Development of an Interpretable Deep Learning System for the Identification of Patients with Alzheimer's Disease

Development of an interpretable deep learning system for the identification of patients with Alzheimer's disease

Selamawet Workalemahu Atnafu [1,2] [0000-0001-7064-8601] and Stefano Diciotti [1][0000-0001-8702-5654]

[1] Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi", University of Bologna, Bologna, Italy
[1]wselame7@gmail.com
[2] Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar

Abstract. Deep learning using convolutional neural networks has shown great promise in analyzing neuroimaging data. Identification of Alzheimer's Disease patients from healthy individuals using structural magnetic resonance data is one of the clinical problems that has been widely explored by employing convolutional neural networks and producing very high classification accuracies. However, in most studies, the results were not supported by explainability tools.

In this study, an interpretable convolutional neural network model derived from pre-trained VGG-16 is proposed to classify Alzheimer's Disease patients vs. healthy subjects using MRI data obtained from an Open Access Series of Imaging studies (OASIS) dataset. The model was trained and validated based on a 5-fold cross-validation loop and produced a classification accuracy of 71.62% on the validation set. Moreover, we incorporated four CNN visualization techniques that highlight important brain regions used by the model to identify AD patients: saliency map, gradient class activation mapping, occlusion mapping, and heatmap generated by Shapley additive explanation (SHAP) method. The potential of these explainability tools in identifying biased models that produce inflated erroneous results is also investigated, and the resulting heatmaps were able to indicate a bias in the model's training procedure.

Keywords: Alzheimer 's disease, deep learning, interpretability, OASIS, MRI

1 Introduction

Alzheimer's disease (AD) is the most commonly occurring neurodegenerative disorder (Selkoe & Lansbury, 1999) that causes memory impairment at its initial stage and advances to a cognitive decline that can affect behavior, speech, visuo-spatial orientation, and motor system (Kelley & Petersen, 2007). Early diagnosis is essential to plan treatment strategies that could slow down the disease progression and enhance

the quality of life (Small, et al., 1997). Diagnosis of AD needs a physician's follow-up of the patient's medical history by performing clinical assessments and neuropsychological tests scores (Small, et al., 1997). Neuroimaging tools, such as structural MRI (sMRI), functional MRI, and positron emission tomography (PET) are also used to confirm that the cognitive decline caused by AD is altering the brain structure.

In the past, traditional machine learning methods have been frequently used to analyze neuroimaging data. Designing a machine learning system becomes a very long process due to the need to extract hand-crafted features. On the contrary, deep learning – a family of machine learning methods that has the ability to automatically extract features from complex data (LeCun, et al., 2015) – overcomes the main limitations of traditional machine learning approaches and hence became the current state of the art technology in medical imaging, including neuroimaging.

Convolutional neural networks (CNNs), are a type of deep learning models tailored for image processing applications (LeCun, et al., 2015). A basic CNN model consists of convolutional layers, pooling layers, and fully connected layers. Numerous studies have employed CNNs for classifying sMRI of AD patients vs. healthy subjects (Liu, et al., 2020; Oh, et al., 2019; Qiu, et al., 2020; Wen, et al., 2020; Feng, et al., 2020; Yagis, et al., 2019). Most of these studies used MRI data obtained from the Alzheimer 's disease Neuroimaging Initiative (ADNI) dataset, and other few studies (Yagis, et al., 2019; Yagis, et al., 2021; Tufail, et al., 2020; Puente-Castro, et al., 2020; Saratxaga, et al., 2021; Mehmood, et al., 2020; Massalimova & Varol, 2021) applied deep learning techniques on the Open Access Series of Imaging studies (OASIS) collection of brain images.

Apart from their success in many applications, deep learning approaches have been criticized for producing highly non-interpretable models (Linardatos, et al., 2020). Interpretability is a requirement in many applications in which crucial decisions are made by users relying on a model's outputs, such as in medical applications (Lipton, 2018).

CNN visualization methods help in understanding the reasoning behind the model's decisions. Several recent neuroimaging studies have integrated explainability tools in their CNN models to classify different neurological disorders (Gao, et al.,

2021; Jimeno, et al., 2022; Zhang, et al., 2021; Qiu, et al., 2020; Tang, et al., 2019; Lu, et al., 2022; Oh, et al., 2019; Iizuka, et al., 2019; Sánchez Fernández, et al., 2020). Regarding AD classification, a few studies (Lu, et al., 2022; Oh, et al., 2019; Qiu, et al., 2020) employed CNN visualization techniques to highlight the features used by the model to make decisions. These studies used a public brain dataset of AD and healthy individuals, namely the ADNI Initiative.

In this study, we propose an interpretable CNN for classifying sMRI scans obtained from the public OASIS dataset. The CNN model is trained based on a transfer learning technique by utilizing the weights of a pre-trained VGG16 network. Unlike the previous deep learning studies classifying the OASIS collection of brain images, our proposed model includes a wide range of visualization methods to confirm that the models focus on clinically defined AD regions.

2 Materials

In this section, the datasets used, the model architecture, training, and validation schemes, and finally, the CNN visualization methods applied to the trained model and their interpretation are discussed.

2.1 Subjects

In this study, we used the OASIS publicly available dataset of AD patients and healthy control (HC) subjects (Marcus, et al., 2007) which is freely available at https://www.oasis-brains.org/. The dataset consists of a cross-sectional collection of MRI scans of 416 right-handed subjects aged between 18 and 96 years. The scans were acquired using a 1.5 T scanner. A hundred AD patients [(59 women and 41 men, age 76.70 ± 7.10 years, mean ± standard deviation (SD)] and 100 HC subjects (73 women and 27 men, age 75.50 ± 9.10 years, mean ± SD) who have been previously selected by other authors (Hon & Khan, 2017) are included in our experiment. The difference in age was not significant ($p = 0.15$ at t-test) across the two groups while, considering gender, a significant difference ($p = 0.04$ at χ^2-test) was obtained. Table 1 lists the demographic information of the dataset used in this study.

In the OASIS dataset, the global Clinical Dementia Rating (CDR) score was used for AD diagnosis, as well as to determine the severity of the disease. CDR score was derived from individual CDR scores for the domains memory, orientation, judgment and problem solving, function in community affairs, home and hobbies, and personal case (Morris, 1993; Morris, et al., 2001). A global CDR score of 0 represents subjects with normal cognition, while scores of 0.5 (very mild), 1 (mild), 2 (moderate), and 3 (severe) have all been labeled as AD.

Table 1: Demographic features of the dataset used in this study.

	Patients	**Healthy controls**
Number of subjects	100	100
Age (range, years)	62 – 96	59 – 94
Age (mean ± SD, years)	76.70 ± 7.10	75.50 ± 9.10
Gender (women/men)	59/41	73/27

2.2 Data pre-processing

The OASIS dataset publicly provides pre-processed data, where gain-field correction, brain masking, and atlas-based co-registration (Han, et al., 2018) were applied to the raw MRI images resulting in a data matrix size of 176 × 208 × 176 and a voxel size of 1 mm × 1 mm × 1 mm (Han, et al., 2018). We performed 2D image processing on such partially pre-processed 3D MRI volumes. The 2D image processing involves slicing the MRI volumes into 2D gray scale images using an axial anatomical plane, and performing slice selection based on entropy values. We followed an entropy-based slice selection similar to our previous paper (Yagis, et al., 2021). Second, data was split into training and validation sets based on a 5-fold CV scheme. In this step, all the slices of a single MRI volume were included in either the training or validation sets (subject-level split) to prevent the problem of data leakage (Yagis, et al., 2021). Lastly, feature scaling was applied to normalize the images in both the training and validation sets based on the training set feature statistics (using mean and SD values). From each MRI scan, 10 slices are selected based on their entropy values (Hon & Khan, 2017), producing a total of 2000 (1000 AD and 1000 HC) gray scale images. As compared to the number of parameters for building a CNN model, the size of the image dataset is insufficient to effectively train a CNN from scratch. To prevent overfitting of a CNN model due to limited training samples, we employed a transfer learning technique by starting from a pre-trained VGG16 model and fine-tuning the model parameters on the MRI dataset. Since VGG16 is trained on colored RGB images, the gray scale MRI slices were converted to three-channel images by repeating the 2D image onto the three channels. By applying these pre-processing operations, we obtain an array of 2000 x 176 x 208 x 3 size.

2.3 CNN model

The CNN model architecture is customized from the pre-trained VGG16 model. The fully connected (FC) layers of VGG16 are removed and replaced by a global average pooling (GAP) layer, and a last FC classification layer with a ‘softmax’ activation is added (Figure 1). During model training, three convolutional blocks were frozen to reduce the number of trainable parameters and to avoid overfitting. The rest two blocks of convolutional layers were fine-tuned along with the newly added FC layer.

Model training was performed based on a 5-fold CV scheme using an Adam optimizer with a learning rate of 1×10^{-4} and a learning rate decay of 0.5. Although Adam optimizer assigns a specific learning rate for each parameter, which is computed using the initial learning rate as an upper limit, to be sure that every learning rate update step does not exceed the initial learning rate, an exponential learning rate decay is assigned that lowers the upper limit. By doing so, the learning rate decay helps to prevent the loss from diverging after it decreases to a point. In our case, different values of decay were tried out, and the value 0.5produced the best performance of the model over the validation set. The ‘categorical_crossentropy’ was used as a loss function. By training the model for a different number of epochs with a batch size of 128 images, 90 epochs have produced the best performance of the model on the validation set. Four classification metrics such as balanced accuracy, sensitivity, and specificity, were used to measure the performance of the model. The final results of the trained model are reported based on the average accuracy computed over the 5 folds on the validation set.

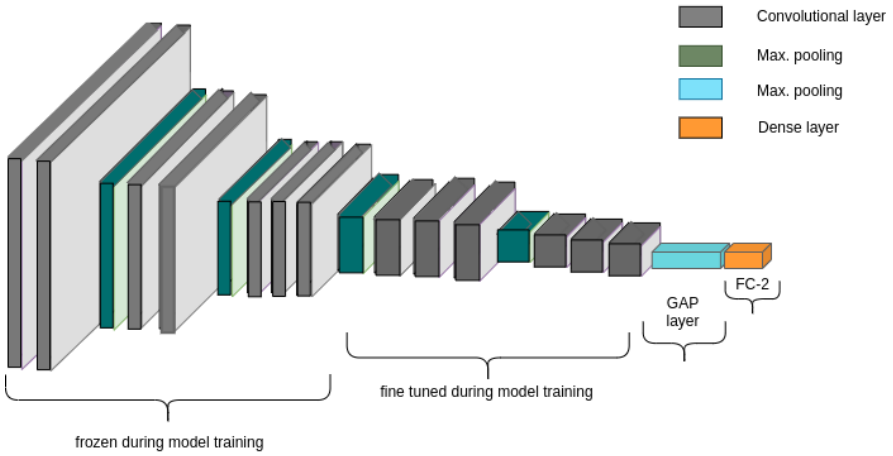

Figure 1: A customized VGG16 model consists of a convolutional that is transferred from the pre-trained VGG16 model, a GAP (global average pooling layer), and two FC layers (FC-256 and FC-2).

2.4 CNN visualization

Model visualization methods enable understanding of the rationale behind a deep learning model's decisions. For a CNN model, these interpretability approaches are applied to a trained model to inspect which image regions or features are given high importance for the prediction analysis. In this study, we employed four attribute-based interpretability techniques (two gradient-based approaches, saliency maps and GradCAM, and two perturbation-based methods, SHAP and occlusion maps) for a classification problem of AD vs. HC subjects. To emphasize the importance of these visualization tools, we performed two experiments. In the first experiment, a model is trained to classify subjects as AD vs. HC, and visualization heatmaps highlight the brain regions that are used by the model to identify AD brain scans from healthy MRI images. This helps to check if our results are in line with the neural correlates of AD, which are defined in the previous AD studies. The aim of the second experiment is to highlight the potential of these visualization tools for identifying biased models producing highly inflated performances. Data leakage caused by slice-level split is

one of the methodological pitfalls of applying 2D CNNs for the classification of 3D MRI data that result in a biased model outputting overestimated performance on the test set (Yagis, et al., 2021; Yagis, et al., 2019; Wen, et al., 2020). Following a similar procedure to the experiment explained in Section 3.2.1, in this study, we also trained two architecturally similar models on the same dataset using two data split methods. While the first model is trained by applying subject-level split, hence without data leakage, the second model is trained on data that is divided based on MRI slices introducing data leakage. Correctly classified AD test samples are then passed through the trained models, and visualization heatmaps generated from the two models are compared to check if reliable features are used by the two CNN models.

3. Results

The results of Experiment 1 and Experiment 2 are presented in this section. In Experiment 1, the performance of our interpretable CNN model as measured by the average accuracy, sensitivity, and specificity values computed over the five folds on the validation set are reported in Table 2. Our model identifies AD MRI slices with an accuracy of 71.62%, sensitivity 71.85%, and specificity score 72.73%.

The learning curve is also shown in Figure 2. An example of the visualization heat map images generated by passing MRI images of AD patients, which are predicted by the model taken from the validation set, can be seen in Figure 3.

In the second experiment, the trained model introducing data leakage achieved a test set accuracy of 95.12% (Table 5.4). Figures 5.8 and 5.10 illustrate the learning curve and the visualization heatmaps of the trained biased model, respectively.

Table 2: Average model's performance computed over the five folds on the validation set.

	sensitivity	specificity	accuracy
training set	0.8146	0.7686	0.7916
validation set	0.7185	0.7273	0.7162

Table 3: Average accuracy of the biased model computed over the five folds on the validation set.

	sensitivity	specificity	accuracy
training set	0.9996	0.9810	0.9923
validation set	0.9592	0.9450	0.9512

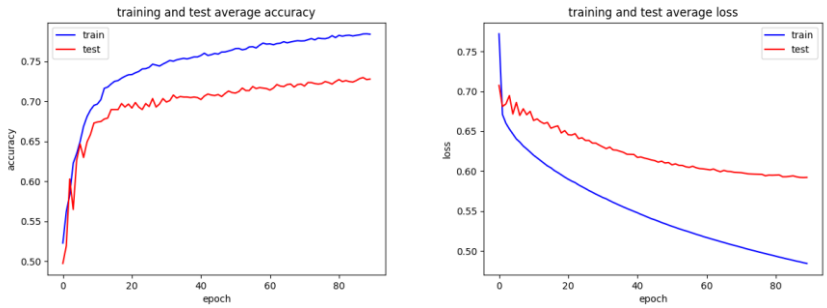

Figure 2: Learning curves of the model on both the training and validation sets.

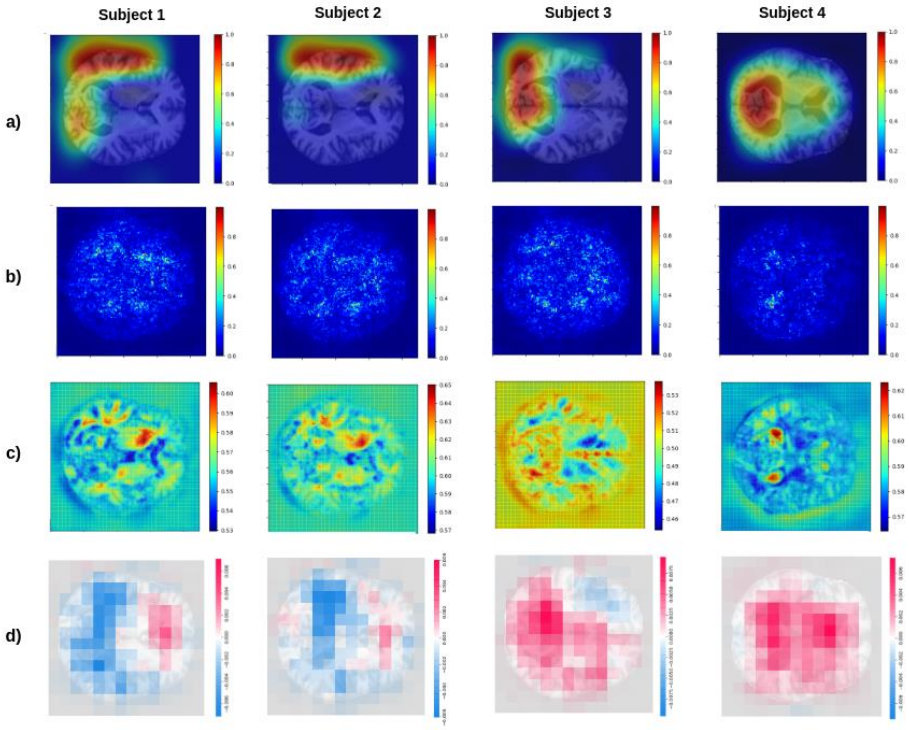

Figure 3: CNN visualization heat maps of MRI slices taken from four representative AD patients, who are correctly classified by the CNN model - row a) represents Grad-CAM images, b) saliency maps, c) occlusion maps, and d) SHAP heat maps.

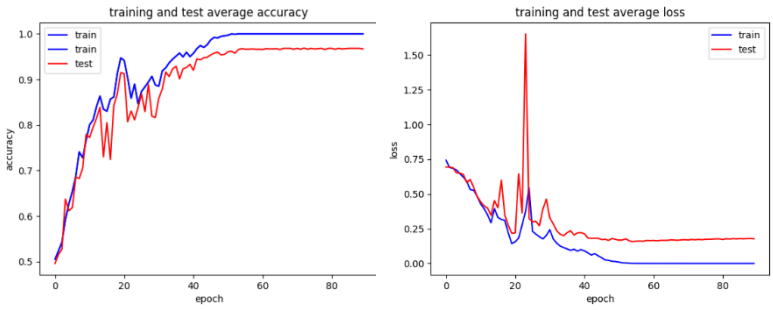

Figure 4: Learning curves of the biased model trained with data leakage.

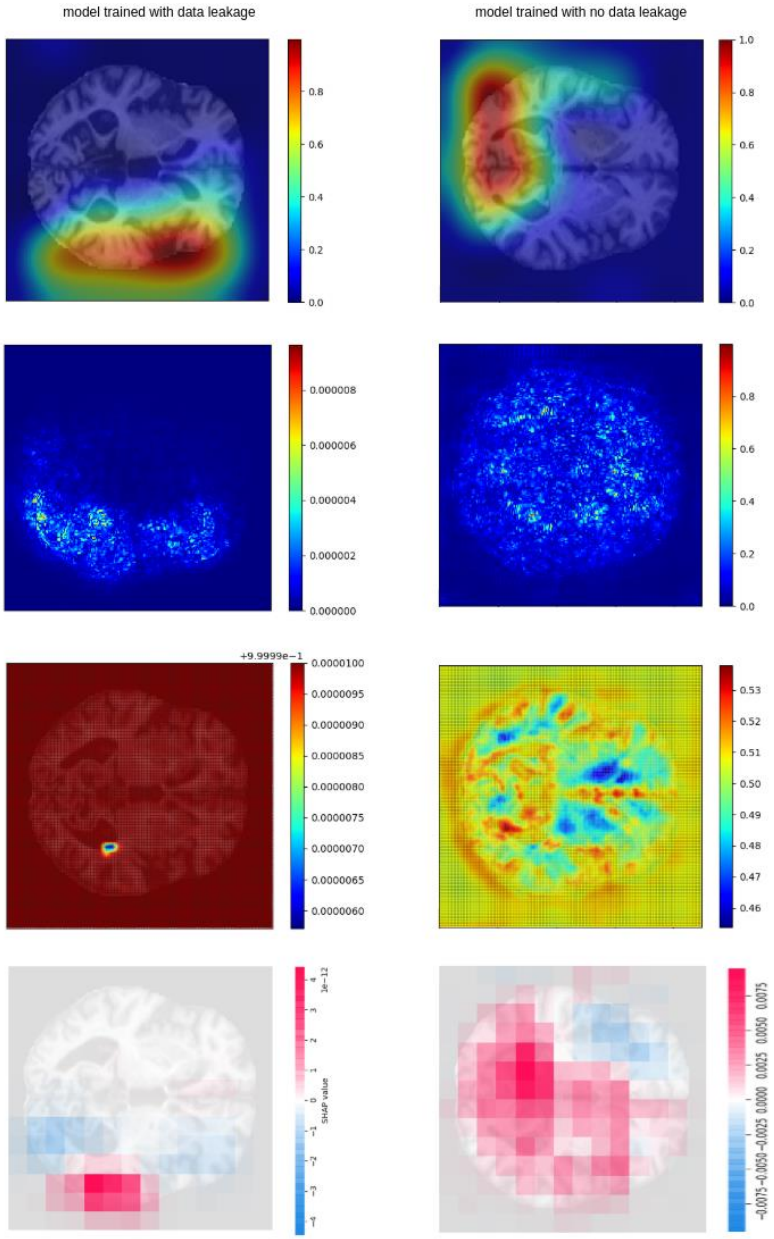

Figure 5: CNN visualization heat maps indicate a model producing a biased performance due to the presence of data leakage. Images on the left side are generated by the model trained on data split based on slices (with data leakage). For CAM, occlusion map, and SHAP, the heatmap represents a very low number (score close to 0), capturing the biased model. Rather, Grad-CAM is seen as less capable of identifying the biased model.

4. Discussion

In this study, a deep learning model customized from VGG16 is proposed for binary classification of AD vs. HC subjects. The proposed CNN model was trained by employing a transfer learning technique to prevent model overfitting caused by the small size of the training data. Instead of training a shallow CNN from scratch, fine-tuning a deeper pre-trained model allows the trained models to achieve excellent results and to have better generalization ability since depth is a crucial parameter in the design of good performing CNN model (Alzubaidi, et al., 2021). Hence a pre-trained VGG16 model is fine-tuned on our MRI dataset. The model was trained on a brain image collection of the OASIS dataset achieving an average accuracy of 71.62% on the test set. Compared to previous studies employing exactly the same T_1-weighted MRI sub-samples taken from the OASIS dataset (Yagis, et al., 2019; Yagis, et al., 2021), our model classifies AD and HC subjects with better accuracy. Although in other few studies (Tufail, et al., 2020; Saratxaga, et al., 2021; Massalimova & Varol, 2021), the authors reported higher accuracies, these results are due to the use of a larger number of subjects, multimodality, and the application of data augmentation to improve the performance of the model. Apart from reporting the model's performance, neither of these studies included model visualization tools to ensure that the models focus on meaningful brain regions to perform the predictive analysis. On the contrary, our proposed model incorporates four different visualization methods, strengthening our system's reliability.

The results also showed that the interpretation techniques highlight features located around the frontal, temporal, and parietal lobe of the cerebral cortex and areas around the thalamus. The SHAP method outperformed the other methods in localizing the frontal lobe, while the cortical atrophy and alterations around the thalamus were better captured by the Grad-CAM method. The visualization outcomes by the CAM-based technique are very much distributed. Instead, the GradCAM method produced a more localized heatmap.

Regarding the role of explainability tools in identifying a biased model, such as a model trained by introducing data leakage, all visualization techniques indicate the bias incurred during the model training procedure. Figure 5 illustrates representative visualization heatmaps produced by the two models by passing correctly classified AD slices. The pixels on the occlusion heatmap represent the classification probability of the input image with respect to the target class (in our case, to be classified as a slice of an AD patient). While the model trained without data leakage produces a heatmap with a probability range of [0.46, 0.53], the biased model generates pixels in a probability range of [$6x10^{-6,}$ $1x10^{-5}$], which is a very small value representing the pixel's insignificance for the model's prediction. Similarly, the heatmap generated by the SHAP method represents the SHAP values that explain the importance of each pixel on the input image. While for the non-biased model, the SHAP values lie between [-0.0075, 0.0075], the values for the biased model range are within [$-4x10^{-12,}$ $4x10^{-12}$], illustrating that all parts of the input image have no importance for the prediction output by the model. Saliency mapping also produces gradient values (gradient of the model's prediction probability with respect to the input images) in the range [0, 0.8] and [0, 0.000008] for non-biased and biased models, respectively, highlighting the reduced importance of input features to determine the model's output in the case of a biased model. GradCAM, in contrast, showed less capability of capturing the biased model as it assigns gradient values between 0 and 0.8 for a biased model and 0 and 1 for the non-biased model.

Since each interpretability approach has its own limitation, incorporating multiple visualization methods helps better understand deep learning-based predictive systems.

The main limitation of this study is that due to the small size of the dataset, the CNN model was fine-tuned insufficiently to achieve higher classification accuracy.

5. Conclusion

In this study, we presented an interpretable 2D CNN model that performs a diagnosis of AD from sMRI data. The model was trained on the OASIS dataset of AD and HC subjects based on a 5-fold CV scheme and achieved a classification accuracy

of 71.62% on the test set. Beyond that, the model is characterized as an interpretable system since it allows for visualizing features or important brain regions that are given the highest importance by the model for the prediction task.

References

[1] Feng, W. et al., 2020. Automated MRI-based deep learning model for detection of Alzheimer's disease process. *International Journal of Neural Systems,* Volume 30(06), p. 2050032.

[2] Gao, J. et al., 2021. Multisite autism spectrum disorder classification using convolutional neural network classifier and individual morphological brain networks. *Frontiers in Neuroscience,* Volume 14, p. 1473.

[3] Han, X. et al., 2018. Brain extraction from normal and pathological images: a joint PCA/image-reconstruction approach. *NeuroImage,* Volume 176, pp. 431-445.

[4] Hon, M. & Khan, N. M., 2017. Towards Alzheimer's disease classification through transfer learning. *In 2017 IEEE International conference on bioinformatics and biomedicine (BIBM),* pp. 1166-1169.

[5] Iizuka, T., Fukasawa, M. & Kameyama, M., 2019. Deep-learning-based imaging-classification identified cingulate island sign in dementia with Lewy bodies. *Scientific reports,* Volume 9(1), pp. 1-9.

[6] Jimeno, M. M. et al., 2022. ArtifactID: Identifying artifacts in low-field MRI of the brain using deep learning. *Magnetic resonance imaging,* 89(., Ogbole, G., & Geethanath, S.), pp. 42-48.

[7] Kelley, B. J. & Petersen, R. C., 2007. Alzheimer's disease and mild cognitive impairment. *Neurologic clinics,* Volume 25(3), pp. 577-609.

[8] LeCun, Y., Bengio, y. & Hinton, g., 2015. Deep learning. *Nature,* Volume 521(7553), pp. 436-444.

[9] Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S., 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy,* Volume 23(1), p. 18.

[10] Lipton, Z., 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue,* Volume 16(3), pp. 31-57.

[11] Liu, M. et al., 2020. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *Neuroimage,* Volume 208, p. 116459.

[12] Lu, P. et al., 2022. A Two-Stage Model for Predicting Mild Cognitive Impairment to Alzheimer's Disease Conversion. *Frontiers in Aging Neuroscience,* Volume 14.

[13] Marcus, D. et al., 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J Cogn Neurosci*, Volume 19, p. 1498–1507.

[14] Massalimova, A. & Varol, H. A., 2021. Input Agnostic Deep Learning for Alzheimer's Disease Classification Using Multimodal MRI Images. *In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2875-2878.

[15] Mehmood, A., Maqsood, M., Bashir, M. & Shuyuan, Y., 2020. A deep Siamese convolution neural network for multi-class classification of Alzheimer disease. *Brain sciences*, Volume 10(2), p. 84.

[16] Oh, K. et al., 2019. Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning. *Scientific Reports*, Volume 9(1), pp. 1-16.

[17] Puente-Castro, A., Fernandez-Blanco, E., Pazos, A. & Munteanu, C. R., 2020. Automatic assessment of Alzheimer's disease diagnosis based on deep learning techniques. *Computers in Biology and Medicine*, Volume 120, p. 103764.

[18] Qiu, S. et al., 2020. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain*, Volume 143 (6), pp. 1920-1933.

[19] Sánchez Fernández, I. et al., 2020. Deep learning in rare disease. Detection of tubers in tuberous sclerosis complex. *PloS one*, Volume 15(4), p. e0232376.

[20] Saratxaga, C. et al., 2021. MRI Deep Learning-Based Solution for Alzheimer's Disease Prediction. *Journal of personalized medicine*, Volume 11(9), p. 902.

[21] Selkoe, D. J. & Lansbury, P. J., 1999. Alzheimer's disease is the most common neurodegenerative disorder. *Basic Neurochemistry: molecular, cellular and medical aspects*, Volume 6, pp. 101-102.

[22] Small, G. W. et al., 1997. Diagnosis and treatment of Alzheimer disease and related disorders: consensus statement of the American Association for Geriatric Psychiatry,. *the Alzheimer's Association, and the American Geriatrics Society.* , Volume 278 (16), pp. 1363-1371.

[23] Tang, Z. et al., 2019. Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nature communications*, Volume 10 (1), pp. 1-14.

[24] Tufail, A., Ma, Y. K. & Zhang, Q. N., 2020. Binary classification of Alzheimer's disease using sMRI imaging modality and deep learning. *Journal of digital imaging*, Volume 33(5), pp. 1073-1090.

[25] Wen, J. et al., 2020. Alzheimer's Disease Neuroimaging Initiative, 2020. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*, Volume 63, p. 101694.

[26] Yagis, E. et al., 2021. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific reports*, Volume 11(1), pp. 1-13.

[27] Yagis, E., De Herrera, A. G. & Citi, L., 2019. Generalization Performance of Deep Learning Models in Neurodegenerative Disease Classification. *in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 1692–1698.

[28] Zhang, Y. et al., 2021. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods* , Volume 353, p. 109098.