IL NUOVO CIMENTO **49 C** (2026) 197
DOI 10.1393/ncc/i2026-26197-9

COMMUNICATIONS: SIF Congress 2025

Development of an explainable AI model for early classification of MCI and Alzheimer's disease using T1 MRI scans

E. MAGGIO(1), C. RUNFOLA(1), M. ROMEO(1)(2)(5), G. COTTONE(1)(5), C. GAGLIARDO(3)(4) and M. MARRALE(1)(5) for the ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE(*)

(1) *Department of Physics and Chemistry "Emilio Segrè", University of Palermo Palermo, Italy*
(2) *Department of Biological, Chemical and Pharmaceutical Sciences and Technologies Palermo, Italy*
(3) *Department of Biomedicine, Neurosciences and Advanced Diagnostics (BiND), University of Palermo - Palermo, Italy*
(4) *Neuroradiology Unit, University-Hospital Paolo Giaccone - Palermo, Italy*
(5) *National Institute for Nuclear Physics (INFN), Catania Division - 95123 Catania, Italy*

received 30 January 2026

Summary. — Alzheimer's disease (AD) is one of the leading causes of dementia. Neuroimaging permits to identify and monitor the disease, but radiological analysis requires expert radiologists. Artificial intelligence (AI) offers high performance in automating this analysis but it is limited in interpretability. In this work, an AI algorithm is used to classify structural magnetic resonance images according to the stage of AD progression. An explainable AI algorithm is employed to highlight the brain regions most relevant to the predictions of the AI algorithm.

1. – Introduction

Alzheimer's disease (AD) is one of the most common neurodegenerative diseases. It is characterized by extracellular deposits of beta-amyloid proteins, hyperphosphorylation of tau protein with fibril formation, and loss of neurons and synapses. Neuroimaging is among the main diagnostic tools, and Magnetic Resonance Imaging (MRI) detects structural changes in the brain and represents a marker of disease progression [1, 2].

(*) Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`.

Radiological image analysis requires time and specialized personnel. AI enables automation of this analysis with high levels of accuracy [3], but the decisions of AI models are not interpretable and transparent, requirements that are essential for applications in the medical field [4, 5]. Hence, the need to develop explainable AI (XAI) techniques to increase robustness of AI models and thereby expand safety and accountability in diagnostic and therapeutic contexts [4, 5].

Convolutional Neural Networks (CNNs) are a staple of image analysis and pathology identification [3]. Leveraging the features extracted from the convolutional layers of a CNN, it is possible to generate a heatmap to highlight which regions in the original image impacted the network's prediction, thus making the latter explainable [3, 5]. Gradient-based Class Activation Mapping (GradCAM) is an algorithm to generate such heatmaps, which is independent of the CNN architecture and can be implemented after training [6].

In this work, we employed two binary CNN classifiers, which take 3D brain MRI images as input and output a classification label indicating the stage of AD progression. The first CNN model (model A) classifies MRI scans as either belonging to Cognitively Normal subjects (CN) or AD subjects. The other CNN model (model B) classifies scans of subjects in an early stage of AD, namely Mild Cognitive Impairment (MCI), from those of AD subjects. We employed a GradCAM algorithm to show which brain regions, according to the models, are involved in the pathologies.

2. – Methods

The data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

For training, we used T1w MP-RAGE and IR-FSPGR images, selecting the scans for the earliest visit for healthy subjects and of the last visit for pathological subjects. A single scan was chosen for each subject.

Data preprocessing included registration into MNI space using FSL [7] and brain extraction from the skull using Freesurfer (7.4.1) [8]. The images were cropped to a size of (159,197,197), to exclude most of the background. The network used for both classifiers is based on the VGG-16 architecture [9], consisting of 13 convolutional layers, one Global Average Pooling layer, and three dense layers, as shown in fig. 1. The models were implemented in Python using Tensorflow [10] (Version 2.10.0) with the Keras API (Version 2.10).

We employed rotations of ± 5 and ± 10 degrees and flipping as data augmentation.

The training dataset consists of 418 subjects (189 with AD) for model A and 419 subjects (190 with AD) for model B. The test datasets, built with ADNI subjects not used for training, include 100 subjects (50 with AD and 50 from the other class) for both models.

A 3D GradCAM algorithm was applied to the test subjects' images for both models, to highlight the regions that affect the predictions.

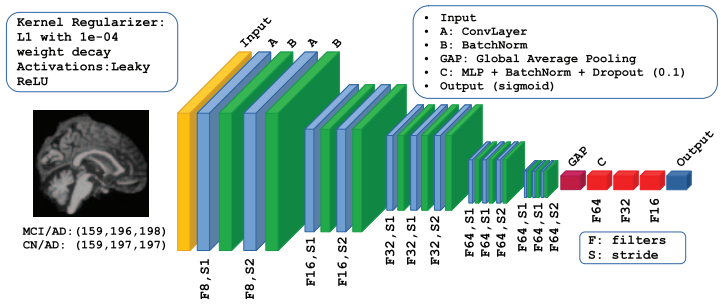

Fig. 1. – The VGG network used for the classifiers.

3. – Results and discussion

Model performance is evaluated using the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as defined in [11]. These quantities define standard metrics: *Accuracy* measures the proportion of correctly classified samples ((TP + TN) / (TP + TN + FP + FN)), *Precision* quantifies the fraction of correctly classified subjects for the positive class (TP / (TP + FP)) and *Recall* reflects the model's ability to detect the positive cases (TP / (TP + FN)).

The *Area Under the Receiver Operating Characteristic Curve* (AUC-ROC) summarizes the model's discriminative ability across all decision thresholds by plotting *Recall* against the False Positive Rate FPR $= FP/(FP + TN)$. Higher AUC–ROC values indicate stronger separation between the positive and the negative classes.

For model A, we obtained Accuracy, Precision for the AD class, Recall, and AUC-ROC metrics equal to 84%, 85.4%, 82%, and 0.91, respectively. For model B, the metrics were 75%, 74.5%, 76%, and 0.84, respectively. The reduction in model performance likely reflects the fact that MCI is a prodromal stage of AD, and consequently the structural differences between MCI and AD subjects are less pronounced than those between CN and AD subjects. These results are comparable with those obtained with other convolutional models [12, 13].

The heatmaps produced by the GradCAM highlighted regions in the frontal lobes, the basal nuclei and the lateral ventricles, as shown in fig. 2, in accordance with the scientific literature [2, 14-17].

These results indicate that the proposed XAI framework can deliver useful, interpretable results suitable for aiding early diagnosis and monitoring of disease progression. Future work should pursue external validation, multi-site cohorts, multi-modal imaging, and methodological refinements to improve MCI discrimination.

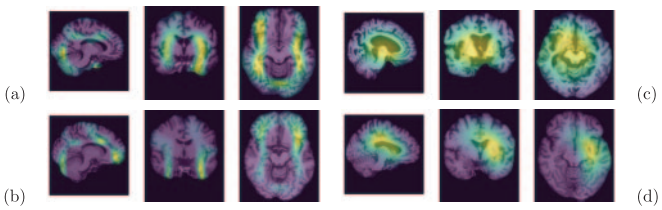

Fig. 2. – GradCAM heatmaps for correctly classified MRI scans (sagittal, coronal and axial view). (a) Model A, AD; (b) Model A, CN; (c) Model B, AD; (d) Model B, MCI.

* * *

This work has received funding from the project "AD-RICAMI" (PRJ1799) within the program "RAISE", Spoke 2 (Smart Devices and Technologies for Personal and Remote Healthcare), Project ID code: ECS00000035. This work has received funding from the Italian Ministry of Health Ricerca Finalizzata 2016 - Giovani Ricercatori, call under grant agreement no. GR-2016-02364526, from the Italian National Institute of Nuclear Physics with the research project Artificial Intelligence in Medicine (next-AIM and AIM_MIA). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California.

REFERENCES

[1] Xu L., Wang Z., Li M. and Li Q., *J. Glob. Health.*, **15** (2025) 04156.
[2] Liu E., Zhang Y. and Wang J. Z., *Transl. Neurodegener.*, **13** (2024) 45.
[3] Takahashi S., Sakaguchi Y., Kouno N. *et al.*, *J. Med. Syst.*, **48** (2024) 84.
[4] Eke C. I. and Shuib L., *Neural. Comput. Appl.*, **37** (2025) 1999.
[5] Katarzyna B. *et al.*, *Eur. J. Radiol.*, **162** (2023) 110787.
[6] Selvaraju R. R., Cogswell M., Das A. *et al.*, *Int. J. Comput. Vis.*, **128** (2020) 336.
[7] Jenkinson M. *et al.*, *NeuroImage*, **62** (2012) 782.
[8] Fischl B., *NeuroImage*, **62** (2012) 774.
[9] Simonyan K. and Zisserman A., *Very Deep Convolutional Networks for Large-Scale Image Recognition, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, 7-9 May 2015* (2015) pp. 1-14.
[10] Abadi Martín *et al.*, *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*, arXiv:1603.04467 (2016).
[11] Parikh R. *et al.*, *Indian J. Ophthalmol.*, **56** (2008) 45.
[12] Lin W. *et al.*, *Front. Neurosci.*, **12** (2018) 777.
[13] Shastry K. A., *SN Comput. Sct.*, **5** (2024) 917.
[14] Johnson J. K. *et al.*, *Arch. Neurol.*, **56** (1999) 1233.
[15] Nestor S. M. *et al.*, *Brain*, **131** (2008) 2443.
[16] Cho H. *et al.*, *J. Alzheimer's Dis.*, **40** (2014) 285.
[17] Hall A. M. *et al.*, *Alzheimer's Dement.*, **4** (2008) 271.