Deep Learning for Individualized Dementia Severity Risk Score and Level Prediction

Alausa Salami Afolabi

***Abstract*— Early and precise assessment of dementia severity is essential for effective clinical management. In this study, we present a deep learning framework that generates individualized risk scores and corresponding severity levels from neuroimaging-derived and clinical features. Our approach employs a fully connected neural network trained via K-Fold cross-validation to ensure robust performance and generalizability. The model outputs probabilistic risk scores that are subsequently mapped to interpretable severity levels, providing actionable insights for patient monitoring and intervention. Experimental results demonstrate that the proposed framework accurately stratifies dementia severity, offering a scalable solution for enhancing diagnostic precision in clinical settings.

Keywords— Dementia, Deep Learning (DL), Neural Networks (NN), K-Fold Cross Validation, Machine Learning (ML).

Introduction

Millions are impacted by dementia [1], [2], a degenerative neurological disorder that requires prompt and precise diagnosis in order to enable effective interventions [3]. Conventional diagnostic methods rely on neuroimaging, clinical evaluation, and cognitive tests, but they frequently lack the predictive ability to accurately measure a person's risk level [4], [5]. In response, deep learning (DL) and machine learning (ML) methods have been used more and more to classify the severity of dementia, taking use of intricate patterns in multimodal data to increase prediction accuracy [6], [7], [8].

Prior research focused largely on single-modality methods, especially when employing neuroimaging data. For example, a study by [9] classified the degree of dementia from MRI scans using convolutional neural networks (CNNs). This work was expanded upon by [10], who improved classification performance by combining volumetric and functional MRI characteristics. By using deep residual networks to improve feature extraction in structural MRI-based dementia severity prediction, Lee et al. (2023) developed these techniques even further. Although these studies showed encouraging outcomes, their relevance in clinical settings with variable patient data was limited by their dependence on a single data modality.

A more recent study by [11] introduced a novel deep learning framework for Alzheimer’s Disease classification that not only identifies disease stages but also estimates classification confidence. Their approach employed leave-one-out cross-validation (LOOCV) to train a CNN on a comprehensive dataset that included cognitive and functional assessments, tau-PET and MRI neuroimaging, medical and family history, demographic factors, and APoE genotype. Also, a softmax confidence metric was used in the study to assess classification reliability based on the model’s output activity highlighting the importance of integrating multimodal data and interpretability mechanisms in dementia classification models.

Building on these advancements, our study proposes a DL-based framework for individualized dementia severity risk score and level prediction. Unlike prior single-modality approaches, we leverage a neural network (NN) model trained on diverse clinical and neuroimaging features to generate probabilistic risk scores, reflecting the likelihood of an individual belonging to a particular dementia severity category. By employing K-Fold cross-validation and incorporating dropout regularization, our model enhances generalizability while mitigating overfitting. This personalized risk assessment aims to support clinical decision-making by providing interpretable severity scores tailored to individual patient profiles.

Related Works

ML and DL techniques have been used to study dementia severity classification, with a particular emphasis on models based on neuroimaging [12], [13]. To distinguish between the severity levels of dementia, early research mostly used structural MRI and PET scans [14], [15]. Recent developments, however, have focused on enhancing model interpretability and integrating diverse data sources [16], [17]. This section discusses the flow of innovative techniques from single-modality imaging models to multimodal deep learning approaches with confidence estimates highlighting the significant studies in dementia classification.

- **Neuroimaging-Based Deep Learning Approaches**

To differentiate between various stages of cognitive decline, [18] created a CNN-based model specifically designed for Alzheimer's disease classification, utilizing deep feature extraction. Their research showed that DL is a useful tool for MRI scan analysis, despite ongoing issues with overfitting and interpretability. Building on this framework, [19] investigated network design enhancements through optimized convolutional kernels and deeper layers. Their method improved the model's classification performance by improving its capacity to detect minute structural alterations in brain shape. However, because model generalizability was impacted by differences in MRI acquisition techniques across several clinical sites, the study also brought attention to the problem of dataset heterogeneity.

The classification of dementia was further refined with the development of residual networks (ResNets). [20] used a ResNet-based model to examine sMRI data, resolving issues with previous CNN designs by maintaining spatial hierarchies in brain pictures. This breakthrough increased classification accuracy in deeper networks illustrating the progression of MRI-based machine learning from basic CNN models to more sophisticated architectures aimed at enhancing feature representation and model robustness.

- **Expanding Beyond Imaging: The Shift to Multimodal Data**

Recognizing the limitations of neuroimaging alone, researchers began integrating clinical and cognitive assessment data. [4] introduced a hybrid model combining MRI scans with neuropsychological test scores, achieving higher predictive accuracy than imaging alone. Similarly, **[5]** incorporated cerebrospinal fluid (CSF) biomarkers into their deep learning pipeline, improving early-stage AD detection. However, these studies relied on manually extracted features, limiting automation and scalability.

A more comprehensive multimodal approach was proposed by [6], who combined MRI, CSF biomarkers, and genetic data (APoE genotype) in a deep learning framework. Their model showed strong predictive performance but suffered from imbalanced datasets, necessitating synthetic data augmentation techniques. [7] addressed this issue by integrating synthetic minority oversampling (SMOTE) into their multimodal framework, improving classification balance and robustness.

- **Enhancing Model Interpretability and Confidence Estimation**

Concerns about the interpretability and reliability of DL models for MRI-based dementia diagnosis have grown as the models have developed. Although early CNN models performed well in classification, clinical trust was limited by their black-box nature. [20] addressed this problem by highlighting important brain areas affecting model predictions using gradient-based saliency maps. This method helped researchers determine if models concentrated on anatomical components important to disease by offering visual explanations for classification decisions.

A layer-wise relevance propagation (LRP) technique was introduced to trace model decisions to specific brain regions, enhancing interpretability [20]. Beyond this, confidence estimation has become vital for model reliability [21], [22].[23] explored Bayesian deep learning, using dropout during inference to generate probability distributions over predictions, enabling clinicians to assess diagnostic certainty.

To further measure model confidence, cross-validation methods have also been used. [24] utilized Leave-One-Out Cross-Validation (LOOCV), excluding each sample individually as a test case. This approach provided insights into model performance on single cases while minimizing reliance on specific training samples. Meanwhile, [25] applied K-Fold Cross-Validation, dividing the dataset into multiple subsets for training and testing across different folds. Greater consistency across folds indicated higher model generalizability, ensuring more reliable predictions in clinical applications.

- **Towards Personalized Dementia Risk Assessment**

To enhance model interpretability and confidence estimation, [11] proposed a DL approach for Alzheimer's disease classification at the individual level. Their method combined a CNN with a softmax-based confidence metric to assess classification certainty. Using a heterogeneous dataset, they employed LOOCV to train the model while concurrently estimating confidence scores. Their approach enables clinicians to assess the reliability of model decisions, fostering greater trust in AI-assisted diagnosis.

Building on these efforts, our study presents a deep learning-based model for individualized dementia severity risk prediction. Inspired by prior work, we integrate multimodal data with softmax-based confidence estimation while enhancing risk scoring through K-Fold cross-validation. Our framework bridges the gap between conventional classification and personalized risk assessment, providing a scalable, interpretable, and clinically viable solution for dementia severity prediction.

METHODOLOGY

A. Data Description

This study leverages the ADNIMERGE-3 dataset from ADNI, focusing on participants with MRI and tau-PET scans to assess tau protein deposition, a key Alzheimer's marker. The dataset integrates 224 features, including 7 sociodemographic/medical history attributes, 40 cognitive and functional assessment (CFA) scores and 177 neuroimaging-derived features from tau-PET co-registered with MRI. After preprocessing, it includes 559 participants: 363 control normal (CN), 137 mild cognitive impairment (MCI), and 59 Alzheimer's disease (AD) cases, which serve as target labels for model training.

B. Data Preparation and Preprocessing

The dataset from ADNIMERGE-3 includes clinical, neuroimaging, and cognitive assessment data [26]. Preprocessing involves structuring the data in a Pandas DataFrame [27], defining features, and setting the AD_LABEL target variable (CN, MCI, AD) with zero-based indexing for categorical classification. To address class imbalance, SMOTE generates synthetic samples for underrepresented classes [28], [29], reducing bias. The features are then standardized using StandardScaler to ensure uniform scaling [30], preventing higher-magnitude attributes from dominating model learning.

C. Model Definition

The DL model is implemented as a feedforward neural network using TensorFlow's Keras API, balancing complexity and computational efficiency. It consists of three fully connected layers with ReLU activation functions to capture non-linear relationships in the data [31]. The input layer processes feature representations, followed by hidden layers with 64 and 32 neurons. To prevent overfitting, dropout regularization (0.3) is applied after each hidden layer, randomly deactivating neurons during training to improve generalization [32]. The output layer employs a softmax activation function [33], producing probability distributions across the three diagnostic classes (CN, MCI, and AD). The model is compiled using sparse_categorical_crossentropy as the loss function, suitable for integer-encoded labels in multi-class classification. The Adam optimizer [34] is chosen for its adaptive learning rate, ensuring faster convergence and improved generalization. Accuracy is used as the primary evaluation metric to assess classification performance during training and validation [35], [36].

D. Model Training

To ensure a robust evaluation of the model's performance, 5-fold cross-validation [37] is used where the data is partitioned into five subsets, with the model being trained on four folds while the remaining fold is used for testing. This process is repeated five times, allowing each data point to be included in both training and validation phases. This approach helps mitigate overfitting to a specific dataset split and enhances the model's generalizability. To prevent overfitting, early stopping is implemented to halt training once validation performance stops improving [38]. Each fold follows an 80/20 training-validation split, with the model training for up to 50 epochs.

Results and Performance Evaluation

The model's performance was assessed using precision, recall, and F1-score, demonstrating strong classification capabilities for Alzheimer's disease severity levels (CN, MCI, AD). With an overall accuracy of 93%, the model records high precision and recall across all classes indicate reliable predictions, ensuring a balanced performance without favoring any particular category.

Class (Label)	Precision	Recall	F1-score
0 (CN)	0.93	0.9	0.9
1 (MCI)	0.89	0.91	0.9
2 (AD)	0.97	0.97	0.97

Figure 1 - Classification Performance Metrics for CN, MCI and AD Cases

The confusion matrix provides deeper insights into the model's classification behavior, illustrating how frequently each class is correctly or incorrectly predicted (see fig.1).

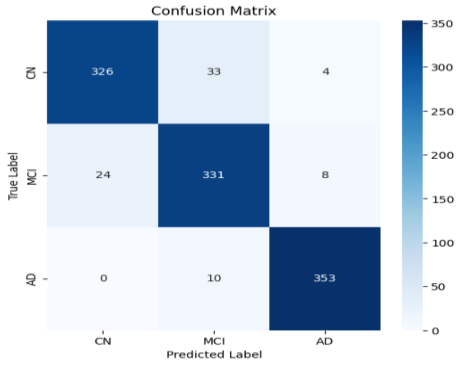

Figure 2 - Confusion Matrix Depicting Model Performance in Classifying CN, MCI, and AD Cases

The confusion matrix confirms the model's strong classification performance, correctly identifying 326 CN, 331 MCI, and 353 AD cases. Misclassifications mainly occurred between adjacent disease stages, with 33 CN misclassified as MCI and 4 as AD, while 24 MCI cases were labeled as CN and 8 as AD. Notably, only 10 AD cases were misclassified as MCI, with no CN cases wrongly labeled as AD. This pattern reflects the challenge of distinguishing between CN and MCI, as well as MCI and AD, due to symptom progression, while reinforcing the model's reliability in identifying advanced AD cases.

Analyzing the learning curves (Fig. 3 & Fig. 4) offers insights into the model's performance evolution during training.

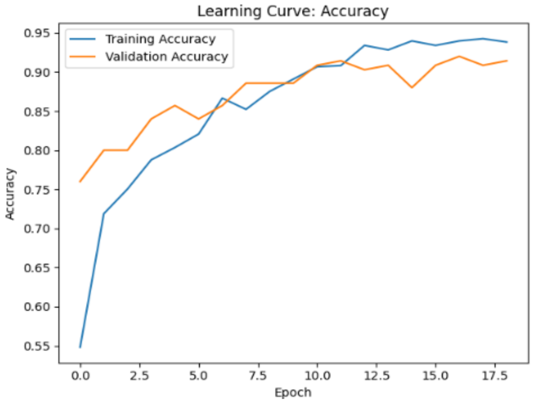

Figure 3 - Learning Curve for Accuracy

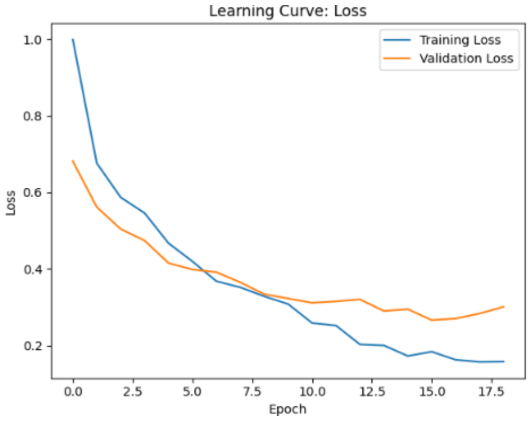

Figure 4 - Learning Curve for Loss

The learning curves illustrate a consistent decrease in training loss, indicating that the model effectively captures patterns from the data. Similarly, the validation loss follows a comparable trend, demonstrating good generalization without significant overfitting. The early stopping mechanism played a vital role by monitoring validation loss and halting training when no further improvement occurred, ensuring optimal model performance while preventing excessive fitting to the training data.

Discussion

With a 93% accuracy rate, high precision, and recall for Alzheimer's disease (AD), the study shows how well deep learning predicts dementia severity using neuroimaging and cognitive assessments. The classification report and confusion matrix validate the model's dependability, especially when it comes to differentiating AD cases, with misclassifications mostly occurring between adjacent stages (CN vs. MCI, MCI vs. AD), reflecting real-world diagnostic challenges.

Conclusion

In conclusion, this study demonstrates that a deep learning model trained on multimodal data can effectively classify dementia severity levels with high accuracy. The results suggest that such models have the potential to assist clinicians in early diagnosis and monitoring of Alzheimer's disease progression. Further validation on independent datasets and real-world clinical applications will be essential for translating this research into practical diagnostic tools. The model's high performance is attributed to its integration of diverse features, including tau-PET imaging, cognitive scores, and genetic markers. Learning curves indicate stable training, with early stopping preventing overfitting, and K-fold cross-validation enhancing generalizability.

Despite its strong performance, the study has certain limitations. The ADNI dataset [39], while comprehensive, may not fully capture the diversity of real-world clinical populations. Additionally, the inherent black-box nature of DL [38] poses challenges in healthcare decision-making. Future research should integrate Explainable AI (XAI) techniques, such as SHAP or LIME, to enhance model interpretability and clinical trust.

References

[1] A. Wimo, L. Jönsson, J. Bond, M. Prince, B. Winblad, and A. D. International, "The worldwide economic impact of dementia 2010," *Alzheimer's Dement.*, vol. 9, no. 1, pp. 1–11, 2013.

[2] A. Wimo *et al.*, "The worldwide costs of dementia in 2019," *Alzheimer's Dement.*, vol. 19, no. 7, pp. 2865–2873, 2023.

[3] G. Livingston *et al.*, "Dementia prevention, intervention, and care," *Lancet*, vol. 390, no. 10113, pp. 2673–2734, 2017.

[4] C. Jacova, A. Kertesz, M. Blair, J. D. Fisk, and H. H. Feldman, "Neuropsychological testing and assessment for dementia," *Alzheimer's Dement.*, vol. 3, no. 4, pp. 299–317, 2007.

[5] E. Pellegrini *et al.*, "Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 10, pp. 519–535, 2018.

[6] S. Grueso and R. Viejo-Sobera, "Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review," *Alzheimers. Res. Ther.*, vol. 13, pp. 1–29, 2021.

[7] A. Javeed, A. L. Dallora, J. S. Berglund, A. Ali, L. Ali, and P. Anderberg, "Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions," *J. Med. Syst.*, vol. 47, no. 1, p. 17, 2023, doi: 10.1007/s10916-023-01906-7.

[8] M. Noroozi *et al.*, "Machine and deep learning algorithms for classifying different types of dementia: A literature review," *Appl. Neuropsychol. Adult*, pp. 1–15, 2024.

[9] N. D. Rezeki, S. Aulia, and S. Hadiyoso, "Severity classification of alzheimer dementia based on mri images using deep neural network," *Rev. d'Intelligence Artif.*, vol. 36, no. 4, p. 607, 2022.

[10] A. Basher, B. C. Kim, K. H. Lee, and H. Y. Jung, "Volumetric Feature-Based Alzheimer's Disease Diagnosis From sMRI Data Using a Convolutional Neural Network and a Deep Neural Network," *IEEE Access*, vol. 9, pp. 29870–29882, 2021, doi: 10.1109/ACCESS.2021.3059658.

[11] A. S. Alausa, J. M. Sanchez-Bornot, A. Asadpour, P. L. McClean, K. Wong-Lin, and A. D. N. I. (ADNI), "Alzheimer's Disease Classification Confidence of Individuals using Deep Learning on Heterogeneous Data," in *UK Workshop on Computational Intelligence*, Springer, 2024, pp. 208–218.

[12] A. Spooner *et al.*, "A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction," *Sci. Rep.*, vol. 10, no. 1, p. 20410, 2020.

[13] M. Tanveer *et al.*, "Machine learning techniques for the diagnosis of Alzheimer's disease: A review," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 16, no. 1s, pp. 1–35, 2020.

[14] G. B. Frisoni, N. C. Fox, C. R. Jack Jr, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nat. Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, 2010.

[15] A. Chincarini *et al.*, "Alzheimer's disease markers from structural MRI and FDG-PET brain images," *Eur. Phys. J. Plus*, vol. 127, pp. 1–16, 2012.

[16] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[17] J. A. B. Moya, "Addressing the gaps in early dementia detection: A path towards enhanced diagnostic models through machine learning," *arXiv Prepr. arXiv2409.03147*, 2024.

[18] S. Bottani *et al.*, "Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse," *Med. Image Anal.*, vol. 89, p. 102903, 2023.

[19] S. Loussaief and A. Abdelkrim, "Convolutional neural network hyper-parameters optimization based on genetic algorithms," *Int. J. Adv. Comput. Sci. Appl*, vol. 9, no. 10, pp. 252–266, 2018.

[20] R. Singh and N. Sharma, "Enhanced Brain Tumor Segmentation in MRI Scans Using ResNet-150," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, IEEE, 2024, pp. 1619–1624.

[21] H. Qu, Y. Li, L. G. Foo, J. Kuen, J. Gu, and J. Liu, "Improving the reliability for confidence estimation," in *European Conference on Computer Vision*, Springer, 2022, pp. 391–408.

[22] V. T. Vasudevan, A. Sethy, and A. R. Ghias, “Towards better confidence estimation for neural models,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 7335–7339.

[23] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, and D. C. M. Saade, “A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer′ s disease and mild cognitive impairment,” *Comput. Biol. Med.*, vol. 51, pp. 140–158, 2014.

[24] J. Zhang, L. Song, Z. Miller, K. C. G. Chan, and K. Huang, “Machine learning models identify predictive features of patient mortality across dementia types,” *Commun. Med.*, vol. 4, no. 1, p. 23, 2024.

[25] M. D. Osterman *et al.*, “Examining the Performance of Polygenic Risk Scores for Alzheimer Disease Within and Across Populations Using k-Fold Cross-Validation,” *Neurol. Genet.*, vol. 10, no. 6, p. e200198, 2024.

[26] R. C. Petersen *et al.*, “Alzheimer’s disease Neuroimaging Initiative (ADNI) clinical characterization,” *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[27] W. McKinney, “pandas: a foundational Python library for data analysis and statistics,” *Python high Perform. Sci. Comput.*, vol. 14, no. 9, pp. 1–9, 2011.

[28] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” *BMC Bioinformatics*, vol. 14, pp. 1–16, 2013.

[29] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[30] S. Vinay, “STANDARDIZATION IN MACHINE LEARNING,” Mar. 2021.

[31] Y. Bai, “RELU-Function and Derived Function Review,” *SHS Web Conf.*, vol. 144, p. 2006, Aug. 2022, doi: 10.1051/shsconf/202214402006.

[32] I. Salehin and D.-K. Kang, “A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain,” *Electronics*, vol. 12, p. 3106, Jul. 2023, doi: 10.3390/electronics12143106.

[33] S. Sharma, S. Sharma, and A. Athaiya, “Activation functions in neural networks,” *Towar. Data Sci*, vol. 6, no. 12, pp. 310–316, 2017.

[34] Z. Zhang, “Improved adam optimizer for deep neural networks,” in *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, Ieee, 2018, pp. 1–2.

[35] M. Hossin and M. N. Sulaiman, “A review on evaluation metrics for data classification evaluations,” *Int. J. data Min. Knowl. Manag. Process*, vol. 5, no. 2, p. 1, 2015.

[36] G. Naidu, T. Zuva, and E. M. Sibanda, “A review of evaluation metrics in machine learning algorithms,” in *Computer science on-line conference*, Springer, 2023, pp. 15–25.

[37] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella, “The’K’in K-fold Cross Validation.,” in *ESANN*, 2012, pp. 441–446.

[38] L. Prechelt, “Early stopping-but when?,” in *Neural Networks: Tricks of the trade*, Springer, 2002, pp. 55–69.

[39] A. D. N. Initiative, “ADNI.” [Online]. Available: https://adni.loni.usc.edu/data-samples/access-data/