# Explainable AI for Medical Image Processing: A Study on MRI in Alzheimer's Disease

Linda M. Duamwan
Department of Computer Science, Nottingham Trent University
Nottingham, United Kingdom
linda.duamwan2021@my.ntu.ac.uk

Jordan J. Bird
Department of Computer Science, Nottingham Trent University
Nottingham, United Kingdom
jordan.bird@ntu.ac.uk

## ABSTRACT

Alzheimer's disease is the most common type of dementia, characterised by memory-related brain changes that impair the patient's cognitive abilities. Early detection of this disease is critical from a clinical point of view to increase the chances of treating a patient at risk of further cognitive degeneration. The rate of development and expansion in the field of deep learning for medical analysis is rapidly increasing alongside the increased incidences of neurodegenerative diseases. Machine learning algorithms are often applied to automate tasks and alleviate issues, but modern methods such as neural networks often present as black boxes. In the field of medicine, it is crucial to understand why a machine learning algorithm has made a prediction. In this study, we initially utilise a CNN-based approach for computer vision in the detection of Alzheimer's disease from the ADNI MRI dataset. On unseen magnetic resonance imaging scans, our algorithm achieves a classification accuracy of 94.96%. We then implement the Local Interpretable Model Agnostic Explanations (LIME) algorithm to reveal visual evidence to support the predictions made by the model, and automatically visualise image segments contributing to predictions via Felzenszwalb's segmentation algorithm. The objectives of explainable AI in this field are to provide medical professionals with specific, easy-to-understand information to support efficient, consistent, and convenient diagnoses.

## CCS CONCEPTS

• **Theory of computation** → **Design and analysis of algorithms**; • **Human-centered computing** → **Visualization application domains**.

## KEYWORDS

Explainable AI, Medical Image Processing, Image Classification

**ACM Reference Format:**
Linda M. Duamwan and Jordan J. Bird. 2023. Explainable AI for Medical Image Processing: A Study on MRI in Alzheimer's Disease. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 5–7, 2023, Corfu, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3594806.3596521


*PETRA '23, July 5–7, 2023, Corfu, Greece*

ACM ISBN 979-8-4007-0069-9/23/07…$15.00
https://doi.org/10.1145/3594806.3596521

## 1 INTRODUCTION

Alzheimer's disease is the most common type of dementia, characterised by memory-related brain changes that impair the patient's cognitive abilities. Socially, Alzheimer's ranks as the second-most feared disease behind cancer [Patterson, 2018]. Diagnoses of the disease are made every three seconds through study of medical history, cognitive tests, clinical testing, and imaging techniques which include Magnetic Resonance Images (MRI) and Positron Emission Tomography (PET) to reveal structural abnormalities in the brain.

Although there is currently no cure for the root causes of the disease, it can be predicted early with the help of machine learning technology, which can identify those most at risk before treatment is administered to slow neurodegeneration [Islam and Zhang, 2018]. State-of-the-art work involves the application of algorithms for multiple brain modalities for early diagnosis. However, machine predictions alone are not enough when lives are at stake, given that deep learning technologies are often inherently black box approaches. Since this notion leads to scepticism, the ability to explain medical predictions is considered a crucial feature for algorithms [Fuhrman et al., 2022]. It was noted in [Ribeiro et al., 2016] that users avoid methodologies and predictions if they are not presented with explanations and rationale for those specific decisions.

The motivation of this study is to discover a continuous pattern that can be recognised when Alzheimer's disease initially manifests itself. This is achieved by analysing brain Magnetic Resonance Imaging (MRI) data with a convolutional neural network for classification. We then implement the Local Interpretable Model Agnostic Explanations (LIME) to reveal visual evidence to support the predictions made by the model. The objective of explainable algorithms in this field is to provide medical professionals with specific, easy-to-understand information to support efficient, consistent, and convenient diagnoses.

## 2 BACKGROUND AND RELATED WORK

In 1901, the German physician Alois Alzheimer saw 51-year-old Auguste Deter who presented symptoms such as hiding belongings, threatening neighbours, accusing her husband of adultery, and difficulty performing everyday tasks [Alzheimer, 1907, McGirr et al., 2020]. This was the first diagnosed case of a form of dementia, now named after the physician, *Alzheimer's Disease* (AD). The disease is the leading cause of dementia today; around 46 million people suffer from Alzheimer's disease worldwide, and this figure is expected

to rise to around 130 million by 2050 [Tufail et al., 2020]. Early detection of this disease is critical from a clinical point of view to increase the chances of treating a patient at risk of increasing cognitive degeneration. Medical image analysis is a method for early identification of neurodegeneration, which can provide methods to improve accuracy and reduce human error due to variation and exhaustion. This section will compare methods of computer vision-assisted diagnosis for Alzheimer's disease proposed by the current literature.

## 2.1 Classification and Computer-Aided Detection

The classification of brain Magnetic Resonance Imaging (MRI) data in medical imaging has been the subject of numerous investigations [Akkus et al., 2017]. MRI brain scan classification methods have also been shown to be promising in the literature [Lu, 2019, Talo et al., 2019]. Typically, one or more images are presented as input to a predictive model, and a single characteristic is output, ie, the presence or progressive stage of the disease [Altaf et al., 2019].

In 2017, Sarraf and Tofighi [Sarraf and Tofighi, 2017] presented a study on fMRI and MRI data, which employed convolutional neutral networks to classify AD. For the two datasets, the constructed network obtained a maximum accuracy of 95.13% and 98.7%, respectively. In [Islam and Zhang, 2018], researchers also used CNNs on Alzheimer's classification based on MRI obtained from the OASIS database. Two baselines on deep CNNs were presented; residual network (ResNet) and Inception-v4 models, modified to process 3-dimensional MRI data to determine if the patient was suffering from Alzheimer's disease, and demonstrate that the proposed model outperforms comparable baselines with 96.4% accuracy. Utilising a transfer-learning approach, [Deepak and Ameer, 2019] proposed a precise and fully automated brain lesion classification model with minimum pre-processing. The suggested approach scored around 98% classification accuracy. In [Achilleos et al., 2020], researchers suggested the use of decision trees and their random forest ensemble counterparts for MRI classification, achieving around 91% accuracy. It is worth noting that this approach, although scoring lower than other methods, is considerably less computationally expensive to both train and infer.

Computer-assisted detection involves detecting feature points in an input image and producing a frame around them. The primary objective here is to find anomalies in patients as early as possible [Rehman et al., 2021]. The majority of published research studies on the region of interest detection or localisation in medical images use CNN models, followed by a variety of post-processing to obtain candidate regions. On a relatively large dataset, [Bäckström et al., 2018] proposed a 3D-ConvNet that showed significant improvements for Alzheimer's disease diagnosis. The researchers evaluated the proposed architecture using the ADNI dataset and obtained an accuracy rate of 98.74%, 100% disease discovery and a 2.4% false alarm rate. By utilising very deep convolutional neural network (VGGNet) architecture [Khan et al., 2019] offered an automated method for Alzheimer's Disease prediction deployed on a double-layer transfer learning in which a predetermined set of layers is trained on MRI images. When compared with other existing methods, their model showed increases in accuracy from 4% to 7%.

## 2.2 Segmentation and Explainable AI

Image Segmentation involves splitting an input image into several portions known as image regions. It is also commonly used in medical image processing using a variety of image formats [Rehman et al., 2021].

In 2018, a complementary segmentation network (CompNet) was trained on brain images [Dey and Hong, 2018]. Compared to a plain U-Net and a dense U-Net, the CompNet achieved an accuracy of 98.27% for normal images and 97.62% for pathological images. By removing some of the restrictions of the Unet standard, [Zhou et al., 2019] suggested an improved design named UNet++. The UNet++ architecture was tested on six different datasets obtained from the public domain. According to the results, the new UNet++ approach outperformed the UNet model with a 4% increase. A study from 2020 shows the preprocessing of the hippocampus through a double-cubic spline approach (LEMS) and an inhomogeneity intensity correction approach [Choi et al., 2020]. The source data were split into three parts, attaining accuracies of 92.3%, 85.6% and 78.1%.

Explainable AI (XAI) is a research field to explore strategies for the explainablilty of predictions made by machine and deep learning approaches [Pawar et al., 2020]. It is important that decisions are visible, explained and trusted in medical image processing [Lakkaraju et al., 2016].

Regarding Alzheimer's disease classification, a double-layered explainable model was proposed in 2021 [El-Sappagh et al., 2021]. The study combined 11 modalities with a Random Forest classifier for multiclass prediction, and was explained through the Shapley Additive Explanations (SHAP) framework. In the second layer of the approach, binary classification was performed on the possibility of cognitive impediment. The first layer showed an accuracy of 93.94% (with an increase of 0.5%), and the second layer achieved a classification accuracy of 87. 08%. Similarly, in the explainable domain, a combination of SpinalNet and Convolutional Neural Network was proposed [Kamal et al., 2021]. In this approach, gene databases and MRI scans were classified and explained using the Local Interpretable Model-agnostic Explanations (LIME) approach. According to the report, support vector classifier outperformed other algorithms when dealing with gene data, while the convolutional neural network had a 97.6% accuracy rate for MRI image categorisation.

It is important to summarise the drawbacks of approaches that were presented by the related literature. These include data availability, imbalance, and the lack of explainability for approaches. For example, a model of 90% accuracy which presents explanations to a medical expert is likely more useful than a superior model of 99% accuracy in the form of a black box which does not explain why such decisions are made. Many researchers focus on improving models without regard for the importance of explainability. A general summary of some of the works covered within this literature review can be found in Table 1.

**Table 1: General summary of several state-of-the-art works in medical image processing.**

| Study | Dataset | Modality | Approach | Proposed model | Accuracy | XAI |
|---|---|---|---|---|---|---|
| [Sarraf and Tofighi, 2017] | ADNI | Multi | Classification | CNN | 95.13 & 98.7 | No |
| [Islam and Zhang, 2018] | OASIS | MRI | Classification | CNN | 96.40 | No |
| [Dey and Hong, 2018] | OASIS | MRI | Segmentation | CompNet | 98.27/97.62 | No |
| [Deepak and Ameer, 2019] | Figshare | MRI | Classification | CNN | 98.00 | No |
| [Achilleos et al., 2020] | OASIS | MRI | Classification | Decision Trees | 91.00 | No |
| [Bäckström et al., 2018] | ADNI | MRI | localization | CNN | 98.74 | No |
| [Zhao et al., 2019] | ADNI | Multi | Registration | RCA | 89.50 | No |
| [Choi et al., 2020] | ADNI | Multi | Segmentation | LEMS | 92.3, 85.6 & 78.1 | No |
| [Kamal et al., 2021] | OASIS | MRI | Classification | SpinalNet & CNN | 97.60 | LIME |
| [El-Sappagh et al., 2021] | ADNI | Multi | Classification | Random Forest | 87.08 to 93.95 | SHAP |
| This Study | ADNI | MRI | Classification | CNN | 94.96 | LIME |

## 3 METHOD

In this section, we will explore the method followed by this study. This covers the data collection and design of a computer vision system for the extraction of features from MRI images prior to predicting the presence and stage of Alzheimer's disease. We will also explore the methods of explaining predictions via the LIME framework.

### 3.1 Data Collection and Preprocessing

Data for this study are initially collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [Petersen et al., 2010]. ADNI contains brain MRI scans gathered from the internet, hospitals, and public sources. It is important to note that each label was then independently validated. The dataset is the result of a partnership between The National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, as well as several private pharmaceutical companies and nonprofit organizations, with the goal of the development of new technologies for diagnosing and treating Alzheimer's disease.

The dataset is made up of preprocessed magnetic resonance imaging (MRI) images that are resized to 128px square images. It is divided into two separate files for training and testing, each of which comprises random samples from the four classes of images (non-dementia, very mild, mild, and moderate). Samples of images from the dataset can be seen in Figure 1.

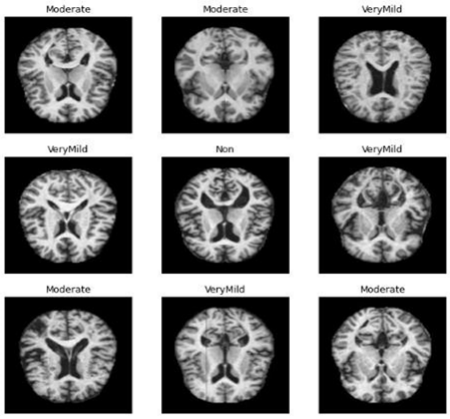


**Figure 1: Nine random samples of images from the ADNI dataset with their respective class labels.**

### 3.2 Machine Learning and Explainable AI

To prevent overfitting to the training set, augmentation is applied for generalisation. Although imaging is uniform, augmentations such as zooming is implemented in order to augment synthetic larger or smaller regions that may be present in brains within the unseen data. The input matrix of pixel values are initially rescaled from 0-255 to 0-1 as floating point numbers to ensure that each value comes from a standard distribution. The class labels are also encoded using the one-hot method to prevent regression. The dataset is also resampled to avoid data imbalance. Brightness ranges are randomly changed within 1% of the original value, randomly flipped horizontally (since CNNs do not generalise mirrored data), and randomly zoomed. Should the augmented image be smaller than the original bounds, a constant fill is used to replace the missing pixels. The model architecture combines three CNN layers of 16, 32, and 64 filters in sequential order, followed by two layers of 128 and 64 rectified linear units. An overview of the proposed model architecture is visualised in Figure 2.

Following machine learning, interpretation of deep learning models and predictions provide further understanding as to which visual features influenced the model's output. To provide this aspect, we implement the Local Interpretable Model-Agnostic Explanations (LIME), and an example is shown in Figure 3.

## 4 RESULTS AND OBSERVATIONS

In this section, we will present and discuss the results of the machine learning model in terms of classification metrics and confusion matrix. Subsequently, we will then explore how explained predictions can be presented to experts via visual masks.

Following implementation of the CNN, initial machine learning metrics are presented in Table 2. As can be observed, the model achieved 94.96% classification accuracy with likewise precision, recall, and F-1 values of 0.95. The categorical cross-entropy loss resulted in 0.24. A further expansion of these metrics is presented

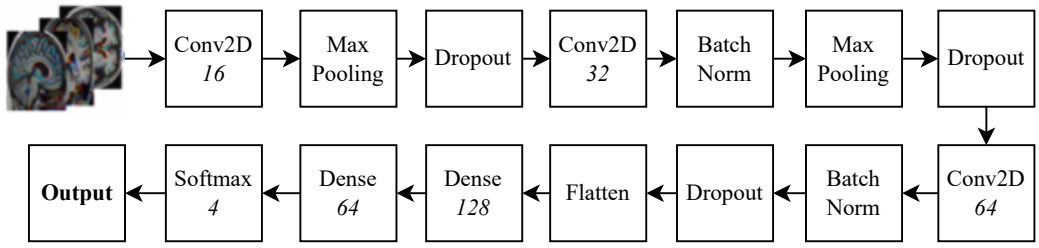


**Figure 2: Model architecture for the Convolutional Neural Network.**

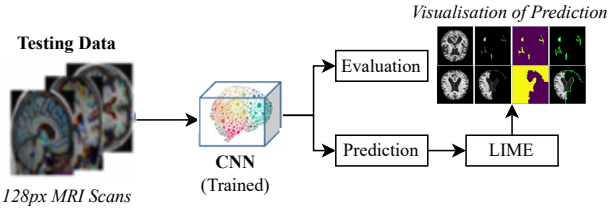


**Figure 3: General pipeline of explainability. Alongside classical machine learning metrics, the system visualises pixels on the input image that have contributed towards the prediction.**

**Table 2: Classifier metrics on unseen data following training.**

| Classifier Metric | Value |
|---|---|
| Accuracy | 94.96 |
| Precision | 0.95 |
| Recall | 0.95 |
| F-1 Score | 0.95 |
| AUC | 99.98 |
| Loss | 0.24 |

**Table 3: Expanded machine learning metrics for the classifier model on unseen data.**

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Non-dementia | 0.98 | 0.93 | 0.96 |
| Very Mild Dementia | 0.9 | 0.96 | 0.93 |
| Mild Dementia | 1 | 1 | 1 |
| Moderate Dementia | 0.97 | 0.94 | 0.95 |
| **Averages** | | | |
| Micro | 0.95 | 0.95 | 0.95 |
| Macro | 0.96 | 0.96 | 0.96 |
| Weighted | 0.95 | 0.95 | 0.95 |
| Samples | 0.95 | 0.95 | 0.95 |

in the complete classification report, which can be found in Table 3. The training accuracy was observed to be around 99.9%, suggesting that some overfitting is occurring. Resampling the data did not completely overcome the issues related to class imbalance, as can be seen from the slight differentiation in the metrics on a per-class basis.

**Table 4: Normalised confusion matrix for the classifier on unseen data.**

| | Non | Very Mild | Mild | Moderate |
|---|---|---|---|---|
| **Non** | 0.93 | 0.07 | 0 | 0 |
| **Very Mild** | 0 | 0.96 | 0 | 0.04 |
| **Mild** | 0 | 0 | 1 | 0 |
| **Moderate** | 0.01 | 0.05 | 0 | 0.94 |

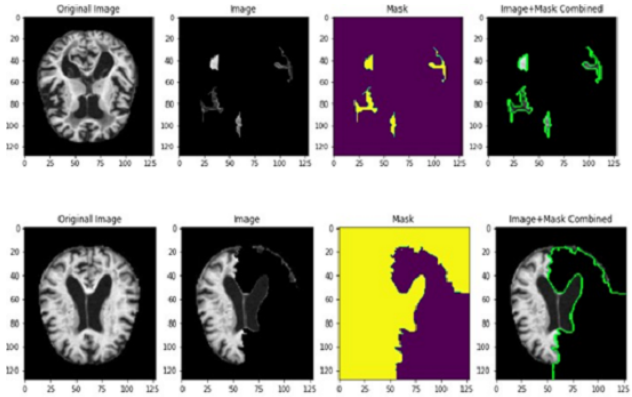


**Figure 4: Explainable image masks on correctly classified dementia. Top: Non-dementia prediction, Bottom: Positive prediction.**

When considering the unseen data, a confusion matrix of predictions can be found in Table 4. When considering medical classification, mistakes are non equal i.e., it is better to erroneously classify a false alarm rather than a missed diagnosis. This suggests that score-based evaluation would be better suited to this problem in future, as can be seen in the moderate dementia classification results. This issue did not occur for very mild and mild dementia, with the mistakes being that very mild was misclassified as moderate. In a clinical context, this issue would cause less of an issue, given that an expert is alerted all the same.

Following the inference of predictions, the LIME explainer is implemented to visualise the brain regions that have contributed the

most to the output. An example of how this could be presented to experts can be found in Figure 4. This goes beyond simply predicting a class value since it can be used to draw attention to regions of the brain for further study by an expert. In this sense, it can contribute to the reduction of time required for diagnoses, and therefore, enhance the expert's ability.

## 5 CONCLUSION AND FUTURE WORK

To conclude this study, we have presented work on the automatic learning and classification from MRI brain scan images as well as examples of how explainable AI can be used to draw expert attention to visual cues within said scans for diagnosis of Alzheimer's disease. As discussed, explainability is crucial for trustworthiness in systems that deal with human health and wellbeing. Prior to presentation of the all-important XAI, we show that our computer vision approach competes with the state of the art on this dataset (with many of those approaches being black boxes). Additionally, our CNN is relatively small and therefore not as computationally expensive as many other methods, e.g., the 138 million parameters of the VGG16 network or the 62.3 million parameters of AlexNet.

The results presented lead to much future work that could be carried out given our findings. For example, we found that upon studying the confusion matrix, some errors are not equal to others when we consider clinical implications. For this reason, it may be better to implement score-based goals for the model in order to prevent situations where sufferers of Alzheimer's are classified as not having the disease. This is due to the implication that a false alarm is better than a patient going undiagnosed. Future work may also explore further model architectures and hyperparameter optimisation to further increase the classification ability beyond our presented results.

## REFERENCES


Kleo G Achilleos, Stephanos Leandrou, Nicoletta Prentzas, Panayiotis A Kyriacou, Antonis C Kakas, and Constantinos S Pattichis. 2020. Extracting explainable assessments of Alzheimer's disease via machine learning on brain MRI imaging data. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 1036–1041.

Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson. 2017. Deep learning for brain MRI segmentation: state of the art and future directions. *Journal of digital imaging* 30 (2017), 449–459.

Fouzia Altaf, Syed MS Islam, Naveed Akhtar, and Naeem Khalid Janjua. 2019. Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access* 7 (2019), 99540–99572.

Alois Alzheimer. 1907. Uber eigenartige Erkrankung der Hirnrinde. *All Z Psychiatr* 64 (1907), 146–148.

Karl Bäckström, Mahmood Nazari, Irene Yu-Hua Gu, and Asgeir Store Jakola. 2018. An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 149–153.

Boo-Kyeong Choi, Nuwan Madusanka, Heung-Kook Choi, Jae-Hong So, Cho-Hee Kim, Hyeon-Gyun Park, Subrata Bhattacharjee, and Deekshitha Prakash. 2020. Convolutional neural network-based MR image analysis for Alzheimer's disease classification. *Current Medical Imaging* 16, 1 (2020), 27–35.

S Deepak and PM Ameer. 2019. Brain tumor classification using deep CNN features via transfer learning. *Computers in biology and medicine* 111 (2019), 103345.

Raunak Dey and Yi Hong. 2018. CompNet: Complementary segmentation network for brain MRI extraction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*. Springer, 628–636.

Shaker El-Sappagh, Jose M Alonso, SM Islam, Ahmad M Sultan, and Kyung Sup Kwak. 2021. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific reports* 11, 1 (2021), 1–26.

Jordan D Fuhrman, Naveena Gorre, Qiyuan Hu, Hui Li, Issam El Naqa, and Maryellen L Giger. 2022. A review of explainable and interpretable AI with applications in COVID-19 imaging. *Medical Physics* 49, 1 (2022), 1–14.

Jyoti Islam and Yanqing Zhang. 2018. Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain informatics* 5 (2018), 1–14.

Md Sarwar Kamal, Aden Northcote, Linkon Chowdhury, Nilanjan Dey, Rubén González Crespo, and Enrique Herrera-Viedma. 2021. Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes. *IEEE Transactions on Instrumentation and Measurement* 70 (2021), 1–7.

Naimul Mefraz Khan, Nabila Abraham, and Marcia Hon. 2019. Transfer learning with intelligent training data selection for prediction of Alzheimer's disease. *IEEE Access* 7 (2019), 72726–72735.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

Yang Lu. 2019. Artificial intelligence: a survey on evolution, models, applications and future trends. *Journal of Management Analytics* 6, 1 (2019), 1–29.

Samantha McGirr, Courtney Venegas, and Arun Swaminathan. 2020. Alzheimers disease: A brief review. *Journal of Experimental Neurology* 1, 3 (2020), 89–98.

Christina Patterson. 2018. World alzheimer report 2018. (2018).

Urja Pawar, Donna O'Shea, Susan Rea, and Ruairi O'Reilly. 2020. Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain.. In *AICS*. 169–180.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, Clifford R Jack, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. 2010. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology* 74, 3 (2010), 201–209.

Aasia Rehman, Muheet Ahmed Butt, and Majid Zaman. 2021. A survey of medical image analysis using deep learning approaches. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 1334–1342.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

S Sarraf and G Tofighi. 2017. for the Alzheimer's Disease Neuroimaging Initiative DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *bioRxiv* (2017).

Muhammed Talo, Ulas Baran Baloglu, Özal Yıldırım, and U Rajendra Acharya. 2019. Application of deep transfer learning for automated brain abnormality classification using MR images. *Cognitive Systems Research* 54 (2019), 176–188.

Ahsan Bin Tufail, Yongkui Ma, and Qiu-Na Zhang. 2020. Multiclass classification of initial stages of Alzheimer's Disease through Neuroimaging modalities and Convolutional Neural Networks. In *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*. IEEE, 51–56.

Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. 2019. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10600–10610.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging* 39, 6 (2019), 1856–1867.


Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009