Alzheimer's Disease Classification Using Brain MRI Based on Combination of Convolutional Neural Network and Random Forest Model

Haoyu Chen*

Physical Sciences Division, University of Chicago, 455 N Cityfront Plaze Dr. Chicago, IL USA 60611

* Corresponding Author Email: hchen99@uchicago.edu

Abstract. Worldwide, someone develops dementia every 3 seconds. Dementia is mostly brought on by Alzheimer's disease (AD). Research has concentrated on diagnosing AD and dementia over the past centuries, and brain Magnetic Resonance Imaging (MRI) has been proved an effective biomarker of AD and other dementias. Throughout the years, many methods, including various forms of Neural Networks, Support Vector Machines, and other machine learning algorithms, have been innovated and applied to the classification of brain MRI scans. This paper aims to propose a model framework that has been rarely used in this field. This novel architecture utilizes a Convolutional Neural Network (CNN) for the feature extracting task and a Random Forest (RF) model for classifying different stages of dementias. The model was evaluated on each label's performance and overall performance. The performance metrics include accuracy, f-1 score, precision, and recall. The comparisons between the proposed model and the other two sodels, a CNN and a combination of Principal Component Analysis (PCA) and RF, were also provided. The implementation of the proposed model resulted in the highest overall accuracy, weighted precision, weighted recall, and weighted f-1 score. It also guaranteed stable and excellent performance across every label.

Keywords: Alzheimer's Disease, CNN, Random Forest, image classfication, dementia.

1. Introduction

Around 55 million people are living with dementia around the world [1], and the forecasted number of patients in 2050 is 152 million [2]. The main factor causing dementia is Alzheimer's Disease (AD), a degenerative and progressive disease worldwide that destroys human memory and other mental functions. About 60 to 80 percent of instances of dementia are caused by AD [3]. There are around 100 billion neurons in a healthy brain, and these neurons are connected by their branches. Neurons and their connections are essential to cognitive functions, including emotions, memories, and skills. Two critical brain abnormalities associated with AD are the phosphorylation of the tau protein inside brain neurons and the buildup of beta-amyloid protein fragments outside the neurons [3]. Tau protein is a microtubule-associated protein that helps maintain neuronal microtubules, which are essential for cell functions, polarity, and intracellular transport. Changes in tau protein may cause neuronal processes to retract, resulting in neuron death [4]. Accumulations of beta-amyloid may also lead to neuron death as it obstructs neuron communications at connections. The brain functions are further damaged as atrophy and inflammation occur due to cell loss [3]. Research has focused on using neuroimaging to aid the diagnosis of AD and other dementias. Correlated with tau deposition and cognitive deficits, MRI-detected brain atrophy or abnormalities is a reliable biomarker of AD and its development [5,6]. Accurate neuroimage classification is thus vital for patients and their families to determine which stages patients are in.

Over the years, deep learning algorithms have been used successfully in AD classification from brain MRI images. The ability to discriminate between AD and healthy aging participants using Support Vector Machine (SVM) has been demonstrated to be effective and generalizable across various centers [7]. Liu et al. combined stacked auto-encoders with a SoftMax output layer to frame a deep learning model that was able to achieve excellent performance [8]. Liang and Gu proposed a weak supervision deep learning framework called ADGNET that consists of an advanced backbone

network that leverages the attention mechanism and two sub-nets: a classification network and a reconstruction network [9]. Gray et al. designed a multi-modal classification architecture based on similarity measures extracted from Random Forest Classifiers, and the results are compatible with other multi-kernel learning studies [10]. Combinations of Principal Component Analysis (PCA) and machine learning models – Neural Networks (NN) and SVM classifiers - also have significant accuracy in brain image classification [11].

Aiming at the research blank of using extracted features from CNN on an RF classifier that could naturally extend to multi-class problems, this paper proposed a cascaded model framework that combines CNN and RF to classify brain MRI. The method was evaluated on 15,260 brain MRI scans separated into four classes. The proposed model was compared with a pure CNN model and a PCA+RF model. The experiment results of the proposed model showed great overall accuracy, precision, recall, and f-1 score, demonstrating promising image classification performance.

2. Data

2.1. Data source

This project used the Alzheimer's Dataset from Kaggle [12]. The dataset includes 6,400 brain MRI scans in total, and they are separated into four classes: Moderate Demented, Mild Demented, Very Mild Demented, and Non-Demented. The numbers of images in each class are 3200, 2240, 896, and 64 accordingly. The images were initially split into 5,121 training images and 1,279 testing images. However, in this project, the images were loaded into Python together and then re-split by a 7-3 train-test ratio.

2.2. Descriptive analysis

Figure 1(a) shows the sample images from the dataset, representing the four classes. The average pixel value for all images in each class was calculated to represent the mean images, as shown in Figure 1(b). The standard deviation pixel value for every image in a class was also obtained to reflect the variability of areas in the brain images. The lighter area of images in Figure 1(c) reflects higher variableness.

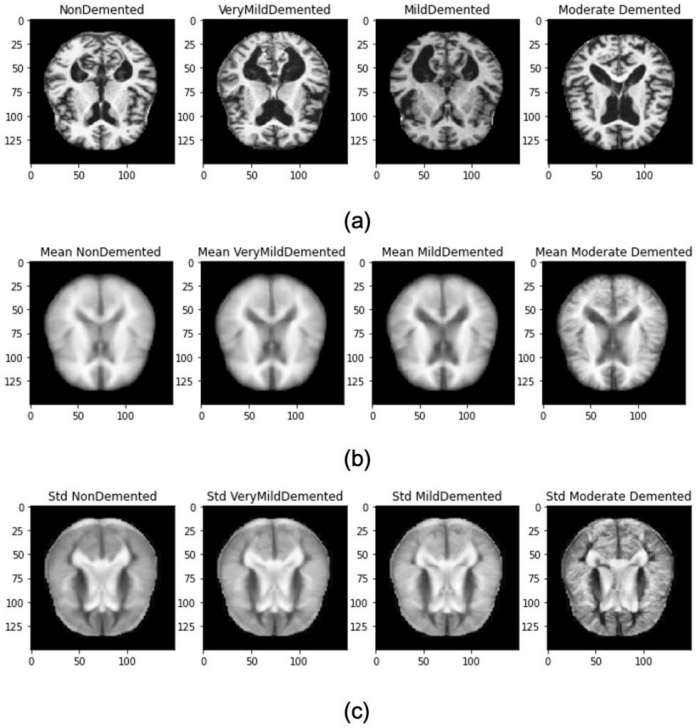


Figure 1. Images of each class in the dataset: (a) sample images (b) average images (c) standard deviation images

3. Methodalogy

3.1. Feature engineering

Each image in the training set was shifted up and down by one pixel individually to expand the dataset. The augmented training set contains 13,440 images in total. Figure 2 shows a sample of preprocessed data.

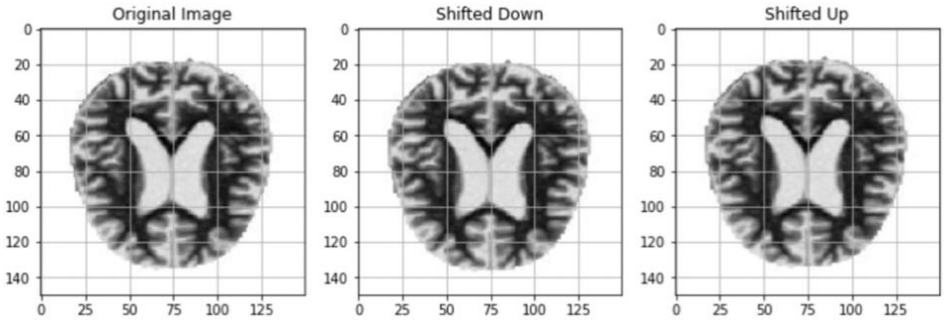


Figure 2. Sample images of data augmentation: original image, shited down image and shifted up image

3.2. Model framework

As a supervised machine learning algorithm, Random Forest has been widely used in regression and classification problems [13]. It is an collection of decision trees and takes the majority vote for classification problems. Each Decision Tree segregates the input data into branches, and these branches divide into other branches until a terminal node is reached. The Random Forest algorithm uses the bagging method for selecting sample data and features to build decision trees and give a weighted final decision. The model architecture is shown in Figure 3 [14].

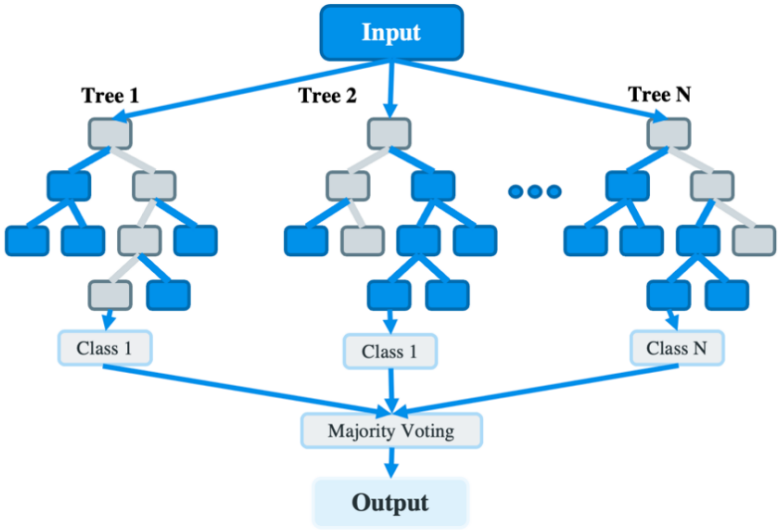


Figure 3. Typical random forest architecture

CNN is a deep learning algorithm typically consisting of convolutional layers, normalization layers, pooling layers, and fully connected layers. CNN is extensively used in Machine Learning and Computer Vision problems, showing excellent results. Different sets of features are included in each convolutional layer for image classification problems. After pooling and normalization, the fully connected layer functions as a classifier and makes the final decision [15]. This characteristic of convolutional layers is utilized in this paper. Without the fully connected layer, the remaining parts of the CNN model are a good candidate for feature extractor, including filtered traits for each class in a given dataset.

The proposed model framework consists of two components: a CNN was implemented on the feature extracting part, and an RF was utilized on the classification task, as shown in Figure 4. The feature extractor contains a 2D convolutional layer that slides over input data and returns the summation of elementwise multiplications, followed by a max-pooling layer and a dropout layer to down-sample and reduce feature dependencies in the network. The network is then flattened into one-dimensional outputs, which can be fed into the RF classifier. Leveraging the bagging methodology, the RF classifier can produce robust and stable classification decisions based on the CNN model's extracted features.

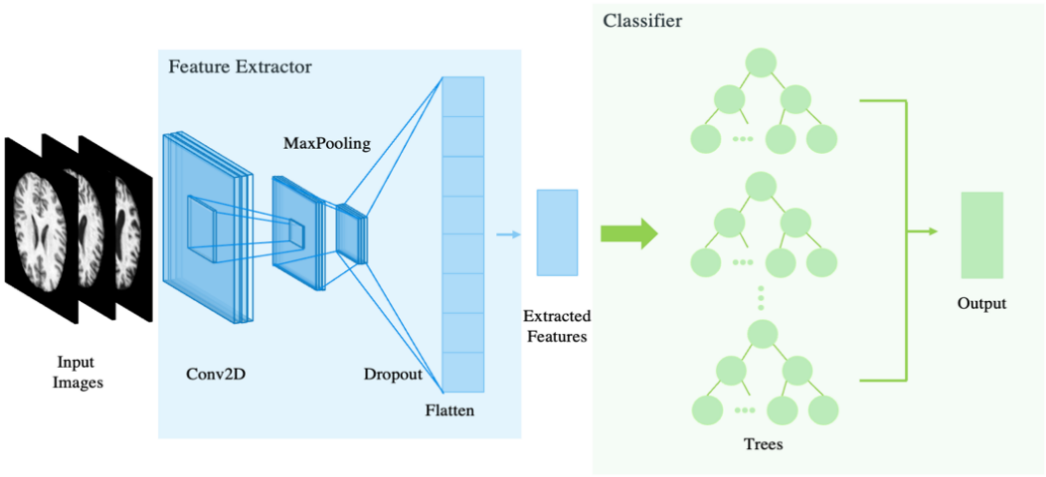


Figure 4. Convolutional Neural Network + Random Forest model framework

3.3. Performance metrics

The proposed model was compared to two baseline models: a CNN model and an RF model using PCA as dimensionality reduction. Moreover, each model was individually trained by the original and preprocessed data. The performance of these models was evaluated mainly on the overall accuracy. The weighted precision, recall, and f-1 score were also calculated. Besides, the study also reported each class's f-1 score, recall, and precision to see differences and variations in classifications of different labels. Table 1 is a confusion matrix that helps explain the definitions of these performance metrics.

Table 1. Confusion Matrix

Actual / Predicted	**Negative**	**Positive**
Negative	True Negative (TN)	False Negative (FN)
Positive	False Positive (FP)	True Positive (TP)

Accuracy is defined as shown in Equation (1).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN.}} \quad (1)$$

Precision is defined as shown in Equation (2).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

Recall is defined as shown in Equation (3)

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

F-1 score is defined as shown in Equation (4):

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}} \tag{4}$$

The weighted recall, precision, or f-1 score is defined as shown in Equation (5). L denotes the set of classes, N_1 denotes the number of the class l, and X_1 is the precision, recall, or f-1 score of the class l.

$$\text{Weighted X} = \frac{1}{\sum_{l\in L} N_l} \sum_{l\in L} N_l X_l \tag{5}$$

Accuracy is simply the ratio of corrected predicted cases to all cases. The precision of each class measures the percentage of true positive out of all the positive predicted, while the recall indicates how many of the total positive are precited to be positive. F-1 score is the harmonic average of recall and precision, which gives the same weights to each metric. Take the Non-Demented class as an example; the precision score tells that out of all the images that are predicted to be Non-Demented, how many of them are true. The recall of this class shows that out of all the Non-Demented images, what percentage are predicted as Non-Demented. The f-1 score is then calculated based on precision and recall, reflecting the average performance of these two scores.

4. Results

4.1. Overall accuracy

Almost all the models have shown excellent accuracy scores, as presented in Figure 5. The CNN+RF model trained using the augmented dataset had the best overall accuracy, with a value of 93.07%. This score means that out of the 1920 images in the test set, this model has correctly classified 1787 images. Each of the CNN+RF models outperformed the other two models using the same data, although the differences are not huge. Besides, models trained using the augmented dataset have shown promising improvement in every model's overall accuracy, and the most significant increase was from 84.64% to 90.42%.

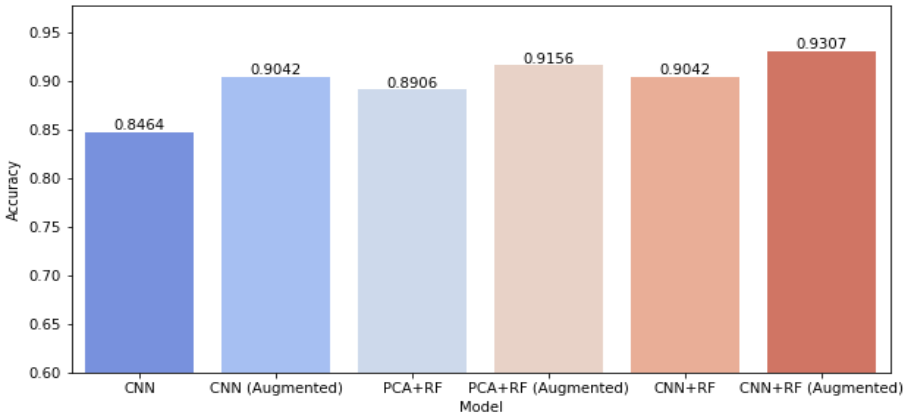


Figure 5. Overall accuracy of each model

4.2. Weighted Precision, Recall, and F-1 score

The weighted precision, accuracy, and f-1 score are shown in Figure 6. The proposed model had very balanced and outstanding performances on all three scores, which were all 93%. The PCA+RF model trained by augmented data had the same weighted precision as the proposed model, but its weighted recall and weighted f-1score were a little bit lower. The CNN+RF model using original data also showed good scores. Even though it did not have the highest scores, it was better than the other models except the PCA+RF model using augmented data and the CNN+RF model using augmented data. The lowest score of each metric came from the CNN model, which was 85%, 85%, and 84%, respectively. Data augmentation also generated positive impacts in weighted metrics, especially for the CNN model.

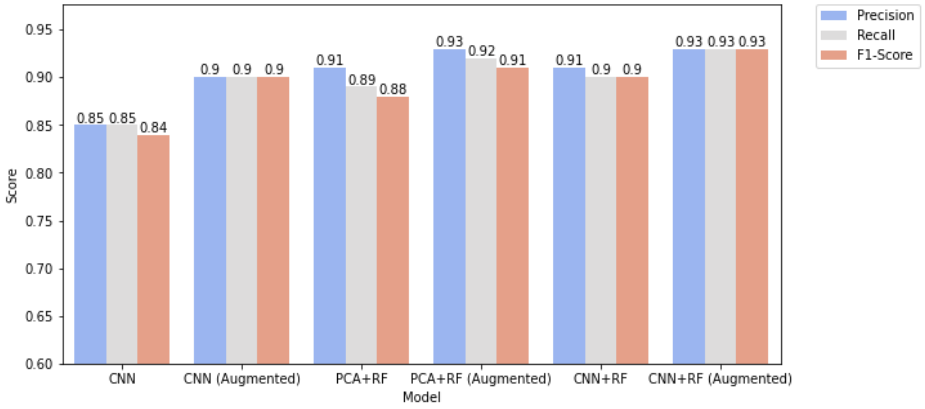


Figure 6. Weighted precision, weighted recall, and weighted f-1 score of each model

4.3. Confusion matrices

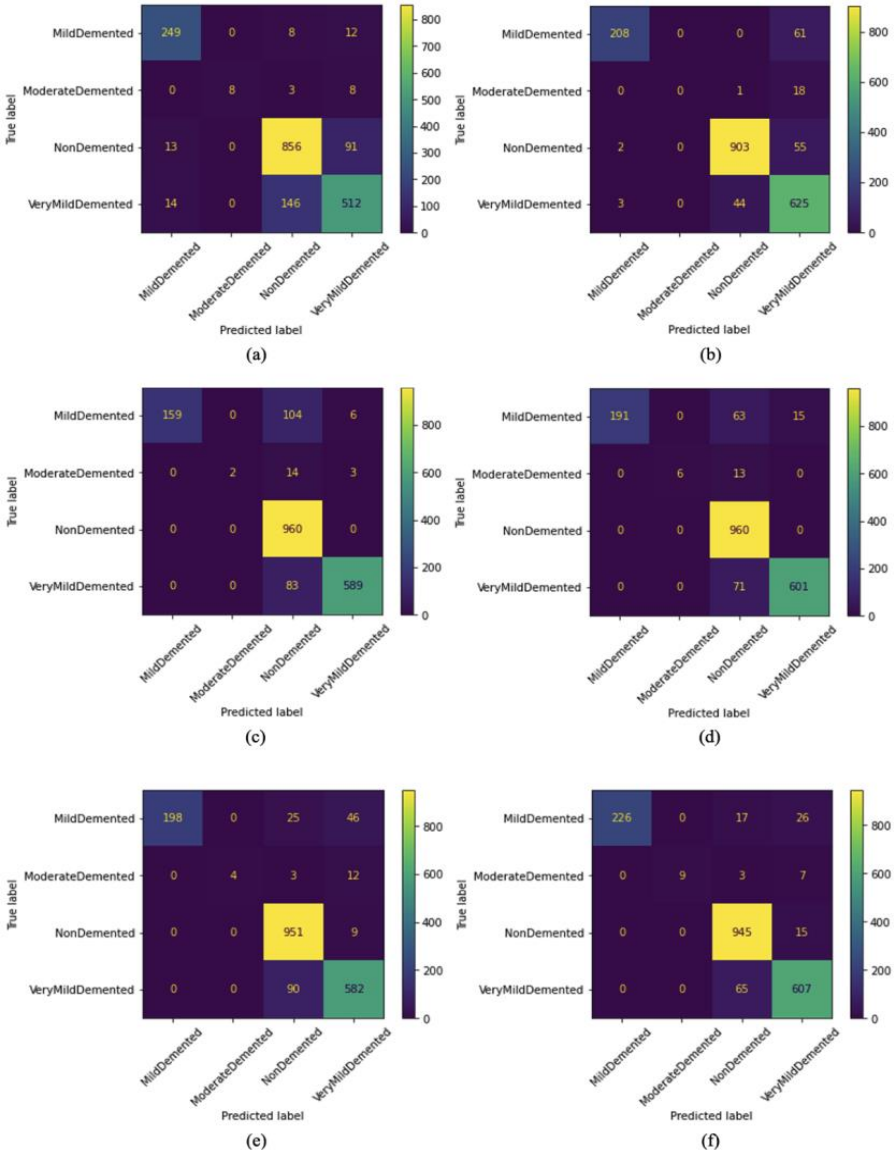


Figure 7. Confusion matrices of each model: (a) CNN, (b) CNN trained by augmented data, (c) PCA+RF, (d) PCA+RF trained by augmented data, (e) CNN+RF, (f) CNN+RF trained by augmented data

The confusion matrices in Figure 7 reflect that performance of the other three labels are better than that of Moderated Demented class, owing to the imbalanced number of classes. Using the augmented dataset has made the CNN model better in classifying Non-Demented and Very Mild Demented classes but worse in terms of Moderate Demented and Mild Demented classes. The PCA+ RF model using augmented data had more correctly predicted labels in Very Mild Demented, Mild Demented, and Moderate Demented classes compared to the PCA+RF model using the un-preprocessed data. The CNN + RF model using preprocessed test data had fewer correctly predicted Non-Demented labels than the original CNN+RF model. However, the former one had fewer misclassified Non Demented labels as well. The performances for the other three categories of the CNN+RF model also improved by fitting into the augmented dataset.

All three kinds of models had problems distinguishing between the Non-Demented images and Very Mild Demented ones. The PCA+RF models misclassified Very Mild Demented scans as Non-Demented; the CNN+RF models had the same problem, and several Non-Demented labels were labeled as Very Mild Demented; the CNN models have similar numbers of misclassification in both ways, and they also categorized some Non-Demented and Very Mild Demented images into Mild Demented labels. Another difficulty that all the models suffered was separating the Moderate Demented class from the Non-Demented and Very Mild Demented classes. Although the Moderate demented stage is closest to the mild demented stage, the models surprisingly have no trouble classifying between the Mild Demented and Moderate Demented classes. In general, the CNN and CNN+RF models had a more severe problem picking out Moderate Demented images from the Very Mild Demented ones. In contrast, the PCA+RF models have misclassified more Moderate Demented images as Non-Demented labels. The Mild Demented scans were often determined as Non-Demented or Very Mild Demented classes, and the problem is worst in PCA+RF models.

4.4. Precision, Recall, and F-1 score by Class

To better compare the performance of each label, Figure 8 consists of the plots of recall, precision, and f-1 score by class. For the Non-Demented class, PCA+RF models had perfect recall scores, meaning that they had successfully found all the Non-Demented images. However, as a trade-off, their precision scores are not very good, which means that they have misclassified relatively more images as Non-Demented. Both the CNN and CNN+RF models using augmented data have the best f-1 scores, which were 95%, so they maintained a better balance between precision and recall. The PCA+RF models did a good job in classifying the Very Mild Demented class, since they had the highest precision and f-1 scores. Even though the CNN model using augmented data had the highest number of Very Mild Demented images picked out from the test data, it mislabeled many other images as Very Mild Demented as well. The CNN+RF model using augmented data did not have the best scores in any of the metrics but had the second-best scores in all three metrics, which were 93%, 90%, and 91%.

In terms of the Mild Demented class, the CNN+RF model using augmented data and the CNN model had the highest f-1 score, which were 91%. The PCA+RF and CNN+RF models had 100% precision scores, so they did not misclassify any other image as Mild Demented, but the recall scores are low as a compromise. On the other hand, the CNN model had the highest recall. As for the Moderate Demented class, all the models' performance dropped beacues of data shortage. The CNN+RF model using augmented data was the absolute best since it had the highest scores in all three metrics, even though the recall and f-1 score were only 47% and 64%. All the models except the augmented CNN model had perfect precision scores, but there were significant drops in the recall scores.

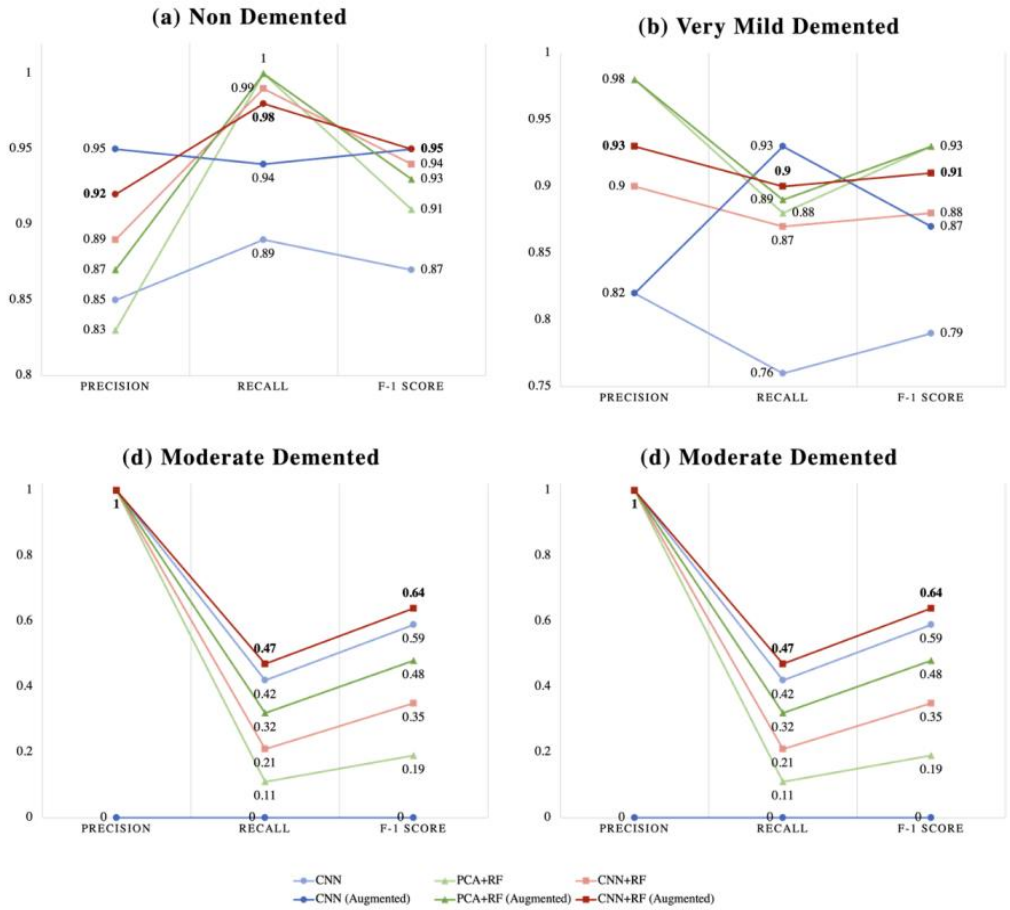


Figure 8. Precision, recall, and f-1 score of every model for each class: (a) Non-Demented, (b) Very Mild Demented, (c) Mild Demented, (d) Moderate Demented

5. Discussion

The proposed model, CNN+RF trained using augmented data, had the highest scores of overall accuracy, weighted recall, weighted precision, and weighted f-1 score. It also performed well in precision, recall, and f-1 score for each class. In particular, it outperformed the other models in classifying the class with the fewest samples, the Moderate Demented class, for all three metrics. The model had the best f-1 scores in the Non-Demented, the Mild Demented, and the Moderate Demented class, and it reached100% precisions for the Mild Demented and the Moderate Demented classes. For the other values that the proposed model did not win, it still had the second-best scores, except for the recall of the Non-Demented class, which was the third-highest. The proposed model generally had the highest overall metric values and guaranteed balanced and stable performances across all categories.

There still are some shortcomings of the proposed model. The proposed model is relatively computationally expensive and requires a large GPU memory to run. The CNN model can be a favorable candidate for saving memory with limited resources. Furthermore, other algorithm optimization tricks and architecture optimization techniques may be introduced in future work to faster the model and reduce the computational complexity. Besides, the parameters settings and model frameworks are flexible; therefore, more advanced hyperparameter tuning can be applied to all the models to generate potential improvements. Also, the dataset used is very imbalanced and not large, which may lead to bias and declines in performance. All the models were tested using only this dataset, and the comparisons are thus limited. Other datasets from diverse fields can be used in the future to test the proposed model and give a more comprehensive analysis.

6. Conclusion

To facilitate the diagnosis of Alzheimer's Disease, researchers have spotlighted the classification of brain MRI images. This paper proposed a novel integrated model framework combining CNN and RF models. The models were compared with a pure CNN model from scratch and an RF model using PCA for dimension reduction. All three models were trained by the original data and the augmented data separately. The proposed model, CNN+RF trained by augmented data, had the best overall accuracy, weighted f-1 score, weighted recall, and weighted precision, which were all 93%. It outperformed other models in addressing the imbalanced problem, resulting in the highest recall, f-1 score, and precision for the Moderate Demented class. It also showed encouraging performance for the other classes, although it was not the all-around winner in terms of every metric in each class. Overall, the proposed model trained by augmented data had the most stable and balanced performance. Besides, this paper has shown that data augmentation has positively impacted the performances of PCA+RF and CNN+RF models in every class but only improved the CNN model in classifying the Non-Demented and Very Mild Demented classes. The number of images was a significant influencer of each category: more images led to better performance; the bigger the difference between numbers of images, the greater the gaps between the performance of classes. Even though the proposed model had the best overall performance, the other models also have other strengths or advantages. The utilization and selection among these models should be based on different purposes and concentrations.

References

[1] World Health Organization, "Dementia," Who.Int/Newsroom/Fact-Sheets, (2021).

[2] Alzheimer's Disease International, "World Alzheimer report 2019," Alzint.org., (2019).

[3] Alzheimer's Association, "2020 Alzheimer's Disease Facts and Figures," Alzheimer's & Dementia, 16(3), 391–460., (2020).

[4] Mandelkow, E. M., and Mandelkow, E., "Tau in Alzheimer's disease," Trends in Cell Biology. (1998).

[5] RadiologyInfo.org, "Alzheimer's Disease," Radiologyinfo.org. (2020).

[6] Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., and Thompson, P. M., "The clinical use of structural MRI in Alzheimer's disease," Nature Reviews Neurology, 6(2), 67–77. (2010).

[7] Kloppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack, C. R., Ashburner, J., and Frackowiak, R. S. J., "Automatic classification of MR scans in Alzheimer's disease," Brain, 131(3), 681–689. (2008)

[8] Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., and Feng, D., "Early diagnosis of Alzheimer's disease with deep learning," Researchers.mq.edu.au; Institute of Electrical and Electronics Engineers (IEEE). (2014).

[9] Liang, S., and Gu, Y., "Computer-Aided Diagnosis of Alzheimer's Disease through Weak Supervision Deep Learning Framework with Attention Mechanism," Sensors, 21(1), (2020). 220.

[10] Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., and Rueckert, D, "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease," NeuroImage, 65, 167–175. (2013).

[11] López, M., Ramírez, J., Górriz, J. M., Álvarez, I., Salas-Gonzalez, D., Segovia, F., Chaves, R., Padilla, P., and Gómez-Río, M., "Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease," Neurocomputing, 74(8), 1260–1271. (2011).

[12] Dubey, S., "Alzheimer's Dataset (4 class of Images)," Kaggle. (2019).

[13] Horning, N., "Random forests: An algorithm for image classification and generation of continuous fields data sets," Semantic Scholar. (2010).

[14] Dahinden, C., "An improved random forests approach with application to the performance prediction challenge datasets," (2009).

[15] Dahinden, C., and Ethz, M. A. T. H, "An improved Random Forests approach with application to the performance prediction challenge datasets," Hands-on Pattern Recognition, Challenges in Machine Learning, 1, 223-230. (2011)

[16] Hussain, M., Bird, J. J., and Faria, D. R., "A study on CNN transfer learning for image classification," SpringerLink. (1970).