OPERATIONALIZING FAIRNESS IN MEDICAL AI ADOPTION: DETECTION OF EARLY ALZHEIMER'S DISEASE WITH 2D CNN

Luca Mira Heising[*]

Tilburg University
Warandelaan 2, 5037 AB
Tilburg, The Netherlands
L.M.Heising@tilburguniversity.edu

Spyros Angelopoulos[†]

Durham University
Mill Hill Lane, DH1 3LB
Durham, United Kingdom
Spyros.Angelopoulos@durham.ac.uk

Contributorship statement: All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. Specifically, L.M.H conceived the idea for the project and S.A oversaw its overall direction and planning. Both L.M.H and S.A worked on the acquisition of the data for analysis. L.M.H designed the computational framework and analysed the data. S.A worked out the technical details for analysis and helped in the interpretation of the results. Both authors discussed the findings and contributed to the writing of the manuscript.

Ethics committee: Durham University Business School, DUBS-2021-10-31T21:05:04-xggj42

Funding statement: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors

Word Count: 2990

[*] Leading author
[†] Corresponding author

OPERATIONALIZING FAIRNESS IN MEDICAL AI ADOPTION: DETECTION OF EARLY ALZHEIMER’S DISEASE WITH 2D CNN

ABSTRACT

Objectives: To operationalize fairness in the adoption of medical artificial intelligence (AI) algorithms in terms of access to computational resources, the proposed approach is based on a two-dimensional (2D) Convolutional Neural Networks (CNN), which provides a faster, cheaper, and accurate-enough detection of early Alzheimer’s Disease (AD) and Mild Cognitive Impairment (MCI), without the need for use of large training datasets or costly high-performance computing (HPC) infrastructures.

Methods: The standardized ADNI datasets are used for the proposed model, with additional skull stripping, using the BET2 approach. The 2D CNN architecture is based on LeNet-5, the LReLU activation function and a Sigmoid function were used, and batch normalization was added after every convolutional layer to stabilize the learning process. The model was optimized by manually tuning all its hyperparameters.

Results: The model was evaluated in terms of accuracy, recall, precision, and f1-score. The results demonstrate that the model predicted MCI with an accuracy of .735, passing the random guessing baseline of .521, and predicted AD with an accuracy of .837, passing the random guessing baseline of .536.

Discussion: The proposed approach can assist clinicians in the early diagnosis of AD and MCI, with high-enough accuracy, based on relatively smaller datasets, and without the need of HPC infrastructures. Such an approach can alleviate disparities and operationalize fairness in the adoption of medical algorithms.

Conclusion: Medical AI algorithms should not be focused solely on accuracy but should also be evaluated with respect to how they might impact disparities and operationalize fairness in their adoption.

SUMMARY

What is already known?

- Most prior studies on early Alzheimer’s Disease (AD) and Mild Cognitive Impairment (MCI) detection have used a three-dimensional (3D) Convolutional Neural Networks (CNN) approach.
- The 3D CNN approach is computationally expensive requiring high performance computing (HPC) infrastructures, and, due to the high number of parameters, it requires larger datasets for training.
- A two-dimensional (2D) CNN needs less parameters, less computational power, and execution time, while requires smaller datasets for training, but has not been applied to date for MCI detection.

What does this paper add?

- The proposed approach based on a 2D CNN operationalizes fairness in the adoption of medical artificial intelligence (AI) algorithms by providing fast, cheap, and accurate-enough detection of early AD and MCI without the need for use of large datasets or costly HPC infrastructures.
- The proposed approach can be extended to other diseases, as well as to other cases where time is scarce, powerful computational resources are not available, and large datasets are out of reach.

OPERATIONALIZING FAIRNESS IN MEDICAL AI ADOPTION: DETECTION OF EARLY ALZHEIMER'S DISEASE WITH 2D CNN

INTRODUCTION

Recent studies show that Artificial intelligence (AI) applications can perform on par with medical experts on Magnetic Resonance Images (MRI) analysis [1]. Such applications to date, tend to oppose the accuracy of AI to the performance of clinicians. For instance, there have been more than 20,000 studies on deep learning (DL) methods for MRI analyses the last decade, which compare the performance of AI to the one of clinicians [2]. Recent work suggests that future studies should focus on the comparison of performance between clinicians using AI, and their performance without an AI aid [3]. The recent global pandemic, however, revealed another urgent need of early disease diagnosis: the ability to make predictions based on a limited number of cases. The AI Computer-Aided-Detection (CAD) frameworks to date, are based on large amounts of data, and require high-performance computing (HPC) infrastructures. To address that lacuna, we propose a synergistic approach, in which clinicians and scientists collaborate for faster, cheaper, and more accurate detection, relying on small datasets to make accurate-enough predictions. A promising frontier where AI can assist clinicians is Alzheimer's Disease (AD) since the release of promising clinical studies for a new drug have unearthed the need for its early detection. As it can take up to 20 years before AD patients show any signs of cognitive decline, it can be challenging to diagnose AD in early stages. We, thus, motivate and implement an AI-CAD framework for the early detection of Mild Cognitive Impairment (MCI) and AD to assist clinicians, while the approach can be extended for the diagnosis of other diseases.

AD is caused by an accumulation of β-amyloid (Aβ) plaques, and abnormal amounts of *tau* proteins in the brain. This results in synapse loss, where the impulse does not reach the neurons, and in loss of structure or function of neurons, including their death, causing memory impairment and other cognitive problems [4]. AD has strong impact on the cognitive and physical functioning of patients, resulting in death. Recent developments in slowing AD decline have increased the relevance of its early detection [5], and

MCI plays an important role in this. MCI is a syndrome where the patients have greater cognitive decline than normally expected, but it does not necessarily affect their daily lives. Although some MCI patients remain stable or return to cognitively normal (CN), there is a 10-15% risk per year of progression to AD [4]. Before the etiology of AD became known, its diagnosis relied on neurocognitive tests. The development of biomarkers improved AD detection. A common method to diagnose AD is hippocampus segmentation, which relates to memory function, and its small volume is an AD biomarker. For a long time, AD diagnosis was done manually by looking at the brain structure and size of the hippocampus on MRI, which requires practice and precision. Prior studies on automated methods for hippocampus segmentation have used DL approaches with promising results [6]. Automated hippocampus segmentation for the diagnosis of AD and MCI, however, requires clinicians' expertise and is sensitive to interrater and intra-rater variability [6].

Convolutional Neural Networks (CNN) can become the foundation of an AI-CAD framework for supporting clinicians in the detection of early AD and MCI, since it is a successful approach for image classification. CNN can improve the performance of image classification [7], and they are becoming increasingly popular in MRI analysis. For instance, recent studies show that CNN can work on par with specialists for classifying MRI of skin cancer patients [1]. Similar approaches with three-dimensional (3D) as well as two-dimensional (2D) CNN have also been used for AD detection with promising results. When it comes to the inner mechanics of these approaches, the classification filter of a 3D CNN slides along all the three dimensions of the input image, resulting in 3D feature maps, whereas in a 2D CNN the classification filter slides along only the height and width of the input image. Thus, the latter results in 2D feature maps, which need less parameters, computational power, and execution time. Most prior studies have used 3D CNN achieving high accuracy [8], while others obtained similar results with 2D CNN [9]. Although previous work on the topic has established that 3D CNN perform better for patch classifications, the results between 2D and 3D approaches for whole image labeling did not differ much [10]. A 3D CNN, however, is more computationally expensive, and, due to the high number of parameters, it requires larger datasets for training [11]. Concurrently, prior studies have not incorporated a 2D CNN approach for detecting MCI. A summary of prior 2D and 3D CNN applications in the literature is presented in Table 1.

Table 1. Performance comparison of 2D and 3D approaches in the literature

Study	**2D CNN**		**3D CNN**	
	AD	**MCI**	**AD**	**MCI**
Basaia et al. [8]	-	-	.99	.87
Feng et al. [12]	-	-	.95	.86
Korolev et al. [13]	-	-	.80	-
Liu et al. [14]	-	-	.85	-
Liu et al. [15]	-	-	.91	-
Senanayake et al. [16]	-	-	.76	.75
Hon and Khan [17]	.96	-	-	-
Sarraf and Tofighi [18]	.99	-	-	-
Sarraf and Tofighi [19]	.97	-	-	-
Wang et al. [9]	.98	-	-	-

We suggest that medical algorithms should not be solely focused on accuracy but should also be evaluated with respect to how they might impact disparities and operationalize fairness in their adoption. Thus, we investigate the extent to which a 2D CNN can detect MCI and early AD.

METHODS

CNN is the most common neural network (NN) architecture for image classification. Fully connected NN take multiple inputs, and hidden layers perform calculations on them, while the neurons in the network connect to each other. Neurons in CNN, however, connect only to those close to them. CNN, therefore, need fewer parameters, which results in benefits such as small risk of overfitting, higher accuracy, and faster processing time. Moreover, in CNN there is no need to transform the input images to one-dimensional, a process which can result in loss of structural information, as the CNN can learn the relationships among the pixels of input by extracting representative features with kernel convolutions [4]:

$$S(i,j) = (I \times K)(i,j) = \sum_{m}\sum_{n} I(m,n)K(i-m,j-n)$$

where I is the input and K is the kernel; the input indices are represented by i and j, and the kernel indices are represented by m and n.

The datasets used in this study were obtained under permission from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership. The primary goal of ADNI has been to test whether MRI, biological markers, and clinical as well as neuropsychological assessment can be combined to measure the progression of MCI and early AD. The ADNI is separated into 3 studies of 5 years, while the first was prolonged by 2 years under the name ADNI-GO. In total, 2517 people of ages 55-90 participated in the study. The ADNI encourages the use of their standardized datasets to ensure consistency in analysis and direct comparison of various methods among studies. We, therefore, used their two standardized datasets 'ADNI1:Complete 2Yr 1.5T', and 'ADNI1:Complete 3Yr 1.5T', which contain MRI that have passed quality control assessment [20].

Our dataset consists of 3312 images, distributed in 828 MRI of CN subjects, 453 MRI of AD patients, and 1203 MRI of MCI patients. The dataset was split into one with CN and AD subjects (1281 MRI), and one with CN and MCI subjects (2031 MRI). Since the participants of the ADNI study returned for more than one check-up, any patient can have up to 12 MRI, which are not identical as they are taken at different moments, and every MRI in the standardized dataset was treated independently. The dataset, thus, refers to 99 AD patients, 212 MCI patients, and a control group of 165 CN subjects. We present the demographic information of the included subjects in Table 2, to enable comparison with other studies.

Table 2. Demographic information of subjects in the dataset

	MCI	**AD**	**CN**
Images	891	412	662
Subjects	212	99	165
Gender	142 M / 70 F	52 M / 47 F	82 M / 83 F
Age	μ = 75.84 σ = 7.02	μ = 76.49 σ = 7.43	μ = 76.93 σ = 5.23

Whilst the datasets are pre-processed, we further performed skull stripping using the Brain Extraction Tool v2 (BET2), which is part of the *NiPype* library. Skull stripping locates the brain in the MRI and removes all surroundings to further remove noise from images. For optimal skull stripping results, neck slices were removed with the *robustfov* function. We used a fraction intensity of 0.3 as an evaluation of BET2 parameters for the ADNI dataset found that this leads to best results. Due to the differences in scanners and techniques used by the ADNI over the years, the MRI used in the datasets were of different sizes, and, therefore had to become uniform. All the MRI in our dataset were resized to: (136, 192, 160) with the *ndimage* zoom function of the *Scipy* library, which zooms the array using spline interpolation. Resizing the MRI results in a different range of pixel values, and, therefore, to assure that the pixel values of all MRI had the same range, z-score normalization was applied, which is defined as follows:

$$z_i = \frac{x_i - \mu(x)}{\sigma(x)}$$

where x is the MRI data and z_i the i^{th} normalized MRI. The dataset was then split into train set, validation set, and test set with a ratio of 60:20:20 respectively.

A NN consists of an input layer, hidden layers, and an output layer. A CNN has hidden layers divided into convolution, pooling, activation, and classification layers. We based our architecture on LeNet-5, which includes 2 convolutional layers, 2 pooling layers, and 2 fully connected layers (Table 3).

Table 3. CNN architecture

Layer	C1	P1	C2	P2	FC1	FC2	FC3
Kernel	3x3	2x2	3x3	2x2	-	-	-
Filter	32	32	64	64	128	64	2

We employ the Leaky Rectified Linear Unit (LReLU) as activation function for all convolutional layers, which allows for a small non-zero gradient [21]. The LReLU activation function in the model, with x being the input data, is described as:

$$y(x) = \begin{cases} x, & ifx < 0 \\ 0.01x, & otherwise \end{cases}$$

A Sigmoid activation function was applied to the dense layer, which outputs the probability of the images' class, with 0 if healthy and 1 if not (AD or MCI). The Sigmoid activation function in the model, with x being the input data, is described as:

$$\sigma(x) = \frac{1}{(1 + e^{-x})}$$

We optimized the model by manually tunning the hyperparameters (see Table 4).

Table 4. Parameter tuning on the AD dataset

PARAMETERS											
Round	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10***	
Learning rate	0.0001	0.0001	0.001	0.001	0.001	0.001	0.001	0.0001	0.01	0.001	0.001
Batch size	32	16	16	8	32	16	8	8	8	8	16
Epochs	50	50	50	30	30	30	30	30	30	30	40
Dropout	-	0.3	0.3	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.2
Batch norm.	-	x	x	x	x	x	-	x	x	x	x
METRICS											
Loss	1.040	0.711	0.637	**0.461**	0.742	0.600	2.292	0.805	0.639	0.600	0.677
Acc	0.833	0.794	0.802	**0.840**	0.833	0.840	0.728	0.767	0.825	0.833	0.837
Precision	0.881	**0.977**	0.891	0.768	0.947	0.949	0.628	0.972	0.788	0.859	. 948
recall	0.628	0.447	0.521	**0.809**	0.574	0.596	0.628	0.372	0.713	0.649	.585

The batch size was set to 16 and we used the Adam optimizer [22] with a learning rate of 10^{-3}. The model showed overfitting, which means that it includes more terms or uses more complicated approaches than necessary [23]. Regularization can control overfitting and drop-out regularization is a commonly used approach because it is computationally inexpensive, and it prevents co-adaptation among feature map units [11]. In drop-out regularization only a fraction of the weights is learned by the NN in each iteration. We added a drop-out layer with a value of 0.2 after each pooling layer (i.e., 80% of the weights were learned in each iteration), leading to better results on all the train, validation, and test sets. To stabilize the learning process, we added batch normalization after every convolutional layer. For each unit in a layer, the value was normalized as follows:

$$\hat{a}_i^{(l)} = \frac{a_i^{(l)} - \mathbb{E}[a_i^{(l)}]}{\sqrt{Var[a_i^{(l)}]}}$$

where a represents the activation vector of the *i*th layer l. Thereafter, the normalized values were scaled and shifted accordingly. After ~40 epochs, the model did not show increment in accuracy or reduction in loss, and overfitting increased, thus, we applied an early stopping at 40 epochs instead of the initial set of 50.

The CNN was built with a *Jupyter Notebook* using *Python 3.6.4*, *Tensorflow 2.4.0,* and *Keras 2.4.0*. To load the data in *NIfTI* format we used the *Nilearn* library, and we used the *scikit-learn* and *SciPy* libraries for data preprocessing. The development, testing, and application of the model took place on the Google Cloud Console, where we used a storage bucket to store the datasets, and three compute engine instances to perform the skull stripping and pre-processing, and to run our model independently as these steps require different computational resources. For skull stripping we used an instance with 8 vCPUs, 52 GB RAM, and two NVIDA Tesla K80 GPUs, for pre-processing we used an instance with 40 vCPUs, and 961 GB RAM. For the CNN we used an instance with 64 vCPUs, 416 GB RAM and four NVIDA Tesla T4 GPUs.

RESULTS

The model was evaluated in terms of *accuracy*, *recall*, *precision*, and *f1-score*. Recall provides sensitivity information on how many patients were correctly identified. Precision expresses how many of the positives that the model returns were actually positive. F1-score is the harmonic mean between precision and recall. A NN adjusts its weights to optimize the loss, which is calculated with the use of binary cross entropy loss:

$$CE = -\sum_{i=1}^{C'=2} t_i log(s_i) = t_1 log(s_1) - (1 - t_1) log(1 - s)$$

where C represents the classes, s_i is the predicted probability value for class i, and t is the true probability for that class. Since the data was unevenly distributed, the accuracy baseline of random guessing was also calculated. The baseline was calculated with respect to the class distribution of the dataset. First, we trained and tested our model on the AD dataset. After passing the baseline of random guessing on the training data (>.548) with an accuracy of .994, we applied the same model on the MCI dataset. The random guessing baseline for the test dataset of the AD model was .536 and for the test dataset of the MCI model was .521. The over-epochs performance of the model is depicted in Figure 1 for AD (left), and for MCI (right).

While the above graphs indicate a normal learning curve, as the performance of the model keeps increasing on the train dataset, the validation performance flattens, which implies overfitting. This appears to be true mainly on the AD dataset. Our model achieved accuracy of .837 on the AD test-set. Irrespective of overfitting, the achieved test accuracy on the AD dataset surpasses the random guessing baseline of .536. The model predicted MCI with accuracy of .735, passing the random guessing baseline of .521. Table 5 presents the performance metrics of the models on the test sets. The model performs better than chance on both sets, with a better predictive performance for the AD dataset than for the MCI dataset. The MCI model, however, seems to perform better on selecting relevant items (i.e., recall, predicted positives relative to all positives). The MCI model shows notably less overfitting than the AD model, which might be due to the size of the dataset, as the dataset used for the MCI was larger (almost double in size) than the AD one.

Table 5. Performance metrics on test data

Data	**Loss**	**Acc.**	**Prec.**	**Recall**	**F1**	**MRI**
AD	0.677	.837	. 948	. 585	. 724	1281
MCI	1.377	.735	.728	.894	.802	2031

By comparing our study to previous ones in the relevant literature (see Table 6), we notice a large difference in the size of the used datasets. Moreover, some of the prior studies only report the number of subjects in the used dataset [8], [12], [14], [15], but the number of images can differ from these since one subject can have up to 12 images in these datasets. As expected, studies with larger datasets, achieved higher accuracy. Furthermore, some of the studies with a 2D approach treated the slices independently [9], [18], [19], thereby enlarging the size of their dataset, however, the MRI was not treated as a whole.

Table 6. Comparison of data and accuracy with previous studies

Study	Subjects	Images	Dimensions	Accuracy	
				AD	MCI
Basaia et al. [8]	645	-	3D	.99	.87
Feng et al. [12]	193	-	3D	.95	.86
Korolev et al. [13]	111	111	3D	.80	-
Liu et al. [14]	193	-	3D	.85	-
Liu et al. [15]	902	-	3D	.91	-
Senanayake et al. [16]	-	322	3D	.76	.75
Hon and Khan [17] *	200	6400	2D	.96	-
Sarraf and Tofighi [18] **	302	62,335	2D	.99	-
Sarraf and Tofighi [19] **	43	367,200	2D	.97	-
Wang et al. [9] **	98	17,738	2D	.98	-
Our	476	3,312	2D	.84	.74

* *accuracy before transfer learning = .74*

** *used MRI slices independently*

DISCUSSION

Whilst AI-CAD frameworks have been thoroughly studied, they have not been proposed as a tool for assisting clinicians. Furthermore, whilst the literature on AI-CAD frameworks is mostly approached from a computer science perspective, clinicians have been shown to lack trust in them [2], [3] [24]. Our work addresses that lacuna by providing a synergistic approach between clinicians and scientists. We contribute to the line of research on using CNN for AD and MCI detection, by applying a 2D approach. Our model predicts AD better than chance by .301, and MCI by .214. As expected, the model performed worse on detecting MCI than AD. The learning process on the MCI dataset, however, was much cleaner than the process on the AD dataset. This might be due to the size of the dataset, which can have a large impact on the process and outcomes of the model. The proposed AI-CAD framework, thus, performs better than chance for AD as well as for MCI and could assist clinicians in the early detection of AD, and MCI.

We suggest that medical algorithms should not be focused solely on accuracy but should also be evaluated with respect to how they might impact disparities and operationalize fairness in terms of computational resources, when it comes to their adoption. Our framework can be further extended to other diseases, and to cases where time is scarce, computational resources are not available, and large datasets are out of reach. Finally, our work is in line with the broader Information Systems research agenda [25], on the adoption of responsible medical AI algorithms [26], and the stewardship of sensitive personal data [27]. Therefore, our work can give rise to new avenues for interdisciplinary research and can become the bedrock for novel methodological advances, as well as ground-breaking empirical findings on the broader topic.

CONCLUSION

Prior studies have used CNN to diagnose MCI and early AD, most of which applied 3D approached. The 3D CNN, however, have drawbacks that relate to needs for HPC infrastructures. Other studies have focused on detecting AD with a 2D CNN, achieving similar results as the 3D approach. Despite the relevance of detecting MCI, prior studies did not investigate how these methods perform on detecting MCI. Our main goal was to determine whether a 2D CNN can be used to diagnose AD and MCI. Our work resulted in an AI-CAD framework that can assist clinicians in the early detection of MCI and AD with high-enough accuracy, based on a relatively small dataset, and without the need of HPC infrastructures. Our work has limitations that need to be acknowledged. First, an important pre-progressing step is image resizing. We used *Scipy ndimage*, which distorts the image and could have a negative effect on the learning process. A better solution for resizing images is needed but to the best of our knowledge is not available. Second, the ADNI datasets consist of more images than participants. If subjects appear in both datasets, the model could learn subject-specific features but the impact on model performance is unknown, as most physical features are removed during skull stripping. Third, the AD model appears to be overfitting, which is a common problem in DL models. To further optimize our model, the overfitting problem needs to be addressed by future research. Future research should also replicate the existing 3D CNN approaches and compare their execution time with the 2D CNN one of our models on the same computational infrastructure.

Such a comparison will further illustrate the merits of our approach. Finally, future research should also evaluate the performance of clinicians using our framework, and their performance without an AI aid.

REFERENCES

[1] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115.

[2] Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;1(6):e271-e297.

[3] Briganti G, Le Moine O. Artificial Intelligence in Medicine: Today and Tomorrow. *Front Med* 2020; 7:27.

[4] Weiner MW, Veitch DP, Aisen PS, et al. 2014 Update of the Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimers Dementia* 2015;11(6):e1-e120.

[5] Selkoe DJ. Alzheimer disease and aducanumab: Adjusting our approach. *Nat Rev Neurol* 2019;15(7):365-366.

[6] Ataloglou D, Dimou A, Zarpalas D, et al. Fast and precise Hippocampus Segmentation through Deep Convolutional Neural Network Ensembles and Transfer Learning. *Neuroinformatics* 2019;17(4):563-582.

[7] Li Q, Cai W, Wang X, et al. Medical image classification with convolutional neural network. *Proc. Int. Conf. Control Automation Robotics Vision*, 2014:844-848.

[8] Basaia S, Agosta F, Wagner L, et al. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage Clin* 2019;21:101645.

[9] Wang SH, Phillips P, Sui Y, et al. Classification of Alzheimer’s disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *J Med Sys* 2018;42(5):85.

[10] Lai M, Deep learning for medical image segmentation, ArXiv:1505.02000 [Preprint], 2015.

[11] Bernal J, Kushibar K, Asfaw DS, et al. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artif Intell Med* 2019;95:64-81.

[12] Feng C, Elazab A, Yang P, et al. Deep Learning Framework for Alzheimer’s Disease Diagnosis via 3D-CNN and FSBi-LSTM. *IEEE Access* 2019;7:63605-63618.

[13] Korolev S, Safiullin A, Belyaev M, et al. Residual and plain convolutional neural networks for 3D brain MRI classification. *Proc IEEE Int Symp Biomed Imaging* 2017:835-838.

[14] Liu M, Cheng D, Wang K, et al. Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer’s Disease Diagnosis. *Neuroinformatics* 2018;16(3):295-308.

[15] Liu M, Zhang J, Adeli E, et al. Landmark-based deep multi-instance learning for brain disease diagnosis. *Med Image Anal* 2018;43:157-168.

[16] Senanayake U, Sowmya A, Dawes L, Deep fusion pipeline for mild cognitive impairment diagnosis. *Proc IEEE Int Symp Biomed Imaging* 2018:1394-1997.

[17] Hon M, Khan NM, Towards Alzheimer’s disease classification through transfer learning. *Proc IEEE Int Conf Bioinformatics Biomed* 2017:1166-1169.

[18] Sarraf A, Tofighi G, DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI. BioRxiv:070441 [Preprint], 2016.

[19] Sarraf A, Tofighi G, Classification of alzheimer’s disease using fMRI data and deep learning convolutional neural networks. ArXiv:1603.08631 070441 [Preprint], 2016.

[20] Wyman BT, Harvey DJ, Crawford K, et al. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimers Dementia* 2013;9(3):332-337.

[21] Lu L, Shin Y, Su Y, et al. Dying ReLU and Initialization: Theory and Numerical Examples. ArXiv:1903.06733 [Preprint], 2019.

[22] Kingma DP, Ba J, Adam: A method for stochastic optimization. ArXiv:1412.6980 [Preprint], 2014.

[23] Hawkins DM, The problem of overfitting. *J Chem Inf Comput Sci* 2004;44(1):1-12.

[24] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform* 2021;113(103655).

[25] Struijk M, Ou CXJ, Davison RM, et al. Putting the IS Back into IS Research, *Inf Syst J*, 2021;32(3), DOI: 10.1111/isj.12368.

[26] Trocin C, Mikalef P, Papamitsiou Z, et al. Responsible AI for digital health: a synthesis and a research agenda. *Inf Syst Front*, 2021;1-19.

[27] Angelopoulos S, Brown M, McAuley D, et al. Stewardship of personal data on social networking sites. *Int J Inf Manage*, 2021;56, 102208.