Detection of Alzheimer's Disease

DATASCI 207 - Fall 2023

Theresa Sumarta, Adeline Chin

Problem Statement

The early and accurate detection of underlying disease pathology by clinicians is fundamental for diagnosing Alzheimer's Disease. Detecting early-stage AD in clinical practice can be challenging and is hindered by several barriers including constraints on clinicians' time, difficulty accurately diagnosing Alzheimer's pathology, and that patients and healthcare providers often dismiss symptoms as part of the normal aging process. Additionally, current AD diagnosis depends on the combination of blood work results, symptom history, and visual analysis of brain scans. As the prevalence of this disease continues to grow, the current model for Alzheimer's disease diagnosis will need to evolve.

Objective

Our main objective is to develop machine learning models that are able to discern patterns in brain MRIs that human eyes cannot. This will be achieved in two parts:

1. Classification of brain MRIs of those with an active Alzheimer's diagnosis and those without.
2. Prediction of brain MRIs that will later develop an Alzheimer's diagnosis.

Our secondary objective is to classify those diagnoses of Alzheimer's or Mild Cognitive Impairment using strictly symptom and medical history data.

Datasets

All datasets are gathered from the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI was a longitudinal, multicenter, multiphase study spanning from 2004 to 2016 for a total of four phases, ADNI-1, ADNI-GO, ADNI-2, and ADNI-3. Over 5,000 participants were recruited from across North America, each with at least one 2D and one 3D brain scan. The main goals of the ADNI study were to detect pre-dementia and track disease progression through biomarkers and to discern when new diagnostic methods would be most effective in Alzheimer's intervention.

Each visit, subjects are evaluated for any neurological changes and their diagnoses are updated accordingly. The labels are Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD).

- *2D Classification Data*

A total of 2,251 2D images were taken from ADNI1 and ADNI-GO phases to be used for the binary classification of Cognitively Normal and Alzheimer's Disease. All images were axial, though they included three different types of MRI sequences, T1, T2, and Flair in order to limit bias in the model. The visual differences in the types of images can be seen in the figure below.

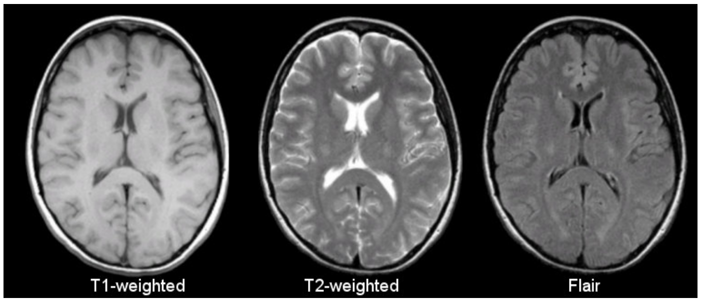


Figure 1:Visual example of different MRI sequences

- *2D Prediction of Conversion to Alzheimer's Data*

For the prediction of conversion to Alzheimer's Disease, a total of 3,258 axial images were gathered from subjects that participated in the ADNI Baseline and ADNI Screening visits. The Diagnosis Summary dataset, provided by the ADNI, was consulted to confirm which subjects had experienced disease progression. The final dataset was composed of subjects with brain scans of labels of Cognitively Normal or Mild Cognitive Impairment at the first visit, who later received an Alzheimer's Disease diagnosis during the study.

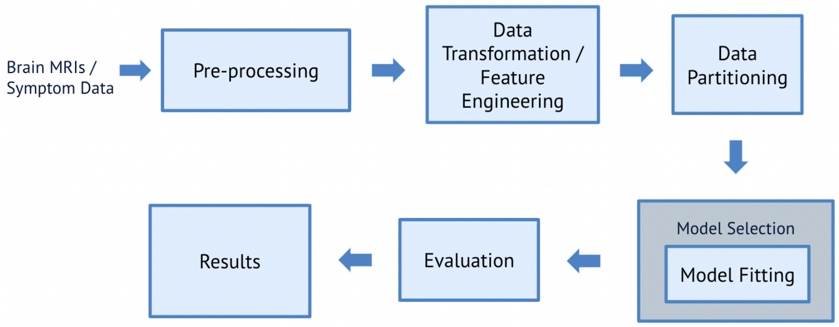
- *3D Image Classification*

A total of 2,000 3D MRI images were used for the analysis. The images used were taken from the ADNI1 study (first phase) which included subjects that were diagnosed with either mild cognitive impairment (MCI), early AD and elderly control subjects. These MRI images are formatted in .NIIfTI and each file contains a single volume of multiple slices that is read as a series of slices, each one resulting in a raster feature. Patient data that includes their patient id, image id and diagnosis were stored in 10 separate .csv files, with each file containing data from different screening phases. Using these datasets, all images have a diagnosis label which was used to predict a diagnosis based on the 3D images.

- *Classification using symptom data*

ADNI also includes multiple datasets with information about diagnoses, symptoms and medical history of consenting subjects. The Diagnostic Summary dataset and the Medical History dataset were used for this analysis. The Diagnostic Summary dataset indicates if subjects are experiencing any of 25 total symptoms, including nausea, vomiting, cough, dry mouth, and chest pain. The Medical History dataset includes patient data that indicates whether patients have or had other diagnoses which are neurologic, cardiovascular or respiratory related. The dataset also has data on whether patients had a history of alcohol or drug abuse. Both datasets include the patients' diagnosis (Normal level, Mild Cognitive Impairment, Alzheimer's disease) and their patient id. These datasets were used to predict a diagnosis based on their symptoms and medical history and predict the probability that patients could develop AD in the future. If health professionals can detect and anticipate the development of AD early, they can prevent the disease from causing significant brain damage to the patients.

Block Diagram



Approach/Methodology

1. *Data cleaning*

2D Images

The 2D scans for each timepoint for each subject contained 48 or 50 horizontal scan “slices” of the brain, where each “slice” was an individual DICOM file. The first image corresponds to the image of the very base of the brain while the last image corresponds to the very top of the brain. Since Alzheimer’s Disease can be physically identified with erosion of brain matter, only slices labeled 22-36 were selected for analysis for each patient, as these scans correspond to the center of the brain. This data cleaning process was applied to the scans for the 2D classification model as well as the 2D conversion prediction model.

3D Images

Each .NIIfTI file was converted into a 3D data array of shape 256 x 256 x 166 with each index in the array representing a voxel (3D cube) in the 3D image. The values of the voxels range from 0 to 2000. It was determined that for all images, some positions share the same value of 0. Since these values will not significantly affect the machine learning model’s ability to make an accurate diagnosis prediction, these voxels were removed from the 3D array. Each data array was sliced to form a 3D array of shape 206 x 72 x 18. With the smaller sized arrays, the CNN models will run more efficiently. 42 of the 2000 images did not have a diagnosis label. By looking at the patient data for the 42 images, images and diagnoses of previous and future visits were found. Since the diagnosis stayed the same for earlier and later screenings, it was assumed that the diagnosis remained unchanged in between those screenings. The assumed diagnoses were inputted into the dataset for the 42 images. The three diagnosis class labels are in the dataset in the form of strings (‘MCI’, ‘NL’, ‘AD’). These class labels were converted into integers (‘NL’ = 1, ‘MCI’ = 2, ‘AD’ = 3) using one hot encoding for ease when running machine learning models.

Symptom dataset

The symptom diagnostic summary dataset and the medical history dataset contain patient data from all phases of the study. At each phase of the study, participants were asked different survey questions and thus one phase was selected for this analysis. The ADNI1 phase data was selected as it had the most number of participants in the survey. Data collected during other phases were filtered out and rows that had the highest confidence of diagnosis rating were selected (DXCONFID = 4). From the symptom diagnostic dataset, only the patient ID, stroke, depression, parkinsonism, MCI and diagnosis columns were used for further analysis. From the medical history dataset, the patient ID, psychiatric, neurologic, cardiovascular, respiratory, hepatic, alcohol abuse, drug abuse and smoking columns were filtered for analysis. These variables were selected based on the preliminary research on Alzheimer's disease that suggests that these variables could contribute to a future AD diagnosis. These two datasets were joined on the patient ID column which resulted in a total of 836 rows.

2. *Data transformation and data augmentation*

2D Images

For the 2D image datasets, all images collected were DICOM image files. The DICOM images were transformed into pixel arrays in order to use one hot encoding for the model. Additionally, the contrast and brightness were adjusted for all images and a random flip was applied to the datasets. The specific values for the contrast and brightness adjustments can be found in the next section.

3D Images

After slicing all 2000 3D MRI images into data arrays of shape 206 x 72 x 18, each of the data arrays were flattened to a 1D array of size 173664. All of the values in the data arrays were normalized to transform all the features to be on the same scale to have better training stability and performance. Using a z-score normalization, the data array valued ranged between 0 and 1. Since the shape of the data arrays and the number of images used in the analysis are large, a principal component analysis (PCA) is used to reduce the dimensionality of the large dataset, while maintaining most of the feature information. Through an analysis of using various numbers of components in the CNN model, it was found that reducing the data to 1000 components optimizes the machine learning algorithm. Thus the final dataset used in the analysis was fitted and transformed to have 729 features.

Symptom Dataset

For the diagnostic summary dataset, the presence of each symptom was indicated by a number that corresponds to it (MCI = 1, Stroke = 2, Depression = 3) in a single column. These categorical variables were transformed using one hot encoding to be represented as integers in the machine learning model. Separate columns were made for each of the symptoms with either the value 0 or 1, based on the symptoms indicated in the single symptom column of the dataset. For the medical history dataset, the columns were one hot encoded which did not require any data transformations. All column values were converted to the numeric data type from the string data type for ease when running machine learning models and rows that contained nan values were dropped. The value 0 in a column indicates that the patient does not have that symptom or medical history and the value 1 indicates that the patient has the symptom or medical history.

3. Creating Machine Learning models

2D Image Classification

The initial model built was a Neural Network with 5 layers. Ultimately, after hyperparameter tuning and layer modification, the test accuracy was capped at 73.04%. The existing NN was converted to a Convolutional Neural Network and the final model only contained 2 convolutional layers. Table 1 below shows the combinations of hyperparameter and data augmentation values that were tested with the CNN model and the resulting training and validation accuracy. The last row indicates the values that were chosen for the final model.

Table 1: Hyperparameter Tuning

Training accuracy	Validation accuracy	Kernel size	Strides	Pool size	Pool	Padding	Brightness Factor	Contrast
0.9060	0.7808	5,5	1,1	2,2	Max	Same	0.3	3
0.9778	0.8014	3,3	1,1	2,2	Max	Same	0.3	3
0.9653	0.8151	3,3	2,2	2,2	Max	Same	0.3	3
0.9492	0.8159	3,3	1,1	3,3	Max	Same	0.3	3
0.9878	0.8190	3,3	1,1	2,2	Max	Valid	0.3	3
0.9630	0.8413	3,3	1,1	2,2	Avg	Same	0.3	3
0.9852	0.7841	3,3	1,1	2,2	Max	Same	0.5	3
0.9762	0.8286	3,3	1,1	2,2	Max	Same	0.1	3
0.9815	0.8254	3,3	1,1	2,2	Max	Same	0.3	2
0.9751	0.7778	3,3	1,1	2,2	Max	Same	0.3	4
0.9175	0.8476	3,3	1,1	2,2	Max	Same	0.3	1

2D Image Conversion Prediction

Using the CNN model from the 2D Image Classification portion as a starting point, the hyperparameters were re-tested. Ultimately, the same values for the hyperparameters resulted in the highest accuracy values. This is reasonable because the images used for the conversion prediction were of the same type as the images used for the strict classification. The layers for this CNN model were also modified. Additional dropout layers were added with a rate of 0.2, however, the layers heavily decreased the accuracy of the model and were ultimately removed. The final CNN model also included additional batch normalization layers.

3D Image Classification

A 3D CNN model of 6 layers was trained without the use of PCA. This model consists of a total 173664 dimensions and ran with a significantly long computation time. After 20 epochs, the model had a training accuracy of 47.92%. PCA was applied to the dataset to determine if reducing the dimensionality of the data will reduce the computation time while retaining a high training accuracy. As shown in Figure 2 below, reducing the dimensionality of the dataset to 792 features optimized the model as the train accuracy was 48%. Different combinations of hyperparameters were used to run the 3D CNN model using 792 features as shown in Table 2. The second row in red indicates the values that were chosen for the final model as this combination of hyperparameters had the greatest training accuracy.

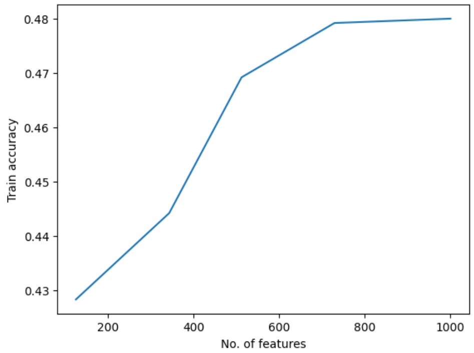


Figure 2: Train accuracy for various number of components using PCA

Table 2: Hyperparameter tuning for 3D CNN model

Training accuracy	Validation accuracy	Kernel size	Strides	Activation	Learning rate	Optimizer
0.4792	0.5075	(3, 3, 3)	(2, 2, 2)	relu	0.01	Adam
0.4800	0.5075	(5, 5, 5)	(2, 2, 2)	relu	0.01	Adam
0.4783	0.5075	(3, 3, 3)	(5, 5, 5)	relu	0.01	Adam
0.4650	0.5075	(3, 3, 3)	(2, 2, 2)	relu	0.001	Adam
0.4783	0.5075	(3, 3, 3)	(2, 2, 2)	softmax	0.01	Adam
0.4758	0.5075	(3, 3, 3)	(2, 2, 2)	relu	0.01	SGD

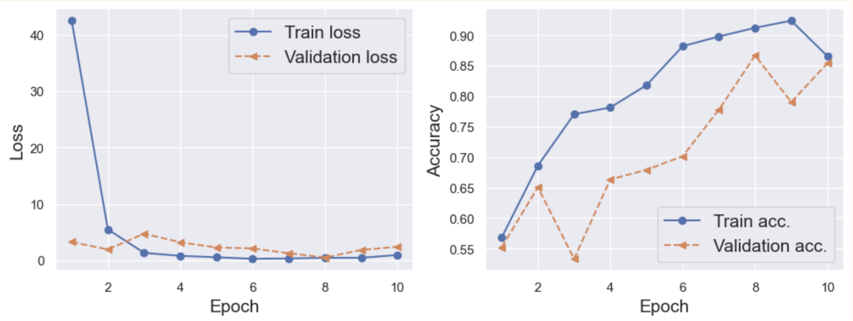
Symptom Classification

The combined dataframe from joining the diagnostic summary and medical history datasets was split into 0.7:0.3, 70% of the data was used for training models and 30% of the data was used to test the accuracy of the models. A total of 7 different machine learning models were run using the diagnostic summary and medical history datasets to determine which model was most suitable in predicting a diagnosis using this data. The models are KNN, Decision Tree, Random Forest, Kernel SVM, Naive Bayes, SVM and Logistic Regression. Various k values were run through the KNN model to determine which k value optimized the machine learning algorithm. In the decision tree classifier, the minimum sample split was set to 10 and the maximum depth was varied to determine the maximum depth value that would optimize the model. Similarly, for the random forest model, the number of estimators was varied and the maximum depth value was kept constant at 6. It was determined that the testing accuracy remained constant when the maximum depth value was varied but the accuracy was dependent on the number of estimators. By running the model with different maximum depth values, the test accuracies for each run were compared to determine the optimum maximum depth value. The metrics for a good parameter value (k value, maximum depth) were that the training accuracy was greater than the testing accuracy which suggests that the model is not underfitting and the test accuracy is within 5% of the train accuracy to ensure that the model is not overfitting. The other 4 models did not require any parameter tunings.

4. *Experiments*

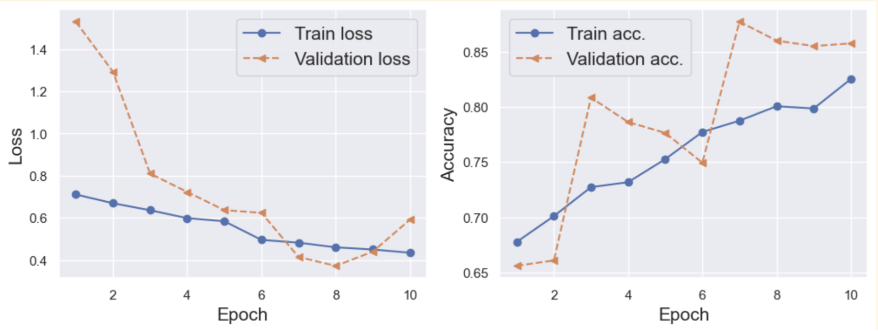
2D Image Classification

The figure below shows the training and validation losses and accuracies for the final Alzheimer's classification CNN model.



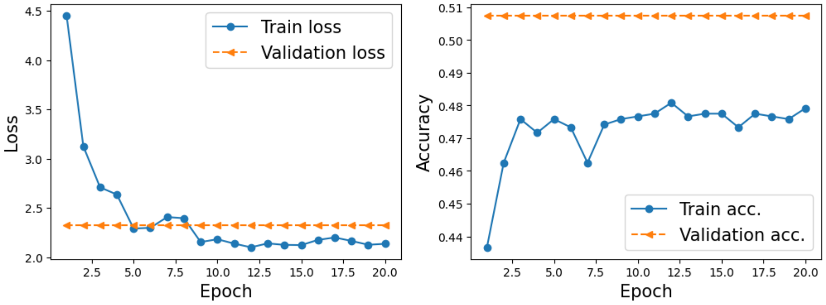
2D Image Conversion Classification

The figure below shows the training and validation losses and accuracies for the final Alzheimer's conversion CNN model. One thing to note is that the validation accuracy is slightly higher than the training accuracy, which might suggest overfitting of the validation dataset. However, the differences in the training and validation accuracy are less than 3.5% and during the experimentation process, the model was tested with different rates for dropout layers, which could also contribute to the higher validation accuracy.



3D Image Classification

The figure below shows the training and validation losses and accuracies for the final Alzheimer's classification 3D CNN model.



Symptom Classification

A correlation matrix was made between all the features in the final dataframe that consists of columns from the diagnostic summary dataset and the medical history dataset. It was concluded that all the features have a low correlation value of less than 0.4 between them as shown in Figure 3 below. Hence none of the features were dropped from the dataframe and all features were run in the machine learning models.

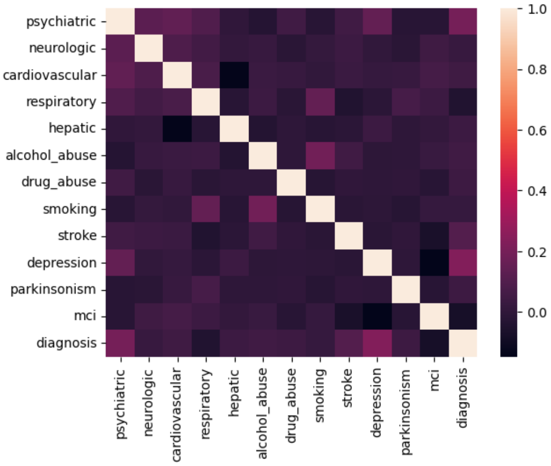


Figure 3: Correlation matrix

K values between 15 and 50 were run through the KNN model to determine which k value would not overfit or underfit the model. Based on Figure 4 below, the optimum k value is 36 as it has a train accuracy of 74.19% and a test accuracy of 73.71%. K values above 36 had train accuracies that are larger than the test accuracies but using larger K values are computationally expensive. Additionally, K values below 36 had train accuracies that are lower than the test accuracies which suggests that the model is underfitting. Using the k value = 36, it is concluded that the KNN model is 73.71% confident in its diagnosis prediction.

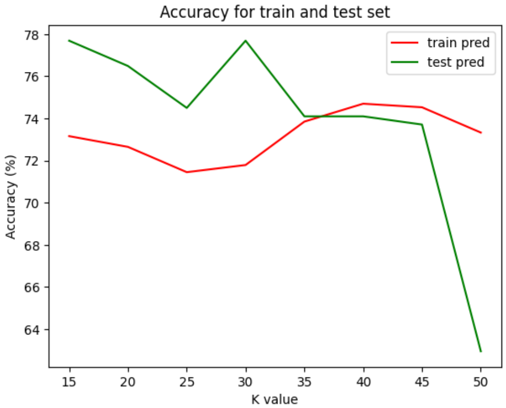


Figure 4: Train and test accuracies of KNN model using various K values.

Maximum depth values between 2 and 15 were run through the decision tree classifier model to determine which value gives the highest testing accuracy. It was found that a maximum depth value of 6 gives the highest accuracy of 78.5%. Figure 5 below shows the visual representation of the decision tree model using max_depth = 6 and min_samples_split=10.

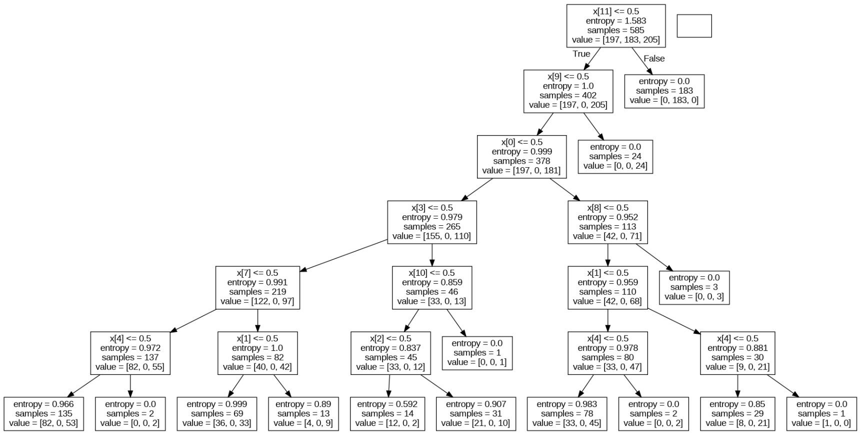


Figure 5: Decision tree model

In the random forest model, the number of estimators is the number of trees in the forest. The number of estimators were varied and ran through the classifier model. Figure 6 shows the results from varying n_estimators. The n_estimator=35 value gives the highest testing accuracy of 79.3%.

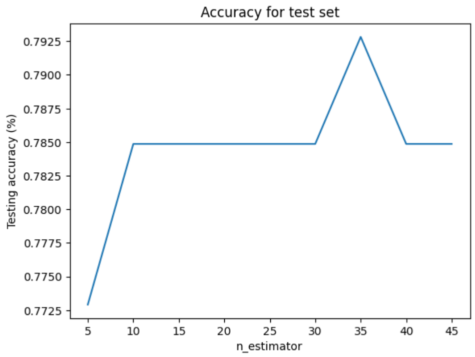
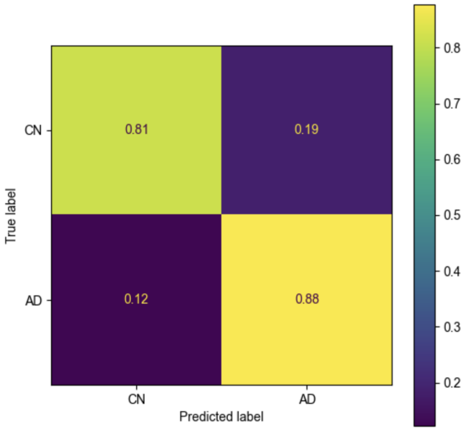


Figure 6: Test accuracies of random forest model using various number of trees.

5. *Model evaluation, analysis, and Results*

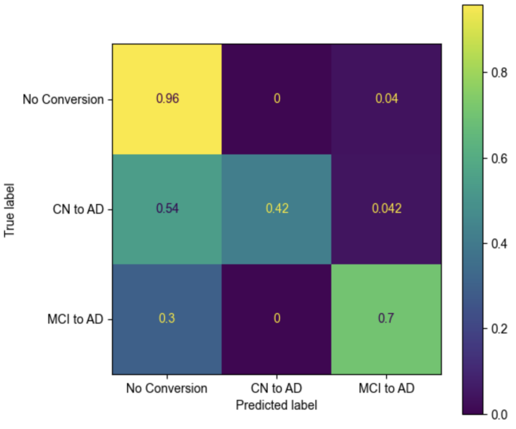
2D Image Classification

The results for the 2D image classification portion can be seen in the confusion matrix below. The values for the success metrics can be found in Table 4.



2D Conversion Prediction

The results for the 2D conversion prediction model can be seen in the confusion matrix below. The values for the success metrics can be found in Table 4.



3D Classification

The 3D CNN model shows a higher validation accuracy than the train accuracy which suggests that the model might be overfitting. Based on the results from the success metrics table below, it is concluded that the neural network is not training and learning properly, thus the test data did not fit well to the model

Symptom Classification

The test accuracies of running all 7 models with optimized parameters are summarized in Table 3 below. The Random Forest model has the highest test accuracy of 79.3% and the Naive Bayes model has the lowest test accuracy of 62.5%. The Naive Bayes model assumes that the features are independent of each other but based on Figure 3 there are small correlations between the features. Thus it is expected that the Naive Bayesian model had a relatively low test accuracy result. The Random Forest model had the greatest test accuracy value of 79.3%. Based on the high dimensionality and the presence of irrelevant features in the dataset, random forest is more suitable than KNN for this particular problem and data. This is because random forest is a more complex algorithm than KNN and performs better with this dataset as compared to KNN which is better at handling smaller and more structured datasets.

Table 3: Summary of test accuracy results

Classifier model	Test accuracy (%)
Random Forest	79.3
Decision Tree	78.5
Kernel SVM	78.5
Logistic Regression	76.5
SVM	75.3
KNN	73.7
Naive Bayes	62.5

Success Metrics

The quantitative success metrics chosen for this study are the test accuracy, the F1 scores, the sensitivity, the specificity, Matthew's Correlation Coefficient. Sensitivity and specificity are only applicable for the binary classification models, while Matthew's Correlation Coefficient is only applicable for the multiclass classification models.

Table 4: Success metric values for all models

	Test Accuracy	F1 Score	Sensitivity	Specificity	Matthew's Correlation Coefficient
2D Image Classification	0.8703	0.8621	0.88	0.81	

2D Image Conversion Classification	0.8428	0.7528			0.7606
3D Image Classification	0.4800	0.1638			0
Classification using symptoms (random forest)	0.793	0.7956			0.6779

Bias

Several methods were used to limit bias in the models. The first was including brain scans of different MRI sequences so the models would focus on the differences of intensity of different parts of the brain, rather than the overall intensity. The second was pre-processing the training image data for contrast and brightness adjustments, as well as random flips, to decrease the probability of the model learning irrelevant patterns. Additionally, the ADNI dataset is imbalanced with 80% of the study participants being over the age of 70 and 93% of the participants are white. This causes an algorithmic bias as the trained model will be able to better predict a diagnosis for patients above the age of 70 and are white, and have lower accuracies for patients below 70 and are not white. It is recommended that the study is conducted with more diverse demographics.

Constraints

One main constraint that we faced was limited computational power. Using thousands of images meant that we occasionally ran into out-of-memory issues and the run time of the models ranged from 30 minutes to >2 hours. In the later stages of model development, tweaking hyperparameters meant that the overall process was very slow.

Challenges and Limitations

For the models using 2D image data, the main challenge was choosing the appropriate subset of horizontal slices from each full set of scans. Using all scans would train the model on irrelevant patterns, but only using a couple slices would mean that the model could not be generalized. PCA was used to reduce the dimensionality of the 3D image data which caused some information to be lost. This lowered the performance of the model's ability to learn from the training data and the test data could not be generalized well to the model. The dataset in the symptoms classification model contained many null values and since rows that contained null values were removed, the final dataset had a small sample size of 836 samples.

Future work

For future analysis, we would like to explore creating models that can take multiple types of data with one label. For example, a model that takes in a brain scan and symptom data as one data point and then predicts whether or not the corresponding subject will receive an Alzheimer's diagnosis. We would also like to look into implementing transfer learning into our model. For example, we can take the knowledge from a 2D image classification model and use it to train the 3D image classification model to improve our model's performance. Additionally, we would like to create models that do not require conversion of medical files to data arrays.

Standards

Python 3.9.5, Python 3.11.5, TensorFlow 2.14.0, NumPy 1.26.1, Pandas 1.4.2

References

1. Al Shehri, Waleed. “Alzheimer’s disease diagnosis and classification using Deep Learning Techniques.” *PeerJ Computer Science*, vol. 8, 20 Dec. 2022, https://doi.org/10.7717/peerj-cs.1177.

2. “Medical Tests for Diagnosing Alzheimer’s.” *Alzheimer’s Disease and Dementia*, Alzheimer’s Association, www.alz.org/alzheimers-dementia/diagnosis/medical_tests#:~:text=Physicians%20use%20diagnostic%20tools%20combined,to%20make%20an%20accurate%20diagnosis.

3. Preston, David C. “Magnetic Resonance Imaging (MRI) of the Brain and Spine: Basics.” *MRI BASICS*, 4 July 2016, case.edu/med/neurology/NR/MRI%20Basics.htm.

4. Shahbaz, Muhammad, et al. “Classification of alzheimer’s disease using machine learning techniques.” *Proceedings of the 8th International Conference on Data Science, Technology and Applications*, 10 Sep. 2019, https://doi.org/10.5220/0007949902960303.