological assessment and structural neuroimaging," *Alzheimer's & Dementia*, vol. 19, no. S17, p. e071340, 2023. https://doi.org/10.1002/alz.071340

[9] Q. Zhao, X. Du, W. Chen, T. Zhang, and Z. Xu, "Advances in diagnosing mild cognitive impairment and Alzheimer's disease using ^{11}C-PIB-PET/CT and common neuropsychological tests," *Front. Neurosci.*, vol. 17, 2023. https://doi.org/10.3389/fnins.2023.1216215

[10] Y.-T. Wang, P. Rosa-Neto, and S. Gauthier, "Advanced brain imaging for the diagnosis of Alzheimer disease," *Curr. Opin. Neurol.*, vol. 36, pp. 481–490, 2023. https://doi.org/10.1097/WCO.0000000000001198

[11] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," *IEEE J. Sel. Top Signal Process*, vol. 14, no. 2, pp. 272–281, 2020. https://doi.org/10.1109/JSTSP.2019.2955022

[12] H. Acharya, R. Mehta, and D. Kumar Singh, "Alzheimer disease classification using transfer learning," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1503–1508. https://doi.org/10.1109/ICCMC51019.2021.9418294

[13] X. Liu *et al.*, "Advances in deep learning-based medical image analysis," *Health Data Science*, vol. 2021, 2021. https://doi.org/10.34133/2021/8786793

[14] X. Xu *et al.*, "A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis," *Bioengineering*, vol. 11, no. 3, p. 219, 2024. https://doi.org/10.3390/bioengineering11030219

[15] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep learning for medical image-based cancer diagnosis," *Cancers (Basel)*, vol. 15, no. 14, p. 3608, 2023. https://doi.org/10.3390/cancers15143608

[16] K. Wong, G. Fortino, and D. Abbott, "Deep learning-based cardiovascular image diagnosis: A promising challenge," *Future Gener. Comput. Syst.*, vol. 110, pp. 802–811, 2020. https://doi.org/10.1016/j.future.2019.09.047

[17] A. Deshmukh, "Artificial intelligence in medical imaging: Applications of deep learning for disease detection and diagnosis," *Universal Research Reports*, vol. 11, no. 3, pp. 31–36, 2024. https://doi.org/10.36676/urr.v11.i3.1284

[18] N. Aziz, N. Minallah, J. Frnda, M. Sher, M. Zeeshan, and A. H. Durrani, "Precision meets generalization: Enhancing brain tumor classification via pretrained DenseNet with global average pooling and hyperparameter tuning," *PLoS ONE*, vol. 19, 2024. https://doi.org/10.1371/journal.pone.0307825

[19] V. G. Buddhavarapu and A. A. Jothi, "An experimental study on classification of thyroid histopathology images using transfer learning," *Pattern Recognit. Lett.*, vol. 140, pp. 1–9, 2020. https://doi.org/10.1016/j.patrec.2020.09.020

[20] Z. Huang, X. Zhu, M. Ding, and X. Zhang, "Medical image classification using a light-weighted hybrid neural network based on PCANet and DenseNet," *IEEE Access*, vol. 8, pp. 24697–24712, 2020. https://doi.org/10.1109/ACCESS.2020.2971225

[21] M. A. Al-masni, D.-H. Kim, and T.-S. Kim, "Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification," *Comput. Methods Programs Biomed.*, vol. 190, p. 105351, 2020. https://doi.org/10.1016/j.cmpb.2020.105351

[22] C. Fjellström and K. Nyström, "Deep learning, stochastic gradient descent and diffusion maps," *Journal of Computational Mathematics and Data Science*, vol. 4, p. 100054, 2022. https://doi.org/10.1016/j.jcmds.2022.100054

[23] G. Mirabnahrazam *et al.*, "Machine learning based multimodal neuroimaging genomics dementia score for predicting future conversion to Alzheimer's disease," *J. Alzheimers Dis.*, vol. 87, no. 3, pp. 1345–1365, 2022. https://doi.org/10.3233/JAD-220021

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

[25] J. Venugopalan, L. Tong, H. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of Alzheimer's disease stage," *Sci. Rep.*, vol. 11, no. 3254, 2021. https://doi.org/10.1038/s41598-020-74399-w

[26] D. Poggiali, D. Cecchin, and S. De Marchi, "Reducing the Gibbs effect in multimodal medical imaging by the Fake Nodes approach," *Journal of Computational Mathematics and Data Science*, vol. 4, p. 100040, 2022. https://doi.org/10.1016/j.jcmds.2022.100040

[27] E. W. Westi *et al.*, "Comprehensive analysis of the 5×FAD mouse model of Alzheimer's disease Using dMRI, immunohistochemistry, and neuronal and glial functional metabolic mapping," *Biomolecules*, vol. 14, 2024. https://doi.org/10.3390/biom14101294

[28] G. Aghakhanyan *et al.*, "PET/MRI delivers multimodal brain signature in Alzheimer's Disease with De Novo PSEN1 mutation.," *Curr. Alzheimer Res.*, vol. 18, no. 2, pp. 178–184, 2021. https://doi.org/10.2174/1567205018666210414111536

[29] M. T. Ferretti *et al.*, "Sex and gender differences in Alzheimer's disease: Current challenges and implications for clinical practice: Position paper of the Dementia and cognitive disorders panel of the European Academy of Neurology," *Eur. J. Neurol.*, vol. 27, no. 6, pp. 928–943, 2020. https://doi.org/10.1111/ene.14174

[30] W. Lin, Q. Gao, M. Du, W. Chen, and T. Tong, "Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data," *Comput. Biol. Med.*, vol. 134, p. 104478, 2021. https://doi.org/10.1016/j.compbiomed.2021.104478

[31] A. Rahman *et al.*, "Sex-driven modifiers of Alzheimer risk," *Neurology*, vol. 95, pp. e166–e178, 2020. https://doi.org/10.1212/WNL.0000000000009781

[32] J. Wu, H. Zhang, X. Zhu, Y. Zhang, X. Ding, and H. Yang, "Ensemble learning-based multimodal data analysis improving the diagnostic accuracy of Alzheimer's disease," in *Proceedings Optics in Health Care and Biomedical Optics XIII*, 2023, p. 1277030. https://doi.org/10.1117/12.2687618

[33] J. Venugopalan, L. Tong, H. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of Alzheimer's disease stage," *Sci. Rep.*, vol. 11, 2021. https://doi.org/10.1038/s41598-020-74399-w

[34] M. Golovanevsky, C. Eickhoff, and R. Singh, "Multimodal attention-based deep learning for Alzheimer's disease diagnosis," *J. Am. Med. Inform. Assoc.*, vol. 29, no. 12, pp. 2014–2022, 2022. https://doi.org/10.1093/jamia/ocac168

[35] H. Wang *et al.*, "Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease," *Neurocomputing*, vol. 333, pp. 145–156, 2019. https://doi.org/10.1016/j.neucom.2018.12.018

[36] Q. Li and M. Yang, "Comparison of machine learning approaches for enhancing Alzheimer's disease classification," *PeerJ*, vol. 9, p. e10549, 2021. https://doi.org/10.7717/peerj.10549

[37] F. M. J. M. Shamrat *et al.*, "AlzheimerNet: An effective deep learning based proposition for Alzheimer's disease stages classification from functional brain changes in magnetic resonance images," *IEEE Access*, vol. 11, pp. 16376–16395, 2023. https://doi.org/10.1109/ACCESS.2023.3244952

[38] B. Song and S. Yoshida, "Explainability of three-dimensional convolutional neural networks for functional magnetic resonance imaging of Alzheimer's disease classification based on gradient-weighted class activation mapping," *PLoS ONE*, vol. 19, p. e0303278, 2024. https://doi.org/10.1371/journal.pone.0303278

[39] G. N. Alwakid, S. Tahir, M. Humayun, and W. Gouda, "Improving Alzheimer's detection with deep learning and image processing techniques," *IEEE Access*, vol. 12, pp. 153445–153456, 2024. https://doi.org/10.1109/ACCESS.2024.3481238

[40] B. S. Rao, M. Aparna, J. Harikiran, and T. S. Reddy, "An effective Alzheimer's disease segmentation and classification using Deep ResUnet and Efficientnet," *J. Biomol. Struct. Dyn.*, vol. 43, no. 6, pp. 2840–2851, 2023. https://doi.org/10.1080/07391102.2023.2294381

[41] M. M. Al-Nawashi, O. M. Al-Hazaimeh, and M. K. Khazaaeleh, "A new approach for breast cancer detection-based machine learning technique," *Applied Computer Science*, vol. 20, no. 1, pp. 1–16, 2021. https://doi.org/10.35784/acs-2024-01

[42] N. Gharaibeh *et al.*, "Swin transformer-based segmentation and multi-scale feature pyramid fusion module for Alzheimer's disease with machine learning," *International Journal of Online & Biomedical Engineering*, vol. 19, no. 4, pp. 22–50, 2023. https://doi.org/10.3991/ijoe.v19i04.37677

8 AUTHORS

Lailil Muflikhah is with the Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia (E-mail: lailil@ub.ac.id).

Galih Restu Baihaqi is with the Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia.

Shafatyra Redhita Shalsadilla is with the Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia.

Achmad Ridok is with the Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia.

Sri Soenarti is with the Department of Internal Medic, Faculty of Medical, Brawijaya University, Malang, Indonesia.