Alzheimer's Classification using OASIS Dataset

Subhrato Som
CCI Drexel University
Philadelphia, USA
ss5654@drexel.edu

Imon Bera
CCI Drexel University
Philadelphia, USA
ib385@drexel.edu

***Abstract*—This project aims to develop and evaluate classification models for Classifying Alzheimer's disease using Brain Magnetic Resonance Image (MRI) scans. In this project, we have performed both multiclass and binary classification to classify Alzheimer's using multiple classification models and different modifications on the images and also ensembled the best-performing multiclass models. Among all the models we implemented the KNN model which utilizes a canny edge, performed third best with an accuracy of 78% and Logistic Regression and LDA using (1 vs 1) techniques for multiclassification were first and second. These three models were then chosen for the ensemble. Using the ensemble technique we found an increase in accuracy. Following model implementation, we computed the precision, recall, and F1 score for the ensemble model.**

***Index Terms*—Alzheimer's, Ensembling, 1vs1 Techniques, Accuracy, KNN**

I. Introduction

Alzheimer's is a neurodegenerative disorder that initially causes people to lose memory and the ability to make decisions and eventually causes them a severe level of dementia, a syndrome that causes the brain not to be able to make cognitive decisions. A few of the common symptoms of Alzheimer's include not being able to make proper decisions, challenges in performing a familiar task, commonly forgetting things, poor judgment making, and some others. One of the things that is very threatening about Alzheimer's is that it is very much progressive in nature. This project aims to classify Alzheimer's disease using Brain's Magnetic Resonance Image (MRI). The dataset for this project is the OASIS Alzheimer's Dataset which we have collected from Kaggle, consisting of multiple Brain MRI images in multi-classes. In this project, we down-sampled our data into two dimensions, which are 128x64 and 64x32.

Several research has been done in this field, majorly focused on using deep learning models. But there are some research done too, which has used some improvements of traditional machine learning models like Logistic Regression and SVM. The approach we aimed to target in this project is, initially we classified the data into Alzheimer's and Not-Alzheimer's using models like Logistic Regression and LDA which provided us with a good prediction accuracy, certainly higher than the highest class prior, we decided on performing multiclass classification. To make the decision on which model to choose, we analyzed and decided from binary classification to perform MultiLDA, and to reduce the time complexity we first performed PCA on our dataset and then implemented MulticlassLDA. However, we found that our results were not so satisfactory, we performed preprocessing on the image data with the implementation of Blur and Edge Detection on to picture. As that did not work as planned either, we came up with to use of label multiclass classification models including Naive Bayes and K-Nearest Neighbours. Using these models on normal images we got an improved accuracy and within them, Naive Bayes performed better than KNN. We then applied blur and edges to the image and performed the classification on them. With blurred data, both the models gave us a similar result, however, when it came to using edge data the accuracy of both the models went opposite way. KNN started performing better, even better than Naive Bayes and Naive Bayes got a lot worse. From these results, we came up with implementing 1vs1 models for each class. Again inspired by binary classification, we tested the concept of the 1v1 approach A binary classifier is trained for every pair of classes in a 1vs1 classification. In this approach, if there are N classes, Nx(N-1)/2 binary classifiers would be trained. Every classifier undergoes training to differentiate between examples of the two classes it is assigned to. 1vs1 Logistic Regression and 1vs1 LDA which gave us a significant improvement in accuracy over all other models. As these two and our KNN model with Edges were our better-performing models, we went on to ensemble them.

II. Literature Survey

In [1] Challis et.al proposed using Bayesian Gaussian Linear Regression by analyzing the pattern that at the time a lot of research used support vector machine (SVM), and their model outperformed and also went above linear classification, by providing 75% accuracy in distinguishing between mild dementia and normal brain, while providing 97% distinguishing mild dementia with Alzheimer's.

In [2] Subramoniam et.al. proposed using Resnet101 for feature extraction and classification for a class-labeled dataset of magnetic resonance image(MRI) of the brain with there labels non-demented, mild, very mild, and moderate for the dataset they had taken from Kaggle provided an accuracy of 95.32% with using CNN along with dense layered Deep Neural Network(DNN).

In [3] Knox et.al used an MRI dataset that is localized to the area of the hippocampus of the brain. They used two differently trained Auto-Encoder for extracting the features of the Images. After extracting features they tried classification out of which Gaussian Naive Bayes performed the best providing

an accuracy of around 80 percent in the extracted features of the image.

In [4] Fulton et.al. proposed using RESNET-50 for finding the presence of Clinical Dementia and using Gradient Boosting they analyzed Alzheimer's with an accuracy of 91 percent on a minimal mental state exam data. While using RESNET-50 with their features, they did 3 class classifications of Dementia with an accuracy of around 99%. Through this research they provided a step process in which first would be Alzheimer's Detection using Gradient Boosting techniques and then using MRI we can identify the level of Dementia.

In [5] Durate et.al. used Transfer Learning Techniques, and as they used Transfer Learning, to justify its outcome they used a small dataset where they used VGG 16 for initial layers and then used SVM and some other approaches for combining prior prediction. In this research, they did binary classification with whether there is Alzheimer's or not. Out of multiple approaches they have applied, FP Priori provided the highest accuracy of around 71%.

III. Method

A. Dataset Description

In this project, we have used OASIS-1 Alzehimer's Dataset which we have taken from Kaggle. It consists of multiple Brain MRI images taken from 461 subjects. As the original Oasis dataset was three-dimensional in size in the Kaggle dataset, they sliced it around the z-axis into 256 pieces and took images from the slice range from 100-160, into four different classes that are (Non Demented; Very Mild Dementia; Mild Dementia; and Moderate Dementia). Deep-diving into the dataset we were able to understand the medical significance in z-slicing the MRI data in such way. The dataset in the Kaggle was designed with respect to substance shrinkage in the brain, which is consistent along these slices. hence each image in the data can be counted as a single sample. The overall size of the Data was around 1.3 GB and it consists of around 86.4k image files out of which 67k(77%) is labeled as non-demented, 13.7k(15%) as very mild dementia, 5k(5%) as mild dementia and almost 500(2%) images as moderate dementia.

B. Data-Preprocessing

Due to class imbalance being so high, we had to reduced our data set as follows:
For binary classification, we picked 488 images (Random sampling) from the non-demented class labeled not Alzheimers, and from all other 3 classes we took 162 images from each and labeled as Alzheimers. Making class prior to be around 50%.
In multiclass as the moderate dementia class had 488 images, so we randomly sampled 488 images from each class. That makes our class prior to be around 25%.
The other task we did was to handle the data, we downsampled the images and worked with two dimensions of images: one being 1/8 of the size of the image which is (64x32) and the other being 1/4 which is (128x64). Fig 2 show the original image, downsampled to 128 and downsampled to 64.

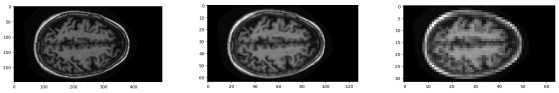

Fig. 1. 1) Original; 2)Down to 128; 3)Down to 64.

After that, we converted the image and modified the image to get blurred images and separately applied canny edge detection to get the edges in the images. Fig 1 shows 1) the Original image in 128x64; 2) Blurred and 3) Edges.

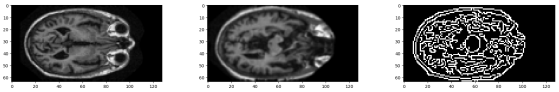

Fig. 2. 1) Original; 2)Blur; 3)Canny.

C. Binary Classification

Binary Classification is the process in machine learning where the aim is to classify into no more than two classes. While there can be many models and ideas to implement this we have used Logistic Regression and Linear Discreament Analysis to classify Alzheimer's and not Alzheimer's.

1) Logistic Regression: Logistic Regression is a binary classification model which can find the probability of a discrete outcome. We have performed binary classification using Logistic Regression for the Classes as Mentioned above for 2 dimensions of the image as mentioned above. We classified the data into Alzheimer's and not Alzheimer's.

2) Linear Discreament Analysis: Another common binary classification algorithm is LDA, Linear Discriminant Analysis. It finds a linear combination of features, that classifies the data into their classes. We performed PCA for LDA to reduce the feature set to 1500 as it helped us reduce the time complexity of the model. After that, we used our LDA model for the classification of Data into Alzheimer's and Not Alzheimer's for both dimensions of the image.

D. Multi-class Classification

Multi-class Classification is the process in machine learning where the aim is to classify into more than two classes. In this project, we aim to classify as described in the dataset that are non-demented, very mild dementia, mild dementia, and moderate dementia. For this, the classification models which we have used are:

1) Multi-Class LDA: Multiclass LDA is an extension to binary LDA which also aims to find differences in the linear combination of features for more than 2 classes. We used Multiclass LDA in our project to classify the data into four classes as mentioned above that are Non Dementia, Very Mild, Mild, and Moderate. Similar to Binary LDA, we also performed PCA to reduce the feature set to 1500 for the same reason of time complexity. The additional experiments that we performed here other than using images in two dimensions are mentioned in the above section, we have taken the blurred version of the image as well as the edges separately for both dimensions.

2) K-Nearest Neighbours: K-Nearest Neighbors (KNN) is a machine learning algorithm, that uses the majority class of a data point's k-nearest neighbors in the feature space to classify it. Similar to MultiLDA we used KNN to classify the data into 4 classes, however as the KNN does not take so much time complexity, so we did not perform PCA in it. Other than that we kept everything similar for the experimentation, with two image dimensions and three different variations in each dimension.

3) Naive Bayes: Naive Bayes is a classification model which uses probabilistic techniques to calculate posteriors. Here also we aimed to perform the multi-classification for the same classes with Naive Bayes in the same way as mentioned above and similar to KNN without using PCA.

4) (1vs1) Techniques: Observing low accuracy around the board for all multi-class classifications even with separate preprocessing of data, we deployed custom one versus one models to boost the accuracy. The one-versus-one approach is based on a binary classifier for each pair of classes. The main idea behind these models was the utilization of the good accuracy for binary classification, to reproduce (or vote) a good multi-class result. Since we were dealing with four different classes, six (N*(N-1)/2) different systems were trained on the same data, validated on the same data, and then results were voted to create multi-class labels. Here we also included blur and Canny edge preprocessing, to observe the effects of these on the system.

E. Ensembling

Ensembling Techniques combine the prediction of multiple models to improve overall accuracy. Here each classifier returns the probability for an unseen sample that belongs to a particular class, and various voting systems can be implemented based on their probabilities to finalize the classification result. This way, one system's result is usually getting verified by other systems available, resulting a much accurate prediction for the entire dataset. For our instance, we combined the results from our most accurate models (1V1 LDA, 1v1 Logistic Regression and KNN with Canny edge) for both of our image dimensions 64x32 and 128x64 pixels.

IV. Evaluation

We began experimenting with binary classification on both the downsampled dimensions of the image data. We used Logistic Regression and LDA for Binary Classification as mentioned above both the models gave us a good accuracy with LDA providing an accuracy of 70% in lower dimensions(64x32) and 78% in higher(128x64), while Logistic Regression performed slightly better in lower dimension with an accuracy of 72% than in higher with 75%. Table 1 shows a clear picture of accuracy comparison in both dimensions. In Figure 3 the confusion matrix represents the performance of two binary models. The top two matrices correspond to logistic regression models with dimensions 128x64 and 64x32, respectively. The bottom two matrices correspond to Linear Discriminant Analysis (LDA) models.

TABLE I
Accuracy Comparision of Binary Models

Binary Model	Accuracy	
	64x32	128x64
LDA (PCA)	0.7046	0.7784
Logistic Regression	0.7229	0.7507

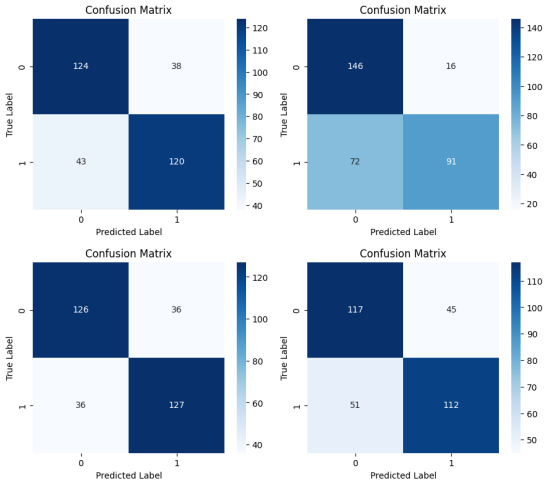

Fig. 3. Confusion Matrix of Binary Models with the upper 2 being logistic regression in 128x64 and 64x32 respectively and the bottom 2 are of LDA respectively. Where 0 refers to Alzheimer's and 1 Not Alzheimer's

As our binary models were performing well, inspired by that we aimed to use LDA to perform multiclass LDA which theoretically we thought would be performing well. However, while it did perform better than the highest class prior, the accuracy results were not so satisfactory. We got an accuracy of 51% and 48% in higher and lower dimensions respectively. To deal with that we tried to use some techniques for preprocessing the data. Figure 4 depicts the accuracy comparisons in both lower and higher dimensions. We first tried to blur the data before downsampling which performed strangely with Multi LDA, i.e. in 64x32 it reduced the accuracy to 44% while slightly increasing the accuracy to 52% in 128x64. We then separately tried to use Canny Edge Detection in the whole data to convert the original data to binary edge data, which showed us a steep decline in accuracy. In Table 2 we have shown the accuracy in the table. We guessed that the reason for getting Lower Accuracy could be because of the fact that LDA assumes the linear relationship between data is quite separable which could have been the reason for LDA not performing better and it could be proven by visualizing the confusion matrix in the figure 5, which is confusion matrix of a blurred version of the image, classified using multi LDA we can clearly see that it is not able to classify properly. As for the class 3 as the difference was very high i.e. which is separable as it contains data from moderate dementia class, it performed well.

As our results were not so satisfactory, we turned our focus to some other label Multiclass Classification Models including KNN and Naive Bayes. The KNN model using the similarity

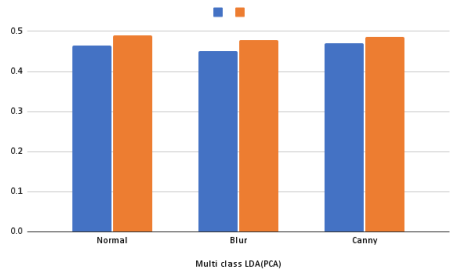

Fig. 4. Accuracy of MultiLDA with blue legend depicting lower dimension and orange depicting higher. 0: Non,1:VMild,2:Mild,3:Moderate

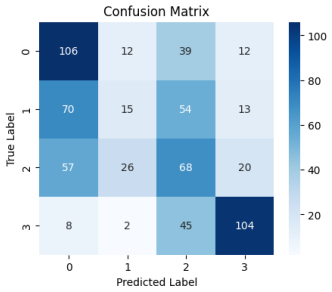

Fig. 5. Sample Confusion Matrix of Multiclass LDA.

function did improve on Multiclass LDA by giving an accuracy of almost 62% in both of the classes. we then performed blur for which it gave us similar results while when we performed Canny Edge Detection, the accuracy improved to76% and 78% respectively as mentioned in table 2. The accuracy gain maybe explained by Canny Edge Detection's capacity to attenuate noise and draw attention to important features. Consequently, this enhanced the ability to differentiate across classes, enabling the model to provide more intelligent conclusions.

TABLE II
MULTICLASS LDA(PCA) VS NAÏVE BAYES VS KNN ON DIFFERENT INPUT SIZES

Multi Model	64x32			128x64		
	Normal	Blur	Canny	Normal	Blur	Canny
LDA (PCA)	0.46	0.45	0.47	0.49	0.47	0.48
Naïve Bayes	0.66	0.68	0.59	0.66	0.69	0.59
KNN	0.62	0.61	0.76	0.62	0.62	0.78

We then performed similar experiments with Naive Bayes: while using the normal and blurred version of image data, it outperformed KNN with the model giving 66.5% accuracy in both dimension, (and sort of similar results when blurred,) but the accuracy saw a steep decline with it giving in higher 50% as given in the table 2 when used with Canny edges. Maybe this is happening because Naive Bayes relies on the assumption of feature independence, which can be challenged when edges are introduced. It might be violated by the introduction of dependencies between characteristics introduced by edges extracted by Canny Edge Detection. We can use the confusion matrix generated by the Canny and Normal to analyze this as given in figure 6 where the left side represents the confusion matrix of the Naive Bayes classifier on normal data in 128x64 and in edge image data in 128x64 in the right.

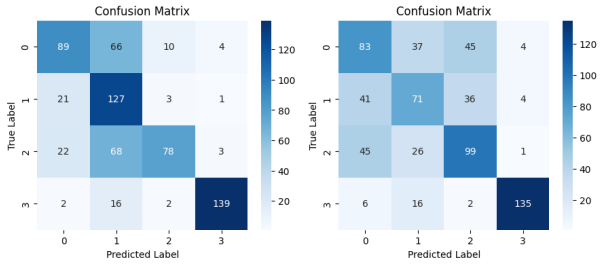

Fig. 6. Naive Bayes classifier on normal data in 128x64 in left and in edge image data in 128x64 in the right.

The confusion matrix kind of verifies our assumption regarding the data to be right, as it is clearly visible that the differentiation of class 0,1,2 (non,vmild, and mild) is higher in normal data than in edge data. figure 7 shows the accuracy graphs of both models (Naive Nayes and KNN) in both dimensions.

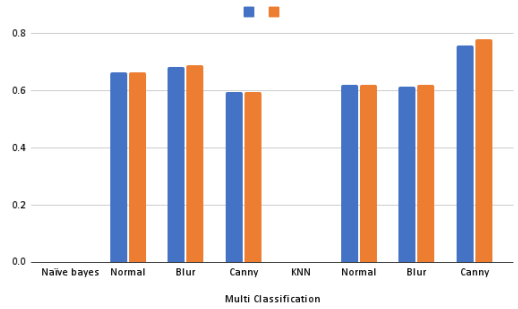

Fig. 7. Accuracy of NB(left) and KNN (right) with orange legend being higher dimension and blue lower

TABLE III
1VS1 MULTICLASS MODEL WITH DIFFERENT DIMENTIONS

1vs1 Model	64x32			128x64		
	Normal	Blur	Canny	Normal	Blur	Canny
LDA	0.75	0.73	0.66	0.77	0.76	0.61
LR	0.82	0.81	0.73	0.83	0.82	0.71

We then went on to implement the 1vs1 approach as mentioned above in methods. As shown in Table 3, these 1vs1 approaches performed excellently on normal preprocessing, this time actually increasing accuracy with dimension increase (64x32 to 128x64). The voting method in the custom system helped deal with the data indifference for various classes (e.g. Non, mild, and very mild data). However, even with this kind of high performing classifier, Canny Edge did not provide good results. One intuition behind this could be, the feature distinction getting suppressed with Canny edge for classification for this kind of model based on binary evaluation.

Using the 1vs1 and all previous experiments we got our three best-performing models which are 1vs1 LDA, 1vs1 Logistic Regression, and KNN(Canny Edge). We them ensembled these three models using the mean ensemble method where we average the prediction of the three selected models in both the dimension. We got an accuracy of 82% for the 64x32 dimension and 83% for 128x64 dimension. We can vizualize both of the ensemble model results for different dimension in figure 8, which shows the confusion matrix of the ensemble method with the left one being the lower dimension and the right being higher.

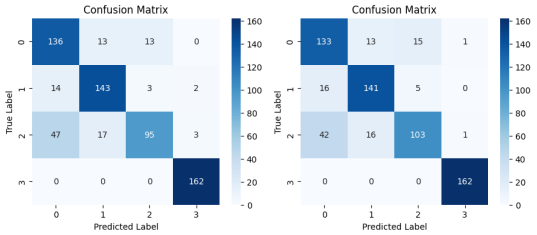

Fig. 8. Confusion Matrix for 64x32 (left) and 128x64 (Right) of Ensemble.

The confusion matrix depicts a clear analysis of prediction we can see that in both dimensions the number of true positives was quite higher and the random scattering was lower. We then calculated the precision, recall, and f1 score in both dimensions. Precision is the quality of a positive prediction made by the model. The recall is the percentage correct positive prediction of a particular class and the F1 score is the harmonic mean of precision and recall. The formula for the discussed matrix is as mentioned below.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F}_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Using these we got a precision score of 0.8271 for lower dimensions and 0.8357 for higher dimensions it is 0.8377. The recall score we got was 0.8271 for lower and f1 scores were 08314 and 0.8347. Figure 9 depicts the bar graph of the 4 outcomes we got using ensembling.

TABLE IV
PERFORMANCE METRICS FOR DIFFERENT IMAGE SIZES

Ensemble Evaluation	Image Size	
	64x32	128x64
Accuracy	0.8271	0.8318
Precision	0.8357	0.8377
Recall	0.8271	0.8318
F1	0.8314	0.8347

V. CONCLUSION

Based on the results obtained from the evaluation of results, we can see that ensembling the three models for dimensions 64x32 (Ensembling KNN(Canny) 76%, 1vs1 LDA 75.6%, and 1vs1 Logistic Regression 81.9%) had a small increase in accuracy with the accuracy increasing to 82.6% and similarly, the accuracy increased to 83.17% in higher dimension. These results show that the ensemble method, which combines the mentioned three models, works well to improve the accuracy of Alzheimer's classification, especially when the image dimensions are raised. This emphasizes how important ensemble modeling is for maximizing performance in a range of model configurations.

VI. FUTURE WORK

Several Different ideas could also be implemented including the merging of data from labels 100 to 160 and then augmenting data (mirroring, slight rotating etc), and performing similar experiments as previously done. Another thing that was out of the scope of this project but was discussed and has been implemented in the related works is the use of deep learning frameworks. We can use auto-encoders for feature extraction as done in some of the previous works and using deep learning-based image classification models could also be done to get better performance. One more concept that can also be visited is implementing all the done experiments and the possibly discussed ideas in the highest resolution (498x258) of the image which could be very computationally heavy.

REFERENCES

[1] Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., & Cercignani, M. (2015). Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. NeuroImage, 112, 232-243.

[2] Subramoniam, M., Aparna, T. R., Anurenjan, P. R., & Sreeni, K. G. (2022). Deep learning-based prediction of Alzheimer's disease from magnetic resonance images. In Intelligent vision in healthcare (pp. 145-151). Singapore: Springer Nature Singapore.

[3] Knox, S. A., Chen, T., Su, P., & Antoniou, G. (2021). A parallel machine learning framework for detecting alzheimer's disease. In Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings 14 (pp. 423-432). Springer International Publishing.

[4] Fulton, L. V., Dolezel, D., Harrop, J., Yan, Y., & Fulton, C. P. (2019). Classification of Alzheimer's disease with and without imagery using gradient boosted machines and ResNet-50. Brain sciences, 9(9), 212.

[5] Duarte, K. T., de Paiva, P. V., Martins, P. S., & Carvalho, M. A. (2019). Predicting the Early Stages of the Alzheimer's Disease via Combined Brain Multi-projections and Small Datasets. In VISIGRAPP (4: VISAPP) (pp. 553-560).

[6] 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.