Alexandria Engineering Journal (2023) 63, 211–221

HOSTED BY

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

An MRI-based deep learning approach for accurate detection of Alzheimer's disease

Marwa EL-Geneedy [a,*], Hossam El-Din Moustafa [a], Fahmi Khalifa [a], Hatem Khater [b], Eman AbdElhalim [a]

[a] Electronics and Communications Engineering Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt
[b] Electrical Department, Faculty of Engineering, Horus University Egypt, New Damietta 34518, Egypt

Received 2 December 2021; revised 25 July 2022; accepted 31 July 2022
Available online 11 August 2022

KEYWORDS

Alzheimer's Disease (AD);
Deep Learning;
Neurodegenerative;
Transfer learning;
Brain MRI

Abstract Alzheimer's disease (AD) is the most prevalent type of dementia of the nervous system that causes many brain functions to weaken (eg, memory loss). Non-invasive early diagnosis of AD has attracted a lot of research attention nowadays as early diagnosis is the most important factor in improving patient care and treatment results. This research develops a deep learning-based pipeline for accurate diagnosis and stratification of AD stages. The proposed analysis pipeline utilizes shallow Convolutional Neural Network (CNN) architecture and 2D T1-weighted Magnetic Resonance (MR) brain images. The proposed pipeline not only introduces a fast and accurate AD diagnosis module but also provides a global classification (i.e., normal vs. Mild Cognitive Impairment (MCI) vs. AD) as well as local classification. The latter deals with an even more challenging task to stratify MCI into a Very Mild Dementia (VMD), mild dementia (MD), and Moderate Dementia (MoD) as the prodromal AD stage. In addition, we compare our approach to cutting-edge deep learning architectures, e.g., DenseNet121, ResNet50, VGG 16, EfficientNetB7, and InceptionV3. The reported results documented the high accuracy and the suggested method's resilience, as evidenced by the overall testing accuracy of 99.68%.

1. Introduction

Alzheimer's disease (AD) is a tough brain ailment that affects the elderly. It is the world's fourth leading cause of mortality, behind heart disease, cancers, and brain hemorrhage [1]. Around 50 million individuals worldwide suffer from dementia, with almost 60% of them residing in low- and middle-income nations. Dementia affects 5–8% of the general population aged 60 and up at any one moment [1]. Dementia is expected to affect 82 million individuals in 2030, rising to 152 million by 2050. The increased number of dementia patients in low- and middle-income nations accounts for a significant portion of this growth[1]. Nearly 10 million new cases are reported every year. In terms of neurology, AD is a

* Corresponding author.
E-mail address: melgeneedy@std.mans.edu.eg (M. EL-Geneedy).
Peer review under responsibility of Faculty of Engineering, Alexandria University.

[1] https://www.who.int/

https://doi.org/10.1016/j.aej.2022.07.062
1110-0168

long-term neurodegenerative illness that causes tissue loss and nerve cell death all across the brain. This gradually results in deterioration of the patient's cognition and memory often referred to as senile dementia. Furthermore, Alzheimer's disease has a negative influence on patients' ability to conduct everyday activities (e.g., handwriting, talking, and trying to read), and also issues identifying friends and family. AD has three progression stages: early, moderate cognitive, and late stage.Patients in the intermediate cognitive stage respond aggressively, and those in the late stage have heart failure and death-causing respiratory system malfunction [2]. Generally, a dementia diagnosis is a difficult task [3]. Routine clinical diagnosis for AD disease follow-up is necessary at three separate stages, and can be carried out by (i) Consulting the general practitioner, (ii) Carrying out rational neuropsychiatric assessments, and (iii) Taking magnetic resonance imaging (MRI) or positron emission tomography (PET) scans. According to the statistics on AD, it was found that 66% of people with dementia have AD, and only 10% of them were diagnosed early, and 90% were not diagnosed in the early stages. Thus, prompt diagnosis of AD is beneficial to mankind as it paves the way for early intervention and development of health care programs, protective precautions, and affordable interventions. Late diagnosis, on the other hand, often discovers patients at too late a stage to benefit from traditional treatment, and existing therapeutic techniques have a high incidence of misdiagnosis. In addition, late diagnosis makes it difficult to cure the disease by addressing amyloid fibrils.

Noninvasive image-based Computer-Aided Diagnostic (CAD) systems, combined with recent advances in Deep Learning (DL), for early diagnosis of diseases have revolutionized performance and demonstrated great clinical benefits over the years. In particular, MRI-based non-invasive analysis for brain diseases provides rich clinical diagnosis and biomedical research knowledge. In recent years, MRI-based CAD systems have shown high potential for the identification of AD subjects from standard controls for the elderly [4]. The structural MRI (sMRI) neuroimaging technique allows for the detection of brain damage (atrophy, tumours, and lesions) and can help rule out alternative causes of dementia other than AD [5]. The utilization of DL, especially CNN, has seen exponential growth in the field of medical imaging diagnosis [6]. CNN uses the 2D or 3D images directly as the input and automatically learns meaningful higher-level local and global features, thus eliminating multiple measurement errors induced by the conventional hand-crafted feature.

Our current work on the structure convolutional neural network (SCNN) is being used to enhance the identification and categorization of several stages of Alzheimer's disease, ranging from no dementia to mild AD. The suggested technique's main goal is to lessen the reliance on huge datasets. We used the OASIS repository to obtain 2-D representations of the human brain dataset and outperformed state-of-the-art performance on small MRI images. The proposed pipeline offers significant improvements and contributions to AD early prediction. This paper's primary contributions are as follows:

1. We propose a supervised DL method to predict AD using various pre-trained CNN models.
2. We provided an efficient model for dealing with data shortages in an unbalanced dataset.
3. We performed extensive experiments and compared our method for AD prediction with several current methods. According to the results of the experiments, our technique outperforms those methods.
4. We built a regularised model that learns from a small dataset but still outperforms other models in diagnosing AD.

2. Related Work

In recent years, a great deal of research work for early AD diagnosis or prognosis has been developed [7]. Recent breakthroughs in machine learning (ML) and deep learning (DL) techniques have backed this notion. For example, a DL-based approach to classifying Alzheimer's brains and healthy brains was proposed by Zhang et al. [1]. In their work, CNN, which is one of the DL network architectures is utilized to produce a trained and predictive model. Their proposed strategy,for the diagnosis of AD, MCI, and its early stages, included a stacked autoencoder, a regression layer of softmax, and minimal labelled learning samples, and thus less prior experience. The methodology is evaluated using neuroimaging data from 311 ADNI participants, including 65 AD subjects, 67 converter MCI subjects, 102 non-converter MCI (ncMCI) subjects, and 77 normal control (NC) subjects. The results indicated that binary classification accuracy was 88.58% utilising both MRI and Positron Emission Tomography (PET) images, while 4-class classification accuracy was 47.42%.

A hybrid multi-class DL framework for early diagnosis of AD was proposed by Bhatkoti and Paul [7]. To find passively degraded brain regions, they used an improved k-Sparse autoencoder (KSA) classification. Experiments were conducted using 150 images for MRI scans as well as CSF and PET images from the ADNI study. The reported results showed that modified KSA outperformed zero-masking strategy and traditional KSA in terms of overall accuracy. Enhanced pair predictions with 100 classifier combinations had an accuracy of 83.143%, compared to 71.327% for 50 classifier combinations. A 3D-CNN framework for AD diagnosis that utilized both MRI and PET images was presented by Chiyu et al. [8]. To improve the results of their method, they used FSBi-LSTM to completely stack bidirectional long short-term memory on the concealed spatial information from deep feature maps. The approach was tested using the ADNI dataset from the AD Neuroimaging Initiative. Classifying AD from NC, pMCI from NC, and sMCI from NC had diagnosis accuracy results of 94.82%, 86.36%, and 65.35%, respectively. MCI is the earlier state of AD, which is divided into two types: progressive (pMCI) and stable (sMCI) (i.e., sMCI). Islam et al. [9] proposed a deep CNN-based pipeline in an effort to identify AD and classify its stages. Their technique is made up of three deep CNNs, each with a little different structure. Technique validation was performed using the OASIS and the accuracy, precision, recall, and F1-score were 93.18%, 94%, 93%, and 92%, respectively. However, the validation database has only 416 sMRI data. More recently, a multi-model deep CNN framework for automatic hippocampus segmentation and classification of AD was developed by Liu et al. [10]. First, a deep CNN model was used for hippocampal segmentation. Then, using the segmented hippocampus area as a starting

point, a 3D DenseNet was created to learn discriminating image features for disease categorization. The evaluation dataset consisted of T1-weighted sMRI data from the ADNI database (97 AD, 233 MCI, and 119 NC subjects). They found that for categorising AD vs. NC participants, their method had an accuracy of 88.9%, an area under the receiver operating characteristics (ROC) curve (AUC) of 92.5%, and an accuracy of 76.2%; and an AUC of 77.5% for detecting MCI vs. NC subjects. Neffati et al. [11] designed an AD classification algorithm based on ML. The foundation of the system is Downsized Kernel Principal Component Analysis (DKPCA) and multiclass Support Vector Machine (SVM). DKPCA validation was performed utilizing generated data to evaluate its dimension reduction efficacy. Then, the OASIS MRI database was used for in vivo evaluation and the DKPCA based technique was compared against other conventional AD classification approaches. A functional MRI (fMRI) study by Sarraf et al. [12] was proposed for the classification of AD. The technique is based on the LeNet DL architecture and utilizes transfer learning. With 96.86%, their technique effectively categorised the AD samples from NC. Despite the high accuracy, their method can only classify AD and NC, but not MCI. To evaluate MRI images to identify AD, an approach based on a more suitable model of the CNN model known as the Inception V3 model was presented by Cui, Zhenyu, et al. [13]. To improve the focus accuracy of the Inception V3 model, this strategy used three more appropriate Inception blocks. The more appropriate Inception V3 mannequin classified patients with ordinary MRI, mild cognitive impairment with an accuracy of 85.7% using a complete set of 662 3D brain MRI images for training and testing. To categorize MRI images and identify patients with AD, MCI, and NC, a technique based on an ensemble of CNNs was suggested by E. Jabason [14]. DenseNet121, DenseNet169, DenseNet201, and ResNet50 were among the CNN architectures used in their approach. The Open Access Series of Imaging Studies (OASIS) data-set was used to evaluate this ensemble using structural MRI. The ensemble method has a 95.23%, according to the results. Shahbaz et al. [15] constructed a traditional ML framework for AD diagnosis using the ADNI dataset. Their study compared six different ML and data mining algorithms to find the most distinguishing feature for AD staging. The latter, include k-Nearest Neighbors (kNN), Decision Tree (DT), rule induction, Naive Bayes, generalized linear model (GLM), and deep learning algorithms. All six classifiers were subjected to a 10-fold cross-validation procedure, and the results showed that the GLM can effectively diagnose the stages of AD with an accuracy of 88.24% on the test dataset. However, the accuracy was not very high, and the work represents a comparative study of ML techniques. A computational approach by Tran et al. [16] was presented for the diagnosis of AD using 3D brain MRI. Their method starts with the tissue segmentation stage, which integrates a CNN module and a Gaussian Mixture Model (GMM). In the second stage, a hybrid classification model is proposed that combines both extreme gradient boosting (XGBoost) and SVM to classify AD depending on the segmented tissues. The evaluation was conducted on both the AD-86 and AD-126 datasets, achieving a Dice of 0.96 for segmentation in both datasets and classification accuracies of 0.88 and 0.80, respectively. Their method's key benefit is the combination of the two classifiers, which helps to overcome each limit. The AD-126 dataset, on the other hand, is made up of difficult-to-classify MR brain images of elderly individuals. This is mainly because of the anatomical adjustments and abnormalities due to normal ageing and disease, respectively. As a result, the AD-126 group's classification accuracy was lower than the AD-86 group's.

The focus of this work is the development of a DL-based framework for the diagnosis of these stages of AD disease. CNN is the most widely used and represented technique in most research work when it comes to brain image processing and analysis. This paper introduces a mechanism that processes data from these accelerometers and a CNN approach for determining the stage of AD based on the patient's mobility patterns. The proposed algorithm uses sequential steps, beginning with the preparation and preprocessing of data sets. This is followed by CNN model building, design, and parameter setting. Finally, the measurement and performance of the model are conducted and compared against other state-of-the-art methods. Due to the similarity of MRI scans between AD and healthy individuals, detecting AD is difficult. Although several studies looked towards AD diagnosis using MRI scans, they concentrated on improving and altering multiple CNN architectures or ensembles of CNN models to provide high-accuracy AD diagnostic predictions. Our method focuses on highlighting the structural similarity of AD image classes (i.e., Non-Demented (ND), Very Mild Demented (VMD), Mild Demented (MD), and Moderated Demented (MoD)) while maximising the variance between classes to achieve robust and accurate predictions for AD diagnosis.

3. Materials and Methods

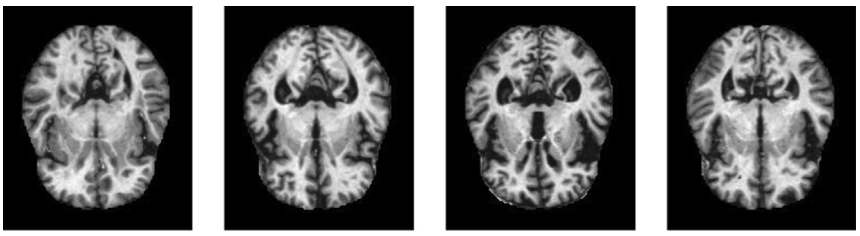
3.1. Patients

This is an experimental research project analyzing publicly available T1-weighted cross-sectional MR brain scans. Both training and testing MR brain images Open Access Series of Imaging Studies (OASIS) provided the data for this study [2] database. OASIS-3 is the most recent version of the OASIS, which aims to make neuroimaging datasets freely available to scientists. For normal ageing and AD, OASIS-3 is a longitudinal neuroimaging, clinical, cognitive, and biomarker dataset. The dataset contains a total of 6400 images and is divided into 4 classes according to the severity of Alzheimer's. Specifically, ND, MoD, MD, and VMD of different patients. The number of images for each class is 3200, 64, 896, and 2240 for ND, MoD, MD, and VMD, respectively. The respective numbers of cases in each category used for testing and training and examples of MR images for the four groups are shown in Fig. 1.

3.2. Preprocssing and Data Augmentation

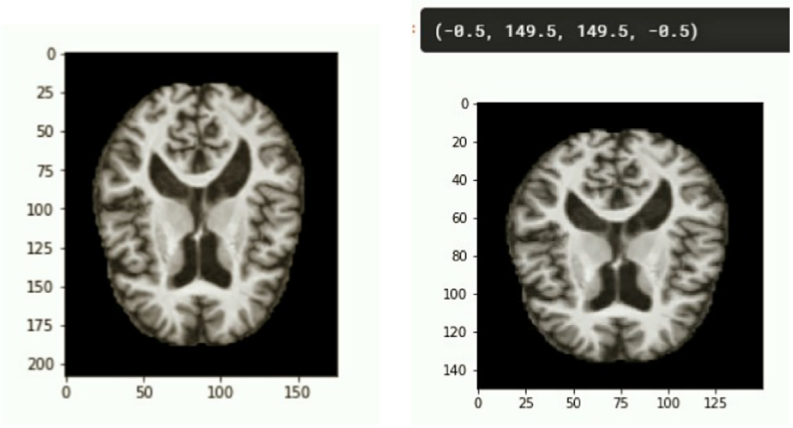
MRI images are deteriorated during data collection, such as low variation owing to the optical equipment's inadequate brightness. Image improvement techniques are usually utilized to correct or enhance the distribution of pixels over a large variety of intensities to solve this problem for the enhancement of MRI scans. Thus, we initially employed image normalization to adjust the image pixel intensity values by reducing

[2] https://www.oasis-brains.org/



(a) (b) (c) (d)

Fig. 1 MR image examples for different Dementia stages: (a) None Demented(ND), (b) Moderated Demented (MoD), (c) Mild Demented (MD), and (d) Very Mild Demented (VmD), from left to right, respectively.



(a) (b)

Fig. 2 Example of an MR image rescalling and resize: (a) before and (b) after resizing.

machinery and impulse noise. For image normalization, the pixel values are rescaled to [-1,1] using a pixel-wise multiplication factor of 0/255 as follows:

$$I\hat{N} = \left(1 - \hat{I}^{T}_{Min}\right) + \frac{\hat{I}^{T}_{Max} - \hat{I}^{T}_{Min}}{Max - Min}\hat{I}^{T}_{Min} \quad (1)$$

where I and $I\hat{N}$ represent input and normalized brain image, respectively, $\hat{I}^{T}_{Max}, \hat{I}^{T}_{Min}$ are the normalised image's intensity range, and Min = 0; and Max = 255 represent the input brain image's pixel intensity range. After normalisation, the grayscale images are resized to 150x150 pixels before training to match the dimensions of the pre-trained model's input layer. Additional channels were established by duplicating the pixel values three times because the pictures were grayscale. The sample MRI images before and after resize and after are shown in Fig. 2.

For a reliable DL model, more training data should be used to avoid model overfitting and provide a more generalized model. However, access to large datasets is troublesome in medical research because of privacy concerns [2]. In particular, for AD, the availability of a large number of scans is a major problem in neuroimaging research. Additionally, a small imbalanced dataset generates overfitting problems that affect the efficiency of the model. Thus, data augmentation is usually employed to overcome data availability and class imbalance problems. In our pipeline, for each accessible MRI image, we employed the augmentation approach to produce new images. Fig. 3 shows samples of data augmentation images. We tried to use more data augmentation techniques such as brightness, zoom, and rotation range, but it was not helpful to our proposed model and led to a decrease in the results. Therefore, we sufficed only with horizontal flipping. All preprocessing steps were conducted in Python (Python 2.3; Python Software Foundation, using libraries Keras (https://keras.io/) and Scikit-learn (https://scikit-learn.org/stable/) and using the TensorFlow backend.

3.2.1. Proposed Framework

The suggested framework's analytical flowchart is shown in Fig. 4 and consists of multiple processing stages. The first stage is the preprocessing and data augmentation. This is followed by the developed CNN architecture and its details, i.e., transfer learning, model training, and parameter setting, and finally the classification. Next, the details of those stages are fully described.

3.3. Proposed Deep Learning Architecture

The suggested DL architecture is depicted in Fig. 5. After preprocessing, the images are 150 x 150 matrix size and were

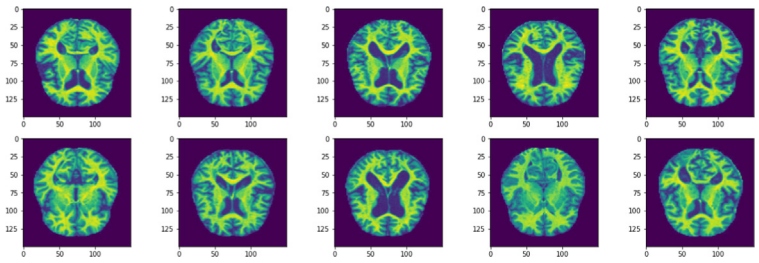


Fig. 3 Samples of data augmentation images.

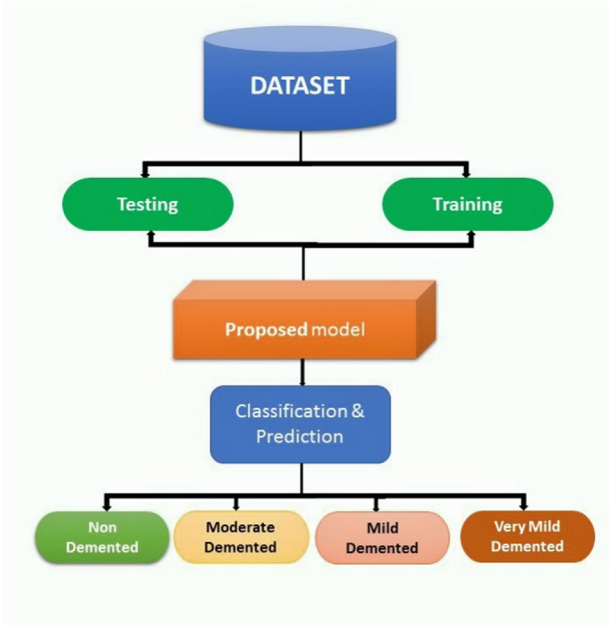


Fig. 4 Flowchart representation for the proposed analysis framework.

supplied to the proposed CNN for training and testing. The proposed CNN model consists of an input, an output, and multiple designed layers. In particular, five convolutional layers in two dimensions (2D) were employed in this study, each of which included a 2D maxpooling layer. Convolution is a linear operation between the input and a kernel (or filter) that acts as a function detector. The filters are taught to extract certain information from an image and have a narrow receptive field. This is how the convolution layer is designated:

$$\mathbf{x}_n^r = \alpha\left(\sum_{m=1}^{k}\mathbf{x}_m^{r-1} * \mathbf{w}_{mn}^r + \mathbf{b}_m^r\right) \tag{2}$$

where, $\mathbf{x}_n^r$ represents n^{th} activation map of the current (r^{th}) layer, $\mathbf{x}_m^{r-1}$ is the n^{th} activation map of the previous layer $(r-1)^{th}$, and k is the number of input activation maps. $\mathbf{w}_{mn}^r$ and $\mathbf{b}_m^r$ are weight and bias vectors. The $*$ operator is used for convolution operation and α denotes the activation function.

The produced activation maps are then sent to the pooling layer after applying the activation function to each activation map. By lowering the resolution of the activation maps, the pooling layer offers translation invariance. The $d \times d$ (e.g. $d = 2$) window of activation maps in convolution layer generates pooling layer activations value.Max pooling is the most commonly used pooling method. Finally, the completely linked layer creates a categorization map using the information from all of the previous layer's activation maps. The optimizer plays a significant role in training the deep CNN model by iteratively changing the parameters of all the layers in the network.

To minimize the loss function $L(\Theta)$, the parameters are updated in the reverse direction of the gradient of the loss function (i.e., $\nabla_\Theta L(\Theta)$) with respect to the parameters. At each iteration, the desired output and predicted output are compared, and the error is back-propagated. One of the most popular performance measurement metrics is cross-entropy. The cross-entropy value is near to zero when the desired output and predicted output are exactly the same, and this is the main aim of any optimization technique.

As CNN's run internally with convolutions in multiple sliding windows, these models will locally distinguish patterns and thus allow a stronger distinction between which each class is represented. The activation function of each convolutional layer is linear rectification (Rectified Linear Unit or ReLU) and the dropout layer has been set as 0.25 for the first and

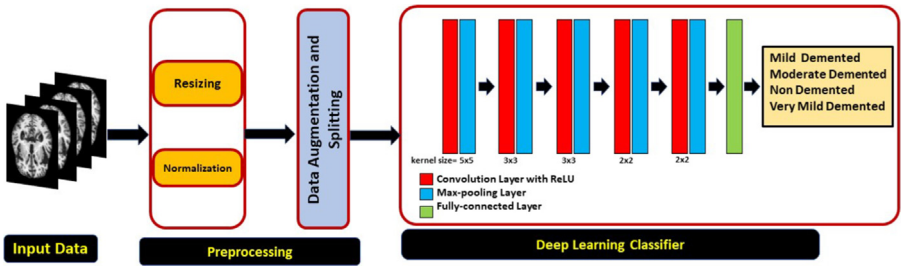


Fig. 5 The CNN model architecture used for AD diagnosis. Note that, ReLU stands for rectified linear unit.

second convolutional layers and it's been set as 0.3 for the following convolutional layers.The model can extract patterns from the input data, and deliver it to the next layers. The output of the previous convolution is then flattened and fed into the last two layers: a fully connected (FC) layer with 0.2 dropout and a softmax layer with four neurons. This layer of the network is in charge of categorization by calculating the likelihood that the input supplied belongs to a specific label. With pooling and increasing filter sizes, this type of multi-layer structure has proven effective for time-series analysis. [17,18].

The model's input is defined as $x = x_1, x_2, \ldots.., x_n$, and the output sequences is indicated as $y = y_1, y_2, \ldots.., y_m$. The output of the network's final layer was modified using the cost matrix (xi). If y is the individual model's output, $(\mathbb{C})$ is the desired class, and (L) is the loss function, (∂^i) is the modified output as follows:

$$\partial^i = Ł\left(\zeta_{\mathbb{C}}, y^i\right), : \partial^i_{\mathbb{C}} \geqslant \partial^i_j \forall j \neq \mathbb{C} \tag{3}$$

The loss function is modified as:

$$Ł = -\sum_n t_n log(\partial_n) \tag{4}$$

where ∂_n incorporates the class-dependent cost (ξ) and is related to the output on (y_n)via the softmax function [19]:

$$\partial_n = \frac{\xi_{\mathbb{C},n} exp(y_n)}{\sum_k \xi_{\mathbb{C},k} exp(y_k)} \tag{5}$$

The weight of a class is determined by the number of samples it contains. The goal is to make one sample of class p as important as t samples of class η, if class η contains t times more samples than class p. As a result, the class weight of p is t times greater than that of η.

In our model as shown in Fig. 5, we use 2D convolutional layers with a kernel size 5×5 for the first block,and 3×3 for the second and third blocks. Also, we use 2×2 for the last two blocks. For all blocks, the ReLU activation function is used and no Batch Normalization (BN) was used. Each block's second convolutional layer used a stride of two to conduct downsampling. The initial block had 64 filters, and each succeeding block had double the number of filters. Following the last convolutional layer, a dropout layer (p = 0.3) was added before being linked to one FC dense layer with ReLU activation values of 1024. Between those thick layers, a dropout layer (p = 0.3) was also placed. Finally, a multi-dense neuron with softmax activation supplied the model output. The training was carried out for a maximum of 100 epochs at a learning rate of 0.001, using the Adam optimizer to estimate model parameters and a batch size of 40, and utilising the categorical-cross-entropy loss function, which is frequently used for multiclass classification issues, The categorical-cross entropy is defined as $H(p,q) = -\sum_x p(x) log(q(x))$, where p is the true distribution and q is the computed distribution. Table 1 summarises all the parameters of the proposed deep learning pipeline.

4. Experimental Results

Detailed information is provided in this section about the experimental setting and the findings. The experimental configuration involves the model and programme platform training knowledge used in the current analysis. The network is trained from data for 100 epochs, with each epoch consisting of 40 batches. All experiments are performed by splitting data into 20%–80% for network testing and training, respectively. From the training set, 16% of the data is used as a validation set. Implementation is done using the Keras framework. Parallel processing is required to train deep neural networks. Thus, we made use of the open-source package python 3.0 and Kaggle to execute out the classifier's training and validation (GPU: NVIDIA TESLA P100 GPUs, 16 GB RAM). To create our suggested model, we used the Keras library from Tensorflow modules, and the run time was 760.3 s.

The purpose of the model evaluation is to help decide how well a given data model generalizes to new data so that we can distinguish between different models. To this end, we need not only an estimation technique such as train-test break or cross-validation but also a metric calculation to determine

Table 1 Summary of our proposed system parameter settings.

Layer	Convolution	Maxpooling	Dropout
First layer	filters = 64, kernel_size = (5,5), padding = 'Same', activation = 'relu'	pool_size = (2,2)	(0.25)
Second Layer	filters = 128, kernel_size = (3,3), padding = 'Same', activation = 'relu'	pool_size = (2,2), strides = (2,2)	(0.25)
Third Layer	filters = 128, kernel_size = (3,3), padding = 'Same', activation = 'relu'	pool_size = (2,2), strides = (2,2)	(0.3)
Fourth Layer	filters = 128, kernel_size = (2,2), padding = 'Same', activation = 'relu'	pool_size = (2,2), strides = (2,2)	(0.3)
Fifth Layer	filters = 256, kernel_size = (2,2), padding = 'Same', activation = 'relu'	pool_size = (2,2), strides = (2,2)	(0.3)
Batch Size		40	
Learning Rate		0.001	
Optimizer		Adam	
No. of Epochs		100	
Total Parameters		4,619,524	
Trainable Parameters		4,619,524	
Non-Trainable Parameters		0	

the performance of various models. A basic metric is the classification accuracy (ACC), which calculates how much an instance class is correctly predicted by the model in the validation set. Additional metrics are used, such as sensitivity (SEN) and specificity (SPE), and are defined respectively as:

$$\begin{aligned} \text{ACC} &= \frac{\text{TN+TP}}{\text{TN+TP+FN+FP}} \\ \text{SEN} &= \frac{\text{TP}}{\text{TP+FN}} \\ \text{SPE} &= \frac{\text{TN}}{\text{TN+FP}} \end{aligned} \tag{6}$$

where TP, TN, FN, and FP represent true positive, true negative, false negative, and false-positive, respectively. the proposed model achieves resulted in a sensitivity and specificity of 100%. In addition to quantitative metrics, network performance is qualitatively evaluated using accuracy/loss graphs against the number of epochs. In the training phase, loss of training and testing and also the loss of validation data were measured as shown in Fig. 6(a). Our proposed CNN model achieved 99.68% accuracy for testing and 100% for training with validation accuracy 99.27% as shown Fig. 7(a).

Another quantitative evaluation is performed using the confusion matrix. The confusion matrix of the proposed systems is shown in Fig. 8 (a). This is very helpful to examine which classes, if any, are being misclassified more. Confusion matrices are not just useful in model evaluation but also in model monitoring and model management. Additional

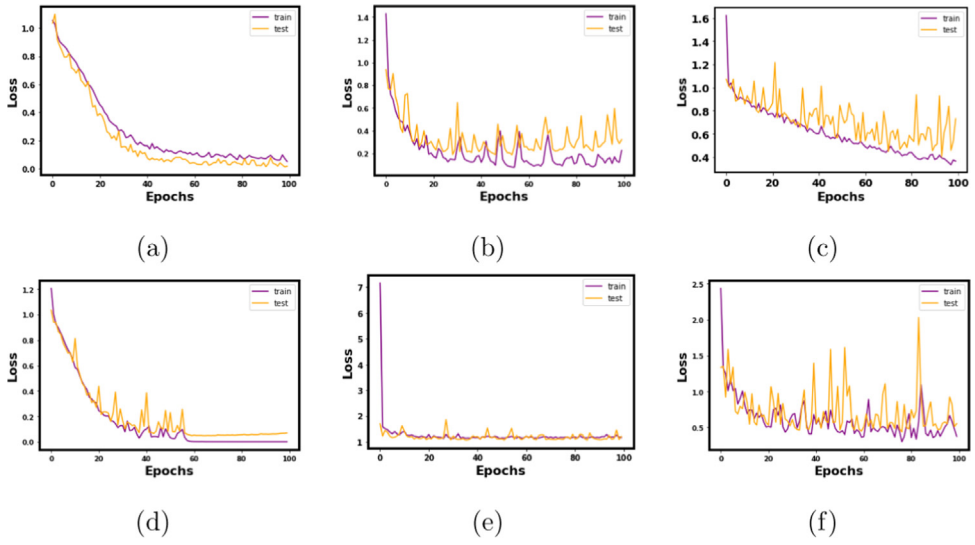


Fig. 6 Training and validation loss against the number of epochs for (a) proposed model (b) DenseNet121 (c) Resnet50 (d) VGG 16 (e) EfficientNetB7, and (f) InceptionV3 models.

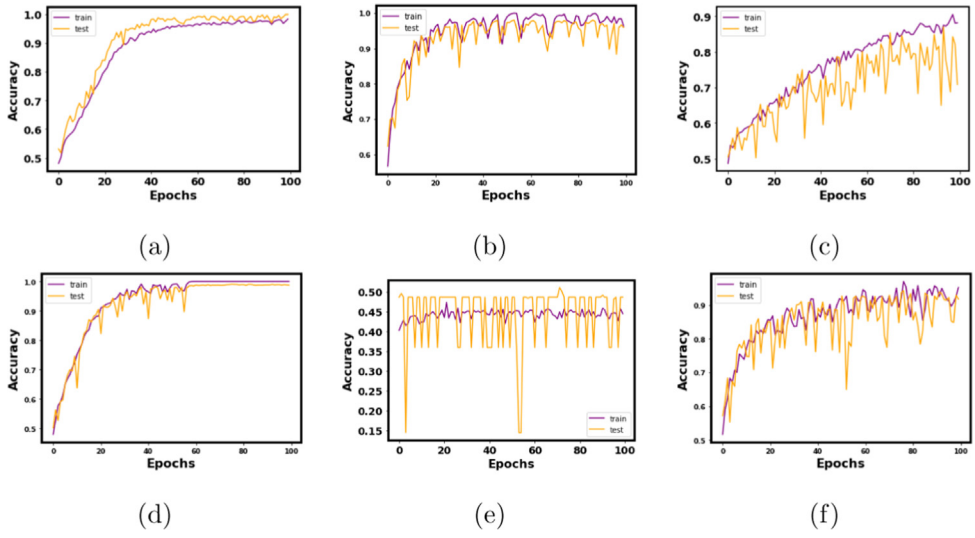


Fig. 7 Training and validation accuracy against the number of epochs for (a) proposed model (b) DenseNet121 (c) Resnet50 (d) VGG 16 (e) EfficientNetB7, and (f) InceptionV3 models.

indices can be measured from the generated confusion matrix that includes the precision, F1-score, and recall. Generally, both the confusion matrix and related metrics are used together for the assessment/evaluation of classification models.

To highlight the advantages of the proposed CNN model, we compared its performance against some of the classification work in brain MR images using CNN models and other ML classifiers. Table 2 below compares the categorization accuracy of our proposed model to that of other STOA results of MRI

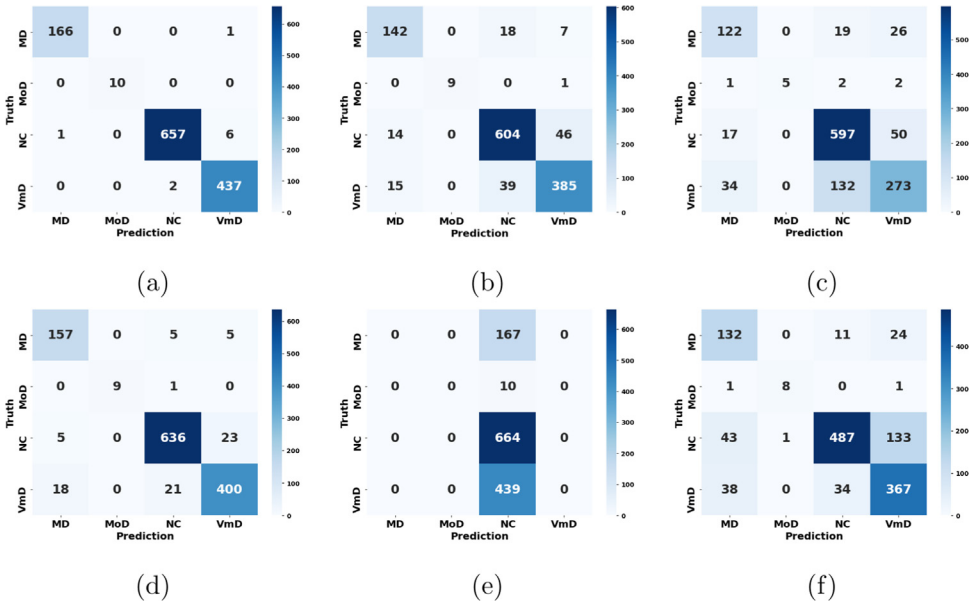


Fig. 8 Confusion matrices for the (a) proposed (b) DenseNet12 (c) ResNet50 (d) VGG 16 (e) EfficientNetB7, and (f) InceptionV3 prediction models.

Table 2 Comparative accuracy of the proposed pipeline with other state-of-the-art literature techniques. Here, "ACC", "SEN", "SPE", "AUC", stand for Accuracy, Sensitivity, Specificity, and Area Under the Curve, respectively.

Methods	Labels	Data Source	ACC	SEN	SPE	AUC
Zhang et al. [1]	MCI vs. AD vs. NC	ADNI	88.58%	-	-	-
Chiyu et al. [8]	AD vs. NC vs. s-MCI[3]	ADNI	94.82%	-	-	-
Liu et al. [10]	MCI vs. AD vs. NC	ADNI	88.9%	-	90.8%	92.5%
Zhenyu et al. [13]	MCI vs. AD vs.NC	ADNI	85.7%	100%	93%	-
Sarraf et al. [12]	AD vs. NC	ADNI	96%	97.39%	84.27%	-
E. Jabason et al. [14]	AD vs. NC vs. MCI	OASIS	95%	-	-	-
Proposed	NC vs. VmD vs. MD vs. MoD	OASIS	**99.68%**	**100%**	**100%**	**100%**

[3] s-MCI: stable- Mild Cognitive Impairment

Table 3 Proposed Model evaluation against other CNN that employs transfer learning. Here, "SEN", "SPE", "AUC", and "NA" stand for Sensitivity, Specificity, Area Under the Curve, and Not Applicable, respectively.

Model	Train Accuracy	Test Accuracy	SEN	SPE	AUC
VGG 16	100	96.39	99.29	100	0.9988
DenseNet121	98.54	96.29	100	100	0.9981
Resnet50	92.46	89.85	100	100	0.9869
InceptionV3	91.38	87.71	99.12	100	0.9804
EfficientNetB7	49.78	48.68	NA	NA	0.7922
Proposed	100	99.68	100	100	1.00

neuroimaging CNN classification. As readily seen by the comparison in the table, we have reached the highest accuracy rate compared to previous work. It is worth mentioning that previous research work used deep-learning models to handle various multistage AD predictions, i.e., AD vs. MCI, or NC vs. MCI, or AD vs. NC vs. MCI. In addition to that, our method deals with even a more challenging task, i.e., categorizing MCI into VMD, MD, and MoD. In addition, we apply the t-test to determine the significance of our findings. With a p-value of.00001, our results indicate statistical significance once more. In addition, we compared the proposed system against other CNNs using transfer learning models. In particular, we compared our method against the VGG16, ResNet50, Inception V3, EfficientNetB7, and DenseNet121 architectures. The compared nets are pre-trained on the ImageNet dataset and were used as pre-trained. We compared the results in Table 3. The

Table 4 Accuracy achieved from CNN across 6 runs.

Run1	Run2	Run3	Run4	Run5	Run6
96.84%	98.53%	98.73%	99.02%	99.22%	99.68%

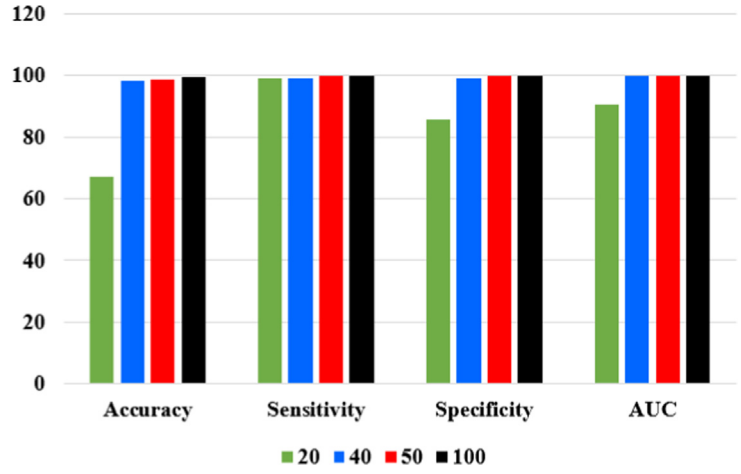


Fig. 9 Performance of proposed framework across different numbers of epochs.

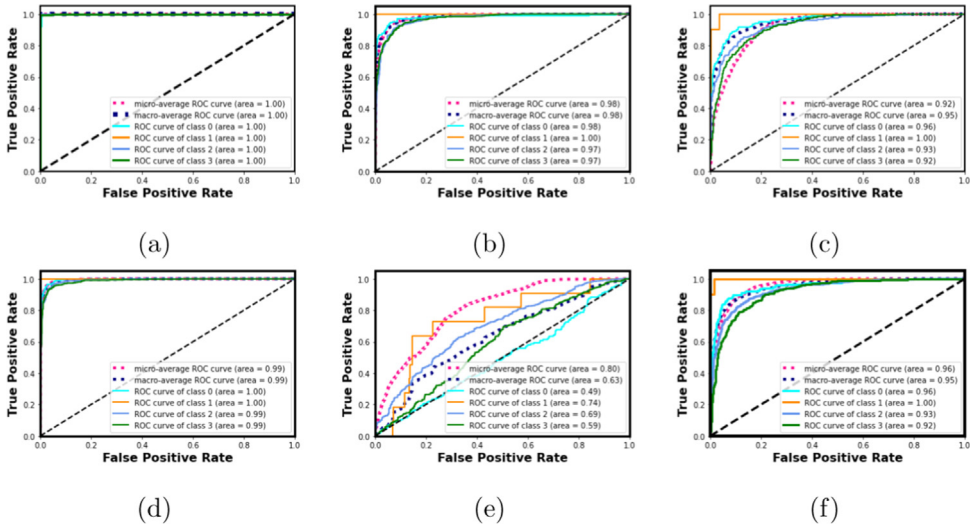


Fig. 10 Receiver Operating Characteristic Curve (ROC) for the (a) proposed (b) DenseNet12 (c) ResNet50 (d) VGG 16 (e) EfficientNetB7, and (f) InceptionV3.

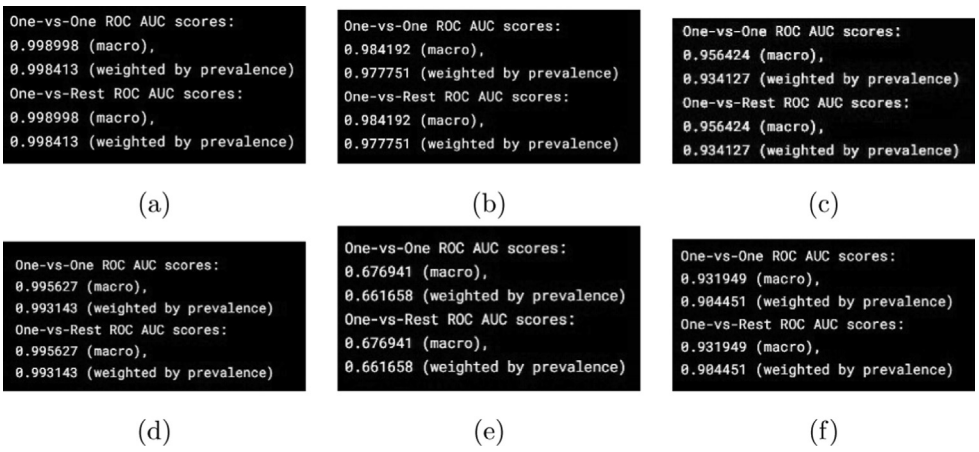


Fig. 11 One-vs-One ROC AUC scores for the (a) proposed (b) DenseNet12 (c) InceptionV3 (d) VGG 16 (e) EfficientNetB7, and (f) ResNet50.

best model is VGG 16 which, achieved a high accuracy of 96.39% with the same hyperparameters we ues to learn our network. The loss, systems accuracy, and confusion matrices are shown in Figs. 6–8, respectively, for the different transfer learning methods in Table 3.

To achieve the robustness and reproducibility of the deep neural network, the changes in hyperparameters for DNN to achieve the best classification are conducted. Table 4 shows the changes in the performance of the proposed model. Figs. 9 additionally shows the achievement of the proposed model against the number of epochs. As we can see in the figure, when the epochs are greater than 50, this margin enhancement in overall system performance is significant, so we set the number of epochs for our experiments to 100 epochs.

The highest performing model's training progress curve and ROC plot are given in Fig. 10 for the datasets. The proposed model reaches the maximum level of accuracy for the datasets in 100 epochs, as shown by the training progress graph. Moreover, the ROC plots with a value equal to 100, 100, 100, 100, and 100 supplement the findings that our system is capable of correctly classifying the samples from the datasets with a high true fraction value and a low false-positive rate. We applied the One-vs-One (OvO) and One-vs-Rest (OvR) schemes. Compare every unique pairwise combination of classes in the OvO scheme. The comparison in the OvR scheme is between one class and the rest of the classes. Next, The estimated ROC AUC scores of three datasets utilizing the OvO and OvR schemes are given in Fig. 11. For each model, we show the macro average and weighted by prevalence.

5. Conclusion

This research has resulted in the development of a DNN-based pipeline to successfully identify multi-class Alzheimer's disease using brain MR images. The proposed pipeline demonstrated an accuracy, sensitivity, specificity, and ROC of 99.68%, 100%, and 100%,100%, respectively. Validation and testing were performed using a data set of 6400 brain MRIs. The method's robustness has been verified using ROC analysis, and higher multi-class classification has been confirmed by comparing our framework against well-known CNNs' performance. The higher accuracy of our approach, using the proper selection of the network architecture, suggests its application to predicting different stages of Alzheimer's disease for multiple age groups. We will use advanced data mining algorithms on different datasets to combine them in future work so that the performance and effectiveness of AD prediction can be enhanced at earlier stages using different datasets and stages.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Zhang, Z. Li, B. Zhang, H. Du, B. Wang, X. Zhang, Multi-modal deep learning model for auxiliary diagnosis of alzheimer's disease, Neurocomputing 361 (2019) 185–195.
[2] A. Mehmood, M. Maqsood, M. Bashir, Y. Shuyuan, A deep siamese convolution neural network for multi-class classification of alzheimer disease, Brain Sci. 10 (2) (2020) 84.
[3] S.B. Shree, H. Sheshadri, Diagnosis of alzheimer's disease using naive bayesian classifier, Neural Comput. Appl. 29 (1) (2018) 123–132.
[4] Y. Zhang, Z. Dong, P. Phillips, S. Wang, G. Ji, J. Yang, T.-F. Yuan, Detection of subjects and brain regions related to alzheimer's disease using 3d mri scans based on eigenbrain and machine learning, Front. Comput. Neurosci. 9 (2015) 66.
[5] A. Abrol, M. Bhattarai, A. Fedorov, Y. Du, S. Plis, V. Calhoun, A.D.N. Initiative, et al, Deep residual learning for neuroimaging: An application to predict progression to alzheimer's disease, J. Neurosci. Methods 108701 (2020).
[6] M.A. Ebrahimighahnavieh, S. Luo, R. Chiong, Deep learning to detect alzheimer's disease from neuroimaging: A systematic literature review, Comput. Methods Programs Biomed. 187 (2020) 105242.
[7] M.P. Bhatkoti Pushkar, Early diagnosis of alzheimer's disease: A multi-class deep learning framework with modified k-sparse autoencoder classification.
[8] F. CHIYU, A. ELAZAB, Y. PENG, T. WANG, F. ZHOU, Deep learning framework for alzheimer's disease diagnosis via 3d-cnn and fsbi-lstm.
[9] J. Islam, Y. Zhang, Brain mri analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks, Brain informatics 5 (2) (2018) 1–14.

[10] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, M. Xu, A.D.N. Initiative, et al, A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease, NeuroImage 208 (2020) 116459.
[11] S. Neffati, K. Ben Abdellafou, I. Jaffel, O. Taouali, K. Bouzrara, An improved machine learning technique based on downsized kpca for alzheimer's disease classification, Int. J. Imaging Syst. Technol. 29 (2) (2019) 121–131.
[12] S. Sarraf, G. Tofighi, Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks, arXiv preprint arXiv:1603.08631.
[13] Z. Cui, Z. Gao, J. Leng, T. Zhang, P. Quan, W. Zhao, Alzheimer's disease diagnosis using enhanced inception network based on brain magnetic resonance image, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 2324–2330.
[14] E. Jabason, M.O. Ahmad, M. Swamy, Classification of alzheimer's disease from mri data using an ensemble of hybrid deep convolutional neural networks, 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), IEEE, 2019, pp. 481–484.
[15] M. Shahbaz, S. Ali, A. Guergachi, A. Niazi, A. Umer, Classification of alzheimer's disease using machine learning techniques, DATA (2019) 296–303.
[16] T.A. Tuan, T.B. Pham, J.Y. Kim, J.M.R. Tavares, Alzheimer's diagnosis using deep learning in segmenting and classifying 3d brain mr images, Int. J. Neurosci. (2020) 1–10.
[17] C.A. Ronao, S.-B. Cho, Human activity recognition with smartphone sensors using deep learning neural networks, Exp. Syst. Appl. 59 (2016) 235–244.
[18] Z. Wang, W. Yan, T. Oates, Time series classification from scratch with deep neural networks: A strong baseline, 2017 International joint conference on neural networks (IJCNN), IEEE, 2017, pp. 1578–1585.
[19] S.H. Khan, M. Hayat, M. Bennamoun, F.A. Sohel, R. Togneri, Cost-sensitive learning of deep feature representations from imbalanced data, IEEE Trans. Neural Networks Learn. Syst. 29 (8) (2017) 3573–3587.