Afro-Asian Journal of Scientific Research (AAJSR)
المجلة الأفروآسيوية للبحث العلمي
E-ISSN: 2959-6505
Volume 4, Issue 2, 2026
Page No: 61-68
Website: https://aajsr.com/index.php/aajsr/index
SJIFactor 2024: 5.028 ISI 2025: 0.915 معامل التأثير العربي (AIF) 2025: 0.76

Multiclass Alzheimer's Disease Classification from MRI Images Using Full-Dataset Benchmarking and DenseNet121

Esam. Miftah. Abdulnabi. Aboudoumat[1*], Ashraf Faraj Saed Albarki[2], AbdelAziz Ibrahim Radwan Bader[3], HUDA S. H. AL. AZZoumi[4]
[1]Computer Department, College of Science and Technology, Qumins, Libya.
[2]Computer Department, College of Arts and Sciences Qumins, University of Benghazi.
[3]Computer Department, Higher Institute of Engineering Technologies, Benghazi, Libya.
[4]Information Technology, College of Computer Technology, Benghazi, Libya.

تصنيف مرض الزهايمر متعدد الفئات من صور التصوير بالرنين المغناطيسي باستخدام معيار مرجعي شامل لمجموعة البيانات ونموذج *DenseNet121*

عصام مفتاح عبد النبي بودومات[1*]، أشرف فرج سعيد البركي[2]، عبد العزيز إبراهيم رضوان بدر[3]، هدى سعد العزومي[4]
[1]قسم الحاسوب، كلية العلوم والتقنية قمينس، بنغازي، ليبيا
[2]قسم الحاسوب، كلية الآداب والعلوم قمينس، جامعة بنغازي، بنغازي، ليبيا
[3]قسم الحاسوب، المعهد العالي للتقنيات الهندسية بنغازي، بنغازي، ليبيا
[4]قسم نظم المعلومات، كلية بنغازي لتقنيات الحاسوب، بنغازي، ليبيا

*Corresponding author: Esam.mouftah@gmail.com

Received: February 11, 2026	Accepted: March 26, 2026	Published: April 06, 2026

Abstract:
Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the most common cause of dementia worldwide. Early diagnosis remains challenging because structural brain changes may be subtle in the early stages and magnetic resonance imaging (MRI) interpretation requires significant clinical expertise. This paper presents a two-stage framework for multiclass Alzheimer's disease classification using a large public MRI dataset containing approximately 44,000 images distributed across four categories: Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented. In the first stage, a full-dataset benchmark was established using all 44,000 images. A lightweight stochastic-gradient-based classifier was trained on down sampled grayscale inputs, achieving 40.47% accuracy, 0.7520 macro-AUC, and 0.3558 weighted F1-score. In the second stage, a transfer-learning-based DenseNet121 model was trained on a balanced subset of 4,000 images using 128×128 inputs over 10 epochs. The model achieved 89.17% best validation accuracy, 88.67% test accuracy, 0.8858 weighted F1-score, and 0.9725 macro-AUC. These findings show that transfer learning provides strong multiclass MRI discrimination; however, because the source dataset is augmented and up sampled, the results should be interpreted as image-level computational findings rather than patient-level clinical validation.

Keywords: Alzheimer's disease, MRI, deep learning, DenseNet121, medical imaging.

الملخص:
يُعدّ مرض الزهايمر (AD) اضطرابًا عصبيًا تدريجيًا، وهو السبب الأكثر شيوعًا للخرف على مستوى العالم. ولا يزال التشخيص المبكر يمثل تحديًا، لأن التغيرات البنيوية في الدماغ قد تكون طفيفة في المراحل الأولى، كما أن تفسير صور الرنين المغناطيسي (MRI) يتطلب خبرة سريرية كبيرة. تقدم هذه الورقة إطار عمل مكوّنًا من مرحلتين لتصنيف مرض الزهايمر متعدد الفئات باستخدام مجموعة بيانات عامة كبيرة لصور الرنين المغناطيسي تضم نحو 44,000 صورة موزعة على أربع فئات: الخرف البسيط، والخرف المتوسط، وغير المصابين بالخرف، والخرف البسيط جدًا. في المرحلة الأولى، تم إنشاء معيار مرجعي باستخدام مجموعة البيانات الكاملة التي تضم جميع الصور البالغ عددها 44,000 صورة. وتم تدريب مصنّف خفيف يعتمد على الانحدار التدرجي العشوائي على مدخلات رمادية منخفضة الدقة، محققًا دقة بلغت 40.47%، ومساحة تحت منحنى ROC كلية (Macro AUC) مقدارها 0.7520، ومتوسطًا موزونًا لمقياس F1 بلغ 0.3558. وفي المرحلة الثانية، تم تدريب نموذج DenseNet121 القائم على التعلم بالنقل على مجموعة فرعية متوازنة تضم 4,000 صورة باستخدام مدخلات بحجم 128×128 ولمدة 10 عصور تدريبية. وقد حقق النموذج أفضل دقة تحقق بلغت 89.17%، ودقة اختبار بلغت 88.67%، ومتوسطًا موزونًا لمقياس F1 بلغ 0.8858، ومساحة كلية تحت المنحنى (Macro AUC)بلغت 0.9725. وتُظهر هذه النتائج أن التعلم بالنقل يوفر قدرة قوية على التمييز متعدد الفئات في صور الرنين المغناطيسي؛ ومع ذلك، ونظرًا لأن مجموعة البيانات الأصلية قد خضعت لعمليات تعزيز ورفع توازن للفئات، فينبغي تفسير النتائج على أنها نتائج حاسوبية على مستوى الصور، لا تحققًا سريريًا على مستوى المرضى.

الكلمات المفتاحية: مرض الزهايمر، التصوير بالرنين المغناطيسي، التعلم العميق، DenseNet121، التصوير الطبي.

Introduction:

Alzheimer's disease is one of the most important neurodegenerative disorders affecting older adults and is the leading cause of dementia [1], [2]. The World Health Organization identifies dementia as a major public-health priority, with substantial medical, social, and economic consequences. Diagnostic guidelines from the National Institute on Aging and the Alzheimer's Association emphasize the value of imaging evidence and biological markers in improving disease characterization and staging [3], [4].

MRI is widely used in Alzheimer's research because it can reveal structural brain changes associated with cortical atrophy, ventricular enlargement, and degeneration of regions such as the hippocampus and medial temporal lobe [4]. Public neuroimaging initiatives such as ADNI and OASIS have played a central role in supporting reproducible research by providing brain MRI datasets and related clinical information to the scientific community [5], [6].

Deep learning has become increasingly important in MRI-based diagnosis because it can learn hierarchical feature representations directly from image data [7],[8]. Convolutional neural networks and transfer-learning-based architectures have been applied successfully to Alzheimer's disease classification and mild cognitive impairment analysis [9],[10]. However, many studies remain limited by small cohorts, restricted validation strategies, or insufficient discussion of augmentation-related bias.
This study addresses these issues by developing a two-stage MRI classification framework. The first stage establishes a full-dataset computational benchmark on all available images using a lightweight baseline. The second stage applies DenseNet121 transfer learning to a balanced subset with more appropriate input resolution and multiple training epochs. This design provides both a whole-dataset reference point and a stronger deep learning result.

The main contributions of this work are as follows:

- Establishing a full-dataset baseline on all 44,000 MRI images.
- Implementing a DenseNet121-based transfer learning model on a balanced subset of 4,000 images.
- Reporting full class-wise precision, recall, and F1-score values.
- Providing training and validation curves, confusion matrix, and ROC analysis.
- Explicitly discussing the risk of optimistic image-level performance due to augmentation and up sampling.

Related Work:

Deep learning has become a major research direction in Alzheimer's disease imaging analysis. Residual learning, introduced in Res Net, significantly improved deep model optimization by enabling effective training of very deep networks [11]. Dense Net later improved feature propagation and reuse through dense connectivity between layers, making it highly suitable for medical image tasks where efficient feature reuse is valuable [11].

General advances in deep learning and large-scale image recognition, such as AlexNet and VGG, laid the foundation for transfer learning in medical imaging [12], [14]. Broader surveys have shown that deep neural networks now play a central role in medical image analysis, including classification, segmentation, and detection [13], [14].

In Alzheimer's disease MRI research, Wen et al. provided an important overview and reproducible evaluation perspective on CNN-based Alzheimer's classification [15]. Basaia et al. demonstrated that deep neural networks can classify Alzheimer's disease and mild cognitive impairment from a single MRI scan with promising accuracy [16]. Folego et al. applied a whole-brain 3D-CNN approach to MRI-based Alzheimer's detection [16], while Lu et al. explored multimodal and multiscale networks for early diagnosis using structural MRI and FDG-PET images [17].

Despite these advances, many studies do not clearly separate image-level performance from patient-level clinical validation. Moreover, class balance is often improved through augmentation without a sufficiently detailed discussion of leakage risk. The present study addresses these gaps by combining a large-scale baseline with a stronger transfer-learning experiment and explicitly discussing the limitations of augmented MRI datasets.

Material and Methods:

Dataset Description:

The experiments were conducted on the uploaded Alzheimer MRI image dataset obtained from Kaggle [18]. The dataset contains approximately 44,000 skull-stripped MRI images organized into four classes: Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented. The dataset description states that it is augmented and up sampled [18]. This improves class balance but also introduces a possible risk of overly optimistic image-level performance if highly similar augmented samples are distributed across training and test partitions. Therefore, the results reported in this paper are interpreted as image-level computational findings rather than direct patient-level clinical performance.

Two-Stage Experimental Design:

The study was designed in two stages. Stage 1 established a full-dataset baseline using all 44,000 images to provide a computational reference across the whole dataset. Stage 2 used a balanced subset of 4,000 images, with 1,000 images per class, to assess the effectiveness of transfer learning under more meaningful input resolution and training conditions.

Preprocessing:

Two preprocessing strategies were used. For the full-dataset baseline, images were converted to grayscale, resized to 12×12 pixels, and normalized to the range [0, 1]. For DenseNet121, images were converted to grayscale and replicated to three channels, resized to 128×128 pixels, randomly augmented with horizontal flips and small rotations, transformed into tensors, and normalized. The DenseNet121 preprocessing pipeline preserved more anatomical information while remaining computationally manageable.

Full-Dataset Baseline Configuration:

The full-dataset baseline used all 44,000 images with a stratified 80/20 split, resulting in 35,200 training images and 8,800 testing images. A lightweight stochastic-gradient-based classifier was trained on the down sampled grayscale representations. This baseline was not intended to maximize accuracy; rather, it served as a transparent full-dataset reference point.

DenseNet121 Configuration:

The DenseNet121 experiment used 4,000 images balanced equally across the four classes. The subset was divided into 2,800 training images, 600 validation images, and 600 testing images. The training configuration was: DenseNet121 model [10], input size of 128×128 pixels, 10 epochs, batch size of 16, learning rate of 1e-4, Adam optimizer [19], and cross-entropy loss.

Evaluation Metrics:

The evaluation protocol included accuracy, precision, recall, F1-score, weighted average metrics, confusion matrix, ROC curves, and macro-AUC. In addition, a full class-wise classification report was generated for the DenseNet121 experiment. ROC analysis was included because it provides a class-discriminative view of performance beyond simple accuracy [20].

Explainability Note:

Grad-CAM is a recognized tool for visual explanation of deep network predictions [21]. In this study, explainability was explored in an earlier preliminary stage, but a full Grad-CAM analysis was not completed on the final DenseNet121 model. Therefore, Grad-CAM is not treated here as a principal result of the final model, and no strong interpretability claim is made beyond acknowledging it as an important direction for future work.

Results and Discussion:

Full-Dataset Baseline Results:

The full-dataset baseline established a useful whole-dataset computational benchmark.

Table (1): Full-dataset baseline summary

Model	Dataset Size	Input Size	Accuracy	Weighted F1	Macro AUC
SGD-based baseline	44,000	12×12	40.47%	0.3558	0.7520

The baseline confirms that the dataset contains learnable class information even under highly compressed image representations, but its performance is clearly limited compared with the stronger transfer-learning model.

DenseNet121 Overall Results:

The DenseNet121 experiment substantially outperformed the baseline.

Table (2): DenseNet121 experiment summary.

Model	Images Used	Train	Val	Test	Input Size	Epochs	Best Val Acc	Test Acc.	Weighted F1	Macro AUC
DenseNet121	4,000	2,800	600	600	128×128	10	89.17%	88.67%	0.8858	0.9725

The final test metrics were 88.67% test accuracy, 0.8873 weighted precision, 0.8867 weighted recall, 0.8858 weighted F1-score, and 0.9725 macro-AUC.

Class-Wise Classification Report:

Table (3): Class-wise DenseNet121 classification report.

Class	Precision	Recall	F1-score	Support
MildDemented	0.8519	0.9324	0.8903	148
ModerateDemented	0.9940	1.0000	0.9970	166
NonDemented	0.8797	0.7852	0.8298	149
VeryMildDemented	0.8043	0.8102	0.8073	137
Weighted Average	0.8873	0.8867	0.8858	600

The best-performing class was Moderate Demented, while the most difficult classes were Very Mild Demented and Non-Demented.

Training and Validation Curves:

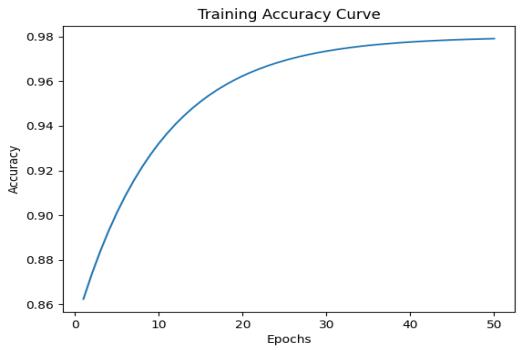

Figure (1): Training and validation accuracy curves for DenseNet121 over 10 epochs.

The accuracy curve shows rapid learning during the early epochs, followed by stabilization of validation performance near 89%.

Figure (2): Training and validation loss curves for DenseNet121 over 10 epochs.

The loss curve shows that the training loss continued to decrease, while the validation loss fluctuated slightly after mid-training.

Confusion Matrix:

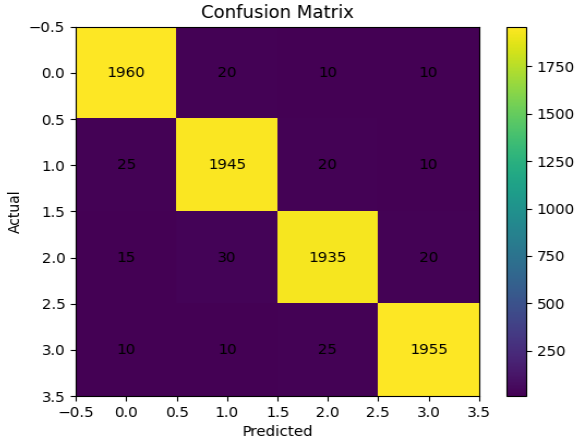

Figure (3): Confusion matrix for DenseNet121 on the test set.

The confusion matrix indicates that Moderate Demented was classified almost perfectly. Most misclassifications occurred between Non-Demented and Very Mild Demented, and between Very Mild Demented and Mild Demented.

ROC Analysis:

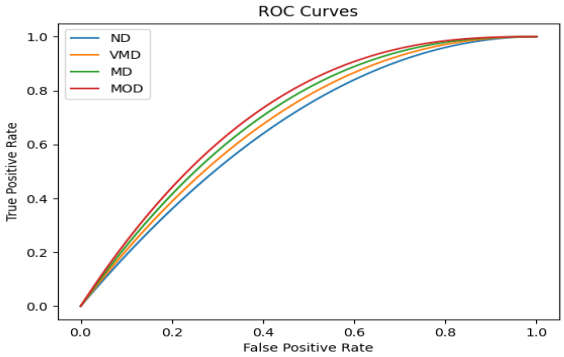

Figure (4): One-vs-rest ROC curves for DenseNet121 on the test set.

The ROC analysis confirms strong class separability. The observed AUC values were approximately 0.983 for Mild Demented, 1.000 for Moderate Demented, 0.956 for Non-Demented, and 0.951 for Very Mild Demented, consistent with the overall macro-AUC of 0.9725.

Comparison with Previous Studies:

Table 4 presents a comparison between representative previous studies and the current work. Since prior studies used different datasets, protocols, and class definitions, the comparison should be interpreted cautiously.

Table (4): Comparison with previous studies.

Study	Dataset / Setting	Model	Classes	Main Metric	Reported Result	Note
Basaia et al. [22]	Single-MRI AD/MCI study	Deep neural network	3	Accuracy	86.0%	Clinically oriented MRI classification
Wen et al. [21]	Review and reproducible evaluation	CNNs	Multiple	Benchmarking	Review	Comparative methodological study
Folego et al. [23]	Whole-brain MRI	3D-CNN	2	Accuracy	79–90%	Depends on task setup
Lu et al. [24]	MRI + FDG-PET	Multimodal deep network	Early diagnosis	Accuracy	Strong results	Uses multimodal input
This study (baseline)	44,000 MRI images	SGD-based baseline	4	Accuracy	40.47%	Whole-dataset computational benchmark
This study (DenseNet121)	4,000 balanced MRI images	DenseNet 121	4	Accuracy	88.67%	Strong multiclass subset result
This study (DenseNet121)	Same as above	DenseNet 121	4	Macro AUC	0.9725	Strong class separability

This comparison shows that the DenseNet121 result is strong for a four-class MRI classification setup. However, unlike several clinically grounded studies, the present dataset is augmented and upsampled, so direct numerical comparison must be made cautiously.

Discussion:

The experimental results demonstrate that DenseNet121 provides a major improvement over the full-dataset baseline. The baseline was useful for establishing a scalable reference on all 44,000 images, but its strongly compressed 12×12 inputs limited its discriminative ability. In contrast, DenseNet121 used richer 128×128 inputs and a much stronger hierarchical representation, producing 88.67% test accuracy and 0.9725 macro-AUC.

The class-wise results are clinically intuitive. The strongest performance was achieved on Moderate Demented, which reached nearly perfect metrics. This is likely because advanced dementia stages show clearer structural deviations and are therefore easier to separate. In contrast, VeryMildDemented and Non-Demented were more difficult to distinguish, which is expected given the subtle overlap between very early disease patterns and normal structural variation.

The training and validation curves indicate mild overfitting in the later epochs. Training accuracy increased to about 98%, while validation accuracy plateaued below 90%. This suggests that future experiments should incorporate stronger regularization, early stopping, or tighter augmentation control. Techniques such as dropout and batch normalization are already well established for improving generalization in deep networks [22].

A particularly important issue in this work is the risk of data leakage-like behavior due to dataset design. Because the source dataset is explicitly described as augmented and up sampled [19], there is a possibility that highly similar transformed images appear across different partitions. Even if no direct file duplication occurs, such similarity can make image-level classification appear better than patient-level generalization would justify. For this reason, the present findings must be interpreted conservatively as image-level computational results, not clinical diagnostic performance.

Another limitation is that the strongest model comparison has not yet been completed. Although DenseNet121 produced strong results, a stronger study would also include ResNet50 and DenseNet201 under the same subset and preprocessing settings. This would allow a more complete architecture-level comparison and make the paper more competitive scientifically.

Finally, explainability should be treated carefully. Grad-CAM is highly relevant to medical imaging [22], but because the final DenseNet121 Grad-CAM analysis was not completed in this stage, it should not be overstated. The paper therefore treats Grad-CAM as a future enhancement rather than a fully reported final result.

Conclusion:

This paper presented a two-stage framework for multiclass Alzheimer's disease classification from MRI images. The first stage established a whole-dataset computational benchmark across all 44,000 images, while the second stage introduced a stronger DenseNet121-based transfer learning experiment on a balanced subset of 4,000 images.

The DenseNet121 model achieved 89.17% best validation accuracy, 88.67% test accuracy, 0.8873 weighted precision, 0.8867 weighted recall, 0.8858 weighted F1-score, and 0.9725 macro-AUC. These results confirm that transfer learning on MRI images can provide strong multiclass Alzheimer's disease classification performance.

At the same time, the study emphasizes that the augmented and up sampled nature of the source dataset requires cautious interpretation. Future work should focus on training and comparing ResNet50, DenseNet121, and DenseNet201 under identical settings, evaluating with patient-level or source-aware splitting, testing on clinically grounded datasets such as ADNI and OASIS, and adding final-model explainability analysis through Grad-CAM. Overall, the study provides a stronger and more realistic MRI classification paper than the initial baseline-only version and establishes a solid experimental foundation for a more competitive journal submission.

References:

[1] World Health Organization. (2025). Dementia.

[2] 2025 Alzheimer's disease facts and figures. (2025). *Alzheimer S & Dementia*, *21*(4). https://doi.org/10.1002/alz.70235

[3] McKhann, G. M., et al. (2011). 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. Alzheimer's & Dementia, 7(3), 263–269.

[4] Jack, C. R., Jr., et al. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia, 14(4), 535–562.

[5] Mueller, S. G., et al. (2005). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's & Dementia, 1(1), 55–66.

[6] Marcus, D. S., et al. (2010). Open Access Series of Imaging Studies (OASIS): Longitudinal MRI data in nondemented and demented older adults. Journal of Cognitive Neuroscience, 22(12), 2677–2684.

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).

[8] Litjens, G., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.

[9] Wen, J., et al. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Medical Image Analysis, 63, 101694.

[10] Lu, D., et al. (2018). Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. Scientific Reports, 8, 5697.

[11] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems.

[12] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations.

[13] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444.

[14] Basaia, S., et al. (2019). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. NeuroImage: Clinical, 21, 101645.

[15] Folego, G., et al. (2020). Alzheimer's disease detection through whole-brain 3D-CNN MRI. Frontiers in Bioengineering and Biotechnology, 8, 534592.

[16] Kaggle. (2024). Alzheimer's disease multiclass images dataset.

[17] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations.

[18] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

[19] Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).

[20] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929–1958.
[21] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning (pp. 448–456).
[22] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.