Neural Computing and Applications (2024) 36:3777–3801
https://doi.org/10.1007/s00521-023-09301-6

ORIGINAL ARTICLE

A multilayered framework for diagnosis and classification of Alzheimer's disease using transfer learned Alexnet and LSTM

Palak Goyal[1] · Rinkle Rani[1] · Karamjeet Singh[1]

Received: 23 March 2023 / Accepted: 14 November 2023 / Published online: 7 December 2023

Abstract
Alzheimer's disease (AD) is the most frequent type of dementia that has no effective cure, except early discovery and treatment that may help patients to include successful years in patient's lives. Currently, mini-mental state examination (MMSE) score and manual examination of magnetic resource imaging (MRI) scan along with machine learning techniques are used to diagnose the disease; however, they possess certain accuracy limits. Therefore, this paper proposes a deep learning-based multilayered framework for AD classification using transfer learned Alexnet and LSTM for multiclass and binary classification of MR images. However, the deep learning models used in the current study necessitate a large training dataset to produce better outcomes. As a result, this work also utilizes generative adversarial network (GAN) as a data augmentation tool to improve the classification results and further to solve the problem of overfitting. The study uses Alzheimer's disease neuroimaging initiative (ADNI) dataset of 60 AD, 73 mild cognitive impairment (MCI) and 67 cognitively normal (CN) patients from which 2 D MR image scans are extracted. Furthermore, the proposed method achieved the classification accuracy on AD–CN at 98.13%, AD–MCI at 99.38% and CN–MCI at 99.37%, respectively. Also, the multiclass classification shows the promising accuracy of 96.83% for the proposed framework. Finally, the proposed model's performance is compared to other state-of-the-art techniques and the experimental results show that the proposed model outperforms in terms of accuracy, sensitivity and hypothesis testing.

Keywords Alzheimer's disease · Alexnet · Generative adversarial network (GAN) · Long short-term memory (LSTM) · Deep learning · Image classification

1 Introduction

Dementia is a term that has been coined to describe the problems, diseases, and situations that result from the death or abnormal functioning of numerous brain cells. Alzheimer's disease (AD) is the most frequent type of dementia that affects a huge section of the world's population, accounting for 60–80 percent of dementia cases [1–3]. It is a multifaceted neurological brain disease that disrupts brain cells because of the accumulation of various amounts of proteins (Aβ and Tau) in plaques of amyloids and neurofibrillary tangles, which causes memory loss and impairs thinking skills [4, 5]. Moreover, MCI patients are more prone than others to acquire AD [6, 7]. Also, the influence of AD on a patient's brain is both extensive and complex, making it difficult to prevent or diagnose this disease [8].

Alzheimer's disease affects more than 47 million people globally, according to research by Alzheimer's Disease International (ADI) [9]. Furthermore, by 2050, this number will have risen to 152 million individuals, implying that one person will develop dementia every three seconds [9, 10]. Figure 1 shows the analysis for the year 2020 of 5.8 million Americans aged 65 and older in the USA with AD. It is predicted to reach 13.8 million by 2050 [6].

As the disease is spreading on a large scale and further, no definite diagnosis is possible [11, 12]; as a result, a solid

✉ Palak Goyal
pgoyal60_phd18@thapar.edu

Rinkle Rani
raggarwal@thapar.edu

Karamjeet Singh
karamjeet.singh@thapar.edu

[1] Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, Punjab 147001, India

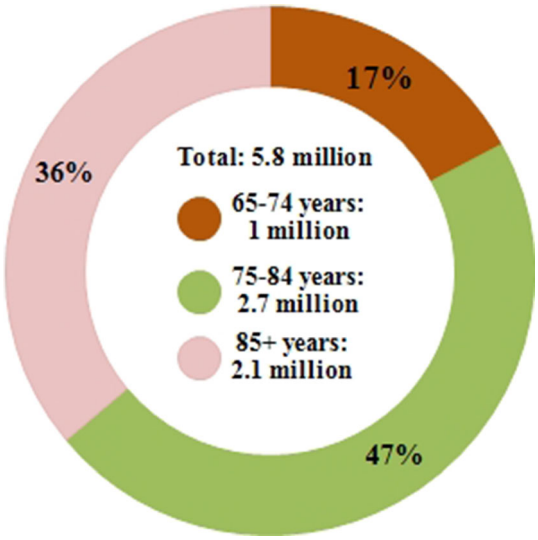

Fig. 1 Age-wise Percentage of people affected with AD in the USA

agreement for the rapid and precise diagnosis of AD is critical since early therapies can slow the progression of the illness, allowing individuals who are affected to live longer [13, 14]. For this, a computer-aided system (CAD) is used to identify Alzheimer's disease in order to reduce expensive care costs, which are projected to skyrocket [15]. Additionally, classic machine learning algorithms for early Alzheimer's disease diagnosis typically use two types of data: ROI-based and voxel-based features [16]. Fundamental assumptions concerning anatomical anomalies in the brain, such as regional cortical thickness and hippocampal volume, are heavily relied upon by these techniques [17, 18].

Also, conventional approaches rely mainly on manual feature extraction, which seems to be a lethargy and subjective process which depends primarily on technical skill and repeated tries. Therefore, deep learning, particularly convolutional neural networks (CNNs) and their variants, offers a viable solution to these issues [19]. CNN can improve efficiency even further and demonstrate excellent performance in AD diagnosis without implementing handmade feature extraction because it does so automatically [20, 21].

This paper proposes a multilayered framework based on deep learning methods and convolution neural networks for AD's early diagnosis and classification. AD is divided into three stages: (i) cognitive normal (CN), (ii) mild cognitive impairment (MCI), and (iii) Alzheimer's disease (AD); thus, there is a need for multiclass classifier [22, 23]. In addition, pairwise image classifications are also implemented between all possible pairs of AD stages. The proposed multilayered framework for AD diagnosis and classification using transfer learned Alexnet and LSTM in this research use convolution layers of pre-trained Alexnet to excerpt the features from 2 D magnetic resource imaging (MRI) brain scans and feed these extracted features into the LSTM layers whose output is used as the input to the transfer learned dense layers of Alexnet.

1.1 Our contribution

The following are the primary contributions of this study:

- A multilayered framework is proposed for the timely identification of AD and multiclass classification of MR images.
- Multiclass AD Classification uses a transfer learned model on pre-trained Alexnet and LSTM. Here, image features are extracted from convolution layers of pre-trained Alexnet, which are then fed into the LSTM layers and further to transfer learned dense and output layers to do the classification.
- The limited size of the dataset is one of the most significant issues for medical images leading to inaccurate results. As a result, a data augmentation technique named GAN is used to increase the dataset's size and avoid the problem of overfitting.
- Three AD stages named AD, MCI, and CN are tested for multiclass as well as binary medical image classification systems.
- The performance evaluation of the proposed model shows promising results in terms of various evaluation metrics such as accuracy, precision, and recall.

The remainder of the paper is organized as follows: Sect. 2 discusses the related research works in this field. Section 3 outlines the background and preliminaries used in this research. Section 4 illustrates how the proposed model works. Section 5 presents the performance evaluation of the model. Finally, Sect. 6 concludes with a conclusion and recommendations for future work.

2 Related work

The identification of AD has been extensively researched, and it entails a number of concerns and challenges, based on which a brief review has been done, from which some of the studies are discussed. Yan et al. [24] developed a deep learning-based method using squeeze and excitation mechanism and pyramid squeeze attention mechanism on the fully convolution network model along with MLP model to classify and diagnose AD. In another study, Zhu et al. [25] developed an advanced deep learning architecture based on self-attention mechanism. This model has the advantage that it integrates representation learning, distilling and classification into the one framework, thus reducing the complexity. Further, Li et al. [26] developed a

method that used 3-D CNN along with multichannel contrastive strategy for AD diagnosis and improving generalization ability due to integration of supervised and unsupervised loss. Divya et al. [27] compared the different feature selection techniques along with different classifiers for selecting the set that produced the best classification results and proved that SVM with radial function kernel along with mini-mental state examination (MMSE) score gave the highest performance. Then, in another study, Kang et al. [28] developed an ensemble approach with 2-D CNNs such as VGG16, ResNet 50 using multimodel and multislice ensemble to classify AD with a limited dataset. Further, Feng et al. [29] proposed a method that used nonsubsampled contourlet sub-band-based networks for feature extraction and the concatenation of node and edge features are used for classification using SVM with radial function. Jain et al. [30] proposed the $P_F S_E C_{TL}$ mathematical model, which is built using transfer learned CNN and VGG-16 inbuilt models. For the classification task, VGG-16 served as a feature extractor. In addition to that, Basheera et al. [31] utilized the segmented gray matter from MR images using enhanced independent component analysis (ICA) and CNN to extract the features and do classification on the extracted features. Choi et al. [32] used deep CNN to identify 139 AD and 182 HC using 3-dimensional PET volumes, with an accuracy of 84 percent. To extract features and incorporate nonlinear features for task-specific classification, Liu et al. [33] used a T1-MRI and FDG-PET to classify with CNN. A 3D CNN was used to extract features, while a 2D CNN was utilized to combine various features. Further, Shi et al. [34] used deep polynomial networks (SDPNs) for classification. Two SDPNs features are learned from MRI and PET data, which are then combined and given to a final stage SDPN, achieving an accuracy of 97%.

Further, Korolev et al. [35] demonstrated that an equivalent result might be achieved, but when the residual network and regular 3D CNN designs were applied to 3D structural images, the results revealed that the diversity of the two networks was very similar but not up to the mark. In addition, Liu et al. [36] retrieved many types of features from the individuals' MRI dataset using multiple selected templates and grouped into tissue density maps. Finally, the subject was classified using an ensemble of support vector machines (SVM). In another study, Zu et al. [37] used regional GM volume and FDG-PET intensity, as well as MKSVM, for the classification of various stages of AD and MCI. Ortiz et al. [38] presented a multiview DBN-RBM to acquire MRI and PET data at the same time. The learned representations were given to a number of rudimentary SVM classifiers, which were then voted together to construct a more powerful, high-level classifier. To improvise the presentation of ROI features, Li et al. [39] developed a dropout method based on a robust deep learning framework for AD/MCI diagnosis. A multimodal technique for extracting neuro-imaging features for AD diagnosis was presented by Liu et al. [40]. The collected characteristics using autoencoders and zero masking method were classified using the SVM classifier, which had an accuracy of 86.86%. Further, a dense encoder and 3D CNNs were employed in the work of Payan et al. [41]. The 3-D auto-encoders had been used to train the convolutional layer, but it was not fine-tuned, due to which the performance may be degraded. To train latent feature representation using ROI characteristics of heterogeneous brain images and cerebrospinal fluid (CSF) features for AD classification, Suk and Shen [42] used SAE in conjunction with the multitask feature selection and multikernel learning (MKL) algorithms. In their another study, they introduced a deep architecture for Alzheimer's disease diagnosis that use sparse multitask learning to choose characteristics in a hierarchical manner [43]. DL has also been used to identify Alzheimer's disease and other brain disorders since it can effectively find implicit or latent representation in neuroimaging data. Suk et al. [44] employed a heterogeneous deep restricted Boltzmann machine (RBM) to train features from 3D data to diagnose AD/MCI. By incorporating CSF biomarkers, regional GM volume, and pixel intensity as features for the binary classification of distinct phases of the disease termed AD, MCI, and CN, Zhu et al. [45] developed a function using SVM. The summary of the research work for AD classification is shown in Table 1.

These are some of the recent studies in the AD field in which several models that can handle AD detection and categorization have recently been proposed in the literature. Most of them used deep learning models such as sparse autoencoders (SPAE) and deep belief networks (DBN), to train the classifier as surveyed in [46]. These classifiers need to be trained from scratch, which increases the complexity of the model. Further, the data size used in most of the studies was limited, introducing the problem of overfitting in the model. Also, most of them do not use transfer learning algorithms, multiclass medical image classification, or a deep learning approach to diagnose AD stages and provide effective treatment to patients. To handle the above issues, the researchers propose various solutions, yet the requirement is not satisfied. Hence, a model based on pre-trained Alexnet and LSTM is proposed in this study to do the early identification and classification of AD and its various stages. The comparison of the proposed work and existing AD classification models is shown in Table 2.

Table 1 Summary of the research work related to AD classification

Author(s)	Modality	Sample size	Region(s) selected	Method(s) used	Strength(s)	Limitation(s)
Yan et al. [24]	MRI	AD: 309 CN: 241	Whole brain	Anisotropic Diffusion Filtering + PSA	Able to distinguish AD patients efficiently and stably	Only AD and NC classes were considered
Zhu et al. [25]	MRI	AD: 319 MCI: 316 CN: 324	Whole brain	Self-attention mechanism	Proposed mechanism able to reduce computational complexity	Performed binary classifications only
Li et al. [26]	MRI	AD: 299 MCI: 299 CN: 330	Whole brain	3-D CNN	Enhanced the generalization ability of the network	Performed binary classifications only
Divya et al. [27]	MRI	AD: 171 MCI: 558 CN: 347	Whole brain	Genetic Algorithm with SVM	Achieved the good accuracy with lesser number of features	More imaging measures may be reviewed
Kang et al. [28]	MRI	AD: 187 MCI: 382 CN: 229	Coronal slices of gray density maps	Ensemble learning framework based on 2-D CNN	Cost-effective and used with the limited dataset	Performed binary classifications only
Feng et al. [29]	MRI	AD: 200 MCI: 280 CN: 200	Regions of Interest	Nonsubsampled Contourlet + SVM	Features extracted had low dimension which reduces the complexity	Sub-band individual networks may be used
Li et al. [28]	MRI, PET	AD: 51 MCI: 99 CN: 52	Regions of Interest	PCA, MDLD, RBM, SVM	Handled the problem of overfitting using the dropout technique	Used limited sized data
Liu et al. [29]	MRI, PET	AD: 80 MCI: 374 CN: 204	Regions of Interest	SAE, SVM, Zeromask	Performed multiclass and multimodal AD diagnosis using deep learning	Used limited sized data
Jain et al. [30]	MRI	AD: 50 CN: 50 MCI: 50	Whole brain	Transfer learned VGG 16	Used Transfer learning to train the classifier	Fine-tuned convolution layers not used
Basheera et al. [31]	MRI	AD = 120 CN = 117 MCI = 112	Segmented Gray Matter using enhanced ICA	CNN	GM tissues segmented and combined with clinical information to train the classifier	
Choi and Jin [32]	MRI, PET	AD: 139 CN: 182 MCI: 171	Whole Brain	MM 3 D CNN	Used Deep CNN to predict future MCI patient outcome	Used limited sized data
Liu et al. [33]	MRI, PET	AD: 93 CN: 100 MCI: 204	Whole brain	2D CNN, 3D CNN	Used deep convolutional learning and data-driven approaches for classification	Hyperparameters not tuned
Shi et al. [34]	MRI, PET	AD: 51 CN: 52 MCI: 99	Whole brain	MM-SDPN, SVM	Performed multimodal neuroimaging-based diagnosis to train the classifiers	CSF features not considered in neuroimaging data
Korolev et al. [35]	MRI	AD = 50 LCI = 43 EMCI = 77NC = 61	Whole brain	3 D CNN based on ResNet and VGGNet	Simplified the MRI classification pipeline using 3D convolution architectures	Preprocessed images increased complexity of convolution layers
Liu et al. [36]	MRI	AD: 97 MCI: 234 CN: 128	Selected Templates, GM	ISML, MVFE, SCFE, SVM	Subclass clustering algorithm and ensemble classification done for feature learning	Multiple templated not registered in a standard space

Table 1 (continued)

Author(s)	Modality	Sample size	Region(s) selected	Method(s) used	Strength(s)	Limitation(s)
Zu et al. [37]	MRI, PET	AD: 51 CN: 52 MCI: 99	GM, WM, CSF	LAMTFS, SVM	Relationship between modalities and subjects considered through discriminative features	Multiclass classification not performed
Ortiz et al. [38]	MRI, PET	AD: 70 CN: 68 MCI: 111	Regions of Interest	DBN, SVM	Ensembled different deep belief networks trained on various brain regions to extract complete information	
Payan et al. [41]	MRI	AD: 755 MCI: 755 CN: 755	Whole brain	SPAE, 3-D CNN	Used 3-D convolutions on the whole brain with deep neural networks	Convolution layers not fine-tuned

Table 2 Feature-based comparison of the proposed and existing AD classification models

Author(s)	Parameters				Image classification pairs			
	Pre-processing done	Data augmentation	Transfer learning	Early stopping mechanism	AD-CN-MCI	AD-CN	AD-MCI	CN-MCI
Yan et al. [24]	✔	✗	✗	✗	✗	✔	✗	✗
Zhu et al. [25]	✗	✗	✗	✗	✗	✔	✗	✔
Jain et al. [30]	✔	✗	✔	✗	✔	✔	✔	✔
Basheera et al. [31]	✔	✗	✗	✗	✔	✔	✔	✔
Shi et al. [34]	✔	✗	✗	✗	✗	✔	✗	✔
Korolev et al. [35]	✗	✗	✔	✗	✔	✗	✗	✗
Ortiz et al. [38]	✔	✗	✗	✗	✗	✔	✔	✔
Li et al. [39]	✔	✗	✗	✗	✗	✔	✔	✔
Liu et al. [40]	✔	✗	✗	✗	✔	✔	✗	✔
Payan et al. [41]	✔	✗	✗	✗	✔	✔	✔	✔
Proposed model	✔	✔	✔	✔	✔	✔	✔	✔

3 Preliminaries and background

3.1 Convolution neural networks

Extraction of features, feature selection, and classification are the three levels of classical machine learning algorithms. In a conventional CNN, all of these phases are integrated. There is no need to do the hand-operated feature extraction method while utilizing CNN. Its initial layer weights are used to extract the features, and iterative learning further improves their values. The basic CNN [47] can be made up of many components depending upon the requirements of a particular problem. Some of the main components used in this research, along with their mathematical operations, are summarized in Table 3. A convolution operation extracts features from an input image of size $m \times n$ and output the activation map after applying the kernel filter of size $f_m \times f_n$, padding 'p' and stride 's' The output activation map is of the size $a_m \times a_n$ where:

$$a_m = 1 + \left(\frac{m - f_m + 2p}{s}\right) \tag{1}$$

$$a_n = 1 + \left(\frac{n - f_n + 2p}{s}\right) \tag{2}$$

If t kernel filters are used for the convolution operation, then the final activation map will be of size $a_m \times a_n \times t$.

Table 3 CNN components

Component	Description	Mathematical operations	Definition of operation symbols
Input layer	Extract the pixel values from the input image This layer acts as the interface between external data and internal computations It doesn't have any learnable parameters, i.e., it just feeds the network with correctly shaped input image	Resizing and normalization	–
Convolution layer	Building blocks of CNN They perform kernel convolution actions on the input and enumerate the output of neurons Extracts features using convolution operations	$o_t^c(m,n) = \sum_r \sum_{u,v} i_r(u,v).k_t^c(x,y)$	$i_r(u,v)$ = Input array $k_t^c(x,y)$ = Convolution kernel of t^{th} layer r = channel index $o_t^c(m,n)$ = Output feature map
Pooling layer	Gathers similar data near the receptive field and produces the dominant response in this area This layer is used to reduce space dimensions keeping important information It also helps in reducing overfitting by lessen the number of parameters and computations in the network	$P_t^c = s_p(O_t^k)$	P_t^c = Pooled feature map of t^{th} layer s_p = Pooling operation O_t^k = Output of convolution
Activation function	Acts as a decision-maker Aids in the recognition of complex patterns It helps to introduce nonlinearity in the network	$A_t^c = s_a(O_t^k)$	A_t^c = Transformed output of t^{th} layer s_p = Activation function O_t^k = Output of convolution
Batch normalization	Optimization Method that handles the issues related to covariance shift Unifies the feature maps to zero mean and unit variance Helps to normalize activations in the layers by introducing regularization	$B_t^c = \frac{O_t^k - \mu_{mb}}{\sqrt{\sigma_{mb}^2 + \varepsilon}}$	B_t^c = Normalized feature map μ_{mb}, σ_{mb}^2 = Mean and variance of feature map for mini-batch O_t^k = Output of convolution
Dropout	Introduces regularization to the network to enhance generalization Bypass some units or connections with a specific probability at random It helps to prevent overfitting in the network	$D_t^c = i_r(u,v)M_r(u,v)$	D_t^c = Output of dropout layer $M_r(u,v)$ = Binary mask
Dense/fully connected layer	Build a nonlinear collection of feature maps to classify the data from the feature extraction stages Analyze the output of all the layers before it	$F_t^c = W * i_r(u,v) + b$	F_t^c = Output of fully connected layer W = Weight matrix b = Bias
Output layer	Predict final class names according to scores Uses activation function or neuron count to perform predictions based on the task	Affine Transformation or Softmax (for Classification)	–

3.2 GAN: data augmentation technique

The augmentation step [48] is a censorious technique for addressing the overfitting problem. Due to the unavailability of a significant number of annotated images, image augmentation [49] plays a crucial role in classification problems. GANs (generative adversarial networks) are generative systems used to generate new artificial images based on two neural networks: a generator and a discriminator.

Here, the two neural networks as shown in Fig. 2 named: (i) generator creates new images from some random noise. The goal is to produce fake data from the input images, (ii) the discriminator uses the input images from the generator and distinguishes between actual and counterfeit data [50]. The following equations represent the mathematical formulas used in GAN in two neural networks:

At generator G:

$$G_{\text{loss}} = \log(1 - D(G(y))) \tag{3}$$

$$\text{Total Cost} = \frac{1}{n}\sum_{k=1}^{n} \log\left(1 - D\left(G\left(y^k\right)\right)\right) \tag{4}$$

At discriminator D:

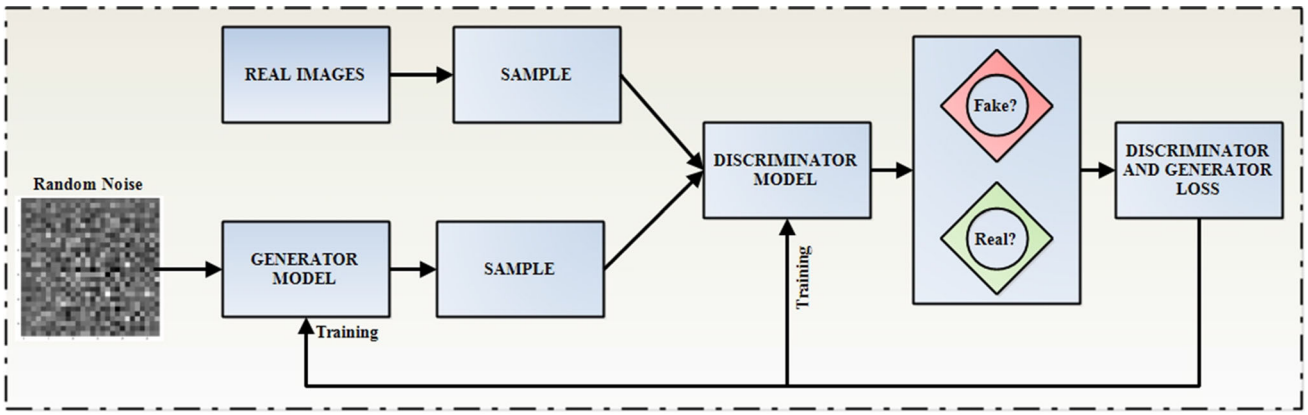

Fig. 2 GAN architecture

$$D_{\text{loss\_fake}} = \log(1 - D(G(y))) \tag{5}$$

$$D_{\text{loss\_real}} = \log(D(r)) \tag{6}$$

$$\text{Total Cost} = \frac{1}{n}\sum_{k=1}^{n} \log\left(D\left(r^k\right)\right) + \log\left(1 - D\left(G\left(y^k\right)\right)\right) \tag{7}$$

Here, y is the random noise image, r is the training image, i.e., real image, $G(y)$ is the generator's output, i.e., fake image, $D(r)$ is the discriminator's output for real image and $D(G(y))$ is the discriminator's output for a fake image. In GAN, our goal is to minimize discriminator's cost and maximize the generator's cost, which is done by updating the cost function at every iteration and, in the end, output the new images for which the optimal values of cost functions are reached.

3.3 Alexnet: CNN model

AlexNet, a deep neural network developed by Alex Krizhevsky and colleagues in 2012, was created to identify images for the ImageNet LSVRC-2010 competition containing 1000 classes to classify [51]. It was also compatible with multiple GPUs. It is an eleven-layered network that can accept an image of size 227 × 227 × 3 [52]. This architecture has 650,000 nodes, 60 million parameters, and 630 million links. The 11 layers of Alexnet as shown in Fig. 3 are: Convolution Layer (96, 11 × 11), Max Pooling Layer (3 × 3), Convolution Layer (256, 5 × 5), Max Pooling Layer (3 × 3), Convolution Layer (384, 3 × 3), Convolution Layer (384, 3 × 3), Convolution Layer (256, 3 × 3), Max Pooling Layer (3 × 3), Fully Connected Layer (4096), Fully Connected Layer (4096), Fully Connected Layer (1000). As the pre-trained Alexnet Model is problem-specific, there is a need to change some of the layers of a pre-trained model to make it convenient according to the problem at hand [23]. Furthermore, CNN requires many MR images to generate and upgrade weights. As a result, changing the weights in the pre-trained model will produce favorable results and speed up convergence [53–55]. The pre-trained Alexnet [56] model built for the ImageNet dataset was used in this model. The five convolution layers and pooling layers, i.e., all the layers before the Flatten layer of pre-trained Alexnet, were used to obtain MR images' features. After the features are extracted, these vectors are flattened and fed into the modified (transferred) dense layers of Alexnet and, finally, to the output layers, which are constructed as per the number of classes to do classification. Furthermore, if there is multiclass classification, the output layer has three neurons with a softmax activation function, whereas, for binary classification, it has two neurons with a sigmoid activation function.

3.4 LSTM: deep learning-based classification model

LSTM networks [64] are a sort of recurrent neural network that may learn order dependency in sequence prediction tasks. Recurrent neural networks are neural networks that repeat themselves. But, RNNs have difficulty with long-term dependencies, which LSTM networks were designed to address. LSTMs are distinguished from more traditional feedforward neural networks by the presence of feedback connections. At any one time, the output of an LSTM is determined by three factors: The network's present long-term memory, often known as the cell state, the previous concealed state is the output at a previous point in time, and at the current time step, the supplied data. The one cell of LSTM containing all the gates is shown in Fig. 4, and its various mathematical operations are given in Eqs. 8–15:

$$f_t = \text{sigm}\left(\left[\left(w_{fh} * h_{t-1}\right) + \left(w_{fh} * x_t\right) + b_f\right]\right) \tag{8}$$

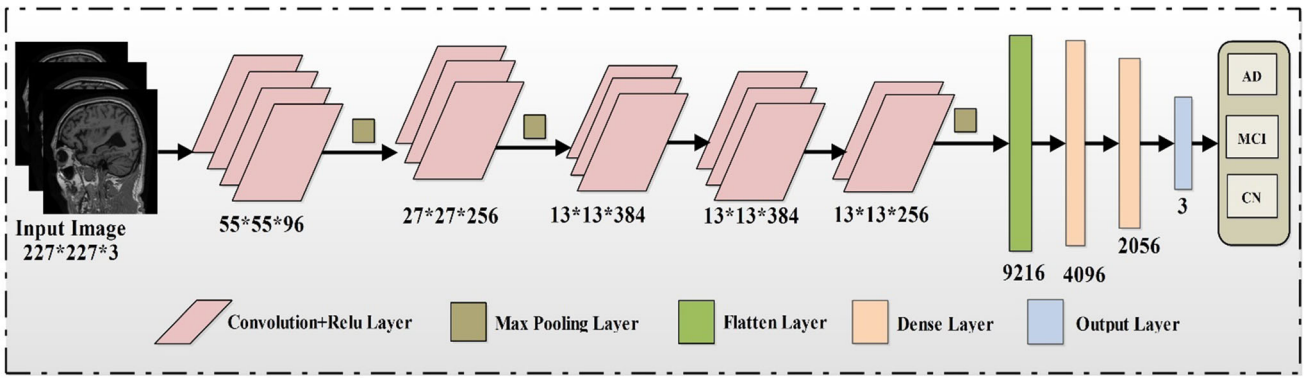

Fig. 3 The Transfer Learned Alexnet Model Architecture

$$C_t^f = C_{t-1} \otimes f_t \tag{9}$$

$$i_t = \text{sigm}([(w_{ih} * h_{t-1}) + (w_{ix} * x_t) + b_i]) \tag{10}$$

$$g_t = \tanh\left(\left[\left(w_{gh} * h_{t-1}\right) + \left(w_{gx} * x_t\right) + b_g\right]\right) \tag{11}$$

$$C_t^i = i_t \otimes g_t \tag{12}$$

$$C_t = C_t^i + C_t^f \tag{13}$$

$$o_t = \text{sigm}([(w_{oh} * h_{t-1}) + (w_{ox} * x_t) + b_o]) \tag{14}$$

$$h_t = \tanh(C_t) \otimes o_t \tag{15}$$

Here, f_t, i_t, g_t, o_t, h_t are the output of forget gate, input gate, input node, output gate, and activation state and C_T^f, C_t^i, C_t is the cell state of these gates, respectively. Also, w represents the corresponding weights of different gates, and b represents their biases. The final output of the LSTM layer having many LSTM cells is given in Eq. 16.

$$y^k = \sum_j w_{ji}^{k-1}\left(\text{relu}\left(h_i^{k-1}\right) + b_i^{k-1}\right) \tag{16}$$

Here, w is the importance of i^{th} node of layer $n-1$ to j^{th} node of layer n and b depicts bias, y is the final outcome of the LSTM layer.

4 The proposed framework

AD can be prevented and controlled more effectively if diagnosed as early as possible. Therefore, this research aims to develop a model for the timely identification and classification of AD stages. The proposed framework, along with its workflow and phases, is discussed in detail in the following sub-sections. The architecture of the proposed classification framework is shown in Fig. 5. Its five phases are as follows:

Phase 1—Data Collection: The data used in this research are downloaded from the Alzheimer's disease neuroimaging initiative (ADNI) website (http://adni.loni.usc.edu/about/). It contains data of 200 patients, which is separated into three classes, namely AD, MCI, and CN where AD class contains 60 patients, MCI class has 73 patients, and CN class comprises 67 patients. The downloaded data were in T1-weighted MRI mode and used the NIFTI standard to store medical images, from which various types of scans such as Coronal, Sagittal, and Axial are extracted. The considered 200 patients lead to a total of 5980 2-D images of size (256 × 256) for all the three

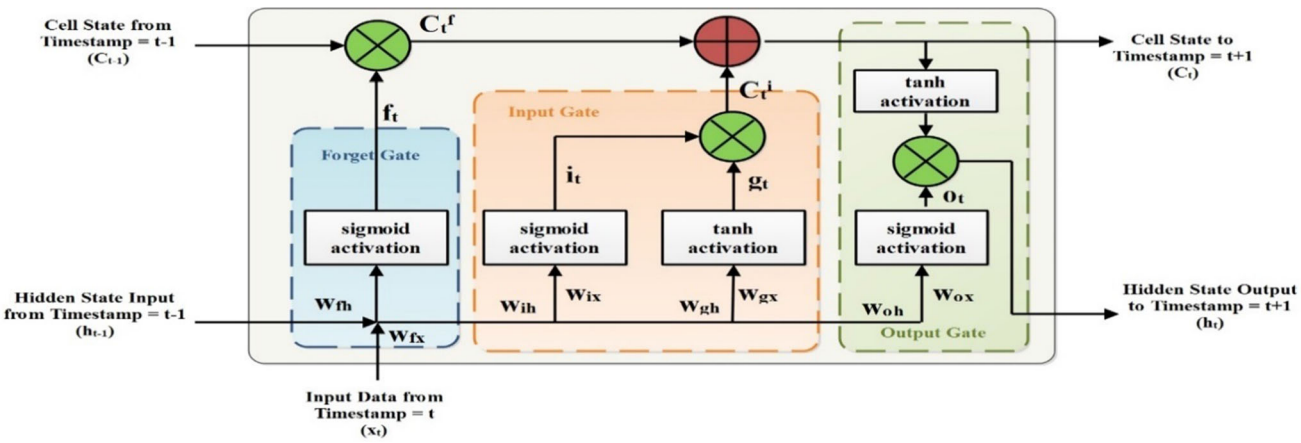

Fig. 4 The single LSTM cell

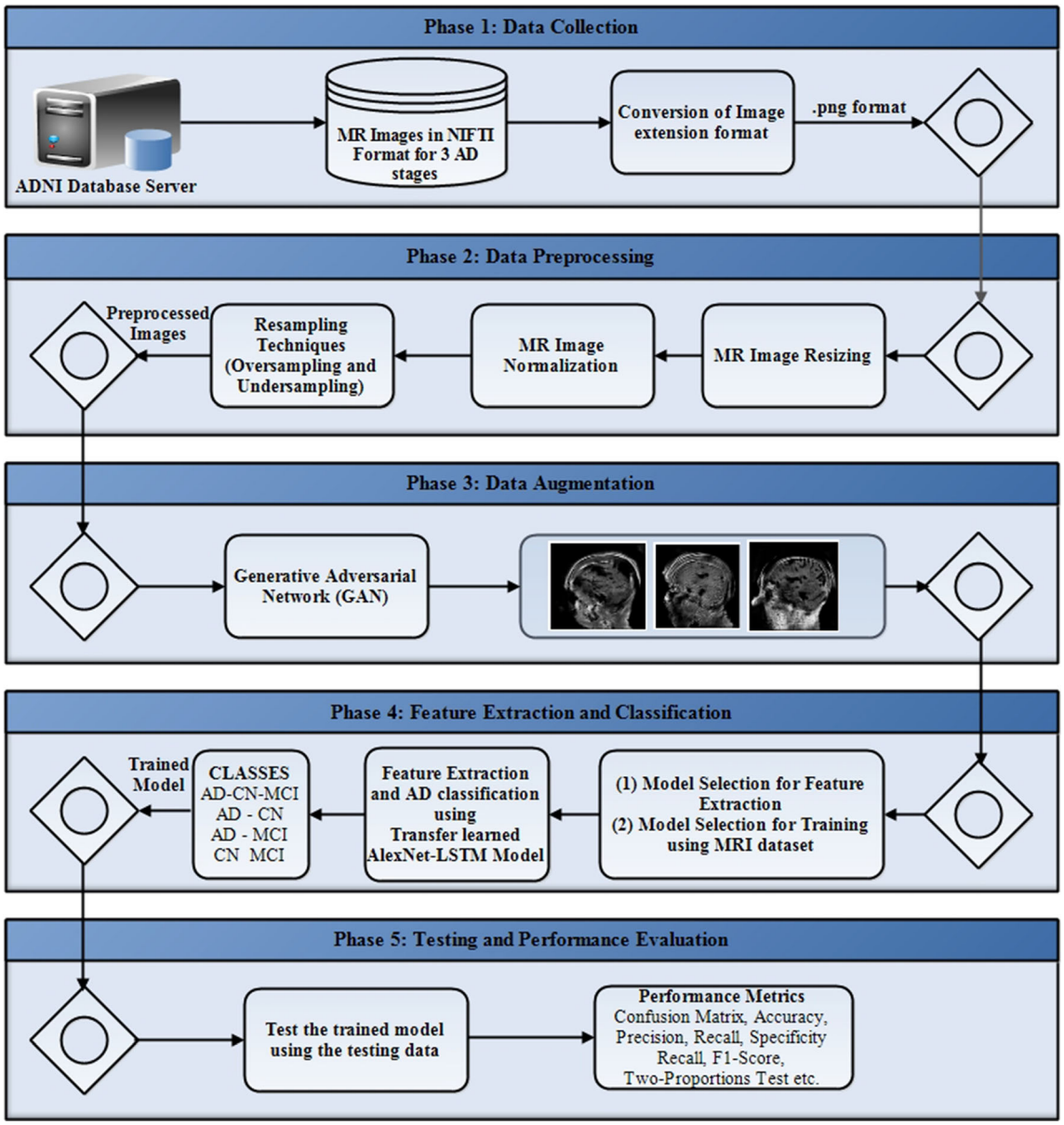

Fig. 5 The proposed Framework

classes. Furthermore, there are 1980 images in the AD class, 1990 in the CN class, and 2010 in the MCI class. Table 4 shows a basic description of the data, including the number of individuals in different classes, the number of males and females, their mean age with standard deviation, etc.

Phase 2—Data Preprocessing: As mentioned in the first step, the obtained data have a problem of unbalanced classes. To handle this problem, we use two resampling methods, i.e., oversampling and undersampling. Undersampling entails eliminating instances from the majority class, whereas oversampling means coping instances for the minority classes. Therefore, using these concepts, oversampling is applied to AD and CN classes, and on the other hand, undersampling in MCI class. As a consequence, the data are evenly distributed, with 1400 images in each group, for a total of 4200 images. Then, as a preprocessing approach, data normalization is used to change the span of pixel values or voxel intensity values. It eliminates significant data variances or brings the data into a particular range so that the value of one or more pixels doesn't get overpowered. There are many types of normalization methods, such as zero mean and unit standard deviation method, [0,1] rescaling, and [− 1,1] rescaling. In this, two methods, namely zero mean and unit standard deviation, are applied. A normalized image has pixel values with zero average, and their standard deviation is one. After this, the data are finally converted to a.png format.

Table 4 Description of the dataset

	AD	MCI	CN
Total number of patients	60	73	67
Male/Female	32/28	38/35	36/31
Age (Mean ± SD)	76.20 ± 7.35	75.25 ± 6.98	75.38 ± 5.01
MMSE (Mean ± SD)	24.83 ± 3.37	27.10 ± 2.2	28.93 ± 1.2
CDR (Mean ± SD)	0.7 ± 0.21	0.5 ± 0.17	0 ± 0

MMSE Mini-mental state examination, *CDR* clinical dementia rating

Phase 3—Data Augmentation: This is one of the necessary steps in the classification framework due to the scarcity of medical datasets and curse of dimensionality problem, i.e., a vast count of features and less amount of data, leading to overfitting of the classification model. A data augmentation technique named GAN is used to enrich the dataset to handle the above problems. Firstly, the original dataset is jumbled and split into three sets named training, validation and testing set according to 70:10:20 split using a random selection method. As a result, the training set has 4200 images, the validation dataset contains 600 images and the testing dataset has 1200 images. Then, the training dataset, after applying data augmentation techniques (GAN), grows to a total of 7500 images in all the three classes. Table 5 summarizes the statistics for training, validation, and testing data for multiclass and binary classifications before and after applying GAN. Furthermore, Fig. 6 shows the sample of a 2-dimensional scan for each class.

Phase 4—Feature Extraction and Classification: This step entails multiclass classification of various stages of AD, namely AD, MCI, and CN. In addition, binary classification between each pair of the classes, i.e., AD vs MCI, AD vs CN, CN vs MCI, is also performed. For this, a comparative analysis was done for selection of best model for feature extraction and in the next step, for training the MRI dataset as follows:

4.1 Model selection for feature extraction

Based on the ImageNet Dataset [57], various deep learning methods such as Alexnet, ResNet, and Inception relied on the transfer learning methodology can be used for image classification. The networks used in the current study are selected based on their respective size/precision ratio and previous applications in the field of AD research [58, 59]. These models are used for feature extraction by taking the inbuilt convolution layers of the models. Then, these layers can further be feed into the transfer learned output layers for classification. The appropriate framework out of five frequently used pre-trained deep learning models named Alexnet [51], VGG 16 [30], VGG 19 [60], Resnet-50 [35] and Inception V3 [61] has been decided by conducting the experimental analysis of the dataset.

4.2 Model selection for training using MRI dataset

After applying and selecting the appropriate model, i.e., Alexnet out of all the pre-trained models analyzed in the aforementioned step, further comparative analysis was done to extract the most suitable model for training our image dataset. Based on the existing literature in the diverse fields and applications of image classification such as crop classifications [62], tissue images classification [63], coronavirus chest image classification [64], CO_2 welding image classification [65] various models such as LSTM, MLP, and transfer learned Alexnet and their hybrid method as used in various studies [65–69] was proposed

Table 5 Description of training, validation, and testing data

Category	Class	Training data size	Validation data size	Testing data size	Total
Without GAN	AD–CN–MCI	4200	600	1200	6000
	AD–CN	2800	400	800	4000
	AD–MCI	2800	400	800	4000
	CN–MCI	2800	400	800	4000
With GAN	AD–CN–MCI	7500	600	1200	9300
	AD–CN	5000	400	800	6200
	AD–MCI	5000	400	800	6200
	CN–MCI	5000	400	800	6200

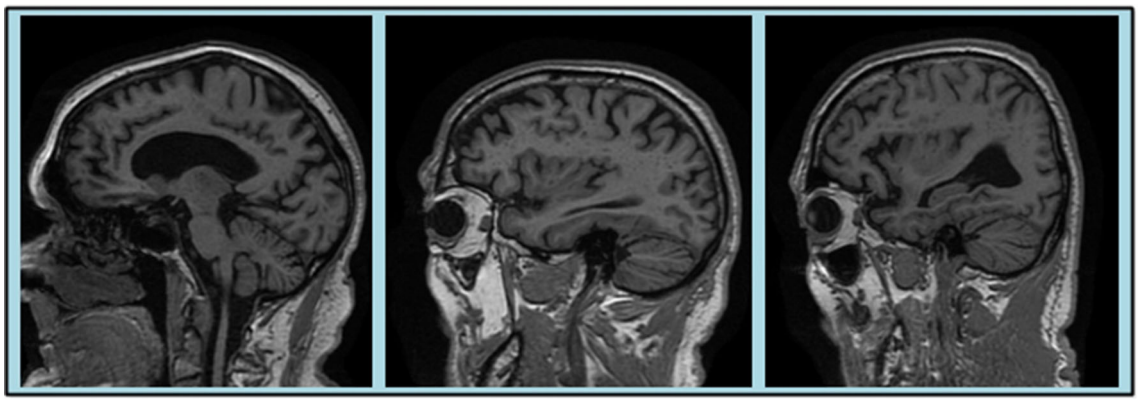

Fig. 6 Sample Images: AD Stage, CN Stage, and MCI Stage (from left to right)

and analyzed to select the best model among all based on the experimental results.

After all the comparative analysis in above two steps, the current research presents a model relied on transfer learned deep learning model, which combines transfer learned Alexnet and LSTM in a single model. It has already been demonstrated that providing relevant features improves the performance of LSTM [64]. Moreover, CNN architecture named Alexnet is well known for its capability to get valuable features from the dataset if the model is reformed corresponding to the necessities of the classification model. As a result, an algorithm that combines CNN and LSTM was suggested in the literature [65–69], and it has produced promising results in a number of disciplines, including text analysis, voice recognition, rainfall prediction, gesture identification, and machine health status prediction. The transfer learned CNN (Alexnet)-LSTM technique is used in this work to predict AD class, either in binary or multiclass classification. Firstly, the relevant features from the MR images are extracted using the Alexnet model's convolution layers. The output of the pre-trained Alexnet's convolution, pooling, and flattening layer is then used as the input layer to the LSTM layer, which models the temporal information to forecast the stage of AD. Let $X_T = [X_1, X_2, \ldots\ldots X_n]$ be the input number of images for Alexnet. Then the model's output is passed to LSTM cells, where each cell consists of various gate units such as forget gate, input gate, input node, and output gate. After the LSTM layer, the last but one layer is a fully connected layer that receives feature vector $h_T = [h_1, h_2, \ldots\ldots, h_n]$. In the end, the output layer is constructed as per the number of classes, i.e., two neurons with sigmoid activation function for binary classification and three neurons with softmax activation function for multiclass classification. Figure 7 depicts the proposed algorithm's complete architecture along with the parameter settings and model tuning hyperparameters.

Phase 5—Testing and Performance Evaluation: The trained model is assessed using test data in this step, and several performance evaluation measures, including accuracy, precision, recall, and others, are calculated. The proposed and existing models are compared based on these performance evaluation parameters.

4.3 Methodology of proposed framework

The model construction in this work was done in Python (3.7.4) with several deep learning packages such as Keras (2.3.0-tf), TensorFlow (2.2.0), and Scikit Learn (0.22.2). The entire implementation was completed online using Google Colab's GPU. Further, Fig. 8 shows a flowchart of the training process for both classification models. As seen in the figure, the MRI data are taken from the ADNI website, in which the brain images are in NIFTI format. According to the requirement, these images can't be directly used as input images, so these are converted into the.png format. To handle the class imbalance problem, resampling methods were applied. Then, the preprocessed data are grouped into training, validation, and testing dataset on the basis of 70:10:20 split. Further, to address the overfitting problem, a data augmentation technique (GAN) was used on the training subset so that there is an appropriate count of images to learn the classifier. Finally, the resizing was done to make the image dataset of the same dimension as needed by the Alexnet model. In the next step, the training data were fed into the convolution layers of the pre-trained Alexnet model for feature extraction. Then these extracted features are flattened and fed into the LSTM layer, from which the final output of LSTM is passed to the transfer learned dense layers optimized on the basis of grid search algorithm so that the best hyperparameters such as learning rate, activation function, and number of neurons can be found. Finally, the optimized features were passed to the output layer which was having 3 neurons for multiclass classification and 2 neurons

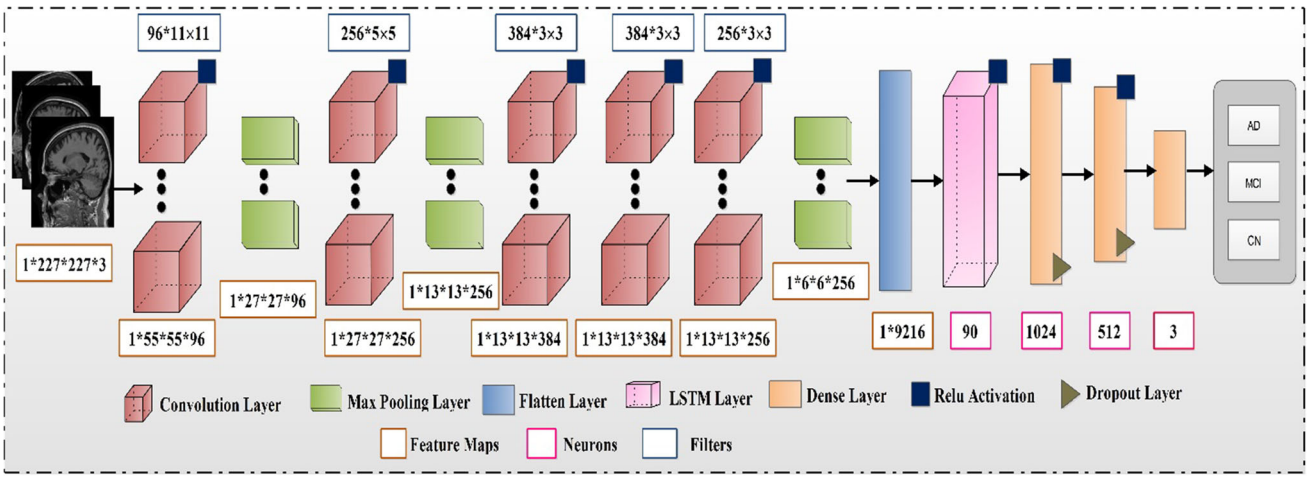

Fig. 7 The multiclass classification model used in the study

for binary classifications. The validation data are given to the trained model at the end of the whole process. The four parameters are calculated: validation accuracy, validation loss, training accuracy, and training loss. If the validation loss computed for *n*th iteration is smaller than the subsequent five iterations, then the model is stopped, and the weights of *n*th iteration is restored for testing. This step further solves the problem of overfitting in the classification model. Finally, testing data are used to validate the optimum model with restored weights.

5 Results and discussion

This research aims to predict and classify the AD stages from medical images of the brain and observe the feasibility of the proposed framework in predicting the AD stage. This section discusses the obtained results thoroughly.

5.1 Performance evaluation parameters for classification model

In order to put the model for the actual implementation, it is required to evaluate its performance. It comprises utilizing performance values to compare estimated model values to actual values to assess how well the proposed model can mimic the actual output. The proposed framework is trained to do binary class and multiclass classification of three stages of AD. The observed/actual result (AD stage) are compared to the predicted outcome using various performance metrics named confusion matrix, accuracy, loss, sensitivity/recall, precision, F1-score, receiver operating characteristics (ROC) curve, area under curve (AUC), true negative rate (specificity), negative predicted value (NPV). Further, Two-Proportions Test is also used to test the classification ability of the proposed framework. The summary of the performance metrics used in the current study is discussed here.

Confusion Matrix: In this matrix, the summary of the classification model's prediction outcomes is represented. The number of right and wrong predictions are gathered and broken down by class using count values. It gives us the TP, TN, FN, and FP values where:

TP (True Positive): both the actual and expected outcomes are correct.
TN (True Negative): both the actual and expected outcomes are false.
FN (False Negative): actual result is true while the predicted is false.
FP (False Positive): actual result is false while the predicted is correct.

Accuracy: It is expounded as the number of accurate labels to the total number of labels represented by the formula as:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{17}$$

Loss: This is elucidated as the prediction error of the model calculated as Eqs. (18) and (19). For binary classification, it is called binary_crossentropy, whereas, for multiclass classification, it is called categorical_crossentropy.

$$\text{loss}_{\text{binary}}(a,p) = -(a \log p + (1-a)\log(1-p)) \tag{18}$$

$$\text{loss}_{\text{multiclass}}(a,p) = -\sum_{n=1}^{N} a_{k,n} \log p_{k,n} \tag{19}$$

Sensitivity/recall/true positive rate: It is calculated using Eq. (20) by dividing the number of TPs by the total of TPs and FNs.

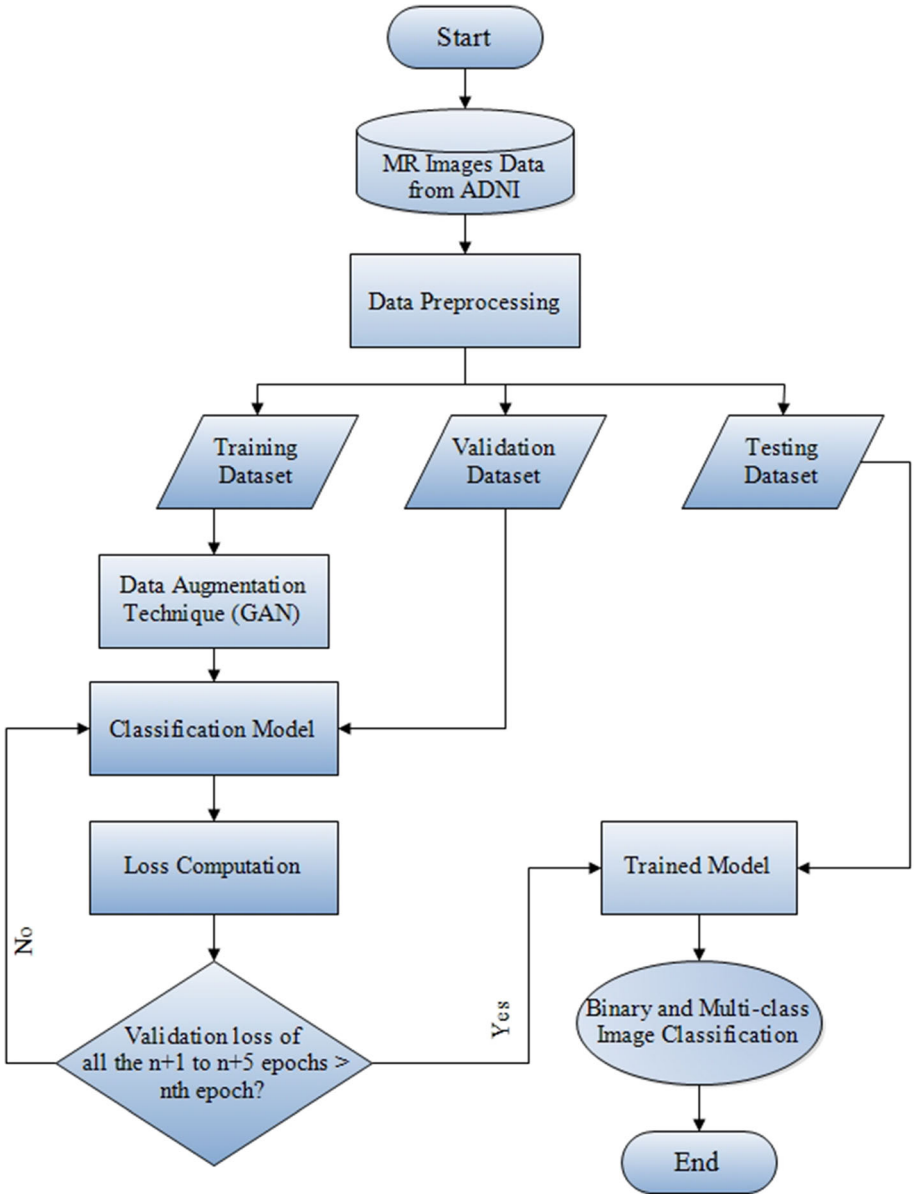

Fig. 8 Flowchart of the Proposed Methodology

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

Precision: It is defined as the percentage of relevant results among the list of all returned search results represented as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{21}$$

F1-score: Equation (22) represents the harmonic mean of precision and recall, called F1-Score. The better the model, the higher the metric's value.

$$F1\text{-Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

ROC curve and AUC: A graph that shows how well a classification model performs across all categorization levels, which is favorable with high value.

Specificity/true negative rate: It is interpreted as the count of correct negative predictions to the total number of negative predictions represented as:

$$\text{Specificty} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{23}$$

Negative predicted value: It can be defined as the percentage of anticipated negatives that turn out to be true negatives, represented by the following formula:

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{24}$$

5.1.1 Two proportions test

The performance of two models can be compared using the common evaluation measures like accuracy, precision, recall, F1-score, etc. These criteria, however, are unable to distinguish between two predictive models' predictions that differ significantly. A statistical inference method known as hypothesis testing uses confidence intervals to assess the significant difference between the two predicted models. In the current study, two proportions test [70] is used to evaluate the effectiveness of the proposed deep learning-based model.

Let p_1 and p_2 denote the accuracies for classifiers 1 and 2, respectively, and N be the number of samples. Further, x_1 and x_2 denote the correctly classified number of samples in classifiers 1 and 2, respectively. Then,

$$p_1 = \frac{x_1}{N}, p_2 = \frac{x_2}{N} \tag{25}$$

In this study, to check the performance comparison of two models, equal accuracy hypothesis is tested. Here, the null hypothesis (H_0) states that the prediction accuracy of the two models is equal, whereas the alternate hypothesis (H_1) states that the prediction accuracy of the models varies; which are given as:

$$\begin{aligned} H_0 &: p_1 = p_2 \\ H_1 &: p_1 \neq p_2 \end{aligned} \tag{26}$$

Hence, the test statistic is formulated as:

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{2\overline{p}(1-\overline{p})}{N}}} \tag{27}$$

where

$$\overline{p} = \frac{(x_1 + x_2)}{2N} \tag{28}$$

The null hypothesis for the current study will be rejected if the test score is not in the range of [− 1.645,1.645] at 95% confidence level.

5.2 Comparative analysis of pre-trained deep learning models for feature extraction

The performance of five frequently used pre-trained deep learning models named Alexnet, VGG 16, VGG 19, ResNet-50 and Inception V3 for multiclass classification, i.e., AD–CN–MCI as well as for binary classification, i.e., AD vs CN, AD vs MCI and CN vs MCI, respectively, on the original dataset without GAN implementation is analyzed and compared in this section. Moreover, the corresponding line plots of the models for different performance metrices such as training loss, training accuracy, validation loss, and validation accuracy are depicted in Fig. 9 for multiclass and binary class classifications. From the plots, it can be seen that the best performance was given by Alexnet Model and the worst performance for the current scenario was given by ResNet-50. Also, the results are in accordance with other studies [59] which justifies the superior performance of Alexnet as compared to other models. In addition, it can be seen that the models get overfitted due to large difference in training, validation and testing accuracy which is solved further by doing data augmentation, i.e., applying GAN technique on the dataset.

5.3 Comparative analysis of deep learning models for classification using training set

The performance of deep learning models named Transfer Learned Alexnet, MLP, LSTM, and their hybrid methods named Alexnet_MLP and Alexnet_LSTM (proposed model) for training using MRI dataset for multiclass classification, i.e., AD–CN–MCI and for binary classification, i.e., AD vs CN, AD vs MCI and CN vs MCI, respectively, are further analyzed on the augmented dataset. Moreover, the corresponding line plots of the models for different performance metrices such as training loss, training accuracy, validation loss, validation accuracy are depicted in Fig. 10 for multiclass and binary class classifications. From the plots, it can be seen that the best performance was given by Alexnet_LSTM and the worst performance for the current scenario was given by LSTM. Also, the results are in accordance with other studies [64, 65] which justifies the poor performance of LSTM when used alone for image classification without feature extraction as compared to other models. In addition, it can be seen that the models do not get overfitted as there is small difference in training, validation and testing accuracy after data augmentation on the dataset.

5.4 Performance evaluation of the proposed model for training set

The proposed model's performance was compared on the basis of four parameters, i.e., training loss, training accuracy, validation loss and validation accuracy. Figure 11 represents the corresponding graphs of the above parameters for the proposed model in multiclass classification, Fig. 12 for binary class classification (AD vs CN), Fig. 13 for AD vs MCI, and Fig. 14 for MCI vs CN. The graphs show that the appropriate epochs to train are 17 for multiclass classification, 12 for AD vs CN, 10 for AD vs MCI, and 12 for CN vs MCI for the proposed model. Further, Table 6 demonstrates the proposed model's performance for multiclass and binary class classification in terms of accuracy and loss parameters for the training and validation

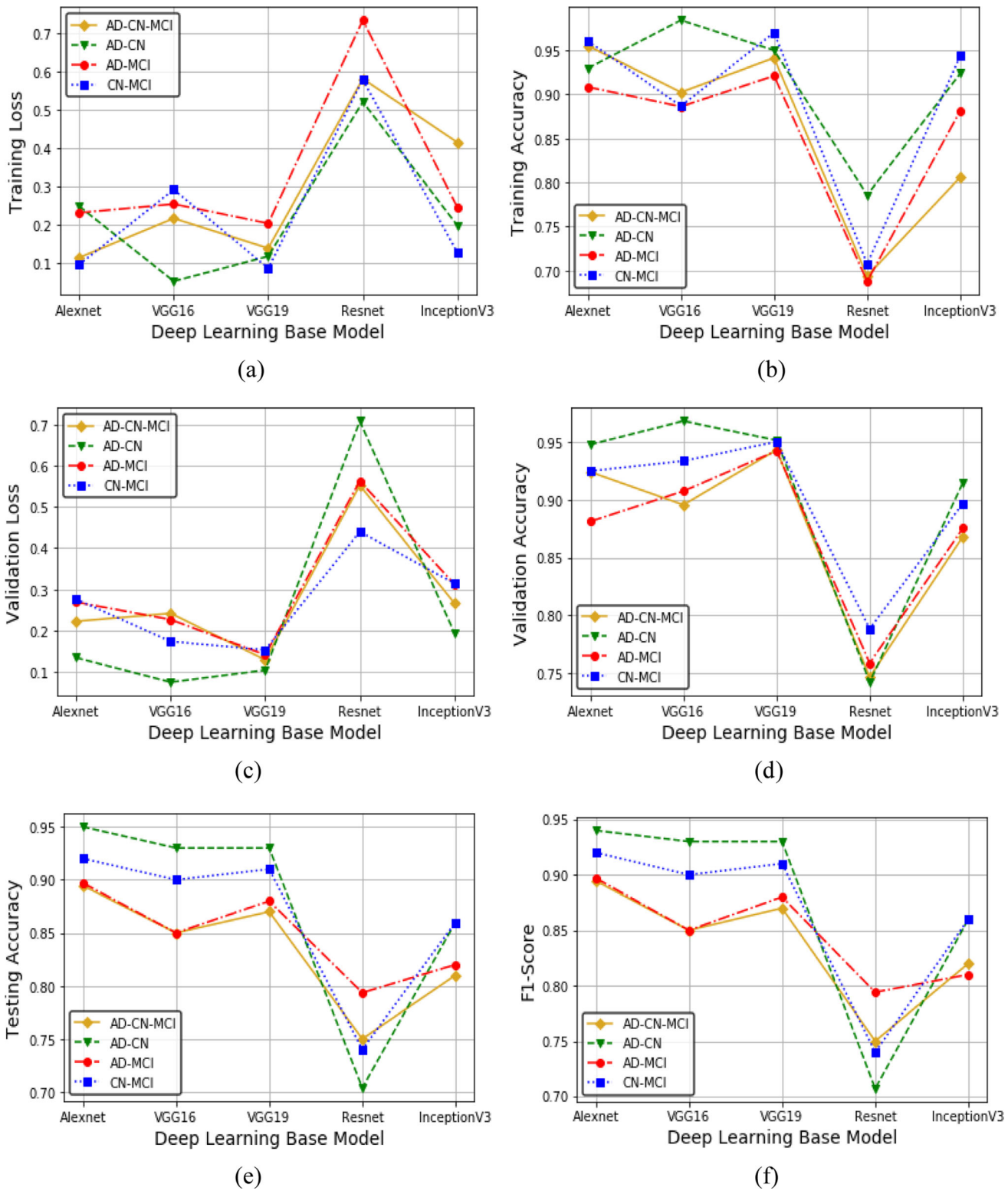

Fig. 9 Line plots for comparison of deep learning pre-trained models on the basis of **a** Training loss, **b** training accuracy, **c** validation loss, **d** validation accuracy, **e** testing accuracy, **f** F1-score

dataset. As evident from the table, it can be said that the proposed model performs superior without getting over-learned as the difference in the training and validation parameters are comparable in some cases and similar in others, so the models don't have an overfitting problem.

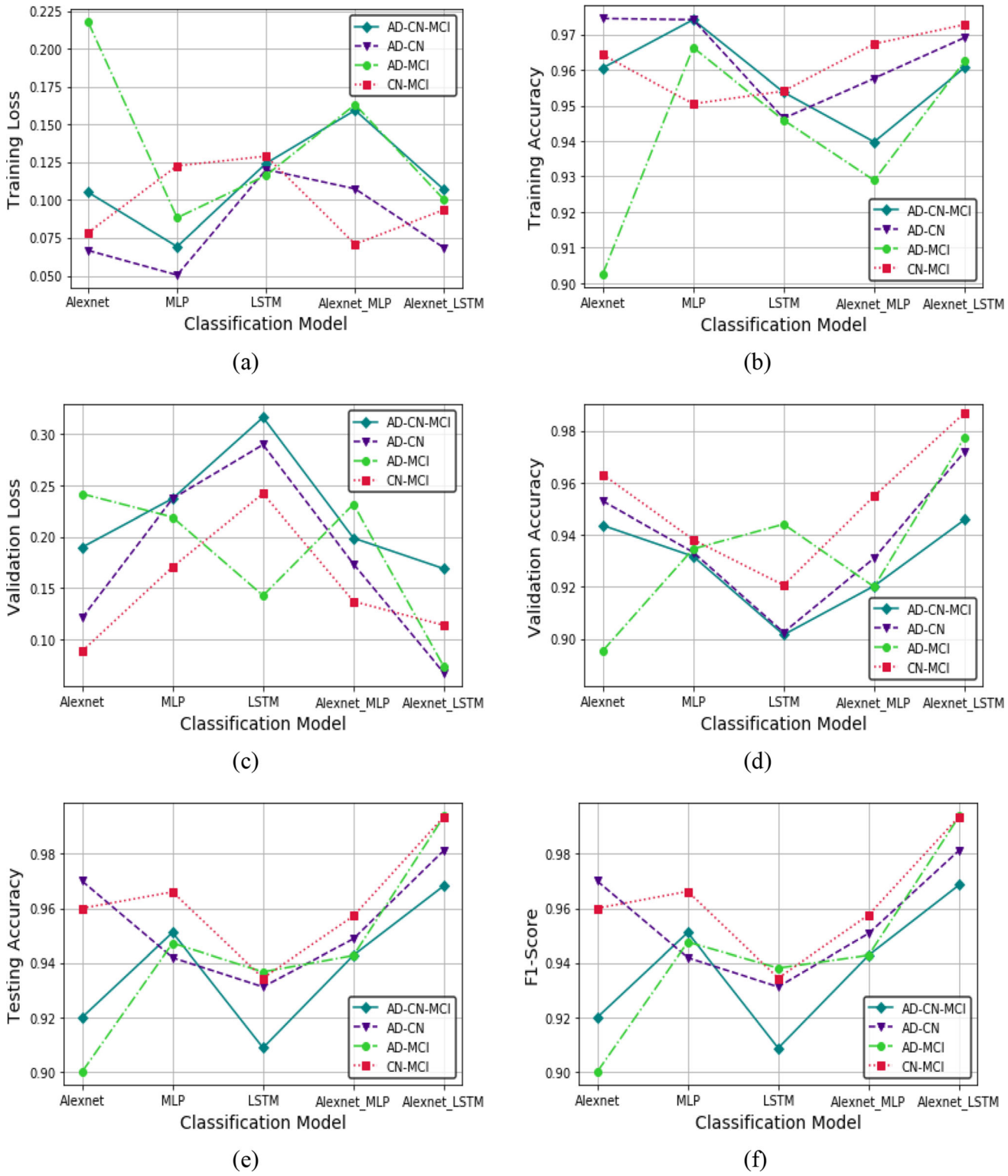

Fig. 10 Line plots for comparison of deep learning classification models on the basis of **a** Training loss, **b** training accuracy, **c** validation loss, **d** validation accuracy, **e** testing accuracy, **f** F1-score

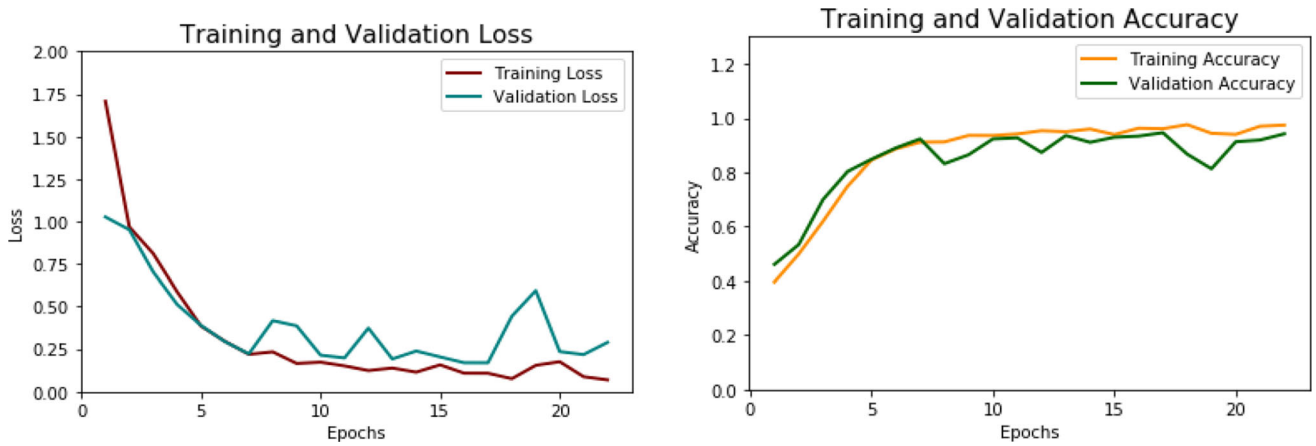

Fig. 11 Plots of training and validation losses and accuracy for proposed model: AD–CN–MCI

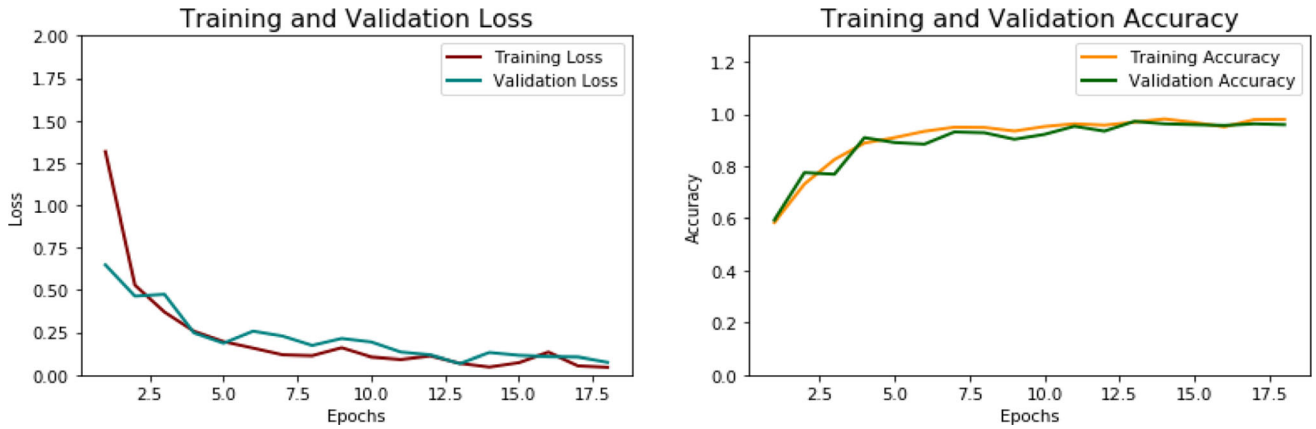

Fig. 12 Plots of training and validation losses and accuracy for proposed model: AD–CN

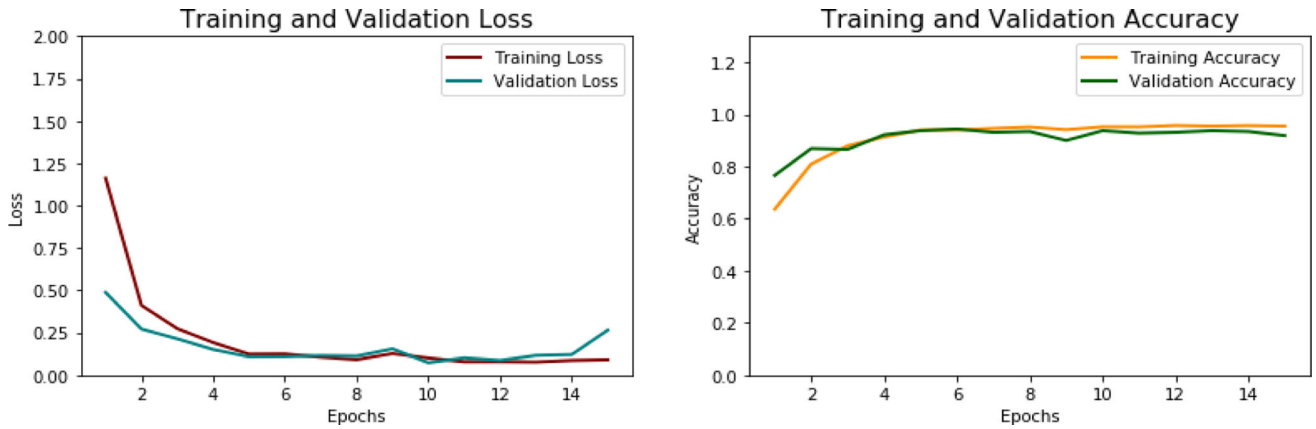

Fig. 13 Plots of training and validation losses and accuracy for proposed model: AD–MCI

5.5 Performance evaluation of the proposed model for testing set

After the proposed model is trained for the required epochs for different categories, the model is evaluated and tested to get the final performance parameters. Based on these parameters, the performance of the models is compared with existing studies and within themselves, i.e., without applying data augmentation technique and after applying data augmentation technique named GAN. Firstly, the foremost parameter of the classification problem, i.e., the confusion matrix, is constructed for the proposed model for

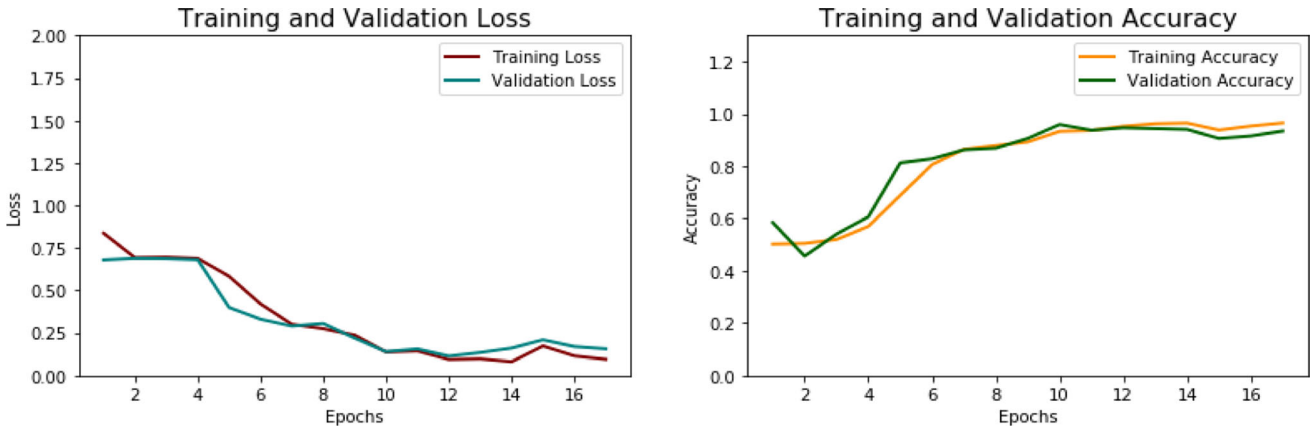

Fig. 14 Plots of training and validation losses and accuracy for proposed model: CN–MCI

Table 6 Training and validation dataset accuracy and loss values (best) for proposed model

Model	Training loss	Training accuracy	Validation loss	Validation accuracy
AD–CN–MCI	0.10685	0.96064	0.16875	0.94583
AD–CN	0.06846	0.96910	0.06686	0.97188
AD–MCI	0.10021	0.95243	0.07270	0.93750
CN–MCI	0.09360	0.95277	0.11344	0.94687

both multiclass and binary classification for the augmented dataset, as shown in Fig. 15. This matrix provides approximate values for true positives, true negatives, false positives, and false negatives, from which various additional metrics like precision and recall can be calculated.

Based on the confusion matrices constructed for the models, the performance metrics like precision and recall, their macroaverage, microaverage, etc., after the appropriate number of epochs for all the combinations are shown in Table 15 for both non-augmented as well as augmented

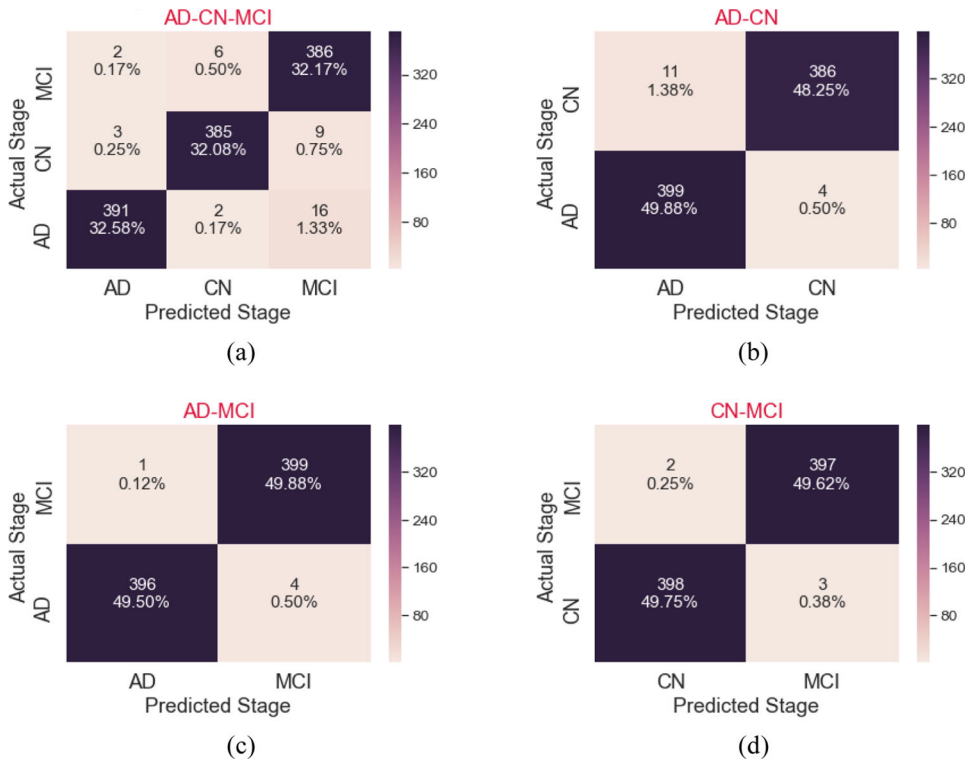

Fig. 15 Confusion matrix for proposed model: **a** AD–CN–MCI, **b** AD–CN, **c** AD–MCI, **d** CN–MCI

dataset to get the better insight to the results. Further, Fig. 16 demonstrates the bar plots of the proposed model for both the datasets to provide the comparison of the results. Also, it can be seen from Table 7 and Fig. 16 that the classification results get improvised by approximately 3–4% after applying the data augmentation on the original dataset.

After the above comparison of the results, another parameter, i.e., the ROC curve, made between true positive rate (TPR) and false positive rate (FPR), is also plotted to evaluate the classification model's performance. The curve tells us the idea of our model performance in terms of training and testing. The higher the AUC score, the more excellent our model is. Figure 17 shows the ROC curves in

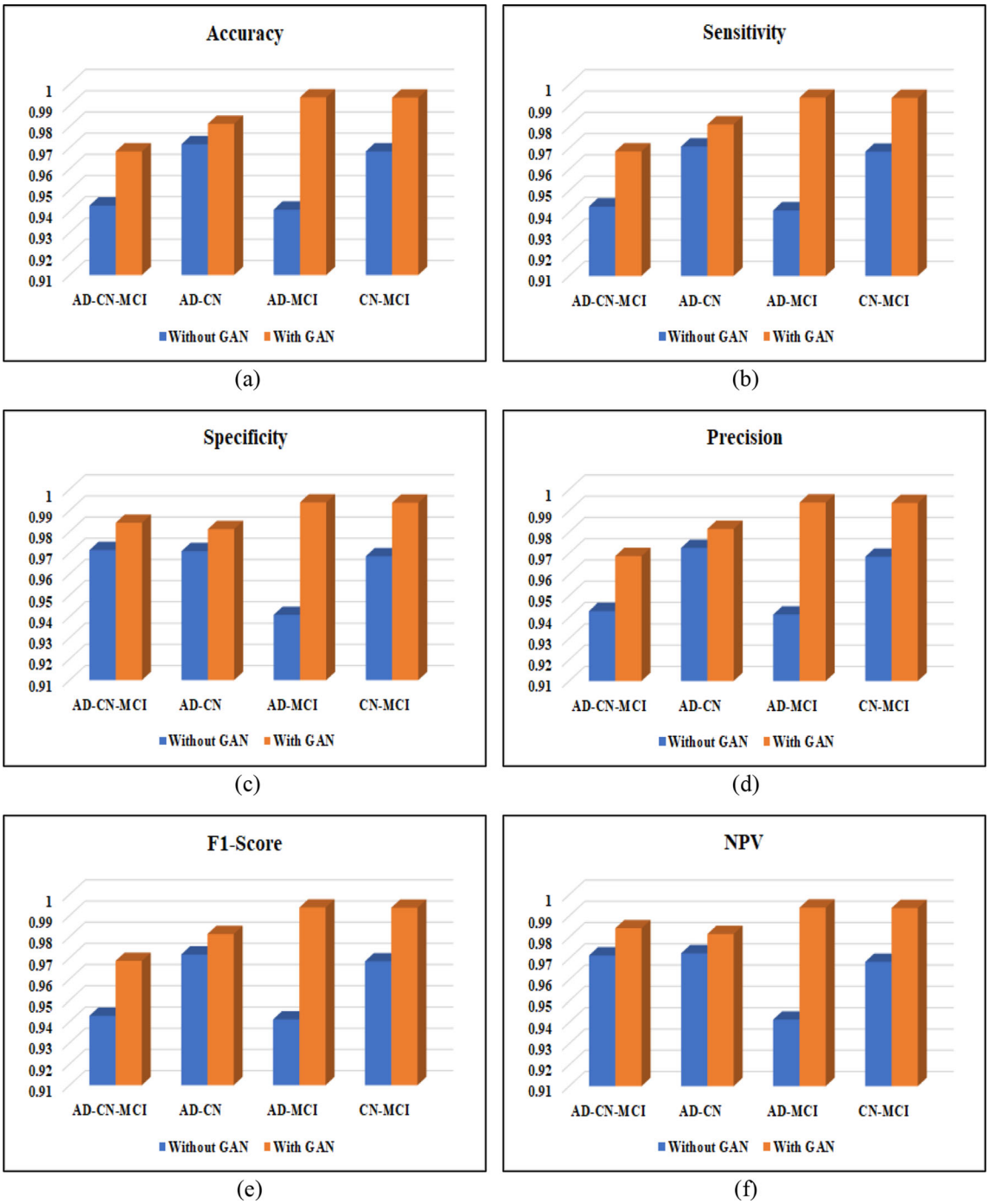

Fig. 16 Bar plots for performance comparison of proposed model for without GAN and With GAN dataset showing: **a** Accuracy, **b** sensitivity, **c** specificity, **d** precision, **e** F1-score, **f** NPV

Table 7 Performance evaluation of proposed model for augmented and non-augmented dataset

Performance metric	Category	AD–CN– MCI	AD–CN	AD–MCI	CN–MCI
Accuracy	Without GAN	0.9427	0.9717	0.9408	0.9683
	With GAN	0.9683	0.9813	0.9938	0.9937
Sensitivity	Without GAN	0.9426	0.9708	0.9408	0.9684
	With GAN	0.9685	0.9812	0.9938	0.9937
Specificity	Without GAN	0.9713	0.9708	0.9408	0.9684
	With GAN	0.9842	0.9812	0.9938	0.9937
Precision	Without GAN	0.9429	0.9725	0.9413	0.9684
	With GAN	0.9687	0.9815	0.9939	0.9937
F1-Score	Without GAN	0.9428	0.9716	0.9411	0.9684
	With GAN	0.9686	0.9813	0.9938	0.9937
NPV	Without GAN	0.9714	0.9724	0.9413	0.9684
	With GAN	0.9842	0.9815	0.9939	0.9937

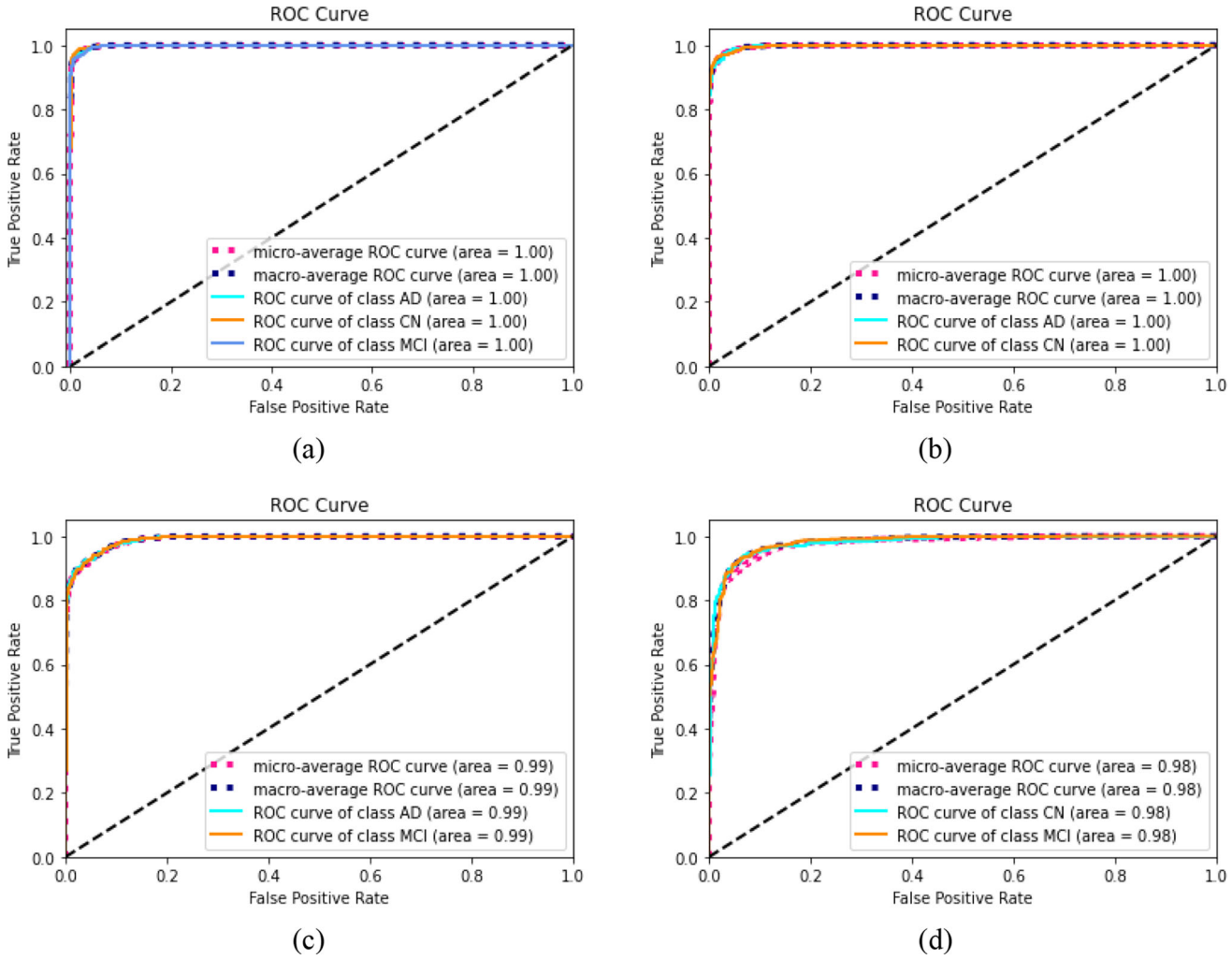

Fig. 17 ROC curves for proposed model: **a** AD–CN–MCI, **b** AD–CN, **c** AD–MCI, **d** CN–MCI

multiclass and binary image classification for the proposed model on the augmented dataset. As seen from the curves, it can be said that the proposed model is good enough to classify as AUC values are greater than 0.95.

5.6 Performance comparison of the proposed model using two-proportions test

The performance of different deep learning models is compared using the common evaluation measures like accuracy, precision, recall, F1-score, etc. Further, to analyze the significant difference between the accuracies of the proposed model with the other deep learning classification models, a statistical test named two proportions test had been used in the current study. The heat maps of test statistic value for each pair of models used in the study had been shown in Fig. 18a–d for multiclass (AD–CN–MCI) and binary classifications (AD–CN, AD–MCI, CN–MCI), respectively. It is observed from the heat maps that the pairs $\{Alexnet\_MLP, MLP\}$ and $\{Alexnet\_MLP, LSTM\}$ for classification AD–MCI have test statistic value in the range of [− 1.645, 1.645]. Thus, for these pairs, the null hypothesis cannot be rejected, i.e., no model is superior to each other. Similarly, the test statistic value for the pair $\{Alexnet\_MLP, Alexnet\}$ for CN–MCI classification and self-pairs for all the models for all classifications have the value in the range of [− 1.645, 1.645] and the null hypothesis for these pairs cannot be rejected. However, all the other pairs involved in the study have test statistic value not in the above range that shows the significant difference in their prediction accuracies. Further, the overall results of two proportions test demonstrate that the prediction results obtained from the proposed framework differ significantly from other deep learning classification models and also, with the positive values which adds on their higher performance as compared to other models.

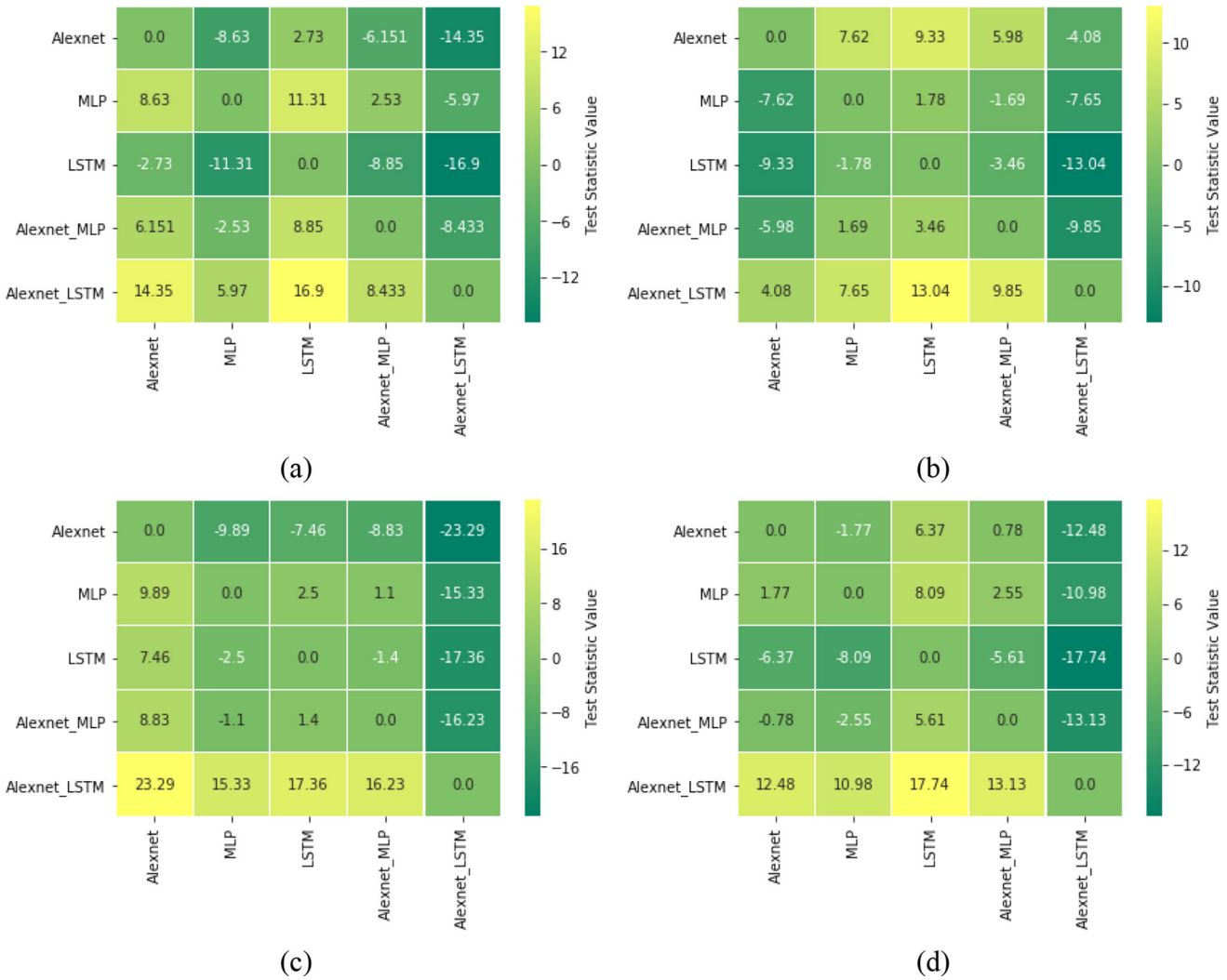

Fig. 18 Two-proportions test: test statistic value for **a** AD–CN–MCI, **b** AD–CN, **c** AD–MCI, **d** CN–MCI

5.7 Comparison of the proposed and existing models in terms of performance

The transfer leaned deep learning model for multiclass and binary classification is proposed in the current study. Table 8 compares the suggested model's performance in terms of accuracy to that of existing state-of-the-art models. The table demonstrates that the proposed framework has achieved the remarkable accuracy of 96.83% for multiclass image classification of AD stages (AD, MCI, CN) in comparison with other existing models. Moreover, the outcomes of the proposed model are comparable with that of Jain et al. [30], but the dataset used in the latter study is minimal, which may lead to an overlearned model. As shown, the suggested model achieves promising accuracy with data augmentation for binary and multiclass classification, which further solves the problem of overfitting compared to previous models. Moreover, Fig. 19 compares the proposed and existing models for multiclass medical image classification as well as for binary classification: AD vs. CN, AD vs. MCI, and CN vs. MCI. Compared to other models, it is clear from Table 8 and Fig. 19 that the suggested model outperforms the existing models.

5.8 Discussion

This study uses the MR Images dataset from the ADNI website to assess the capacity of a deep neural network to perform multiclass and binary class classification. The framework used in the study is built using CNN models which integrates these models for feature extraction and classification in a single architecture. Further, to solve the problem of overfitting, data augmentation technique named GAN has applied on the limited dataset. The proposed model uses the various deep learning algorithms such as Alexnet, VGG 16, and VGG 19 and perform the analysis on them to select the best model according to our dataset for feature extraction and Alexnet had outperformed all the other deep learning algorithms. Further, Alexnet with other frequent deep leaning algorithms such as MLP, LSTM and their fusion with Alexnet are considered and compared to finally select the best model for our study that provide the highest results for multiclass and binary class classifications. Out of all the comparisons, Alexnet with LSTM performed best and this algorithm had used for our proposed study. In addition, results using GAN and without using GAN are compared to see the effect of data augmentation technique which leads to solve the problem of overfitting. In the nutshell, this study can be able to demonstrate that using fusion of deep learning models along with data augmentation techniques can be suitable for achieving the high performance when compared with conventional techniques for multiclass and binary class classifications while handling the problem of overfitting; which is the main drawback of deep learning algorithms. Another significant strength is the accuracy, robustness, and validity of the proposed study in predicting the various stages of AD compared to conventional algorithms, as mentioned in the results section. However, one of the study's weaknesses is that it only takes the MR images for multiclass and binary image classifications. As a result, the investigation of PET, DTI, and clinical features can be conducted to notice the generalized conclusion of the performance of the presented algorithms. Further, other data augmentation techniques such as DCGAN and stacked GAN can be implemented to supplement the dataset which may improve classification results.

Table 8 Accuracy comparison of proposed and existing similar models on ADNI (MRI) dataset

Research paper	Classification accuracy (%)			
	AD–CN	AD–MCI	CN–MCI	AD–CN–MCI
Yan et al. [24]	98.85	–	–	–
Zhu et al. [25]	98	–	91.9	–
Li et al. [26]	93.2	–	80.4	–
Divya et al. [27]	96.82	90.40	89.39	–
Kang et al. [28]	90.36	77.19	72.36	–
Feng et al. [29]	94.21	90.03	84.64	–
Jain et al. [30]	99.14	99.30	99.22	95.73
Basheera et al. [31]	100	96.2	98	86.7
Liu et al. [33]	93.26	–	74.34	–
Shi et al. [34]	97.13	–	87.24	–
Korolev et al. [35]	–	–	–	88
Payan et al. [41]	95.39	86.8	92.1	89.47
Proposed model	98.13	99.38	99.37	96.83

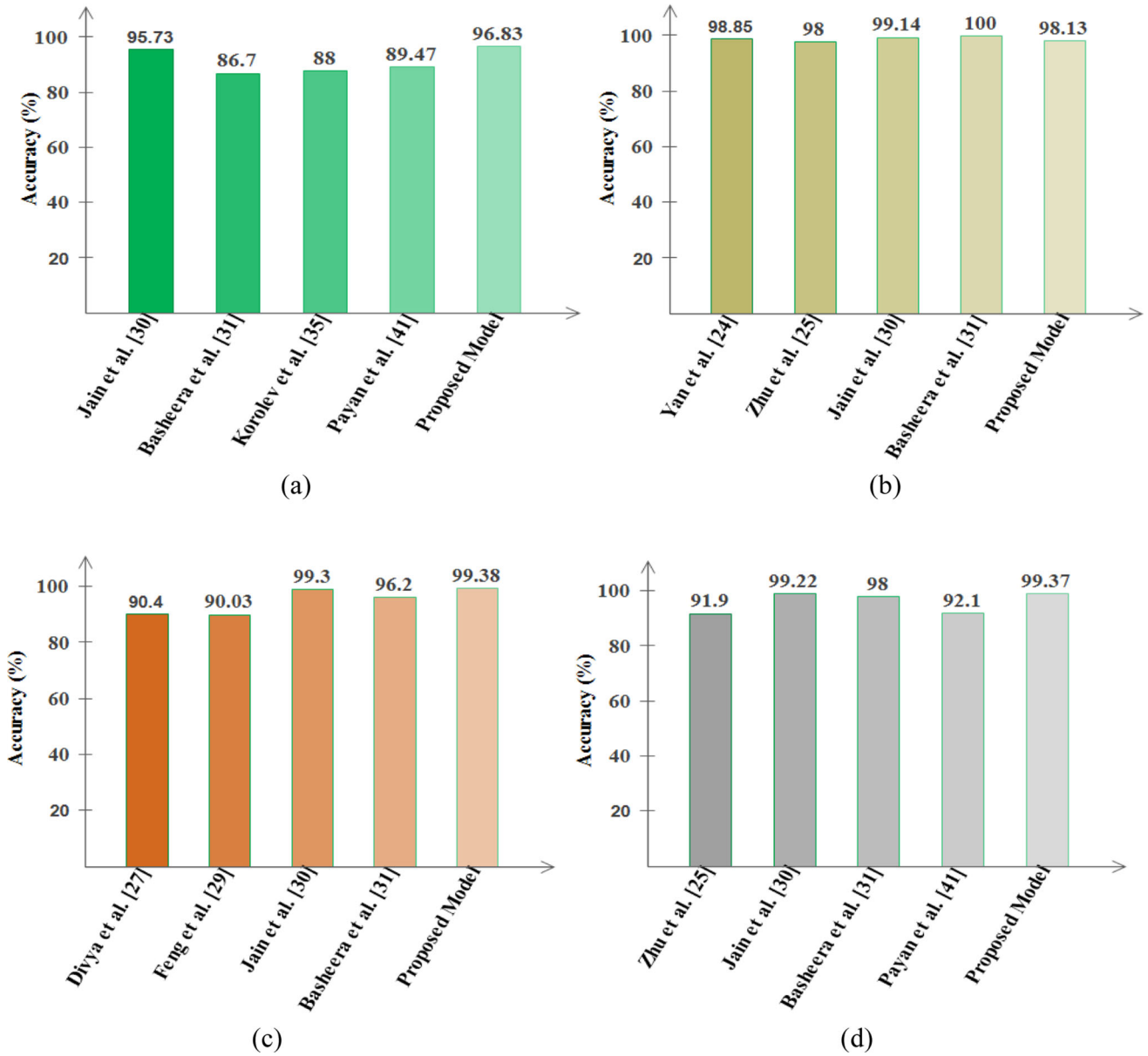

Fig. 19 Comparison of top 5 models in terms of accuracy for: **a** AD–CN–MCI, **b** AD–CN, **c** AD–MCI, **d** CN–MCI

6 Conclusion

The ability of a deep neural network to do multiclass and binary class classification is evaluated in this study utilizing MR images dataset from the ADNI website. The proposed framework in this research is used for AD classification and detection. The suggested framework is built using CNN models that use deep learning to classify three different stages of AD. The proposed model uses the classification algorithm based on a transfer learned deep learning model on pre-trained Alexnet to extract MR image features from convolution layers of pre-trained Alexnet. These features are then fed into the LSTM layers, transfer learned fully connected layers, and finally, to the output layer to do classification. Furthermore, the data augmentation technique (GAN) and early stopping method can get the appropriate number of epochs to prevent the model from getting overtrained and achieve reasonable accuracy. Moreover, the proposed model, with an accuracy of 96.83%, surpasses the other algorithms proposed in the literature. Experiments demonstrate that the proposed design is a suitable, simple structure that reduces overfitting.

In the future, it is intended to test the achievement of other pre-trained models, such as MobileNet and ShuffleNet, for multiclass AD stage classifications. Further, additional data augmentation techniques such as DCGAN

and stacked GAN can also be used to supplement the dataset, which may improve the results. In addition, MRI segmentation will be used to highlight Alzheimer's characteristics before AD classification.

Acknowledgements The authors express their gratitude to the Alzheimer's Disease Neuroimaging Initiative (ADNI) for providing the standardized MR Images Dataset.

Data availability statement The data used to support the findings of the study are made available on Alzheimer's Disease Neuroimaging Initiative (ADNI) at http://adni.loni.usc.edu/about/.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Singh SP, Wang L, Gupta S, Goli H, Padmanabhan P, Gulyás B (2020) 3D Deep learning on medical images: a review. Sensors 20(18):5097
2. Jo T, Nho K, Saykin AJ (2019) Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. Front Aging Neurosci 11:220
3. Cilia ND, D'Alessandro T, De Stefano C et al (2022) Deep transfer learning algorithms applied to synthetic drawing images as a tool for supporting Alzheimer's disease prediction. Mach Vis Appl 33:49
4. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-Gonzalez J, Routlier A, Bottani S, Initiative ADN (2020) Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med Image Anal 63:101694
5. Altinkaya E, Polat K, Barakli B (2020) Detection of Alzheimer's disease and dementia states based on deep learning from MRI images: a comprehensive review. J Inst Electron Comp 1(1):39–53
6. Physicians PC (2020) Alzheimer's disease facts and figures. Alzheimer's Dementia 16(3):391–460
7. Yang Y, Li X, Wang P, Xia Y, Ye Q (2020) Multi-Source transfer learning via ensemble approach for initial diagnosis of Alzheimer's disease. IEEE J Transl Eng Health Med 8:1–10
8. Wang X, Zhen X, Li Q, Shen D, Huang H (2018) Cognitive assessment prediction in Alzheimer's disease by multi-layer multi-target regression. Neuroinformatics 16(3–4):285–294
9. Nanni L, Brahnam S, Salvatore C, Castiglioni I, Initiative ADN (2019) Texture descriptors and voxels for the early diagnosis of Alzheimer's disease. Artif Intell Med 97:19–26
10. Adelina C (2019) The costs of Dementia: advocacy, media, and stigma. Alzheimer's Dis Int World Alzheimer Rep., pp 100–1.
11. Pulido MLB, Hernández JBA, Ballester MAF, González CMT, Mekyska J, Smékal Z (2020) Alzheimer's disease and automatic speech analysis: a review. Expert Syst Appl 150:113213
12. Irankhah E (2020) Evaluation of early detection methods for Alzheimer's disease. Bioprocess Eng 4(1):17–22
13. He Y et al (2007) Regional coherence changes in the early stages of Alzheimer's disease: a combined structural and resting-state functional MRI study. Neuroimage 35(2):488–500
14. Vemuri P, Jones DT, Jack CR (2012) Resting-state functional MRI in Alzheimer's disease. Alzheimer's Res Therapy 4(1):1–9
15. Bron EE, Smits M, Van Der Flier WM, Vrenken H, Barkhof F, Scheltens P, Initiative ADN (2015) Standardized evaluation of algorithms for computer-aided diagnosis of Dementia based on structural MRI: the CAD Dementia challenge. Neuroimage 111:562–579
16. Allioui H, Sadgal M, Elfazziki A (2020) Utilization of a convolutional method for Alzheimer disease diagnosis. Mach Vis Appl 31:25
17. Segato A, Marzullo A, Calimeri F, De Momi E (2020) Artificial intelligence for brain diseases: a systematic review. APL Bioeng 4(4):041503
18. Yamanakkanavar N, Choi JY, Lee B (2020) MRI segmentation and classification of the human brain using deep learning for diagnosis of Alzheimer's disease: a survey. Sensors (Switzerland) 20(11):1–31
19. Noor MBT, Zenia NZ, Kaiser MS, Al Mamun S, Mahmud M (2020) Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's disease. Parkinson's disease, and schizophrenia. Brain Informatics 7(1):11
20. Li F, Tran L, Thung K-H, Ji S, Shen D, Li J (2015) A robust deep model for improved classification of AD/MCI patients. IEEE J Biomed Heal informatics 19(5):1610–1616
21. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. (2018) Deep learning for computer vision: a brief review. Computational Intell Neurosci, pp 1–13.
22. Goyal P., Rani R., & Singh K. (2021) State-of-the-art machine learning techniques for diagnosis of Alzheimer's disease from MR-images: a systematic review. Arch Comput Methods Eng, 1–44.
23. Maqsood M, Nazir F, Khan U, Aadil F, Jamal H, Mehmood I, Song OY (2019) Transfer learning assisted classification and detection of Alzheimer's disease stages using 3D MRI scans. Sensors 19(11):2645
24. Yan B, Li Y, Li L, Yang X, Li TQ, Yang G, Jiang M (2022) Quantifying the impact of pyramid squeeze attention mechanism and filtering approaches on Alzheimer's disease classification. Comput Biol Med 148:105944
25. Zhu J, Tan Y, Lin R et al (2022) Efficient self-attention mechanism and structural distilling model for Alzheimer's disease diagnosis. Comput Biol Med 147:105737
26. Li J, Wei Y, Wang C, Hu Q, Liu Y, Xu L (2022) 3-D CNN-based multichannel contrastive learning for Alzheimer's disease automatic diagnosis. IEEE Trans Instrum Meas 71:1–11
27. Divya R, Shantha Selva Kumari R, Alzheimer's Disease Neuroimaging Initiative (2021) Genetic algorithm with logistic regression feature selection for Alzheimer's disease classification. Neural Comput Appl 33(14):8435–8444
28. Kang W, Lin L, Zhang B, Shen X, Wu S, Initiative ADN (2021) Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis. Comput Biol Med 136:104678
29. Feng J, Zhang SW, Chen L, Xia J, Alzheimer's Disease Neuroimaging Initiative (2021) Alzheimer's disease classification using features extracted from nonsubsampled contourlet subband-based individual networks. Neurocomputing 421:260–272
30. Jain R, Jain N, Aggarwal A, Hemanth DJ (2019) Convolutional neural network-based Alzheimer's disease classification from magnetic resonance brain images. Cogn Syst Res 57:147–159
31. Basheera S, Ram MS (2019) Convolution neural network-based Alzheimer's disease classification using hybrid enhanced independent component analysis based segmented gray matter of T2 weighted magnetic resonance imaging with clinical valuation. Alzheimer's Dementia: Transl Res Clin Intervent 5(1):974–986
32. Choi H, Jin KH (2018) Predicting cognitive decline with deep learning of brain metabolism and Amyloid imaging. Behav Brain Res 344:103–109

33. Liu M, Cheng D, Wang K, Wang Y (2018) Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. Neuroinformatics, pp 1–14.
34. Shi J, Zheng X, Li Y, Zhang Q, Ying S (2018) Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J Biomed Health Inform 22(1):173–183
35. Korolev S, Safiullin A, Belyaev M, Dodonova Y (2017) Residual and plain convolutional neural networks for 3d brain MRI classification. In: IEEE 14th International Symposium Biomedical Imaging, pp 835–838.
36. Liu M, Zhang D, Adeli E, Shen D (2016) Inherent structure-based Multiview learning with multi-template feature representation for Alzheimer's disease diagnosis. IEEE Trans Biomed Eng 63(7):1473–1482
37. Zu C, Jie B, Liu M, Chen S, Shen D, Zhang D, the ADNI (2016) Label-aligned multitask feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment. Brain Imaging Behavior 10(4):1148–1159
38. Ortiz A, Munilla J, Gorriz JM, Ramirez J (2016) Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. Int J Neural Syst 26(07):1650025
39. Li F, Tran L, Thung K-H, Ji S, Shen D, Li J (2015) A robust deep model for improved classification of AD/MCI patients. IEEE J Biomed Health Inform 19(5):1610–1616
40. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, Feng D, Fulham MJ (2015) Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans Biomed Eng 62(4):1132–1140
41. Payan A., Montana G (2015) Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks, arXiv Prepr. arxiv.org/abs/1502.02506.
42. Suk H-I, Shen D (2013) Deep learning-based feature representation for AD/MCI classification. In: International conference on medical image computing and computer-assisted intervention, pp 583–590.
43. Suk HI, Lee SW, Shen D, ADNI (2015) Deep sparse multitask learning for feature selection in Alzheimer's disease diagnosis. Brain Struct Funct 221:1–19
44. Zhu X, Suk H-I, Shen D (2014) A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. Neuroimage 100:91–105
45. Suk H-I, Lee S-W, Shen D, ADNI (2014) Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. Neuroimage 101:569–582
46. Ebrahimighahnavieh MA, Luo S, Chiong R (2020) Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. Comput Methods Programs Biomed 187:105242
47. Yamashita R, Nishio M (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9:611–629
48. Shams S, Platania R et al (2018) Deep generative breast cancer screening and diagnosis. Springer Nature, Berlin
49. Wang D, Lu Z et al (2019) Cellular structure image classification with small targeted training samples. IEEE Access 7:148967–148974
50. Weng Y, Zhou H (2019) Data augmentation computing model based on generative adversarial network. IEEE Access.
51. Kokil P, Sudharson S (2019) Automatic detection of renal abnormalities by Off-the-shelf CNN features. IETE J Educ 60:14–23
52. Saranyaraj D, Manikandan M, Maheswari S (2018) A deep convolutional neural network for the early detection of breast carcinoma concerning hyperparameter tuning. Multimedia Tools Appl 79(15):11013–11038
53. Talo M, Baloglu UV, Yıldırım O, Acharya UR (2018) Application of deep transfer learning for automated brain abnormality classification using MR Images. Cogn Syst Res 54:176–188
54. Kaur T, Gandhi TK (2020) Deep convolutional neural networks with transfer learning for automated brain image classification. Mach Vis Appl 31(3):1–16
55. Wang SH, Xie S et al (2019) Alcoholism identification based on an AlexNet transfer learning model. Front Psych 10:205
56. Sakr GE, Mokbel M et al (2016) Comparing deep learning and support vector machines for autonomous waste sorting. In: IEEE international multidisciplinary conference on engineering technology, pp 207–212.
57. Deng J, Dong W et al (2009) Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp 248–255.
58. Shin HC, Roth HR et al (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35(5):1285–1298
59. Loddo A, Buttau S, Ruberto CD (2022) Deep learning based pipelines for Alzheimer's disease diagnosis: a comparative study and a novel deep-ensemble method. Comput Biol Med 141:105032
60. Mateen M, Wen J, Song S, Huang Z (2018) Fundus image classification using VGG-19 architecture with PCA and SVD. Symmetry 11(1):1
61. Ghaffari H, Tavakoli H, Pirzad Jahromi G (2022) Deep transfer learning–based fully automated detection and classification of Alzheimer's disease on brain MRI. Brit J Radiol, 20211253.
62. Foody GM (2004) Supervised image classification by MLP and RBF neural networks with and without an exhaustively defined set of classes. Int J Remote Sens 25(15):3091–3104
63. Lai Z, Deng H (2018) Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron. Comput Intell Neurosci 2018:2061516
64. Naeem H, Bin-Salem AA (2021) A CNN-LSTM network with multi-level feature extraction-based approach for automated detection of coronavirus from CT scan and X-ray images. Appl Soft Comput 113:107918
65. Liu T, Bao J, Wang J, Zhang Y (2018) A hybrid CNN–LSTM algorithm for online defect recognition of CO2 welding. Sensors 18(12):4369
66. Tsironi E, Barros P, Weber C, Wermter S (2017) An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition. Neurocomputing 268:76–86
67. Oehmcke S, Zielinski O, Kramer O (2018) Input quality aware convolutional LSTM networks for virtual marine sensors. Neurocomputing 275:2603–2615
68. Zhao R, Yan R, Wang J, Mao K (2017) Learning to monitor machine health with convolutional bi-directional LSTM networks. Sensors 17(2):273
69. Nunez JC, Cabido R, Pantrigo JJ, Montemayor AS, Velez JF (2018) Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recogn 76:80–94
70. Isaac ER (2015) Test of hypothesis-concise formula summary. Anna University, Tamil Nadu, pp 1–5