01 Jan 2023

An Explainable Deep Learning Model For Prediction Of Severity Of Alzheimer's Disease

Godwin Ekuma

Daniel B. Hier
Missouri University of Science and Technology, hierd@mst.edu

Tayo Obafemi-Ajayi

Recommended Citation

G. Ekuma et al., "An Explainable Deep Learning Model For Prediction Of Severity Of Alzheimer's Disease," *CIBCB 2023 - 20th IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, Institute of Electrical and Electronics Engineers, Jan 2023.
The definitive version is available at https://doi.org/10.1109/CIBCB56990.2023.10264880

2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) | 979-8-3503-1017-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CIBCB56990.2023.10264880

An Explainable Deep Learning Model for Prediction of Severity of Alzheimer's Disease

Godwin Ekuma
Computer Science Dept.
Missouri State University
Springfield, MO, USA
goe948s@missouristate.edu

Daniel B. Hier
Electrical & Computer Eng. Dept
Missouri University of Science & Technology
Rolla, MO, USA
hierd@mst.edu

Tayo Obafemi-Ajayi
Engineering Program
Missouri State University
Springfield, MO, USA
tayoobafemiajayi@missouristate.edu

Abstract—**Deep Convolutional Neural Networks (CNNs) have become the go-to method for medical imaging classification on various imaging modalities for binary and multiclass problems. Deep CNNs extract spatial features from image data hierarchically, with deeper layers learning more relevant features for the classification application. Despite the high predictive accuracy, usability lags in practical applications due to the black-box model perception. Model explainability and interpretability are essential for successfully integrating artificial intelligence into healthcare practice. This work addresses the challenge of an explainable deep learning model for the prediction of the severity of Alzheimer's disease (AD). AD diagnosis and prognosis heavily rely on neuroimaging information, particularly magnetic resonance imaging (MRI). We present a deep learning model framework that integrates a local data-driven interpretation method that explains the relationship between the predicted AD severity from the CNN and the input MR brain image. The deep explainer uses SHapley Additive exPlanation values to quantify the contribution of different brain regions utilized by the CNN to predict outcomes. We conduct a comparative analysis of three high-performing CNN models: DenseNet121, DenseNet169, and Inception-ResNet-v2. The framework shows high sensitivity and specificity in the test sample of subjects with varying levels of AD severity. We also correlated five key AD neurocognitive assessment outcome measures and the APOE genotype biomarker with model misclassifications to facilitate a better understanding of model performance.**

Index Terms—**Alzheimer's Disease, MRI, Deep learning, Explainability, Prediction models**

I. Introduction

Deep neural networks have demonstrated remarkable success in creating predictive models for diverse applications, including precision medicine [1]. Despite their high predictive accuracy, usability lags in practical applications due to the perception that they operate as black boxes. Model explainability and interpretability are essential for successfully integrating artificial intelligence (AI) into healthcare practices [2]. In this paper, we adopt the definition of Explainable AI from [2] as 'techniques and methods to build AI applications that end users can understand and interpret.' These can be categorized as ante-hoc (i.e., how is the model producing its result?) or post-hoc (i.e., why is the model producing this result?) approaches [3]. Four characteristics of explainability have been recommended for precision medicine: interpretability, understandability, usability, and usefulness [3]. To build deep learning prediction models that are useful (i.e. are of practical worth), we need to promote the understandability and interpretability of the models. Understandability deals with how the system is working, while interpretability focuses on whether the results make sense. To uncover the black-box model that hides the operations of deep neural networks, Liang et al. [4] suggest that local interpretable approaches that analyze individual cases may be more suitable than global ones, given the advantages of easier implementation and lower computational complexity. In this work, we address the challenge of creating an explainable deep learning model for predicting the severity of Alzheimer's disease based on a local interpretation method that focuses on enhancing the understandability and interpretability of the neural network.

Alzheimer's disease (AD) is a common neurodegenerative disorder of aging populations [5]. It is slowly progressive, leading to the death of brain cells, which results in memory loss and cognitive decline [6]. Research on AD is a national priority, with 5.5 million Americans affected at an annual cost of more than $250 billion and no definitive cure available [5]. Early detection allows for more precise treatment, enhances patient outcomes, and reduces unnecessary hospitalization. The diagnosis of AD relies heavily on neuroimaging, especially magnetic resonance imaging (MRI) [7]. MRI is the preferred neuroimaging method, providing abundant information on anatomical structures at high spatial resolution. Dementia is at the most severe end of the spectrum of cognitive impairment seen in aging. Cognitive impairment may be minimal as observed in cognitively normal (CN) aged persons, progress to more severe cognitive impairment without dementia known as mild cognitive impairment (MCI), and evolve into the full-blown dementia of AD [8]. For the accurate diagnosis and classification of Alzheimer's patients, a systematic analysis of multiple types of biological information (including neuroimaging) is needed [9]. This paper is an initial step in that direction as we attempt to incorporate neurocognitive assessments into the MRI analysis.

Deep learning models can perform automated detection and classification of AD severity. Machine learning analysis of extracted neuroimaging data features for AD has been conducted extensively with support vector machine models and obtained high levels of performance (see [5] for a detailed review). Deep

979-8-3503-1017-7/23/$31.00 ©2023 IEEE

learning models hold the promise to analyze whole scans, not just extracted regions. Hence, multiple convolutional neural networks (CNN) models [6], [10]–[12] have been applied to entire brain images. The goal of adding interpretability to these models is to maintain accuracy while enhancing reliability and understandability. We present a deep learning model framework that integrates a local data-driven interpretation method to explain the relationship between the predicted AD severity from the CNN and the inputted MR brain image. We conduct a comparative analysis of three CNN models. The data-driven interpretation model utilizes a deep explainer based on SHapley Additive exPlanation (SHAP) values [13] to quantify the contribution of different brain regions utilized by the CNN to predict outcomes. We correlate five key AD neurocognitive assessment outcome measures and the APOE genotype biomarker with model misclassifications to better understand model performance.

II. Background

A. Convolution Neural Network Architectures

This section briefly describes each CNN model utilized in this work. Dense convolutional network (DenseNet) [14] is characterized by a dense connectivity pattern in which all layers are directly connected with each other (see Figure 1). The intent is to maximize information flow between layers in the network by combing features using concatenation rather than summation. The l^{th} layer has l inputs from the feature maps of all the preceding blocks [14]. In turn, its own feature maps are passed on to all the $L-l$ subsequent layers, resulting in $\frac{L(L+1)}{2}$ connections in an L-layer network. DenseNet consists of a series of dense blocks of varying filter sizes. From one dense block to another, a transition layer takes care of the downsampling via convolution and pooling operations [14]. Each dense layer implements a composite function of operations: batch normalization, rectified linear units (ReLU) and convolution (see Figure 1). DenseNet can have very narrow layers due to the collective knowledge possessed at each point in the network, given the dense connectivity. This is determined by the hyperparameter k, which denotes the growth rate of the network. It regulates how much new information each layer contributes to the global state. Variants of DenseNet can be derived by varying the depth, which is a function of the number of layers in the dense blocks. In this work, we use DenseNet121 and DenseNet169.

Inception-residual network (Inception-ResNet) is a hybrid neural network that leverages the strength of residual connections with the inception architecture [15]. Residual connections enable the training of very deep networks as they utilize skip connections that short-circuit shallow layers to deep layers, enabling the layer to adjust the input rather than attempting to learn the entire function from scratch [16]. The inception architecture utilizes sparse connectivity patterns to decrease the number of connections without compromising network efficiency. It consists of multiple stacked inception module layers with parallel convolutions of various filter sizes and pooling operations that are concatenated to create a single output passed to the next layer [17]. Since inception networks tend to be very deep, Inception-ResNet replaces the filter concatenation stage with residual connections [15].

The Inception-ResNet, as illustrated in Figure 2, consists of three main types of modules: stem (initial layer), Inception ResNet, and reduction blocks [18]. The Inception ResNet blocks (A, B and C) have identical convolutions but varying filter sizes. It aids the network in learning robust representations from the input image. The reduction blocks use pooling along with convolution paths for feature reduction to improve computational efficiency. Each of the three blocks (A, B, and C) differs in the number of max-pooling and convolution paths. In this work, we utilize the Inception-ResNet-v2 [15] variant, a costlier hybrid Inception version with significantly improved recognition performance.

Both DenseNet and Inception-ResNet networks aim to improve accuracy by reducing the number of parameters and promoting feature reuse. The primary difference lies in how they connect the layers in their architectures. Inception-ResNet uses residual connections to add the output of one layer to another layer deeper in the network, while DenseNet uses dense connections to concatenate the output of each layer to the input of all subsequent layers.

B. Neurocognitive Assessments and Genotype Biomarker

To provide context for the clinical relevance discussion (Sections V and VI), this section provides a brief description of commonly used AD outcome measures and the APOE genotype biomarker utilized in this work.

The Mini-Mental State Examination (MMSE) [19] is a cognitive function test that involves a pen-and-paper assessment of various cognitive abilities such as orientation, concentration, attention, verbal memory, naming, and visuospatial skills. The test has a maximum score of 30 points. Although exact cutoff points for the MMSE have not been established, scores of 28-30 are generally considered normal, 26-27 are indicative of MCI, and scores below 25 are suggestive of AD. The Clinical Dementia Rating (CDR) [19] is a global rating scale for staging patients diagnosed with dementia. The CDR evaluates cognitive, behavioral, and functional aspects of Alzheimer's disease and other dementias. The CDR is based on a scale of 0–3: no dementia (CDR = 0), questionable dementia (CDR = 0.5), MCI (CDR = 1), moderate cognitive impairment (CDR = 2), and severe impairment (CDR = 3).

The Functional Activities Questionnaire (FAQ) [20] is a 10-item measure of instrumental activities of daily living functional status. FAQ items are rated on a 3-point ordinal scale of 0–3 (0 = normal or never did but could do now; 1 = has difficulty but does by self or never did but would have difficulty now; 2 = requires assistance; 3 = dependent). Total scores range from 0 to 30, with higher scores indicating greater impairment. Digit Span score [21] includes digits forward and digits backward. Digits forward require the subject to repeat numbers in the same order as read aloud by the examiner, while for digits backward, the subject repeats the numbers in the reverse order. The Alzheimer Disease Assessment Scale

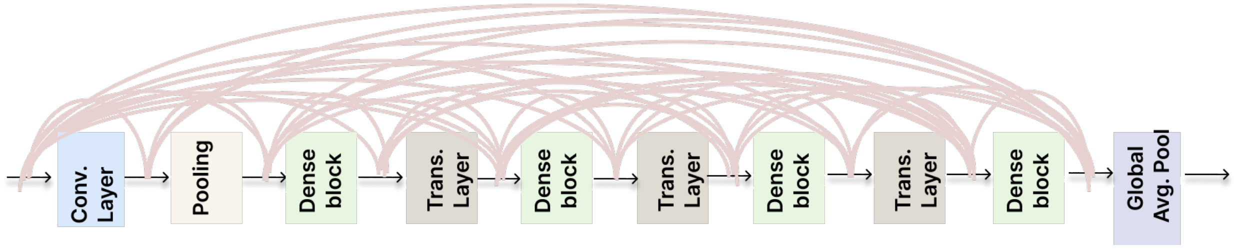


Fig. 1. DenseNet architecture. There is a connection between each layer and all subsequent layers. Differences in the variants such as DenseNet121 and DenseNet169 lies in the varying number of dense layers in the blocks. Conv.: Convolution; Trans.: Transition.

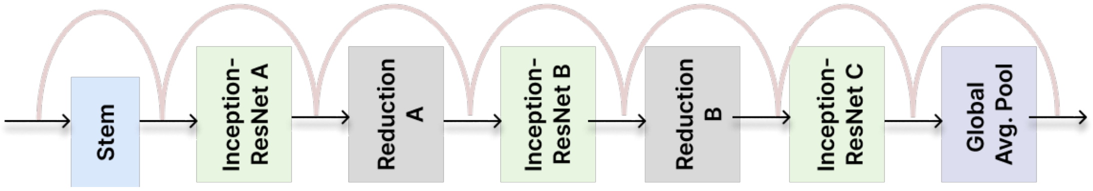


Fig. 2. Inception-ResNet architecture. There is a residual connection between each layer and all subsequent layers.

(ADAS) comprises cognitive and noncognitive sections [22]. Its cognitive section (ADAS-Cog) includes standard tests of language, comprehension, memory, and orientation, tests of visual-spatial ability, such as drawing geometric figures, and physical tasks that reflect ideational praxis, such as folding a paper into an envelope. Patients obtain scores of 0 to 70, with higher scores indicating poorer performance.

The apolipoprotein E (APOE) is a gene that provides instructions for making a protein called apolipoprotein E [23]. There are three versions of the APOE gene, called alleles: $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$. Each person has two copies of the gene (one inherited from each parent) so the APOE genotype represents the combination of inherited alleles. Having two copies of the APOE $\epsilon 4$ allele is strongly linked to an increased risk of developing AD [24]. Following that, individuals with the combination of one APOE $\epsilon 3$ allele and one APOE $\epsilon 4$ allele ($\epsilon 3/\epsilon 4$) also have an elevated risk compared to those with two copies of $\epsilon 3$. Conversely, individuals with two APOE $\epsilon 2$ alleles ($\epsilon 2/\epsilon 2$) tend to have a lower risk of developing AD.

III. Methods

A. Data Acquisition and Preprocessing

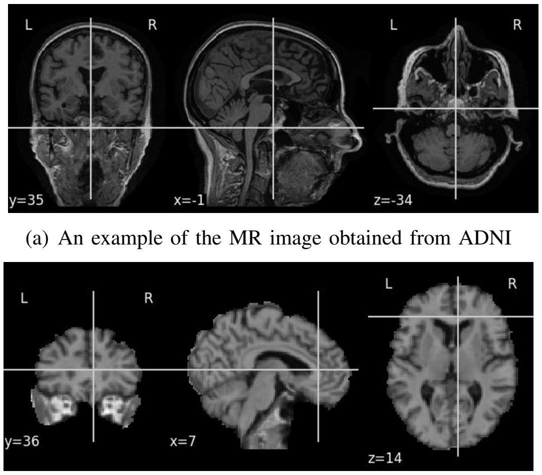
The patient sample analyzed in this work is drawn from the Alzheimer's Disease Neuroimaging Initiative database [8], a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD. There are multiple patient cohorts within ADNI: ADNI-1, ADNI-2, ADNI-3, ADNI-GO. Each consists of neuroimaging data with diverse imaging types (3D and 2D) that underwent certain acquisition and preprocessing phases. We selected a subset of T1-weighted MR images from the ADNI-1 cohort, as it had the largest sample of images in 3D format, processed with N3 correction standard, in the same image acquisition plane (sagittal). The sample consisted of 325 AD, 595 CN, and 1024 MCI images, a total of 1944 from 488 unique patients. Some of the images were scans from obtained at different visit dates for the same patients. All participants were between the ages of 50 and 80 years. A brief demographic description of the 1944 images by class used is presented in Table I.

TABLE I
Sample demographic and neurocognitive characteristics.

(n)	AD (325)	MCI (1024)	CN (595)
Male/Female	149/176	588/436	276/316
Age (mean/std)	72.5 ± 5.4	72.1 ± 5.5	74.7 ± 3.7
Education (Yrs)	14.2 ± 3.1	15.6± 3.0	15.9 ± 2.9
MMSE	21.1 ± 5.4	25.7 ± 3.7	29.2 ± 1.0
CDR Global	17.3 ± 7.8	7.2± 7.0	0.2 ± 1.0
FAQ	1.0 ± 0.5	0.6± 0.3	0 ± 0.1

MMSE: Mini Mental State Examination; CDR: Clinical Dementia Rating
FAQ: Functional Assessment Questionnaire

Certain preprocessing steps were applied to each ADNI-1 image before it could be used in the deep learning model, described as follows. Skull stripping was applied to remove non-brain tissue voxels from images, including skin, fat, muscle, neck, and eyeballs. We utilized the Functional MRI software library (FSL) brain extraction tool [25]. Subsequently, spatial normalization is applied to standardize image orientation and voxel spacing, as these factors can differ among images even if they are acquired from the same scanner. It minimizes variations in image positioning, orientation, shape, and size in the dataset. We employed the FSL's linear image registration tool [26] to linearly align all scans to the T1 MNI 152 template with 2mm isotropy. This yielded 3D images of a standard size of 91x109x109. Figure 3 illustrates the effect of skull stripping and spatial normalization on the raw ADNI-1 images. Each 3D image was resliced into three distinct 2D image sets (sagittal, coronal, and axial orientations) using the Nibabel image slicer [27]. Note that subsequent analyses were conducted only on the axial image slices. Visual examples of



(a) An example of the MR image obtained from ADNI

(b) Brain image after skull stripping and spatial normalization.

Fig. 3. Effect of skull stripping and spatial normalization on MR image.

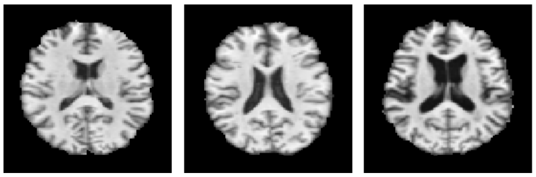


Fig. 4. Images of center axial slices in order of severity: CN, MCI and AD.

the axial slices obtained from MR images of varying severity are presented in Figure 4. To ensure that all the pixels in each image are within a normalized range, we also performed data intensity (or voxel-based) normalization using the zero-mean unit-variance method [28].

B. Class Balancing using SMOTE

In medical data collection, a known issue is the skewed distribution of samples between healthy controls and cases. Though the unique patient distribution in the entire ADNI-1 cohort is relatively balanced, the 1944 sample of MR images obtained were highly skewed (30.6% CN, 52.6% MCI, and 16.7% AD). To avoid possible overfitting and bias of the learning model due to the skewed distribution, we applied the synthetic minority oversampling technique (SMOTE) sampling approach [29], [30]. SMOTE generates synthetic examples in feature space by over-sampling the minority class. Synthetic examples are derived by introducing examples along the line segments joining any/all of the k nearest neighbors from the minority class. Depending on the desired amount of over-sampling, neighbors from the k nearest neighbors are randomly selected. A feature vector X_0 from the sample under consideration is chosen from the minority class. One of its k nearest neighbors X, also belonging to the minority class, is randomly selected. The difference between X and X_0 is then computed, and new synthetic data is generated on a random point in the line segment by connecting the feature vector and the selected neighbor: $Z = X_0 + w\,(X - X_0)$, where w is a uniform random variable in the range [0,1]. Examples of the SMOTE generated images for both AD and CN classes are illustrated in Fig. 5.

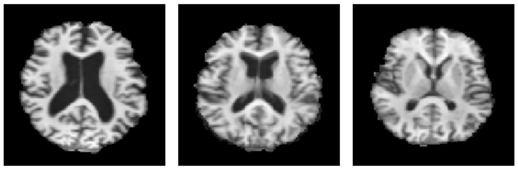


Fig. 5. SMOTE outcome images: the first two are AD while the third is CN.

C. Construction of CNN Learning Model Framework

The overall explainable deep learning model for prediction of the severity of AD from MR brain images is illustrated in Figure 6. It consists of the CNN learning model and the explainability extension. We compare the performance of three CNN architectures: DenseNet121 [14], DenseNet169 [14], and Inception-ResNet-v2 [15]. (See Section II-A for a brief description of each CNN model). A drawback of CNN models on images is the long training times. As a result of the readily available varied pre-trained CNN architectures on Imagenet that is implemented in Keras with Tensorflow backend [31], researchers can integrate transfer learning to enhance efficiency of training times. The basis is that models trained on one problem (Imagenet [32]) can be used as a starting point for training new models on a related problem. Transfer learning allows the use of weights from pre-trained models developed from standard computer vision benchmark data into new models. Transferring the weights (and network parameters) from a pre-trained generic network to train on a specific data set tends to perform better than random weight initialization of the network [33].

The base model is constructed from the pre-trained network and initialized with the Imagenet training weights. We unfreeze some of the higher layers to allow the CNN model to encode more subtle features from the brain MR images geared towards AD severity distinction. The fine-tuning layers consist of the global average pooling layer, two dense layers of varying sizes (256 and 128 neurons), and the fully connected layer. The dense layer included a ReLU activation function, L1 regularization, and a dropout layer to increase the resilience of the network and reduce overfitting. The fully connected layer utilized the softmax activation function for the three classes. We employed the RMSprop optimizer to minimize the categorical cross-entropy loss function.

D. Model Explainability using SHAP Deep Explainer

SHAP [13] is implemented as a data-driven local interpretation method to unravel the black box deep learning model [4]. The SHAP values are computed using the Deep explainer to identify the contribution of each feature in the image in the prediction of AD severity. The number of unique prediction classes determines the number of images SHAP generates when explaining a prediction. In this three-class AD severity context, SHAP calculates the feature importance per pixel and generate 3 explainable images per class (CN, AD, and MCI)

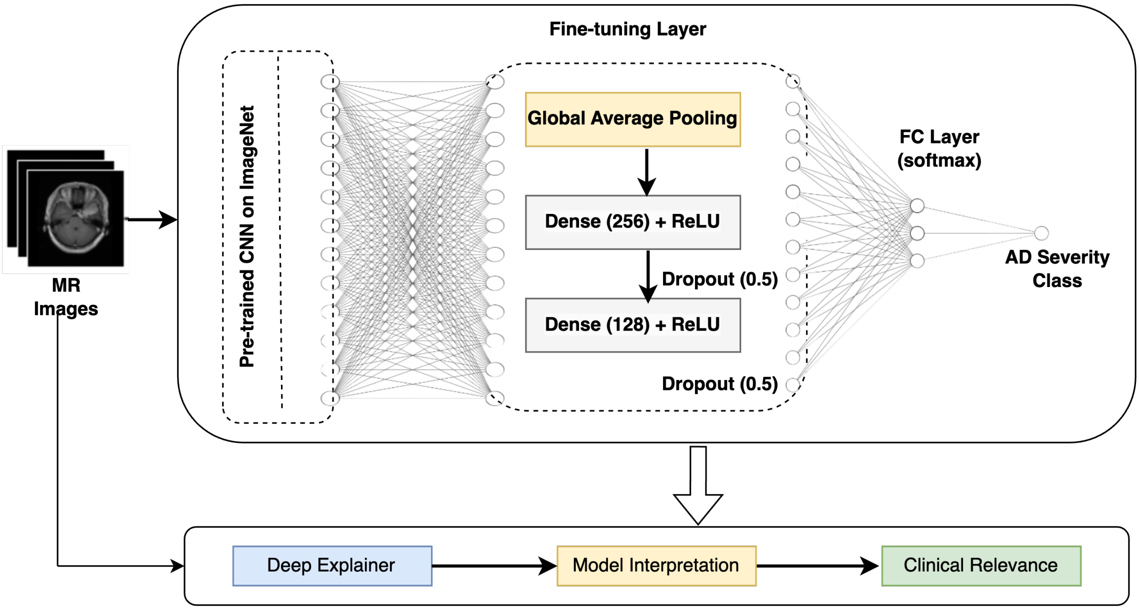


Fig. 6. Explainable learning framework for prediction of AD severity. FC: Fully connected layer. ReLU: Rectified linear unit activation function

It denotes pixel importance by color. Red indicates a positive correlation, while blue indicates a negative correlation on the predicted value. The intensity of the color reflects the level of impact a specific feature has on the prediction. Both red and blue are important, but they have opposite effects on the output. The intensity of the color is crucial as any deeply colored area represents a feature that contributes to the 'feature importance' in the model's classification. Therefore, the SHAP plot can provide valuable insights into how different features contribute to the neural network's output and identify the key regions that seemed to influence the predictions.

In the model interpretation phase, we analyze the correctly classified and incorrectly classified results of the AD and MCI predictions to examine what could have hindered the performance of the CNN model on those tasks. The key question is does the image prediction class line up with the other biological information available? A domain expert assessed the SHAP plot results and the neurocognitive outcome measures to determine the clinical relevance of the results.

IV. Experimental Results and Analysis

All experiments were implemented using Python and TensorFlow ML API on a NVIDIA GEFORCE GTX 1050 GPU machine. In Tensorflow, individual operations and modules in a network are considered as separate layers, including convolutional blocks and auxiliary layers such as activation, batch normalization, pooling, and global average pooling. The DenseNet121, DenseNet169, and Inception-ResNet-v2 architectures have 433, 601, and 785 layers, respectively. To fine-tune the models, we unfroze the later blocks, with DenseNet121 at the 313th layer, DenseNet169 at the 369th layer, and Inception-ResNet-v2 at the 616th layer.

Applying SMOTE on initial skewed distribution of 1944 images yielded a balanced size of 1024 for each severity class, a total of 3072 images. The sample was split into training (80%) and testing (20%) subsets. Note that the SMOTE generated images were all intentionally constrained to the training data so the test set consisted of only real images. During training, the data augmentation techniques (horizontal flipping, height and width shifts within ranges of 0.05 and 0.1, rotation within a range of 5 degrees, and zoom within a range of 0.15) were automatically applied to improve the model's robustness by increasing variability of data. A stratified 5-fold cross-validation was applied for training with a batch size of 32, and learning rate of 0.00001. The best performing fold training model was employed on the test set.

The following metrics were employed to evaluated the model's learning performance: classification accuracy, sensitivity/recall, specificity, and the area under the curve receiver operating characteristic curve (AUC-ROC). Let TP denote true positive, TN true negative, FP false positive, and FN false negative. Accuracy (ACC) assesses how well the model can predict true positives and negatives within the classes. It is defined as: $\text{ACC} = \left(\frac{TP+TN}{(TP+TN+FP+FN)}\right)$. Sensitivity (SEN), also known as true positive rate, indicates the model's ability to locate all positive samples. It is computed as $\text{SEN} = \left(\frac{TP}{TP+FN}\right)$. Specificity (SPE), also known as false positive rate, is the complement of sensitivity and represents the model's ability to correctly identify all negative samples. It is given by $\text{SPE} = \left(\frac{TN}{TN+FP}\right)$.

To conduct a comparative analysis of the three CNN models, we performed multiple experiments by varying the number of epochs (100, 250, 500, 1000). The training performance across the 3 models and 4 epoch levels is shown in Figure

TABLE II
PERFORMANCE OF THE CNN MODELS ON THE AD SEVERITY PREDICTION OF THE TEST IMAGE SUBSET.

		AD		MCI		CN	
Epoch	ACC	SEN	SPE	SEN	SPE	SEN	SPE
			DenseNet121				
100	54.15	15.61	99.76	78.54	49.02	82.44	68.29
250	67.48	30.24	100	92.68	62.2	89.02	79.51
500	77.56	54.63	99.76	89.27	77.32	89.27	88.78
1000	76.91	49.27	100	95.61	72.44	92.93	85.85
			DenseNet169				
100	52.36	19.02	99.76	98.05	30.49	98.29	40
250	67.15	35.61	99.51	95.61	56.59	94.63	70.24
500	75.93	58.54	99.51	95.12	68.29	96.1	74.15
1000	75.12	55.12	98.54	97.56	64.88	99.27	72.68
			Inception-ResNet-v2				
100	84.07	78.05	96.83	95.12	80.98	98.29	79.02
250	91.87	89.27	98.05	91.22	93.17	96.59	95.12
500	93.66	92.2	98.78	96.1	92.93	98.78	92.68
1000	92.03	89.76	98.05	96.1	91.22	98.78	90.24

ACC: **Testing** Accuracy; SEN: Sensitivity SPE: Specificity

7. Table II illustrates the performance of the best performing model per epoch on the test set not seen during training. From the comparison of the overall prediction accuracies of the validation set (training) to that of the test set in Figure 8, we can observe that the DenseNet121 model (blue) appear to overfit, as the accuracies diverge from each other while the converge for both the DenseNet169 and the Inception-ResNet-v2 models. From Table II, we observe that the Inception-ResNet-v2 is the best performing model. In terms of the AUC values, Inception-ResNet-v2 has the highest AUC scores across all three classes (AD: 0.95, MCI: 0.95, and AD 0.96). The AUC scores for DenseNet121 (AD: 0.75, MCI: 0.84, CN: 0.89) and DenseNet169 (AD: 0.77, MCI: 0.81, CN: 0.86) are relatively close to each other, with DenseNe121 having slightly better performance. Overall, Inception-ResNet-v2 is most effective model for classifying new data accurately, as it has the highest test prediction accuracy, AUC, and is least likely to overfit during training.

V. Model Explanation and Clinical Relevance

Since the sensitivity and specificity analysis of the 3 models suggests that the Inception-Resnet-v2-based model outperforms the other two models, we focus on explaining only the Inception-ResNet-v2 model using the SHAP visualization plot.

As Alzheimer's disease progresses, the interhemispheric fissure of the brain widens, the cortical sulci of the brain enlarge, the ventricles enlarge, and the corpus callosum thins [34]. There may also be subtle changes in the white matter and grey matter of the parenchyma. Patients with more severe AD are expected to have larger ventricles and larger cortical sulci. The distribution of SHAP values (red pixels) around the ventricles and near the cortical sulci (see Figures 9 and 10), suggests that the neural networks may have relied on ventricular enlargement and cortical sulci enlargement to identify subject images with likely Alzheimer's disease.

TABLE III
CLINICAL RELEVANCE OF INCEPTION-RESNET-V2 MODEL OUTCOMES FOR CORRECTLY PREDICTED AD VS. AD PREDICTED AS MCI OR CN

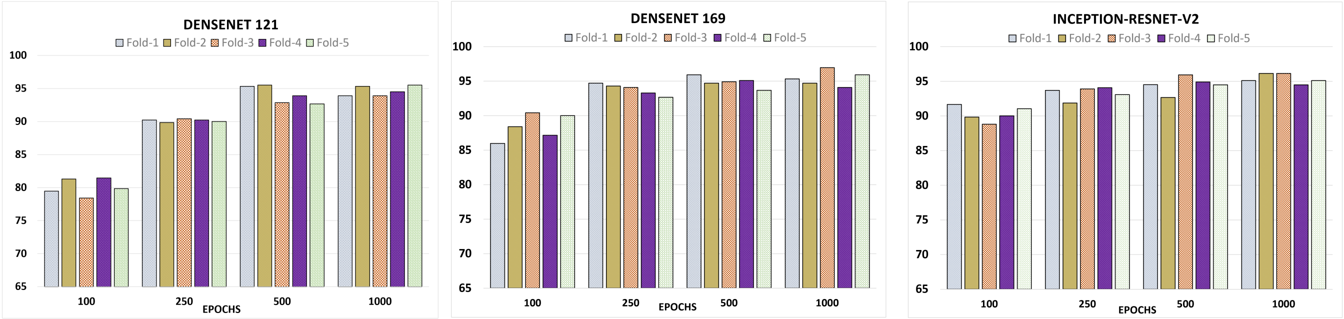
	Correct AD	Incorrect MCI	Incorrect CN
N	193	11	1
Gender (Male/Female)	86/107	4/7	1
Age	72.2 ± 5.75	72.1 ± 4.5	79.1
FAQ total (↑)	17.1 ± 7.6	19.2 ± 11.61	0
MMSE (↓)	21.00 ± 5.5	22.9 ± 3.1	26
Digit span total (↓)	22.5 ± 14.72	21.0 ± 14.5	0
CDR Global (↑)	0.97 ± 0.49	1.0 ± 0.81	0.5
ADAS-Cog (↑)	24.3 ± 11.7	21.4 ± 15.8	0
APOE high severity*	79.6% of 49	50% of 6	0% of 1

*APOE high severity: percentage of patients with genotype of (3/4,4/3, or 4/4).
MMSE: Mini Mental State Examination; CDR: Clinical Dementia Rating
FAQ: Functional Assessment Questionnaire

A neurology domain expert (D. H.) inspected the SHAP plots (Figures 9 and 10). As expected, SHAP intensity in the plots was highest for the class that was selected by the model. When the model correctly classified an AD as AD class, SHAP values were highest for the AD class SHAP plots. The same held true for correctly classified MCI images. Spatial distribution of SHAP values in the images gives insight into those areas of the brain that the NN model utilized most intensively in making the classification. Inspection of Figures 9 and 10 suggests that the NN model depended on examining the size of the ventricles within the brain and the enlargement of the cortical sulci over the surface of the brain. Fewer SHAP intensities were noted over either the white matter or the grey matter within the brain parenchyma. Tables III and IV provide some insight into the images misclassified by the model. In general, cases of AD that were misclassified as either MCI or controls had milder disease as evidenced by higher MMSE scores and lower ADAS-cog scores (Table IV). Even the one case of AD incorrectly classified as control had a relatively high MMSE of 26 and a low CDR of 0.5 (Table IV). Majority of accurately classified AD images also had high severity APOE genotype alleles. Interestingly, about half of the incorrectly classified AD images also demonstrated less severity, as quantified by the lower APOE alleles. This suggests that while the model may have misclassified a few correct samples, some of the misclassified samples might have been labeled incorrectly, as indicated by the lower severity of APOE. The same trend applies to both correctly and incorrectly classified MCI group samples.

VI. Conclusion

This paper presents an explainable framework for prediction of AD severity from MR brain images. We present a comparative analysis of 3 high performing deep CNN models of which Inception-ResNet-v2 outperformed DenseNet121 and DenseNet169. Integrating SHAP into the learning framework provided a better understanding of how the best performing model derived the results. Examination of the correctly classified AD severity groups to the incorrectly classified groups based on neurocognitive outcome measures and APOE



(a)

(b)

(c)

Fig. 7. **Accuracy** of CNN model on the validation set during training for varying epoch levels across the 5-fold cross validation (a) DenseNet121 (b) DenseNet169 (c) Inception-ResNet-v2.

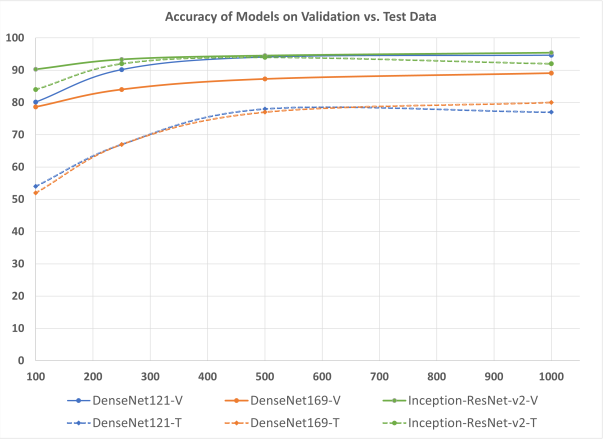


Fig. 8. Accuracy of models on validation vs. test data.

TABLE IV
Clinical Relevance of Inception-Resnet-V2 Model Outcomes for correctly predicted MCI vs. MCI predicted as AD or CN

	Correct MCI	Incorrect AD	Incorrect CN
N	196	5	4
Gender (Male/Female)	105/91	3/2	1/3
Age	72.0 ± 5.4	76.3 ± 2.5	67.2 ± 4.7
FAQ total (↑)	7.2 ± 6.8	2	10
MMSE V	25.4 ± 4.1	25.7 ± 1.5	28.3 ± 0.6
Digit span total	36.5 ± 13.9	56.0	33.0
CDR Global (↑)	0.58 ± 0.26	0.5 ± 0.0	0.5 ± 0.0
ADAS-Cog (↑)	14.6 ± 8.2	15.0	12.7
APOE high severity*	52.63% of 38	50% of 2	50% of 2

*APOE high severity: percentage of patients with genotype of (3/4,4/3, or 4/4).
MMSE: Mini Mental State Examination; CDR: Clinical Dementia Rating
FAQ: Functional Assessment Questionnaire

genotype further increased confidence in the interpretability of the model. Limitations of this study includes using only the center axial slice, which may not have the complete discriminant information for AD severity as well as a relatively small and unbalanced dataset which necessitated the use of SMOTE in estimations of images.

References

[1] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, and B. Kim, "Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments," *Pattern Recognition*, vol. 120, p. 108102, 2021.

[2] Z. Sadeghi, R. Alizadehsani, M. A. Cifci, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhawaldeh, S. Hussain *et al.*, "A brief review of explainable artificial intelligence in healthcare," *arXiv preprint arXiv:2304.01543*, 2023.

[3] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, and J. H. Holmes, "A manifesto on explainability for artificial intelligence in medicine," *Artificial Intelligence in Medicine*, vol. 133, p. 102423, 2022.

[4] Y. Liang, S. Li, C. Yan, M. Li, and C. Jiang, "Explaining the black-box model: A survey of local interpretation methods for deep neural networks," *Neurocomputing*, vol. 419, pp. 168–182, 2021.

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

[6] Y. AbdulAzeem, W. M. Bahgat, and M. Badawy, "A CNN based framework for classification of Alzheimer's disease," *Neural Comput. Appl.*, vol. 33, no. 16, pp. 10 415–10 428, 2021.

[7] P. Kalavathi and V. B. S. Prasath, "Methods on skull stripping of MRI head scan images—a review," *J. Digit. Imaging*, vol. 29, no. 3, pp. 365–379, 2016.

[8] M. W. Weiner, P. S. Aisen, C. R. Jack Jr, W. J. Jagust, J. Q. Trojanowski, L. Shaw, A. J. Saykin, J. C. Morris, N. Cairns, L. A. Beckett *et al.*, "The Alzheimer's disease neuroimaging initiative: progress report and future plans," *Alzheimer's & Dementia*, vol. 6, no. 3, pp. 202–211, 2010.

[9] J. Zhou, L. Hu, Y. Jiang, and L. Liu, "A correlation analysis between snps and rois of alzheimer's disease based on deep learning," *BioMed Research International*, vol. 2021, pp. 1–13, 2021.

[10] S. Murugan, "DEMENT: A deep learning model for early diagnosis of Alzheimer's diseases and dementia from MR images," *IEEE Access*, vol. 9, pp. 90 319–90 329, 2021.

[11] R. Jain, N. Jain, A. Aggarwal, and D. J. Hemanth, "Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images," *Cogn. Syst. Res.*, vol. 57, pp. 147–159, 2019.

[12] H. A. Helaly, M. Badawy, and A. Y. Haikal, "Deep learning approach for early detection of Alzheimer's disease," *Cognit. Comput.*, vol. 14, no. 5, pp. 1711–1727, 2022.

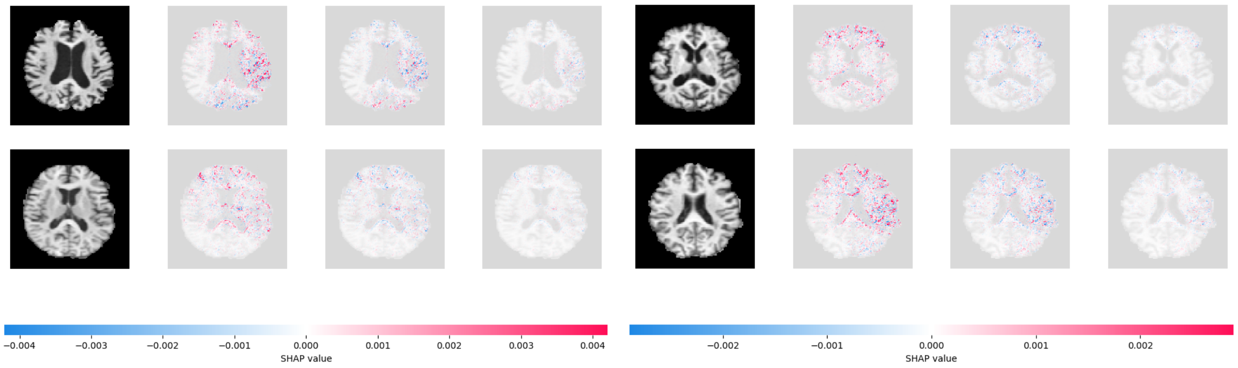
[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, pp. 4765–4774, 2017.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[15] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," *Proc. Conf. AAAI Artif. Intell.*, vol. 31, no. 1, 2017.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

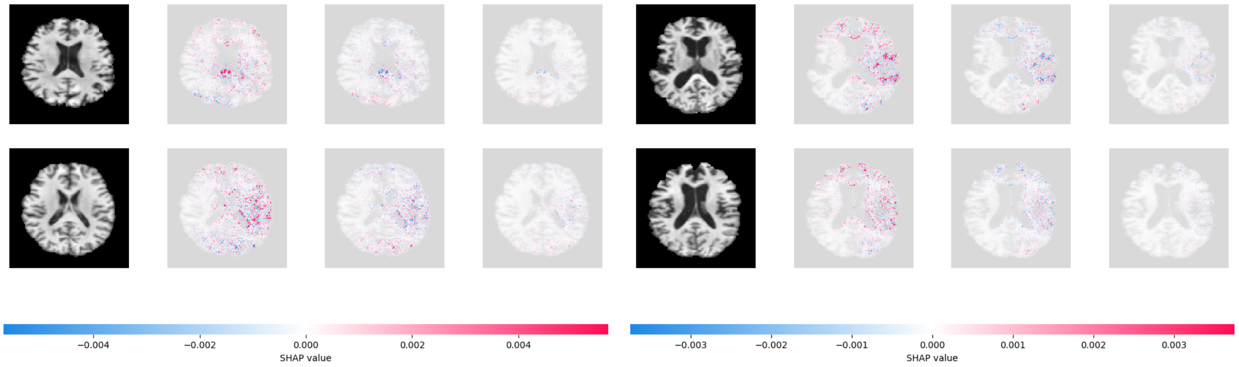
[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions,"



(a) Correctly predicted AD

(b) AD misclassified as MCI

Fig. 9. Predictions for 4 AD images are shown, 3 outputs per image. Column (a) are those correctly predicted as AD while (b) were misclassified as MCI. Red pixels represent positive SHAP values that increase the probability of the class, while blue pixels represent negative SHAP values that reduce the probability of the class. For the correctly classified images, note that SHAP values were highest for the correct class in contrast to the misclassified images.



(a) Correctly predicted MCI

(b) MCI misclassified as AD

Fig. 10. Predictions for 4 MCI images are shown, 3 outputs per image. Column (a) are those correctly predicted as MCI while (b) were misclassified as AD. Red pixels represent positive SHAP values that increase the probability of the class, while blue pixels represent negative SHAP values that reduce the probability of the class. Note that SHAP values are higher for the MCI class for the correctly classified images in contrast to the misclassified images.

in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[18] S. Kokkalla, J. Kakarla, I. B. Venkateswarlu, and M. Singh, “Three-class brain tumor classification using deep dense inception residual network,” *Soft Comput.*, vol. 25, no. 13, pp. 8721–8729, 2021.

[19] M. F. Mendez, *The Mental Status Examination Handbook E-Book*. Elsevier Health Sciences, 2021.

[20] G. A Marshall, A. S Zoller, N. Lorius, R. E Amariglio, J. J Locascio, K. A Johnson, R. A Sperling, D. M Rentz, Alzheimer’s Disease Neuroimaging Initiative *et al.*, “Functional activities questionnaire items that best discriminate and predict progression from clinically normal to mild cognitive impairment,” *Current Alzheimer Research*, vol. 12, no. 5, pp. 493–502, 2015.

[21] L. G. Weiss, D. H. Saklofske, J. A. Holdnack, and A. Prifitera, *WISC-V assessment and interpretation: Scientist-practitioner perspectives*. Academic Press, 2015.

[22] D. M. Kaufman, H. Geyer, and M. J. Milstein, *Kaufman’s Clinical Neurology for Psychiatrists*. Elsevier Inc., 2016.

[23] A. L. Lumsden, A. Mulugeta, A. Zhou, and E. Hyppönen, “Apolipoprotein E (APOE) genotype-associated disease risks: a phenome-wide, registry-based, case-control study utilising the uk biobank,” *EBioMedicine*, vol. 59, p. 102954, 2020.

[24] “Alzheimer’s disease genetics fact sheet,” https://www.nia.nih.gov/health/alzheimers-disease-genetics-fact-sheet, accessed: 2023-6-1.

[25] S. M. Smith, “Fast robust automated brain extraction,” *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[26] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, “Improved optimization for the robust and accurate linear registration and motion correction of brain images,” *Neuroimage*, vol. 17, no. 2, pp. 825–841, 2002.

[27] M. Brett, C. J. Markiewicz, M. Hanke, M.-A. Côté, B. Cipollini, P. McCarthy, D. Jarecka, C. Cheng, Y. Halchenko, M. Cottaar *et al.*, “nipy/nibabel: 3.2. 1,” *Zenodo*, 2020.

[28] N. Pawlowski, S. I. Ktena, M. C. Lee, B. Kainz, D. Rueckert, B. Glocker, and M. Rajchl, “DLTK: State of the art reference implementations for deep learning on medical images,” *arXiv preprint arXiv:1711.06853*, 2017.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[30] D. Elreedy and A. F. Atiya, “A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance,” *Inf. Sci. (Ny)*, vol. 505, pp. 32–64, 2019.

[31] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin *et al.*, “Tensorflow: Large-scale machine learning on heterogeneous systems, version 2,” 2020.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[33] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, “Deep learning for brain MRI segmentation: state of the art and future directions,” *Journal of Digital Imaging*, vol. 30, no. 4, pp. 449–459, 2017.

[34] K. A. Johnson, N. C. Fox, R. A. Sperling, and W. E. Klunk, “Brain imaging in alzheimer disease,” *Cold Spring Harbor Perspectives in Medicine*, vol. 2, no. 4, p. a006213, 2012.