A Comparison Between Modern Day CNNs and Data Augmentation Techniques in Alzheimer's Disease 4-stage Classification

Akathian Santhakumar
Computer Science Graduate Department
Toronto Metropolitan University
Toronto, Canada
asanthakumar@torontomu.ca

***Abstract*—Alzheimer's disease (AD) is a progressive neurodegenerative disorder, with no cure. It is often beneficial to diagnose the disease early to slow its progression and mitigate symptoms. Good clinicians and medical imaging are often required for rapid diagnosis of the disease. Recent studies have attempted to apply AI techniques to classify brain images into different AD stages. An ensemble deep CNN was proposed, showing better performance than other state-of-the-art models. The following study investigates its performance against other popular models on an AD dataset available on Kaggle, in addition to investigating different data augmentation techniques, including Gabor-based data augmentation. We find that traditional data augmentation techniques hinder model learning, but adding Gabor filtered images improve classification. We suggest the use of Gabor filtering for data augmentation in future AD studies. We also find that the ensemble model underperforms on this dataset, compared to the other models investigated.**

Keywords—Alzheimer's Disease, CNN, Gabor Filter, Ensemble, Kaggle, Data Augmentation, Multiclass Classification

I. Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder affecting cells in the brain [1]. It is the primary cause of dementia and is characterized by amyloid plaques and neurofibrillary tangles [1]. These changes can be observed on a microscopic level and are also visible on brain imaging. In these images, a brain with AD might have a smaller cerebral cortex and hippocampus region, with enlarged ventricles [1]. Although there is no cure for AD, its progression can be slowed with appropriate treatment. It is therefore important to diagnose the disease as soon as possible, and as accurately as possible.

There have been four identified clinical phases of AD: pre-clinical, mild or early stage, moderate, and severe [1]. Physiologically, AD is easier to diagnose in imaging at the later stages. Magnetic resonance imaging (MRI) has been key in the diagnosis of AD. Previous research has found significantly reduced hippocampal and entorhinal cortex volumes in AD patients from structural MRI imaging [2].

With the rise of AI in the field of medical imaging, diagnosis of AD with MRI images have been of recent interest, especially given the physiological differences observed in MRI in between the clinical stages of AD. Approaches for diagnosis of AD with AI have either been machine learning (ML) or deep learning (DL) techniques [3].

Some traditional ML techniques used in AD diagnosis via brain imaging include Support Vector Machines (SVM) [4]. An ensemble model of SVM with radial basis function and linear kernels achieved around 55% accuracy in a multi-class classification problem of predicting AD clinical stage based on structural MRI imaging [4]. The data used in the paper consisted of 60 images from each clinical stage of AD. Another paper investigates a multistage classifier using a combination of a Naive Bayes classifier, an SVM and a K-nearest neighbor classifier [5]. Using around 150 samples of MRI images augmented with neurological examination data for each class, the model was able to classify data into 3 stages of AD with a combined accuracy of 90%.

In the history of AD diagnosis research via AI, the idea of Gabor filters was investigated. Gabor filters may be used to extract features from images, or simply highlight the most relevant features [10]. Gabor filters are simply linear filters, that can be applied to images. It has many hyperparameters, all of which must be manipulated to achieve desired results. One specific paper uses Gabor filters on all coronal, axial and sagittal views of the brain to extract information about the hippocampal region. This information was then fed into an SVM, resulting in an accuracy of 80%

More recent DL techniques in AD classification mostly use convolutional neural networks (CNN) at some basic level. Different architectures involve different arrangement of CNN layers, in-between other traditional neural network layers such as dense layers or normalization layers [1]. Comparing the performance of these models is difficult, due to differences in input data types, classification labels and evaluation metrics, among others [7]. There are many papers investigating the performance of various 2D CNNs (CNNs using 2D images) architectures in AD classification.

Notably in one paper, researchers propose an ensemble system of deep CNNs to classify 2D MRI images into 4 stages of AD [8]. They use the Open Access Series of Imaging Studies (OASIS) dataset [9] which contains structural MRI (sMRI) scans of 416 subjects, each of them having 3-4 images. From the 416 subjects, 100 patients have either mild or moderate AD. Each images contained axial, coronal, and sagittal views of the brain. The model they described reached an accuracy of 93%, which outperformed state-of-the-art image classification models that were fine-tuned on the AD imaging data. State-of-the-art CNNs used in image classification include ResNet101, DenseNet169, InceptionV3, MobileNetV2 and VGG19. Given their performance, much research has gone into fine-tuning these models onto medical images for diagnosis of the disease of interest.

The following paper investigates the performance of such state-of-the-art models and the ensemble deep CNN model on the Alzheimer's Dataset found on Kaggle. We expect similar results as in [9], as we have the same classes and are using similar models to classify the images. We also compare the effects of traditional data augmentation and augmentation with Gabor filters.

II. METHODS

A. Environment

We use the Pro version of Google's Colaboratory, which provides access to either NVIDIA V100 or A100 cards, based on availability. It also provides RAM access up to 52GB, and an Intel Xeon CPU @ 2.30GHz.

B. Data

The dataset can be found on Kaggle for free. It is comprised of 6400 axial MRI images. The images are pre-separated into train and test sets. The data has 4 classes: mild demented (MiD), moderate demented (MoD), non demented (ND), and very mild demented (VMD). The test set is split 179 MiD :12 MoD :640 ND :448 VMD, while the training set is split 717 MiD: 52 MoD: 2560 ND: 1792 VMD. These 4 classes correspond to the 4 clinical stages of AD. The dataset does not have much background information; source, number of patients and axial plane number among others are missing. Images are png's of size 176x208.

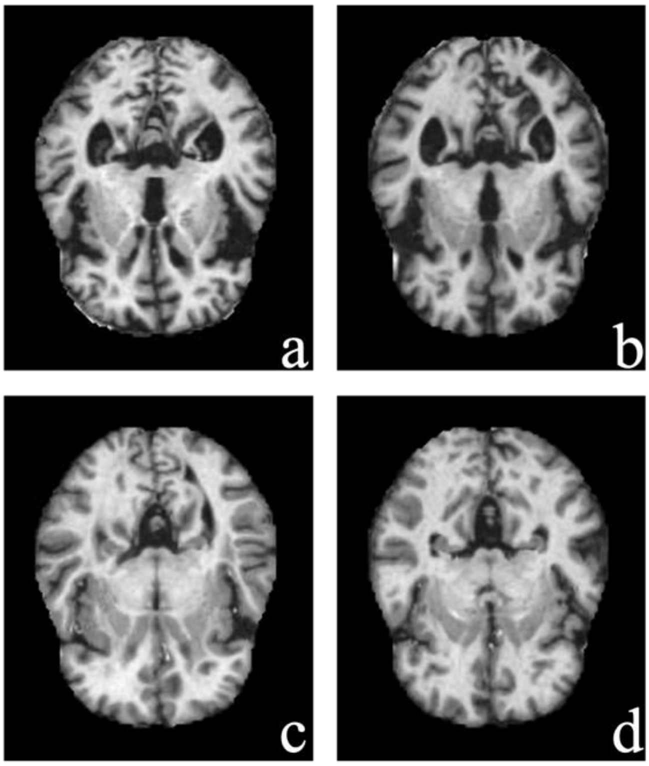

Fig 1. Example images from dataset. Axial MRI image of (a) MiD brain. (b) MoD brain. (c) ND brain. (d) VMD brain.

C. Data Loading

We load the data using TensorFlow's ImageDataGenerator module. The data is scaled by a factor of 1/255 and resized to 224x224. Twenty percent of the training data was used as a validation set.

D. Data Augmentation

We follow (11)'s data augmentation techniques. The researchers in the paper use a combination of traditional data augmentation, Gabor filter augmentation, and GAN-based augmentation. Due to time constraints, we were restricted to traditional and Gabor filter-based data augmentations. We implement traditional data augmentation using TensorFlow's ImageDataGenerator module, again. We allow for horizontal & vertical flips, translations of up to 20%, rotations of up to 5 degrees, and crops of up to 20%. During training, a random combination of these augmentations is presented. Thus, for each epoch, we are seeing much more data than we are provided with. We also augment the data by running a Gabor filter on the training data, using CV2's getGaborKernel function. We optimize the Gabor parameters by doing a hyperparameter search with a traditional CNN. We use 30 for the kernel size, 1 for sigma, a combination of 16 different theta values from 0 to π, 14 for lambda, 0.8 for gamma and 0.8 for psi. We run the images through the Gabor filter to create filtered images, effectively doubling the number of training instances (original images + Gabor images).

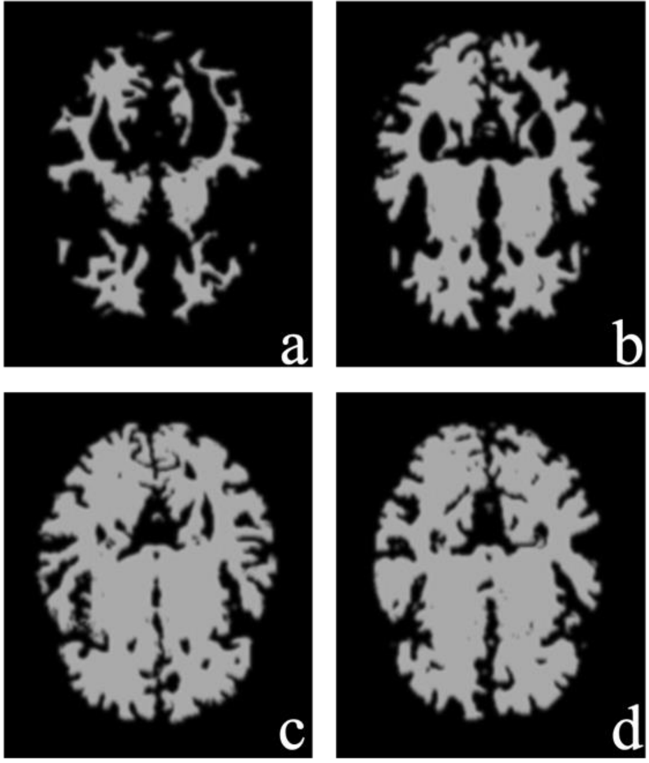

Fig 2. Gabor filtered images. Axial MRI image of (a) MiD brain. (b) MoD brain. (c) ND brain. (d) VMD brain.

To assess the necessity of data augmentation, we start by running three experiments with the ensemble deep CNN model. We run the first experiment without any data augmentation. For the second experiment, we run the model against augmented data, as described earlier. For the final experiment, we generate Gabor images on the regular data, and decided whether to use traditionally augmentation techniques based on the results of the previous two experiments.

E. *Network Architectures*

We load ResNet101, DenseNet169, InceptionV3, MobileNetV2 and VGG19 models from TensorFlow's application module. We build the ensemble deep CNN model from TensorFlow's layers library, as described in the original paper (Fig 3.). We connect convolution blocks, dense blocks, and transition layers to recreate the model. Convolution blocks are made up of a Conv2D layer (filter size of 12, kernel size 7, stride 2 and padding with zeros), a BatchNormalization layer, a ReLU layer, and a MaxPool2D layer. Dense blocks are made up of *n* number of repeats of 2 Conv2D layers, the first with a kernel of size 1, and the second with a kernel of size 3. Transition layers are made up of a BatchNormalization layer, a Conv2D layer with a kernel size of 1 and a 2x2 stride, and an AveragePooling2D layer with a pool size of 2x2 and a stride of 2. Three such models were created, with varying repeats in their dense blocks. Each model had 4 dense blocks. Model 1 had (6, 12, 24, 16) repeats, model 2 had (6, 12, 32, 32) repeats, and model 3 had (6, 12, 36, 24) repeats. Each model was allowed to make its own prediction, but the majority vote was counted as the final classification.

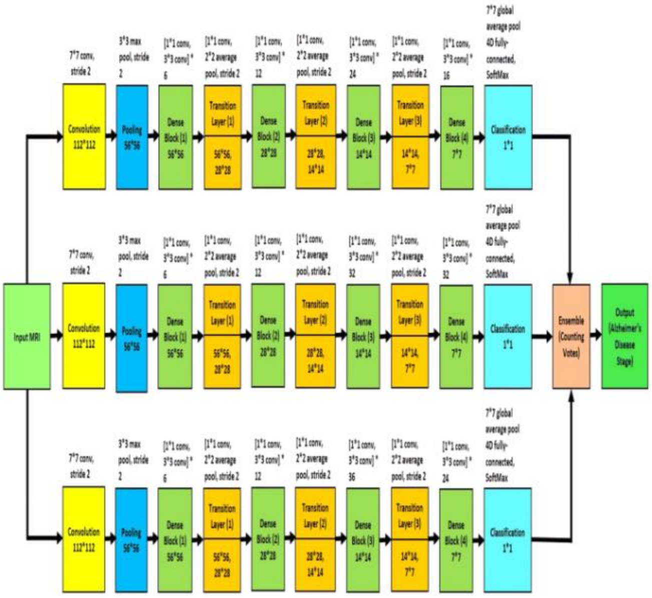
7*7 conv, stride 2
3*3 max pool, stride 2
[1*1 conv, 3*3 conv] * 6
[1*1 conv, 2*2 average pool, stride 2
[1*1 conv, 3*3 conv] * 12
[1*1 conv, 2*2 average pool, stride 2
[1*1 conv, 3*3 conv] * 24
[1*1 conv, 2*2 average pool, stride 2
[1*1 conv, 3*3 conv] * 16
7*7 global average pool 4D fully-connected, SoftMax
Convolution 112*112
Pooling 56*56
Dense Block (1) 56*56
Transition Layer (1) 56*56, 28*28
Dense Block (2) 28*28
Transition Layer (2) 28*28, 14*14
Dense Block (3) 14*14
Transition Layer (3) 14*14, 7*7
Dense Block (4) 7*7
Classification 1*1
[1*1 conv, 3*3 conv] * 32
[1*1 conv, 3*3 conv] * 32
Input MRI
Ensemble (Counting Votes)
Output (Alzheimer's Disease Stage)
[1*1 conv, 3*3 conv] * 36
[1*1 conv, 3*3 conv] * 24

Fig 3. Diagram of ensemble of deep CNNs

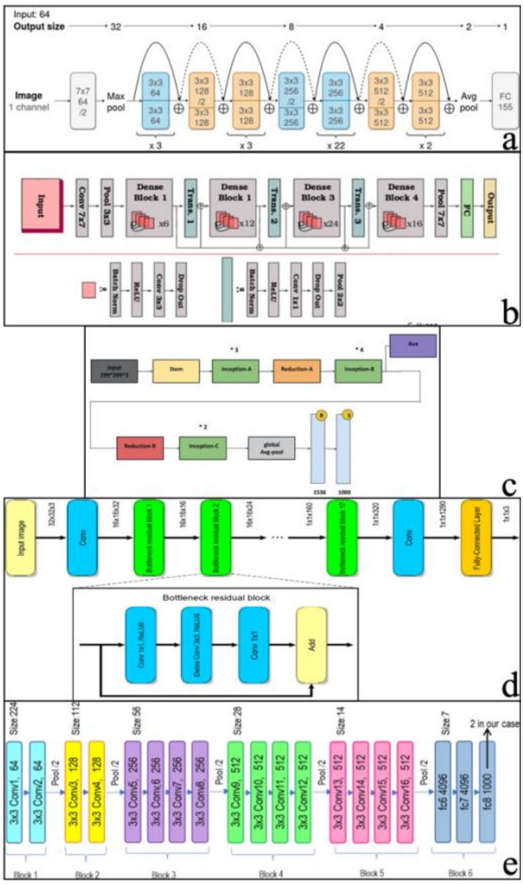

Fig 4. Diagrams of state-of-the-art models (a) ResNet101 (b) DenseNet169 (c) InceptionV3 (d) MobileNetV2 (e) VGG19

F. *Model Training*

The models were compiled with TensorFlow's stochastic gradient descent (SGD) optimizer, with a learning rate of 0.01 and a Nesterov momentum of 0.9. The loss to be measured was the categorical cross entropy loss. A learning rate scheduler was also used, based on an exponential decay function. The accuracy, AUC, precision, and recall for training and validation sets were measured during training. Given the nature of the problem, we focus mostly on the recall score, since we want to reduce the number of false negatives. All models were set to train for 50 epochs but allowed for early stopping the model training after 10 epochs of no significant improvement in loss. Batch size was set to 8.

G. *Model Testing*

In terms of model testing, we use the evaluate method built-in the TensorFlow model with the testing dataset.

III. Results

Following the three experiments described in the methods section to assess the necessity of data augmentation, we find that the model performs best overall on the testing set without traditional data augmentation (Fig 5). Although the AUC and precision scores for the model that used the traditionally augmented data are higher, the model that did not use augmented data scored much better recall. We decide to drop the use of traditional data augmentation as a result. Comparing the non-augmented data against data augmented with Gabor filters, the performances are comparable, with the AUC of the Gabor augmented model being slightly lower. We choose to continue with augmenting the data with Gabor filtering following the marginal gain in recall.

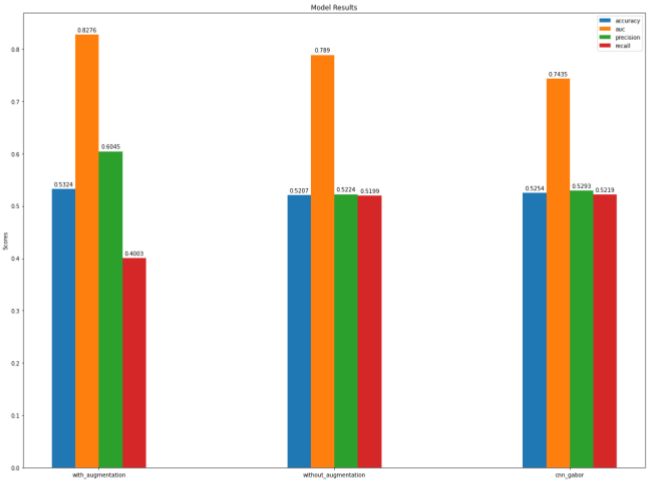

Fig 5. Comparison of accuracy, AUC, precision and recall on the test data between different data augmentation techniques, using the ensemble deep CNN model.

Given this finding, we move on to fine-tune the state-of-the-art models using the same training set augmented with Gabor filtered brain images. These models had greater loss values during training than the ensemble deep CNN model had (Fig 6.). Most models reached a near-perfect accuracy and AUC at the end of training, with precision and recall values nearing 1. The only model that fell short was the ensemble deep CNN model that whose input data was augmented using traditional methods. The VGG19 model showed the best performance when evaluating with the validation dataset after each epoch (Fig 6.).

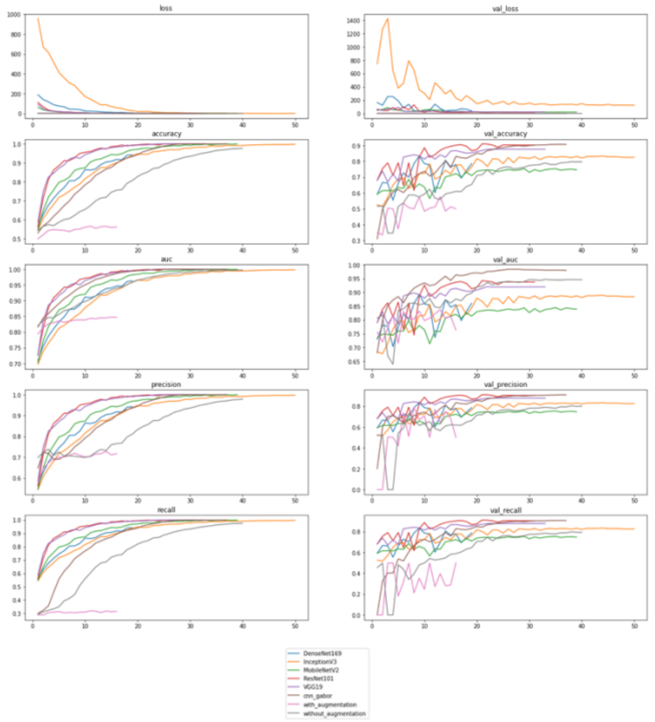

Fig 6. Changes in loss, accuracy, AUC, precision and recall on the training and validation datasets during training over a maximum of 50 epochs.

Following testing of these models, it was found that DenseNet169 had the best overall performance in all metrics, with ResNet101 having the second-best performance (Fig 7).

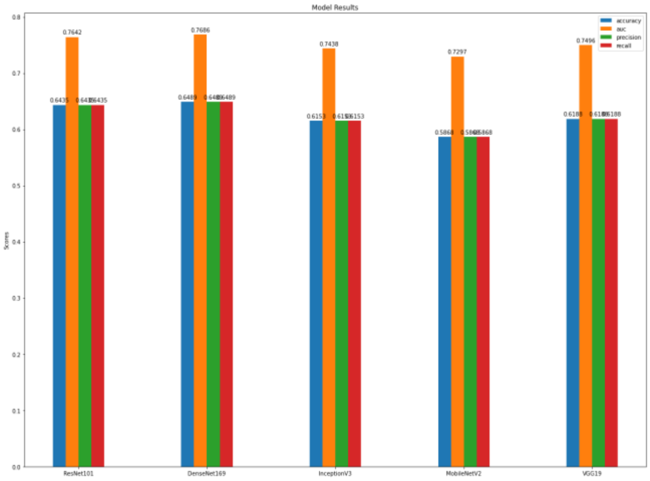

Fig 7. Comparison of accuracy, AUC, precision and recall on the test data between different state-of-the-art models

Comparing the ensemble deep CNN and DenseNet169, we observe the vastly superior performance of DenseNet169 in this task. We report the ensemble deep CNN to have a test accuracy of 52.5%, an AUC of 0.74, a precision of 0.53 and a recall of 0.52. The DenseNet169 model was found to have an accuracy of 64.9%, an AUC of 0.77, a precision of 0.65 and a recall of 0.65 as well.

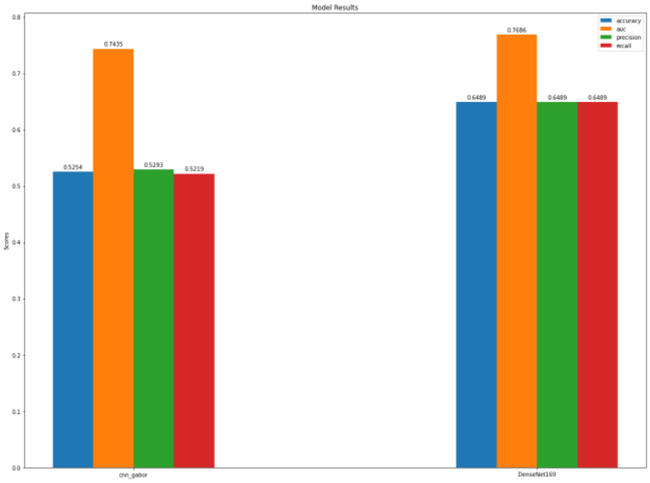

Fig 8. Comparison of accuracy, AUC, precision and recall on the test data between the ensemble deep CNN and Densenet169.

IV. Discussion

The original authors of the ensemble deep CNN model [9] report better performance of this model over InceptionV4 [11], ResNet [12] and ADNet [13]. The authors explain that the observed better performance was due to the fact that their model was not as deep as the aforementioned models. It has been previously shown that deeper models need larger datasets in order to train correctly. In the current paper, we observe the opposite. It is possible that our dataset (6400 images), which is much larger than the dataset used in [9] (OASIS [8]) was enough to train the other models sufficiently to be able to outperform the model described in [9]. Our current dataset is also much more imbalanced that the OASIS dataset [8]. We have very little number of instances of MoD brains, which may affect the recall and precision scores. Such class imbalance can lead to poorer model performance with shallower architectures [14]. The result of poorer performance with data augmentation was not expected. We observed the model having a worse recall score, which we do not want when solving a medical diagnosing problem. The Gabor filters worked well in improving metrics across all models, but given more time, even better hyperparameters could have been explored in order to further improve metrics.

Further directions of this study can involve finding a good Gabor kernel (better hyperparameters) such that more relevant features are highlighted in the resulting images. GAN-based data augmentation should also be researched in AD brain images, whether it aids the performance of the investigated models.

V. References

[1] Breijyeh, Z., & Karaman, R. (2020). Comprehensive review on Alzheimer’s disease: Causes and treatment. *Molecules*, *25*(24), 5789. https://doi.org/10.3390/molecules25245789

[2] J Chandra, A., Dervenoulas, G., & Politis, M. (2018). Magnetic resonance imaging in Alzheimer’s disease and mild cognitive impairment. *Journal of Neurology*, *266*[6], 1293-1302. https://doi.org/10.1007/s00415-018-9016-3

[3] Shastry, K. A., Vijayakumar, V., V, M. K., B A, M., & B N, C. (2022). Deep learning techniques for the effective prediction of Alzheimer’s disease: A comprehensive review. *Healthcare*, *10*[10], 1842. https://doi.org/10.3390/healthcare10101842

[4] Sørensen, L., & Nielsen, M. (2018). Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination. *Journal of Neuroscience Methods*, *302*, 66-74. https://doi.org/10.1016/j.jneumeth.2018.01.003

[5] Kruthika, K., Rajeswari, & Maheshappa, H. (2019). Multistage classifier-based approach for Alzheimer's disease prediction and retrieval. *Informatics in Medicine Unlocked*, *14*, 34-42. https://doi.org/10.1016/j.imu.2018.12.003

[6] Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., & Colliot, O. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, *63*, 101694. https://doi.org/10.1016/j.media.2020.101694

[7] Islam, J., & Zhang, Y. (2018). Brain MRI analysis for Alzheimer’s disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain Informatics*, *5*[2]. https://doi.org/10.1186/s40708-018-0080-3

[8] Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, Nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, *19*[9], 1498-1507. https://doi.org/10.1162/jocn.2007.19.9.1498

[9] Keserwani, P., Pammi, V. S., Prakash, O., Khare, A., & Jeon, M. (2016). Classification of Alzheimer disease using Gabor texture feature of hippocampus region. *International Journal of Image, Graphics and Signal Processing*, *8*(6), 13-20. https://doi.org/10.5815/ijigsp.2016.06.02

[10] Barshooi, A. H., & Amirkhani, A. (2022). A novel data augmentation based on Gabor filter and convolutional deep learning for improving the classification of COVID-19 chest X-ray images. *Biomedical Signal Processing and Control*, *72*, 103326. https://doi.org/10.1016/j.bspc.2021.103326

[11] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-V4, inception-resnet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.11231

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.90

[13] Islam, J., & Zhang, Y. (2017). A novel deep learning based multi-class classification method for Alzheimer’s disease detection using brain MRI data. *Brain Informatics*, 213-222. https://doi.org/10.1007/978-3-319-70772-3_20

[14] Harliman, R., & Uchida, K. (2018). Data- and algorithm-hybrid approach for Imbalanced data problems in deep neural network. *International Journal of Machine Learning and Computing*, *8*(3), 208-213. https://doi.org/10.18178/ijmlc.2018.8.3.689