Using CNN AI Model For Classification of Alzheimer's Disease in MRI Brain Scans

Yazan Daker, Dean Francis Tolero, Rami Alrajei
School of Engineering, Applied Sciences and Technology
Canadian University Dubai
Email: 20220001825@students.cud.ac.ae 20210001275@students.cud.ac.ae 20220001125 @students.cud.ac.ae

***Abstract*—This project investigates the detection of dementia using brain MRI scans through a deep learning pipeline utilizing slices. 3D MRIs are converted into 2D slices to construct a training set suitable for convolutional neural networks (CNNs). A standard 2D CNN is trained to classify patients neural imaging as cognitively normal (CN) or Alzheimer's disease (AD). To improve transparency, Integrated Gradients are applied to generate pixel level attribution maps to identify the pivotal brain regions for each prediction [7]. The quality of these explanations is quantitatively assessed using the Deletion metric on the test set. The results demonstrate the use a simple CNN combined with explainability techniques to support early dementia detection from MRI imaging [3], [4].**

***Index Terms*—data mining, preprocessing, classification, open data, machine learning**

I. Introduction

Dementia is a growing global health concern with significant medical implications [4]. Early detection plays a crucial role in improving patient outcomes, enabling timely fixes, and assisting clinics in taking the decisions to help patients in need. MRIs provide structural brain information that can reveal abnormalities associated with neurodegenerative processes [3], [4]. Recent advancements in deep learning have shown strong potential in automatically learning disease-related patterns from medical images [5], [6].

This project aims to build a deep learning pipeline to classify brain MRI scans into cognitively normal (CN) and Alzheimer's disease (AD). The approach converts MRI volumes into 2D slices to reduce computational requirements while retaining diagnostic information. The model is then paired with Integrated Gradients to highlight important brain regions contributing to predictions [7], and the Deletion metric is used to verify explanation reliability. The overall goal is to develop a practical and effective method for MRI based dementia detection.

II. Domain Background

Dementia refers to a set of conditions characterized by cognitive decline, memory loss, and deteriorating functional mechanisms in the brain [2]. Alzheimer's disease is the most common form, associated with progressive atrophy of key brain regions, which are displayed in modern MRI imaging [3], [4]. MRI's are widely used in medical settings to observe these structural changes.

Convolutional neural networks (CNNs), has become a standard tool for analyzing medical images due to its ability to automatically extract spatial features [5], [6]. However, the explainability of these models remains a challenge, especially in sensitive healthcare applications where clinical trust and transparency are essential. Explainable AI methods, such as Integrated Gradients, have therefore become important for visualizing and validating model decisions [7]. This project combines neurology imaging, machine learning, and explainable AI to support diagnostic understanding in dementia detection.

III. Dataset Description

This paper uses the OASIS-1 dataset, which contains MRI scans of 412 subjects in which: "For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 100 of the included subjects over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD)" [1]. The data set includes metadata for each subject, such as the aforementioned age and sex, the CDR value determining the severity of dementia, and additional parameters such as the Mini-Mental State Examination (MMSE) score and normalized Whole Brain Volume (nWBV).

Table 1 shows the demographic data that was extracted from the data set.

CDR	Total	Male	Female	Right Handed	Left Handed
No CDR	201	89	112	201	0
0.0	135	38	97	135	0
0.5	70	31	39	70	0
2.0	2	1	1	2	0
1.0	28	9	19	28	0

TABLE I
DEMOGRAPHIC BREAKDOWN BY GROUP.

As shown in the table, all subjects were right handed, there was a fair split between male and female subjects, and almost half of the subjects did not have a CDR label. CDR values of 0 represent normal brain function, while values 0.5 and above represent mild to severe dementia. For the rest of the paper, CN

will be used to refer to cognitively normal individuals (CDR = 0) and AD to refer to subjects with Alzheimer's Disease (CDR ¿ 0).

In addition to the demographic breakdown in Table I, several visual summaries help illustrate the structure and characteristics of the OASIS-1 dataset. Figure 1 shows the age distribution of all subjects. The histogram reveals that the dataset is heavily concentrated in the 60–85 age range, with a pronounced peak around the mid-70s, indicating that most participants fall within the age group at highest risk for cognitive decline.

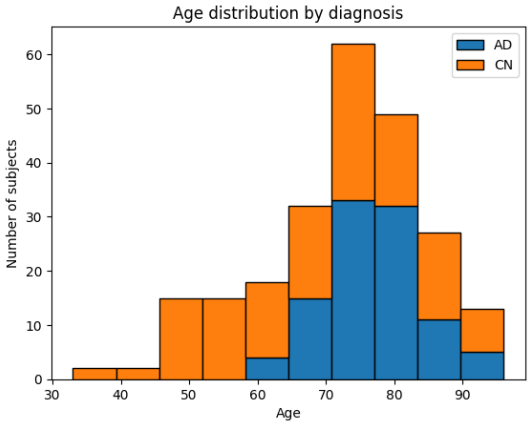

Fig. 1. Age distribution of subjects. The dataset is dominated by participants aged 60–85, with a peak around 75.

To further examine the data set, Figure 2 visualizes the relationship between MMSE scores and normalized Whole Brain Volume (nWBV). This scatter plot separates CN subjects and AD disease (AD) subjects, showing a clear trend in which AD subjects exhibit both lower MMSE scores and lower nWBV values. The key takeaway is that lower brain volume is broadly associated with reduced cognitive performance, supporting its use as a structural biomarker.

It is also evident that the dataset does not contain equal representation of the different severalties of Alzheimer's Disease, as Figure 3 demonstrates. Considering that this paper is not concerned with identifying the severity of the disease, only classifying it, this should not be a significant issue.

Together, these visualizations complement the demographic summary by illustrating the age structure of the dataset and highlighting clinically meaningful relationships between cognitive scores and brain volume. Two additional visual summaries—presented later in this section—provide further insight into class-specific slice counts and the distribution of MRI biomarkers, ensuring comprehensive characterization of the dataset.

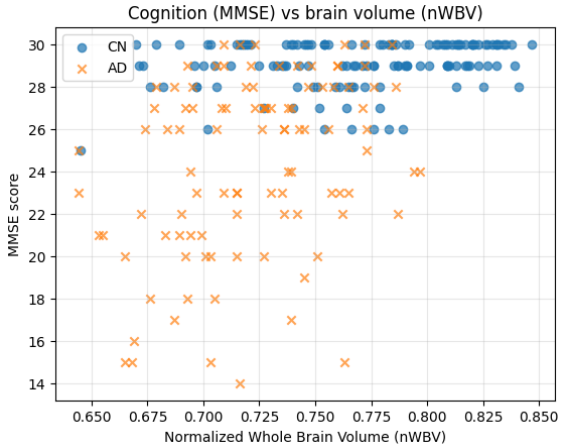

Fig. 2. Cognition (MMSE) vs. Normalized Whole Brain Volume (nWBV). AD subjects cluster toward lower MMSE and lower nWBV values, reflecting greater cognitive impairment and atrophy.

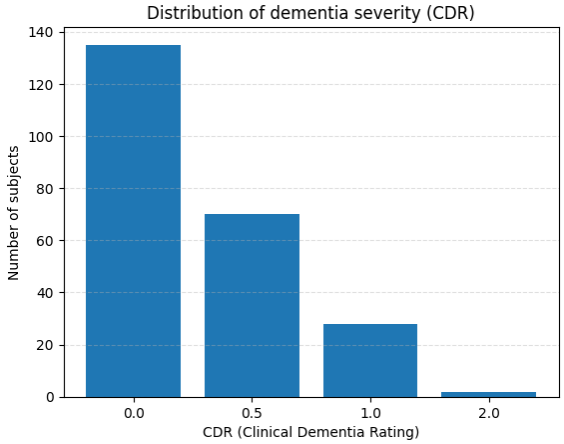

Fig. 3. CDR Distribution. AD CDRs are dominated by 0.5 mild Alzheimer's rating. 2.0 rating is very underrepresented

IV. Data Preprocessing and Methods

A. Data Quality Profiling

The original dataset consists of 3D brain imagines, as well as metadata files containing their respective information. Before preprocessing, a data quality assessment was conducted to ensure that all data points were valid and usable. This profiling revealed that not all subjects' metadata information included a CDR value, which is necessary for our model.

B. Cleaning Procedures

This issue of missing values was handled by ignoring the subjects that did not contain a valid CDR value. This reduced our dataset size but was necessary in order to ensure model accuracy.

C. Preprocessing

To prepare the MRI data for training a 2D convolutional neural network, several preprocessing steps were applied:

3D to 2D Slice Conversion Each 3D MRI volume was decomposed into individual sagittal slices. Slices containing no brain tissue (e.g., extreme ends of the 3D volume) were removed.

Normalization All slices were intensity-normalized using min–max normalization to ensure consistent brightness and contrast across scans, as the raw MRI images would often include hot-spots of high intensity that would otherwise dominate the rest of the image.

Label Assignment (Feature Creation) Each slice was labeled CN or AD according to the range that CDR value of the subject that the slice was extracted from belonged.

Subject-Level Train/Validation Split An 80/20 split of the data was applied at the to prevent data leakage. No slices from a single subject appeared in both sets.

D. Methods

Binary Classification: Binary classification is a type of supervised machine learning task where the model decides between two possible classes based on patterns it sees in labeled examples.

Convolutional Neural Network (CNN): Convolutional Neural Networks are a type of Neural Network AI model that is optimized for handling image data that can be used for classification data mining tasks. Like a traditional neural network, it includes an input, hidden, and output layer. Its hidden layer contains convolution layers which apply convolution operations to the data at each layer.

Integrated Gradient (IG): Integrated Gradient was used as part of XAI, to help give an understanding of what part of the brain is actually affected and whether they are cognitively normal or are classified as AD patients.

The method computes gradients along a path from a baseline image to the input, producing a heatmap highlighting influential brain regions.

Deletion Metric: A way to evaluate the importance of input features or regions in machine learning models that use images by measuring how model performance changes when parts of the input are removed.

V. Experiments and Metrics

This project implemented an end-to-end machine learning pipeline for dementia classification using brain MRI scans from the OASIS dataset. Binary classification was employed to distinguish between cognitively normal (CN) patients and those with Alzheimer's disease (AD) based on their Clinical Dementia Rating scores. A custom Convolutional Neural Network was designed to automatically learn and extract spatial features from the MRI images, progressively identifying patterns from simple edges to complex brain structures across four convolutional blocks. To make the model's decisions interpretable and trustworthy, Integrated Gradients was applied to generate attribution maps that highlight which brain regions the CNN focused on when making its predictions, creating heatmaps that show pixel-level importance. Finally, the Deletion Metric was used to validate the quality of these explanations by systematically removing the most important regions identified by Integrated Gradients and measuring the resulting drop in model performance, ensuring that the highlighted regions were genuinely influential to the classification decisions rather than arbitrary patterns.

Data results and Quality Profiling

- Checking for corrupted MRI files

- Removing non-informative slices

- Normalizing values to one single interval

- Standardizing image dimensions

- Deleting duplicate slices

Data Transformation

- 3D to 2D transformation

- Min–max normalization

- Data augmentation (rotation, flipping, cropping)

- Dimensionality reduction, through image slicing. (ReLu, padding)

Evaluation Metrics

Accuracy measures the number of correct predictions made by a model. To represent it's performance.
Precision measures the model's capability to avoid false positives and measures how many of the positive predictions are actually true. If it is high, it shows that the model is trustworthy.
Recall measures how capable the model is to catch positive instances. High recall misses only a few values.
F1-Score a single metric that compares Precision and Recall, using a harmonic mean. (F1 = 2 * (Precision * Recall) / (Precision + Recall). It displays the trade-off between Precision and Recall, where a high F1-score (closer to 1.0) shows that those metrics are high, whereas a low F1-score (closer to 0.0) shows that the metrics are lower and therefore model being less reliable.
Loss is the measure of how wrong the model's predictions are. Showing how close the predictions are to the actual results. The primary goal of a model is to minimize this loss after training.

VI. Results

These results display our final model's findings after running on one large dataset split in 12 subsections, using the first 11 for training, and the twelfth for model testing.

A. Training and Validation Loss

Figure 4 shows the progress of the model every five epochs, displaying the training loss (the loss value in the training set) and the validation loss (the loss value in the validation set). The figure displays the training loss steadily decreasing every epoch, while validation loss spikes around epoch 15. This could be due to the model overfitting to the training set initially. As it approaches epoch 25, the training and validation loss are more closely aligned; however, training loss is still significantly lower than validation loss, indicating that the problem wasn't entirely eliminated.

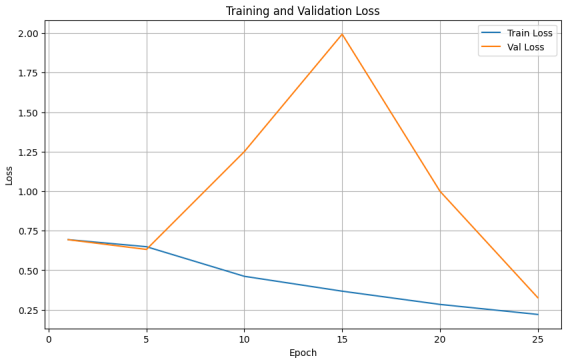

Fig. 4. Graph of Training and Validation Loss Over Epochs

B. Validation Accuracy and F1 Score

Figure 5 demonstrates the accuracy and F1 values of the model throughout training. Initially, the validation accuracy as greater than the F1 score. However, on epoch ten, accuracy decreases much more than F1 score, giving the model a higher F1 score. This could be due to the model initially focused on the class with the higher weight, initially inflating accuracy and F1, then once it start learning to identify the lower weight class, the accuracy went down as it made mistakes, but F1 remained fairly consistent as it is more forgiving. Eventually though, F1 score takes a sharp downward trend, before eventually intersecting with accuracy at around 0.85, which is a significant score.

C. Post-Training Results

As shown in Table II, after training, the model has an accuracy of 80%. The precision shows that the model can predict (AD) correctly, 97% of the time, which implies that when it predicts that a subject is AD, it has a low false positive rate. The model only catches 60% of the AD cases. Approximately 4 out of 10 patients who actually have AD,

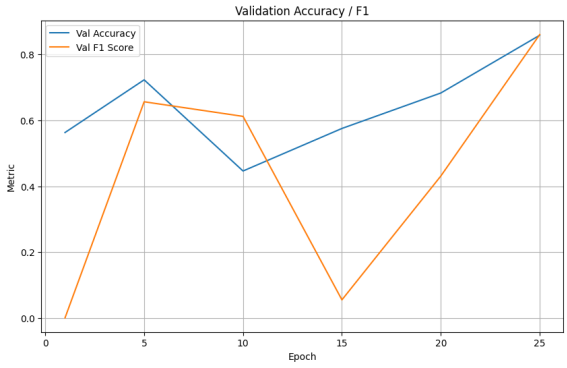

Fig. 5. Enter Caption

TABLE II
Classification Performance Metrics

Metric	Value
Accuracy	0.8012
Precision	0.9694
Recall	0.6051
F1-Score	0.7451

end up classified as CN, which means that the model's biggest weakness is flagging false negatives. Hence the balanced F1-Score noting the caution that the model has with its precision, despite it missing many true cases due to the lower recall.

D. Evaluation Metrics

The model then ran integration gradient and deletion metric calculations for the dataset. This resulted in deletion Score of 0.7895, which implies that if the attributes that were identified by the integrated gradients were to be removed from an AD subject's slice, the model would then identify it as CN around 80% of the time, which means that the attributes that were identified are closely related to what the model utilizes to judge the image slices.

Figure 6 illustrates the results in which a sample brain slice was used to visually display all the outputs of the model from prediction to deletion metric.

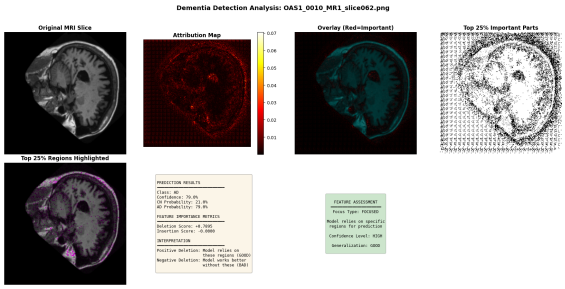

Fig. 6. Sample Analysis Summary

VII. Discussion and Limitations

Creating the model through a 2-Dimensional approach, where we take the 3-Dimensional data and slice it into 2D image slices, was the approach chosen to conduct this experiment. There exists a model *conv3d*, which also could have been applied and might have given different results. The dataset provided by the Open Access Series of Imaging Studies (OASIS) had more or less 50% missing *CDR* values, which made the model train with a smaller dataset. It was discussed that the missing values could be assumed as *CN* (Cognitively Normal). The dataset notes that "100 of the included subjects over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD)."[1] The model was trained under different hardware devices, some training the model significantly faster. Devices that did not come with a GPU (Graphics Processing Unit) ran significantly slower compared to devices that were trained with a GPU.

VIII. Conclusion

This project developed and evaluated a deep learning pipeline for the binary classification of Alzheimer's disease from brain MRI scans. By converting 3D MRI volumes into 2D slices, we demonstrated that a standard 2D CNN architecture can effectively learn discriminative features for this task, achieving a final accuracy of 80.12%. The model exhibited high precision (96.94%), indicating that its positive AD predictions are highly reliable, though the moderate recall (60.51%) suggests a tendency to miss some true AD cases, an important consideration for clinical deployment.

A core contribution of this work was the integration of explainable AI (XAI) techniques to address the "black box" nature of deep learning models in a medical context. The application of Integrated Gradients provided attribution maps at the pixel level, visually highlighting the brain regions—such as those affected by atrophy that the model used to decide its class. The following use of the Deletion metric offered a quantitative assessment, confirming that the regions highlighted by the XAI method were genuinely influential to the model's performance, thereby validating the reliability of the explanations.

Despite these promising results, several limitations must be acknowledged. The preprocessing step, which involved discarding subjects with missing CDR values, reduced the effective dataset size. Furthermore, the conversion of 3D volumes to 2D slices, while computationally efficient, may discard potentially valuable spatial information that a 3D model could have utilizied.

For future work, employing 3D convolutional neural networks could capture the full spatial context of the brain anatomy. Expanding the dataset, potentially by estimating missing labels or using additional data sources, would improve model robustness and generalizability. Finally, exploring more advanced techniques to hand the class imbalance, could help decrease the number false negatives and improve recall, making the model more sensitive for early-stage detection. In summary, this pipeline presents a practical and explainable approach to MRI-based dementia detection, establishing a foundation for building more complex and trustworthy diagnostic tools.

References

[1] Apr. 2024. URL: https://sites.wustl.edu/oasisbrains/home/oasis-1/.

[2] Alzheimer's Association. "2018 Alzheimer's disease facts and figures". In: *Alzheimer's & Dementia* 14.3 (2018), pp. 367–429.

[3] Giovanni B. Frisoni et al. "The clinical use of structural MRI in Alzheimer disease". In: *Nature Reviews Neurology* 6 (2010), pp. 67–77.

[4] Clifford R. Jack, David S. Knopman, William J. Jagust, et al. "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade". In: *The Lancet Neurology* 9.1 (2010), pp. 119–128.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521 (2015), pp. 436–444.

[6] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88.

[7] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *International Conference on Machine Learning (ICML)*. 2017.

D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," Journal of Cognitive Neuroscience, vol. 19, no. 9, pp. 1498-1507, Sep. 2007. doi: 10.1162/jocn.2007.19.9.1498.