Title

Alzheimer's Disease Prediction from Handwriting using Machine Learning Algorithms

Permalink

https://escholarship.org/uc/item/7mt2h559

Author

Chen, Xinyue

Publication Date

2024

UNIVERSITY OF CALIFORNIA

Los Angeles

Alzheimer's Disease Prediction from Handwriting using Machine Learning Algorithms

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Xinyue Chen

2024

ABSTRACT OF THE THESIS

Alzheimer's Disease Prediction from Handwriting
using Machine Learning Algorithms

by

Xinyue Chen
Master of Science in Statistics
University of California, Los Angeles, 2024
Professor Yingnian Wu, Chair

Alzheimer's disease is a type of neurodegenerative disease that is common among the elderly. Although there is no cure, early diagnosis allows for treatments that can manage and delay the symptoms. We will employ machine learning algorithms, such as logistic regression, random forest, and extreme gradient boosting, to predict Alzheimer's disease in two experiments. In the first experiment, each model is applied to all 450 features. In the second experiment, each model is applied to 25 different feature sets, with one set corresponding to each task. Predictions are based on the DARWIN (Diagnosis AlzheimeR WIth haNdwriting) dataset, and model performance is measured using accuracy, ROC curves, and AUC. The results indicate that the random forest model applied to all 450 features is the best performing model in predicting Alzheimer's disease, achieving a model accuracy of 91.43% and an AUC of 0.9441.

The thesis of Xinyue Chen is approved.

Hongquan Xu

Mark S. Handcock

Yingnian Wu, Committee Chair

University of California, Los Angeles

2024

To my family and Bagel,

for their boundless love and support

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Alzheimer's disease is a type of neurodegenerative disease that causes disability in cognitive and memory functions and is common among elderly people. Unfortunately, there is no cure to Alzheimer's disease and its symptoms tend to become more persistent as the disease progresses. However, medication can help manage and delay Alzheimer's symptoms. Therefore, early diagnosis of Alzheimer's disease is crucial for slowing the progression of the disease. Nevertheless, early detection of the disease can be tedious and costly due to extensive data collection and advanced tools required [1].

Machine learning algorithms have been a dominant method in disease prediction and aiding clinical assessments. Most related studies have utilized machine learning algorithms such as classification on MRI images and support vector machine models to predict Alzheimer's disease. Rana et al. developed a model called MudNet, which is trained and validated using both clinical data and structural MRIs to predict the conversion of mild cognitive impairment to Alzheimer's disease [2]. The model achieves an accuracy of 69.8% in conversion prediction and an accuracy of 66.9% in risk classification predictions [2]. Huang et al. achieved a model accuracy of 80.0% with support vector machine classifier on the Alzheimer's Disease Neuroimaging Initiative dataset, which integrates both clinical and MRI data [3].

Given that cognitive functions are associated with coordinating and executing actions, and since handwriting necessitates coordination between the brain and the body, analyzing handwriting behaviors could serve as a cost-effective approach in monitoring the progression of the disease [4]. To evaluate which set of features can effectively distinguish between

Alzheimer's patients and healthy individuals, we will implement logistic regression, random forest, and extreme gradient boosting. We chose these models because they are widely used and represent a variety of classification algorithms. We will assess their performance using model accuracy, ROC curve analysis, and AUC interpretation.

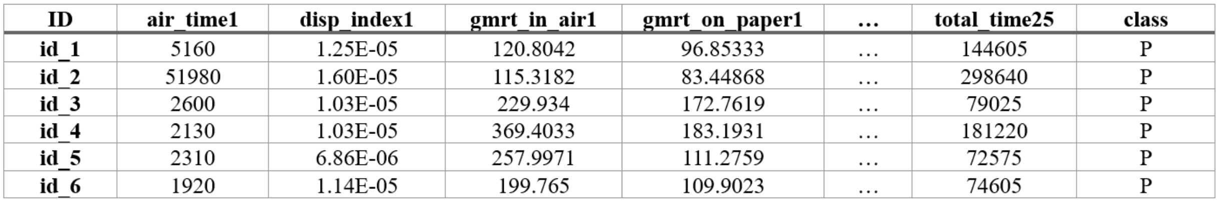
CHAPTER 2

Data

2.1 Data Overview

The data used for this project was obtained from the UCI Machine Learning Repository. The name of this dataset is the DARWIN dataset, which stands for Diagnosis AlzheimeR WIth haNdwriting [5]. The first 6 rows of the raw data is shown in Table 2.1.

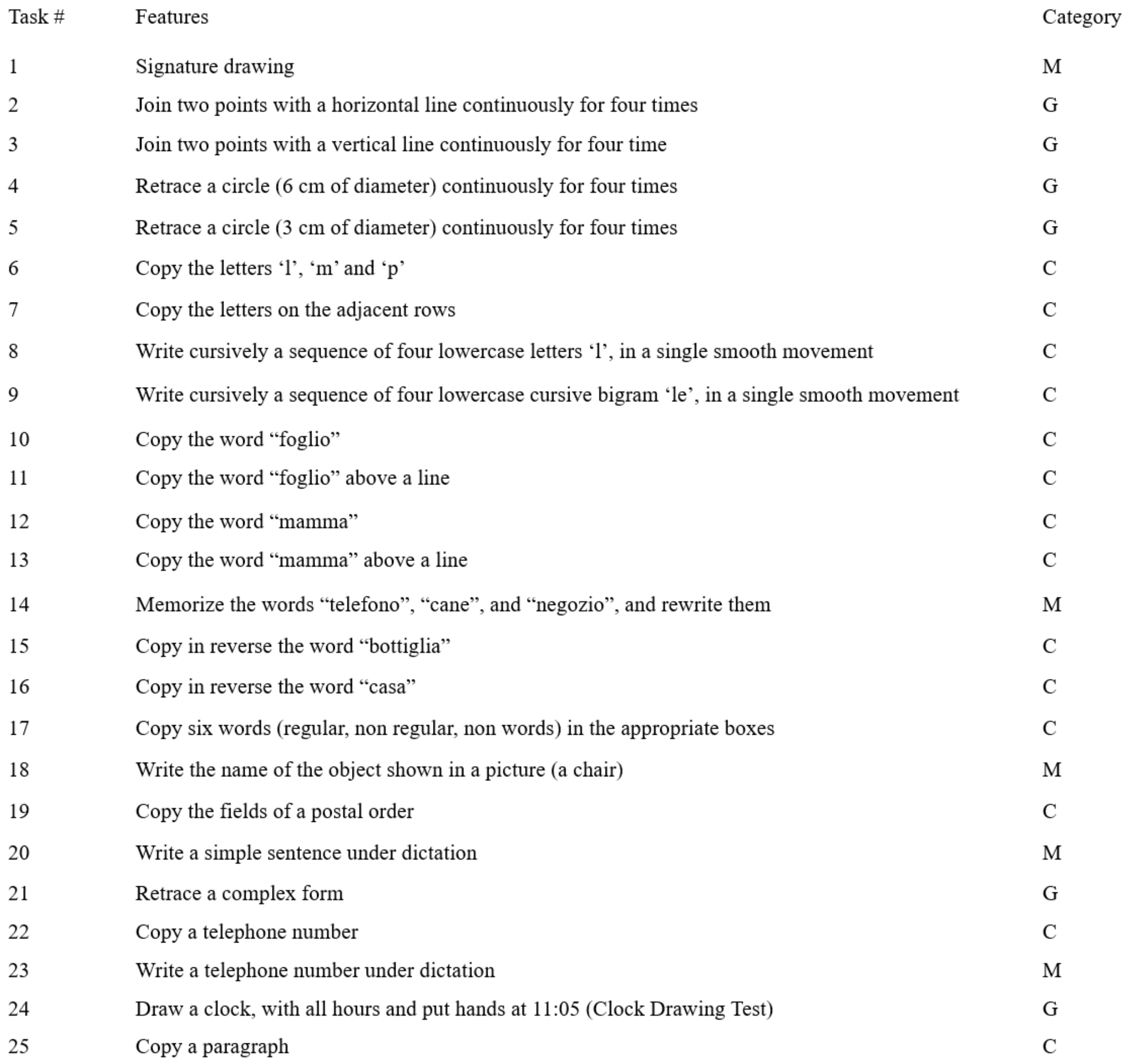
Table 2.1: First 6 rows of the raw data



ID	air_time1	disp_index1	gmrt_in_air1	gmrt_on_paper1	…	total_time25	class
id_1	5160	1.25E-05	120.8042	96.85333	…	144605	P
id_2	51980	1.60E-05	115.3182	83.44868	…	298640	P
id_3	2600	1.03E-05	229.934	172.7619	…	79025	P
id_4	2130	1.03E-05	369.4033	183.1931	…	181220	P
id_5	2310	6.86E-06	257.9971	111.2759	…	72575	P
id_6	1920	1.14E-05	199.765	109.9023	…	74605	P

The dataset contains 174 observations, where each observation represents a participant. Out of the 174 participants, 89 are Alzheimer's patients and 85 are healthy people. To avoid any bias, the participants were recruited to ensure that the two groups matched in terms of age, level of education, work, and gender [4]. Each participant is asked to perform 25 tasks, with each task belonging to one of the following categories: graphic, copy, or memory (Table 2.2).

Table 2.2: List of 25 tasks



Task #	Features	Category
1	Signature drawing	M
2	Join two points with a horizontal line continuously for four times	G
3	Join two points with a vertical line continuously for four time	G
4	Retrace a circle (6 cm of diameter) continuously for four times	G
5	Retrace a circle (3 cm of diameter) continuously for four times	G
6	Copy the letters ‘l’, ‘m’ and ‘p’	C
7	Copy the letters on the adjacent rows	C
8	Write cursively a sequence of four lowercase letters ‘l’, in a single smooth movement	C
9	Write cursively a sequence of four lowercase cursive bigram ‘le’, in a single smooth movement	C
10	Copy the word “foglio”	C
11	Copy the word “foglio” above a line	C
12	Copy the word “mamma”	C
13	Copy the word “mamma” above a line	C
14	Memorize the words “telefono”, “cane”, and “negozio”, and rewrite them	M
15	Copy in reverse the word “bottiglia”	C
16	Copy in reverse the word “casa”	C
17	Copy six words (regular, non regular, non words) in the appropriate boxes	C
18	Write the name of the object shown in a picture (a chair)	M
19	Copy the fields of a postal order	C
20	Write a simple sentence under dictation	M
21	Retrace a complex form	G
22	Copy a telephone number	C
23	Write a telephone number under dictation	M
24	Draw a clock, with all hours and put hands at 11:05 (Clock Drawing Test)	G
25	Copy a paragraph	C

For each task, 18 features were extracted:

- total_time: Total time spent to perform the entire task

- air_time: Time spent to perform in-air movements

- paper_time: Time spent to perform on-paper movements

- mean_speed_on_paper: Average speed of on-paper movements

- mean_speed_in_air: Average speed of in-air movements

- mean_acc_on_paper: Average acceleration of on-paper movements, where acceleration is the variation of speed with respect to time

- mean_acc_in_air: Average acceleration of in-air movements

- mean_jerk_on_paper: Average jerk of on-paper movements, where jerk is the variation of acceleration with respect to time

- mean_jerk_in_air: Average jerk of in-air movements

- pressure_mean: Average of the pressure levels exerted by the pen tip

- pressure_var: Variance of the pressure levels exerted by the pen tip

- gmrt_on_paper: Generalization of the Mean Relative Tremor (MRT) computed for on-paper movements, where MRT measures the amount of tremor in drawing spirals and meanders

- gmrt_in_air: Generalization of the Mean Relative Tremor computed for in-air movements

- mean_gmrt: Average of GMRT on-paper and GMRT in-ai

- num_of_pendown: Counts the total number of pendowns recorded during the execution of the entire task

- max_x_extension: Maximum extension recorded along the X axis, which is calculated from the difference between its farthest/nearest points to the origin on the X axis

- max_y_extension: Maximum extension recorded along the Y axis, which is calculated from the difference between its farthest/nearest points to the origin on the Y axis
- disp_index: Measurement of how the hand-written trace is "dispersed" across the entire piece of paper.

All features are numerical variables. In addition to these features, there is also an "ID" column, which contains a list of consecutive numbers from 1 to 174, and a "class" column, where "P" indicates Alzheimer's patients and "H" indicates healthy people. Thus, the total number of columns for this dataset is 251.

2.2 Data Cleaning

The first step in data cleaning is to drop the "ID" column because it is irrelevant in evaluating model performance and classifying Alzheimer's disease. Next, we search for and remove rows containing missing values and duplicate rows. It appears that the dataset does not contain any missing values or duplicate rows. Lastly, the outcome variable "class" is converted to a binary variable, where 0 corresponds to healthy people and 1 corresponds to Alzheimer's patients.

2.3 Exploratory Data Analysis

In this section, we will analyze the relationship between all the features and Alzheimer's disease using numerical measures and graphical representations, such as barplot, histograms, and correlation map. Since the goal of the project is about classification of Alzheimer's disease, it is crucial that we first check if the dataset is balanced. From the barplot in Figure 2.1, we can see that the dataset is fairly balanced, comprising 85 healthy people and 89 Alzheimer's patients. Given the small difference between the two groups, it appears that resampling of the training data is not necessary.

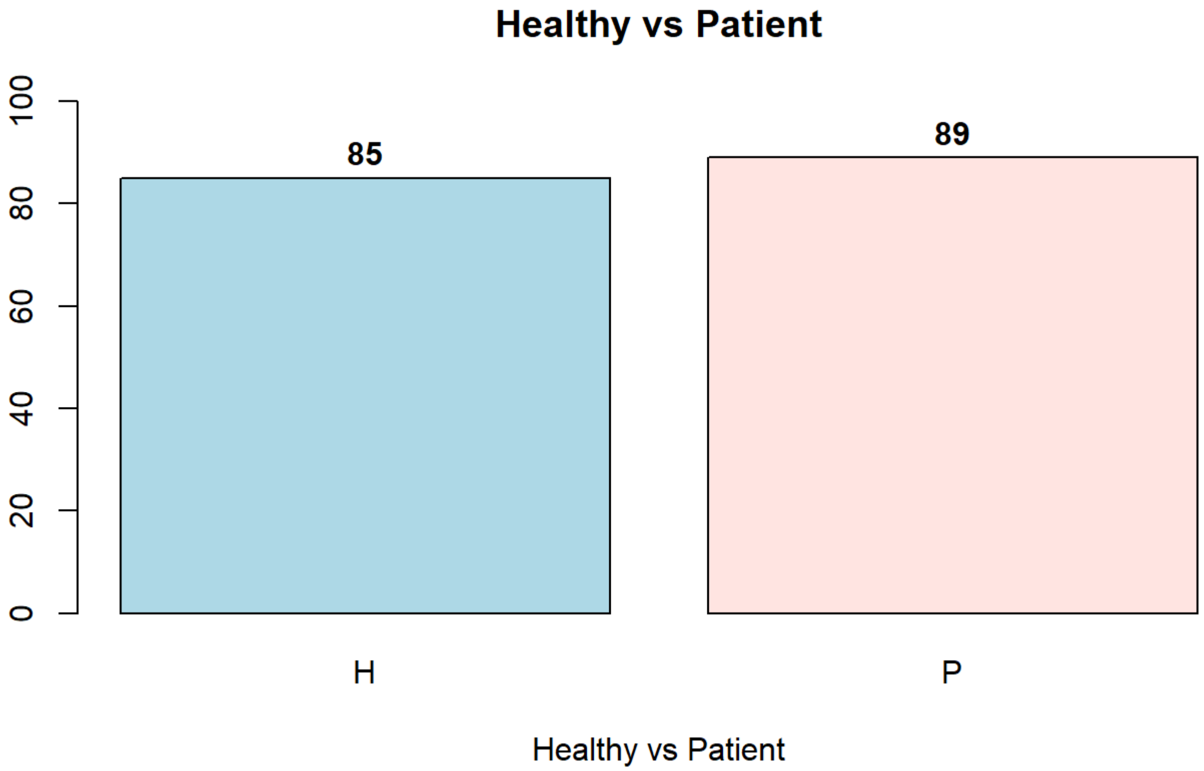
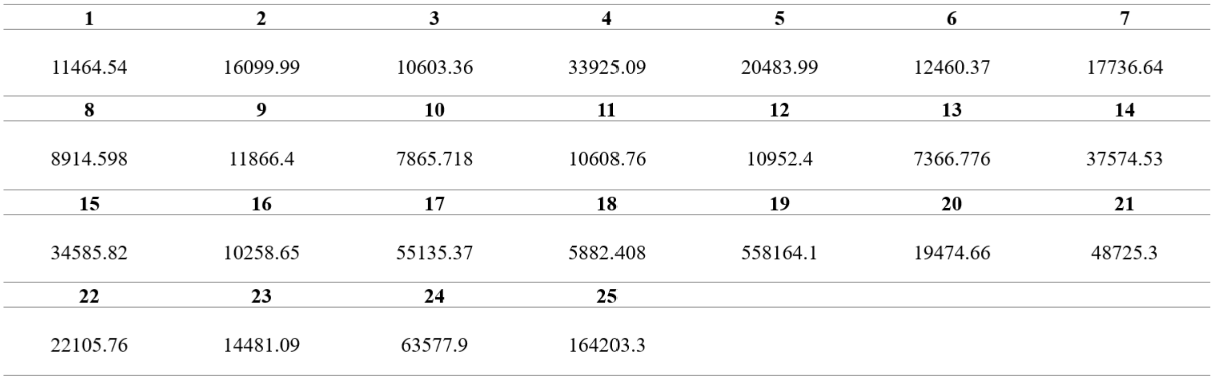


Figure 2.1: Distribution of groups

We can compare the difficulty of different tasks by analyzing and averaging the variable total_time for each task. From Table 2.3, we observe a noticeable difference in the average of total_time for each task, suggesting that there are substantial differences in task difficulty. Task 19 has the longest average total time of 558,164.1, while task 18 has the shortest average total time of 5,822.408. This can be useful in our analysis because tasks with different levels of difficulty target different parts of the brain, making them more effective in discriminating between Alzheimer's patients and healthy individuals.

Table 2.3: Mean of total_time



1	**2**	**3**	**4**	**5**	**6**	**7**
11464.54	16099.99	10603.36	33925.09	20483.99	12460.37	17736.64
8	**9**	**10**	**11**	**12**	**13**	**14**
8914.598	11866.4	7865.718	10608.76	10952.4	7366.776	37574.53
15	**16**	**17**	**18**	**19**	**20**	**21**
34585.82	10258.65	55135.37	5882.408	558164.1	19474.66	48725.3
22	**23**	**24**	**25**			
22105.76	14481.09	63577.9	164203.3			

Next, we will use histograms to visualize the distribution of the features and check for potential outliers. Since there is a large set of features, we will focus on the histograms of total_time (Figure 2.3) and disp_index (Figure 2.4) for all 25 tasks. From the histograms, the distributions of total_time for all 25 tasks appear to be heavily right-skewed and it appears that there exists large outliers for the majority of the tasks. These large outliers are likely Alzheimer's patients who take longer time to complete the required tasks compared to healthy individuals. Since they represent natural variations within the population, these large outliers should be retained and not removed. Unlike the distribution of total_time, the majority of the distributions of the disp_index appear to be roughly symmetric, with a few exhibiting right-skewness. Those exhibiting right-skewness appear to have large outliers, which again are likely Alzheimer's patients who tend to add unnecessary spacing in their handwriting.

Lastly, we will use a heat map to visualize the correlation matrix and identify the relationship between variables. We will again use total_time as an example. The color scale ranges from dark blue (indicating a strong negative correlation) to dark red (indicating a strong positive correlation), with lighter shades indicating weaker correlations. Figure 2.2 indicates that there are no strong negative correlations among the 18 features with a few strong positive correlations present. Majority of the features have a correlation coefficient between -0.5 and 0.5.

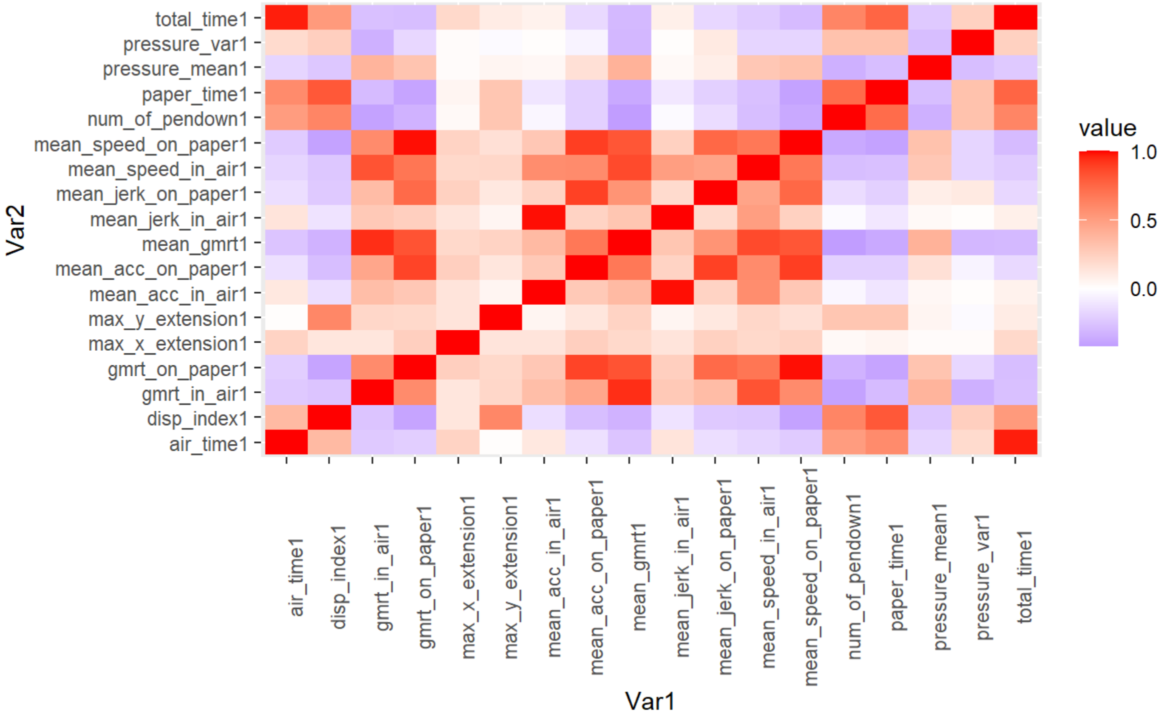


Figure 2.2: Heat map of total_time

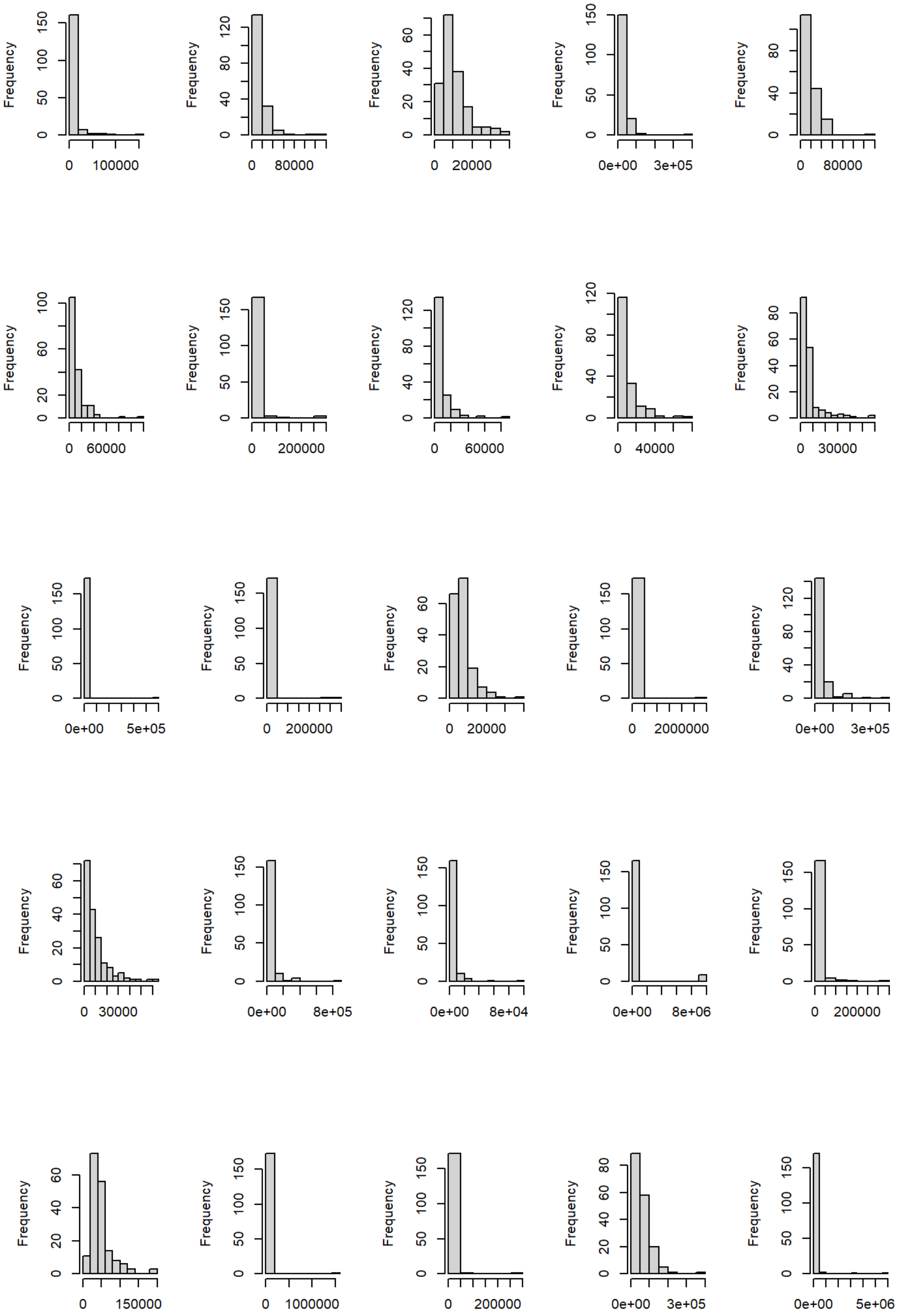


Figure 2.3: Histograms of the distribution of total_time

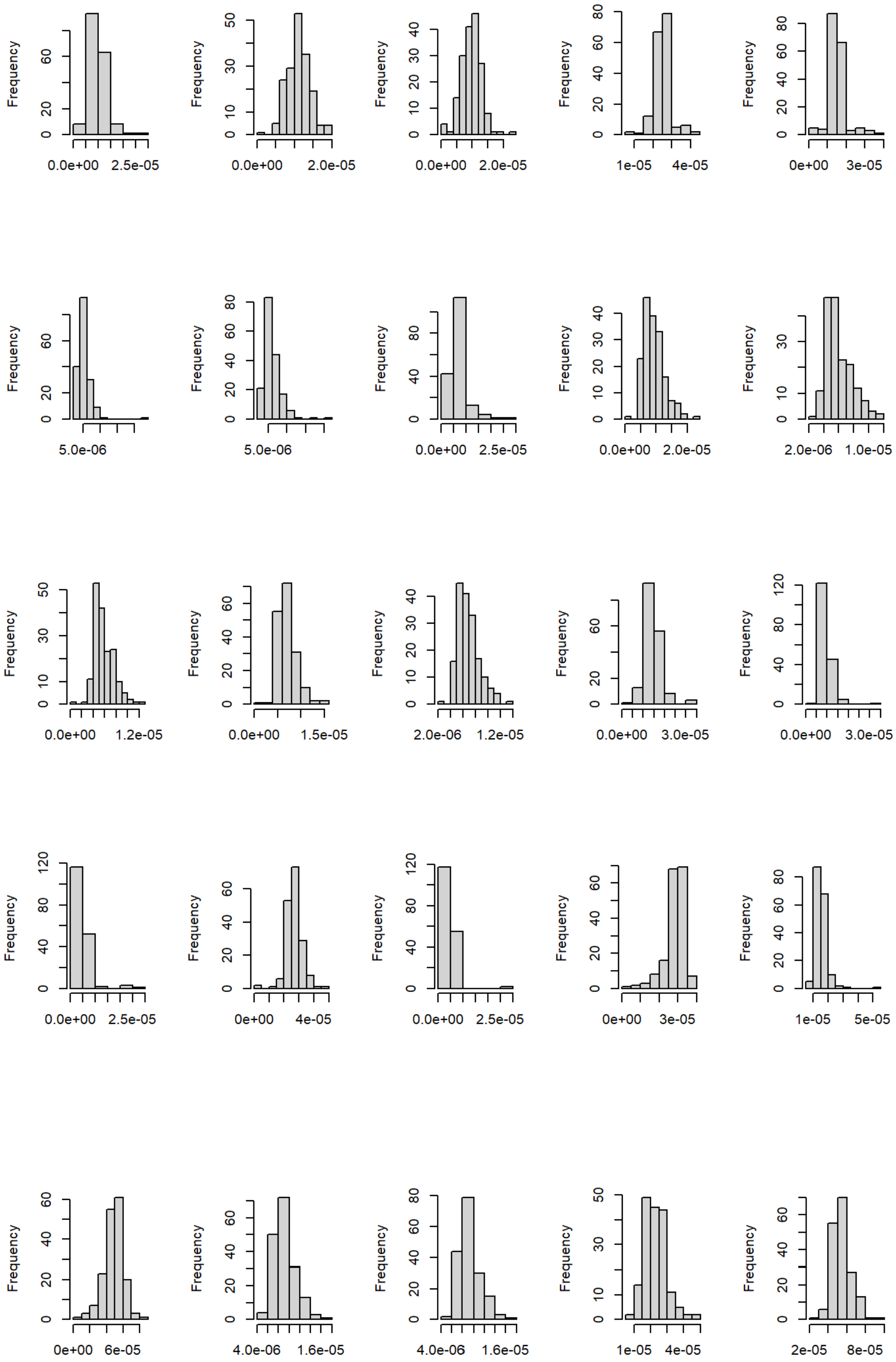


Figure 2.4: Histograms of the distribution of disp_index

CHAPTER 3

Methods

The first step is to split the data into training and testing sets. This ensures that the data used to train the model is not used to make predictions of Alzheimer's disease. In this project, we choose the size of the training set to be 80% of the data size, which is 139, and the size of the testing set to be 20% of the data size, which is 35. Since there are a total of 450 features and only 174 rows, there exists potential issues in model fitting and analysis, such as overfitting and lack of generalization. To mitigate these problems, we will run two experiments. In the first experiment, we will evaluate the performance of each model by considering all 450 features. In the second experiment, we will fit each model on 25 different feature sets, with one set corresponding to each task. The classification models used are logistic regression, random forest, and extreme gradient boost. We chose these models because they are widely used and that they represent a variety of classification algorithms. The techniques that we will use to assess model performance are model accuracy, ROC (receiver operating characteristic) curve, and AUC (area under the ROC curve).

3.1 Logistic Regression

Since the outcome variable "class" is a binary categorical variable that takes values of 0 and 1, a common and simple method to use is logistic regression. Logistic regression is a supervised machine learning binary classification algorithm [6]. The outcome follows the logistic sigmoid function:

$$P(x) = \frac{1}{1 + e^{-x'\beta}},$$

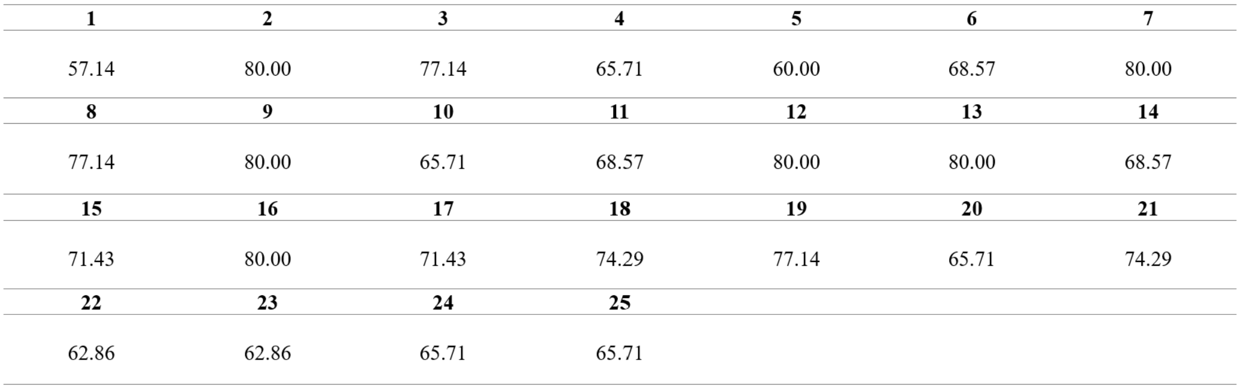
where P(x) is the probability function that returns a value between 0 and 1, e is Euler's number, x is the set of features, and β is a vector of unknown parameters. The advantages of logistic regression include its ease to use, taking less time to train, and tendency to yield low variance, while the disadvantage is that it does not work well with highly correlated attributes [7].

In the first experiment, we fit the data with logistic regression using all 450 features and calculate model accuracy using the formula

$$\text{Accuracy} = \frac{\text{\# of Correct Predictions}}{\text{Total \# of Predictions}}$$

The resulting model accuracy is 57.14%. This accuracy suggests that logistic regression may not be an effective model for all 450 features, as its performance is only slightly better than random guessing, which is 50%. Therefore, we evaluate whether model accuracy can be further improved by fitting logistic regression on each of the 25 tasks separately in the second experiment. The resulting 25 accuracies are shown in Table 3.1. The lowest accuracy is observed in task 1, which is the same accuracy as if we fit all 450 features together. The highest accuracies, 80.00%, are observed in tasks 2, 7, 9, 12, 13, and 16. Notably, tasks 7, 9, 12, 13, and 16 are all copy tasks, suggesting that logistic regression may be more effective at using copy tasks to predict Alzheimer's disease. By averaging the 25 model accuracies, we obtain a mean accuracy of 71.20%. This suggests that, on average, logistic regression performs better when applied separately to individual tasks compared to using all features together.

Table 3.1: Accuracy (in percentage) achieved by logistic regression on each task



1	**2**	**3**	**4**	**5**	**6**	**7**
57.14	80.00	77.14	65.71	60.00	68.57	80.00
8	**9**	**10**	**11**	**12**	**13**	**14**
77.14	80.00	65.71	68.57	80.00	80.00	68.57
15	**16**	**17**	**18**	**19**	**20**	**21**
71.43	80.00	71.43	74.29	77.14	65.71	74.29
22	**23**	**24**	**25**			
62.86	62.86	65.71	65.71			

ROC curves depict the trade-off between sensitivity (true positive rate) and specificity (false positive rate), and are used extensively in clinical assessments and in classification of diseased individuals from the healthy individuals [8]. Optimal performance is indicated by a curve closer to the top-left corner, while inferior performance is indicated by a curve closer to the diagonal line. We will use AUC values in comparison with ROC curves to assess model performance and the model's ability to discriminate Alzheimer's patients from the healthy individuals. The AUC value ranges between 0.5 and 1, with higher values indicating a better ability to discriminate between patients and healthy individuals [9]. The model fitted on task 2 yielded the highest AUC value of 0.8421, while the model fitted on task 22 yielded the lowest AUC value of 0.602. By considering model accuracy, ROC curves, and AUC values, the logistic regression model fitted on task 2 is the top performing model.

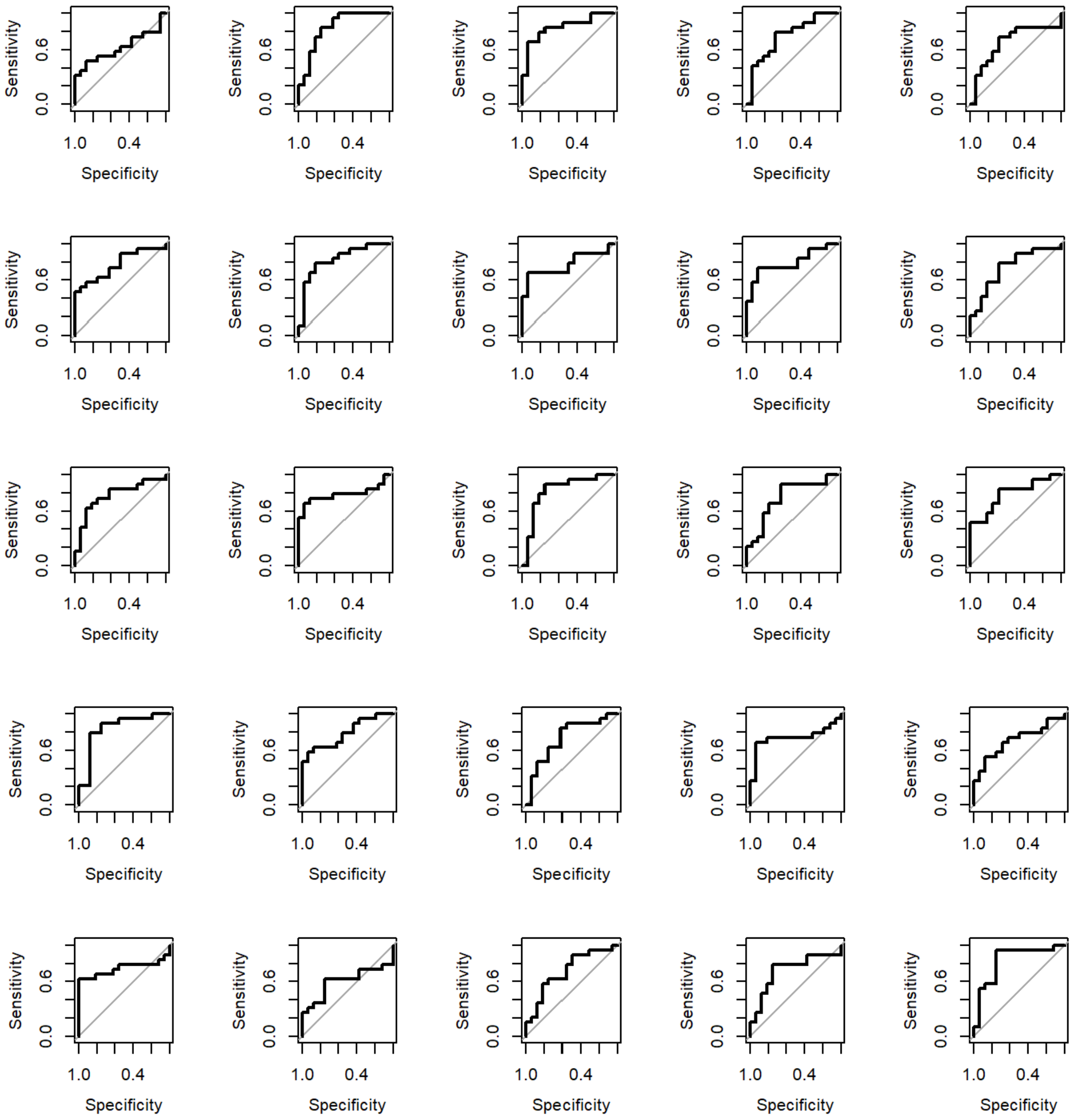
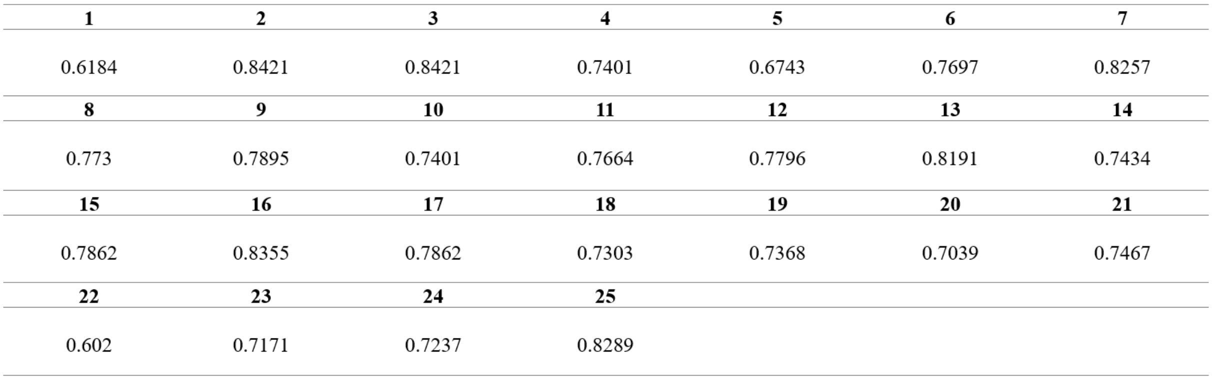


Figure 3.1: ROC curves for logistic regression on 25 tasks

Table 3.2: AUC for logistic regression on 25 tasks



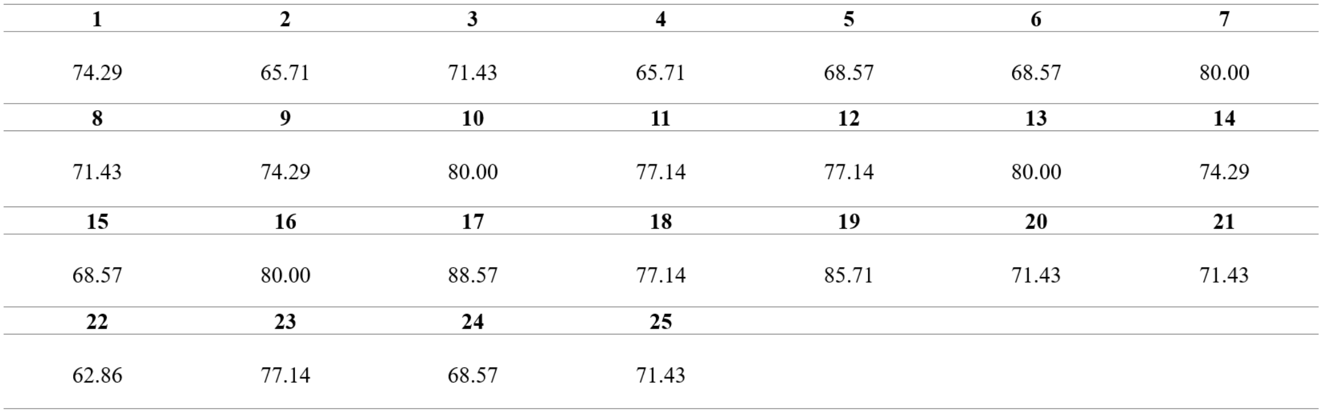
1	**2**	**3**	**4**	**5**	**6**	**7**
0.6184	0.8421	0.8421	0.7401	0.6743	0.7697	0.8257
8	**9**	**10**	**11**	**12**	**13**	**14**
0.773	0.7895	0.7401	0.7664	0.7796	0.8191	0.7434
15	**16**	**17**	**18**	**19**	**20**	**21**
0.7862	0.8355	0.7862	0.7303	0.7368	0.7039	0.7467
22	**23**	**24**	**25**			
0.602	0.7171	0.7237	0.8289			

3.2 Random Forest

Random forest is a supervised machine learning algorithm that is efficient in classification problems by combining multiple decision trees to produce more accurate predictions [10]. Since these decision trees were trained on different subsets of the data, random forest is more resistant to the problem of overfitting. Other advantages of random forest include high flexibility and high accuracy, while the disadvantages are that it is time consuming and requires a lot of computation [7].

In the first experiment, we first fit the data with a random forest using all 450 features. The resulting model accuracy is 91.43%, indicating that random forest is an effective method in predicting Alzheimer's disease. For analysis purposes, we will proceed to the second experiment in performing random forest on each of the 25 tasks separately. The resulting accuracies are shown in Table 3.3. The lowest accuracy, 62.86%, is observed in task 22, while the highest accuracy, 88.57% is observed in task 17. By averaging the model accuracies of the 25 tasks, we obtain a mean accuracy of 74.06%. Despite the high accuracy observed in task 17, the model accuracy is still lower than the model accuracy obtained by considering all 450 features. This suggests that random forest performs better when applied to all features together compared to when it is applied to individual tasks separately.

Table 3.3: Accuracy (in percentage) achieved by random forest on each task



1	**2**	**3**	**4**	**5**	**6**	**7**
74.29	65.71	71.43	65.71	68.57	68.57	80.00
8	**9**	**10**	**11**	**12**	**13**	**14**
71.43	74.29	80.00	77.14	77.14	80.00	74.29
15	**16**	**17**	**18**	**19**	**20**	**21**
68.57	80.00	88.57	77.14	85.71	71.43	71.43
22	**23**	**24**	**25**			
62.86	77.14	68.57	71.43			

We proceed to the analysis of ROC curves and AUC interpretations. From Figure 3.2, we observe that the ROC curve for task 17 is extremely close to the top-left corner, which aligns with its extremely high AUC value of 0.9803 in Table 3.4. This suggests that random forest fitted on task 17 is the top performing model, given that random forest is fitted on each task separately.

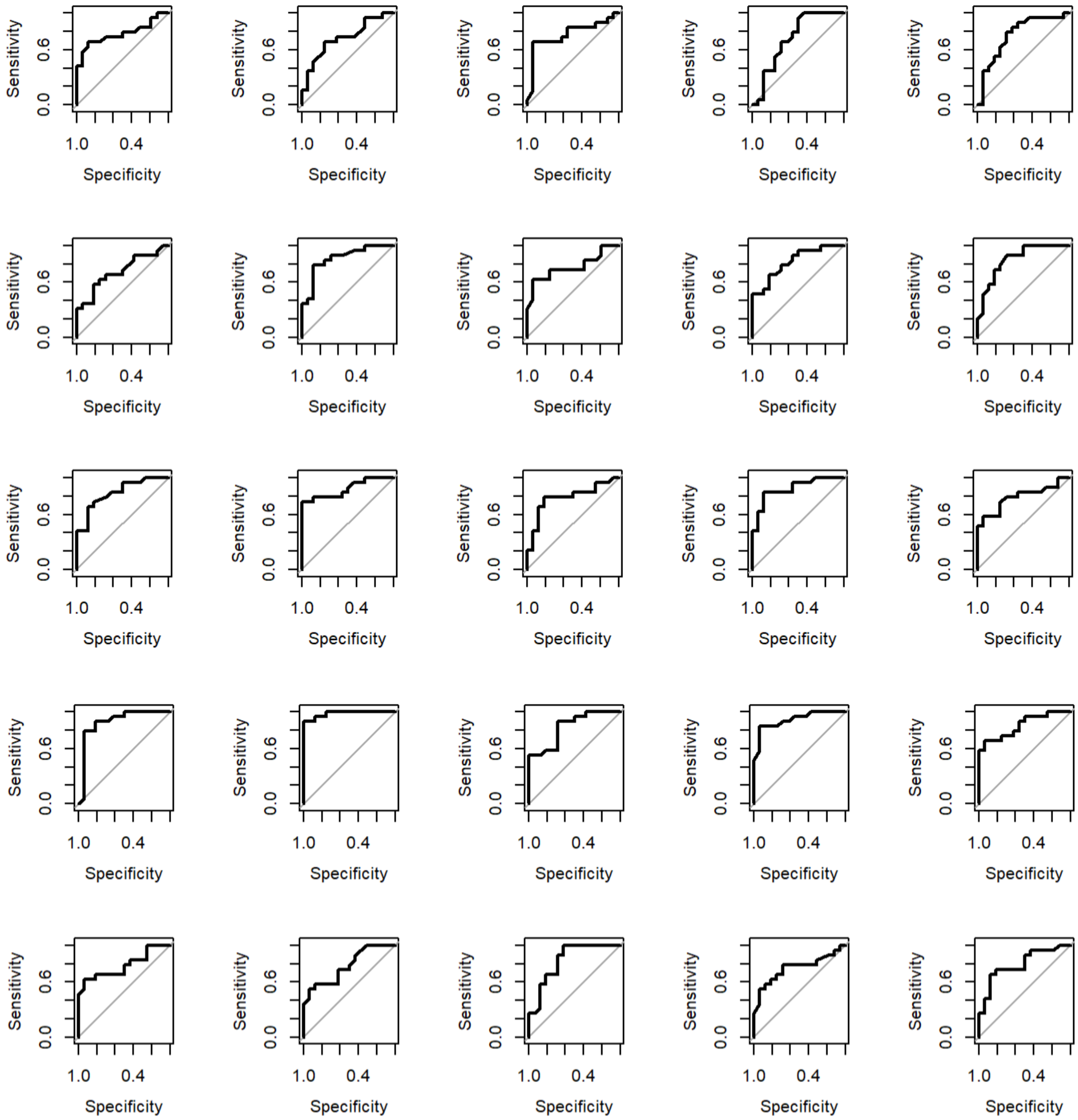
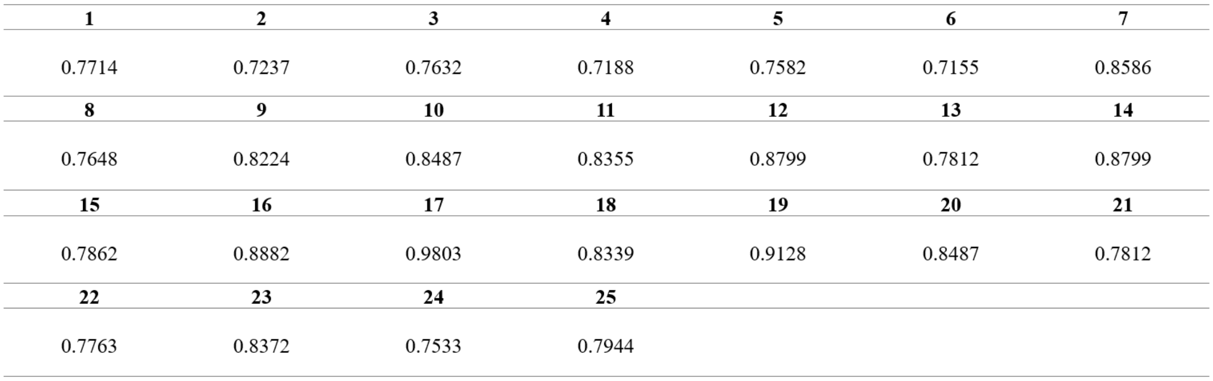


Figure 3.2: ROC curves for random forest on 25 tasks

Table 3.4: AUC for random forest on 25 tasks



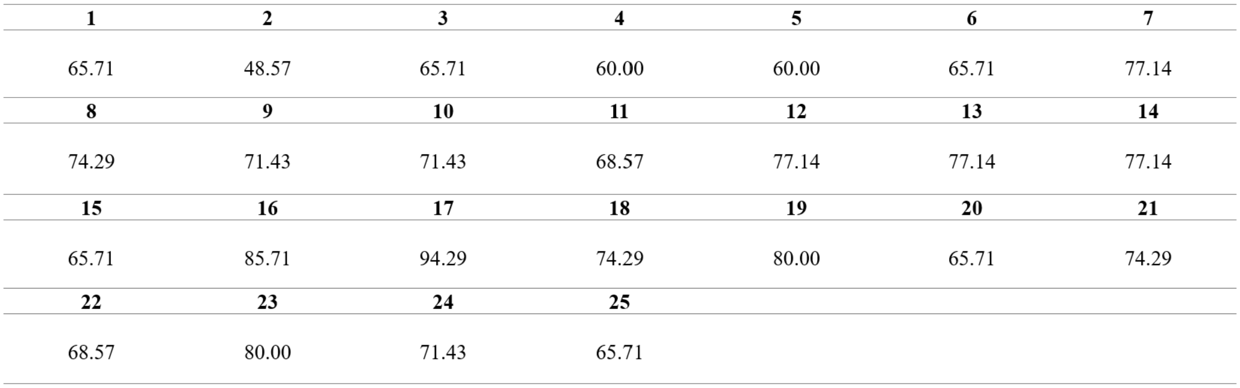
1	**2**	**3**	**4**	**5**	**6**	**7**
0.7714	0.7237	0.7632	0.7188	0.7582	0.7155	0.8586
8	**9**	**10**	**11**	**12**	**13**	**14**
0.7648	0.8224	0.8487	0.8355	0.8799	0.7812	0.8799
15	**16**	**17**	**18**	**19**	**20**	**21**
0.7862	0.8882	0.9803	0.8339	0.9128	0.8487	0.7812
22	**23**	**24**	**25**			
0.7763	0.8372	0.7533	0.7944			

3.3 Extreme Gradient Boosting

Extreme gradient boosting (XGBoost) is a decision tree-based machine learning algorithm. It is more efficient than other algorithms because the output of the model is decided by previous trees rather than the majority [11]. Other advantages of XGBoost include ability to prevent overfitting for clean data and ability to handle data with missing values, while the disadvantage is that XGBoost is more difficult to understand [7].

In the first experiment, we first fit the data with XGBoost using all 450 features. The resulting model accuracy is 85.71%, which suggests that XGBoost is an efficient method in predicting Alzheimer's disease. We can check if there is more room for improvement by fitting XGBoost on each of the 25 tasks separately in the second experiment. The resulting model accuracies for the 15 tasks are shown in Table 3.5. The lowest accuracy, 48.57%, is observed in task 2, while the highest accuracy, 94.29% is observed in task 17. To determine whether XGBoost fitted on task 17 is the top performing model, we need to employ ROC curves and AUC analysis. By averaging the accuracies of the 25 tasks, we obtain a mean accuracy of 71.43%, which is lower than the model accuracy obtained by considering all 450 features.

Table 3.5: Accuracy (in percentage) achieved by XGBoost on each task



1	**2**	**3**	**4**	**5**	**6**	**7**
65.71	48.57	65.71	60.00	60.00	65.71	77.14
8	**9**	**10**	**11**	**12**	**13**	**14**
74.29	71.43	71.43	68.57	77.14	77.14	77.14
15	**16**	**17**	**18**	**19**	**20**	**21**
65.71	85.71	94.29	74.29	80.00	65.71	74.29
22	**23**	**24**	**25**			
68.57	80.00	71.43	65.71			

We then proceed to the analysis of ROC curves and AUC interpretations. From Figure 3.3, we see that the ROC curve for task 17 is extremely close to the top-left corner, which aligns with its extremely high AUC value of 0.9803 in Table 3.6. By considering model accuracy, ROC curves, and AUC, the top performing model is XGBoost fitted on task 17, given that XGBoost is fitted on each task separately.

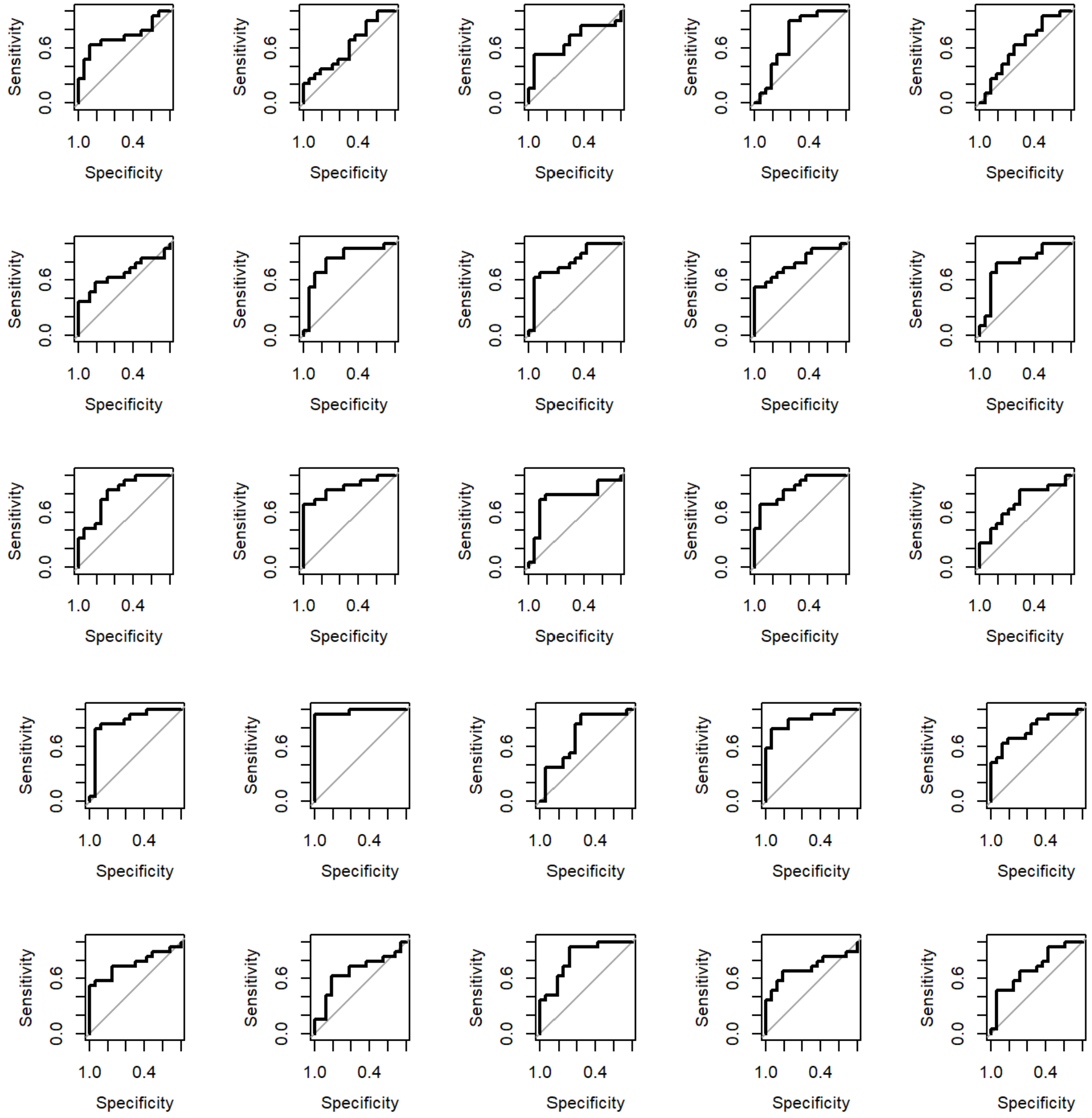
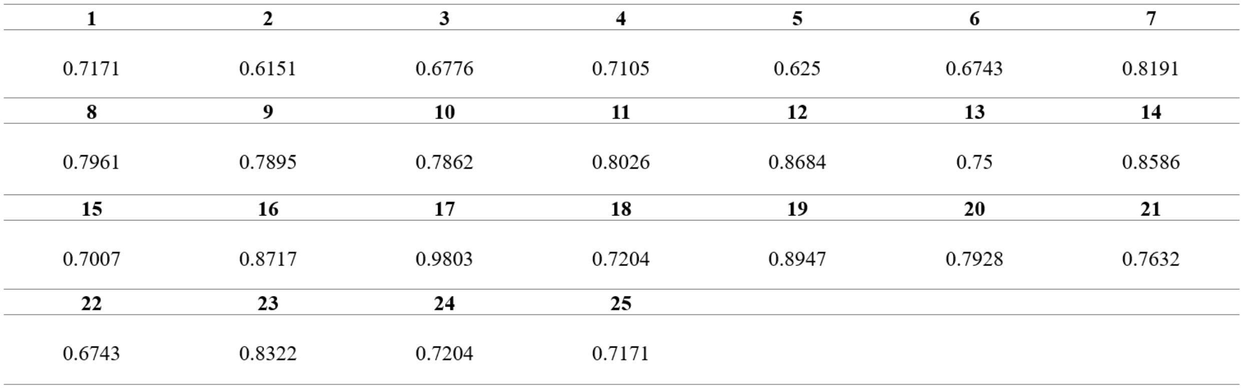


Figure 3.3: ROC curves for XGBoost on 25 tasks

Table 3.6: AUC for XGBoost on 25 tasks



1	**2**	**3**	**4**	**5**	**6**	**7**
0.7171	0.6151	0.6776	0.7105	0.625	0.6743	0.8191
8	**9**	**10**	**11**	**12**	**13**	**14**
0.7961	0.7895	0.7862	0.8026	0.8684	0.75	0.8586
15	**16**	**17**	**18**	**19**	**20**	**21**
0.7007	0.8717	0.9803	0.7204	0.8947	0.7928	0.7632
22	**23**	**24**	**25**			
0.6743	0.8322	0.7204	0.7171			

CHAPTER 4

Results

By analyzing all 450 features together in the first experiment, the model accuracy of logistic regression, random forest, and XGBoost are 57.14%, 91.43%, and 85.71%, respectively. We observe that the model accuracies from the three models vary significantly, with the model accuracy obtained from logistic regression being only slightly better than random guessing. A comparison of the ROC curves of the three methods is shown in Figure 4.1. We observe that the ROC curve for logistic regression is very close to the central diagonal line, suggesting its performance is only slightly better than chance. ROC curves for both random forest and XGBoost are close to the top-left corner, indicating good model performance.The AUC values for logistic regression, random forest, and XGBoost are 0.5312, 0.9441, and 0.9276, respectively. The results from model accuracy, ROC curves, and AUC all indicate that random forest is the top performing model for considering all 450 features.

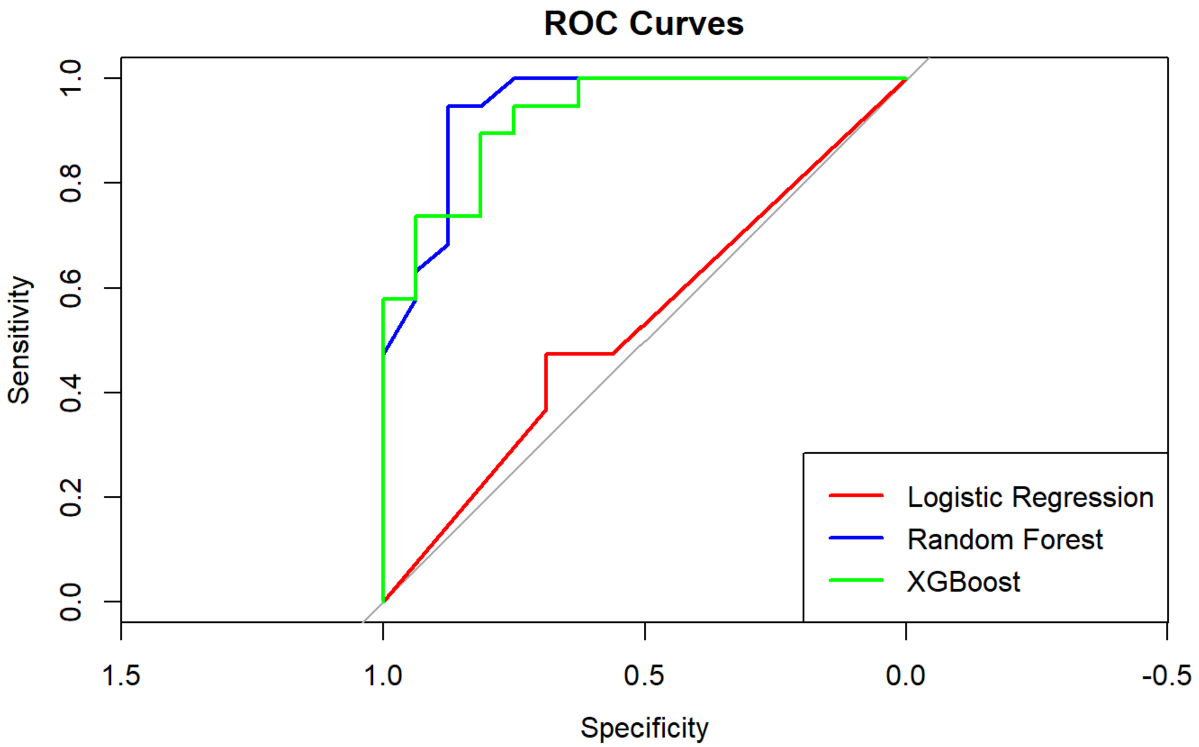


Figure 4.1: ROC curve of methods applied to all 450 features

Since different classification models achieved varying performance, we designed a second experiment where we applied three models on each task separately and combined the results. In the second experiment, the mean accuracy over the 25 tasks for logistic regression, random forest, and XGBoost are 71.20%, 74.06%, and 71.43%, respectively. On average, random forest is the best performing model. Unlike in the first experiment, where the differences between model accuracies were significant, the differences between model accuracies fitted on each of the 25 tasks separately were small. The mean accuracies over the 25 tasks suggest that all three methods are good performing models, but there is potential room for improvement. One notable result observed in Chapter 3 from the AUC tables is that both random forest and XGBoost models fitted on task 17 yielded the same AUC value of 0.9803. This suggests that the two methods, on average, have the same ability in distinguishing the two classes using task 17. The remarkably high AUC indicates that both random forest and XGBoost models fitted on task 17 have excellent power in discriminating between Alzheimer's patients and healthy individuals.

Moreover, the results from the predictive models suggest which sets of features are effective in distinguishing Alzheimer's patients with healthy individuals. The remarkably high accuracy and AUC of task 17 suggests that task 17 is significant in predicting Alzheimer's disease. We can draw a connection to the average total_time for task 17 mentioned in Section 2.3, where task 17 has the second longest average total_time. Task 19, which has the longest average total_time among the 25 tasks, is the only other task with an AUC over 0.9 besides task 17. These results suggest that tasks with higher levels of difficulty are likely to be more effective in distinguishing between Alzheimer's patients and healthy individuals. Another notable result is that the model accuracy obtained from XGBoost on task 2 is 48.57%, which is lower than the accuracy obtained from random guessing. This result is surprising since XGBoost is known for its high accuracy. The low accuracy suggests that task 2 may not be significant in predicting Alzheimer's disease using XGBoost and therefore could be eliminated when constructing the final model.

In determining the best classification model for each task, we will prioritize AUC over model accuracy due to its robustness against prior class probabilities [12]. Consider the abbreviation for logistic regression, random forest, and XGBooost as LR, RF, and XGB, respectively. The best classification methods by task are RF, LR, LR, LR, RF, LR, RF, XGB, RF, RF, RF, RF, LR, RF, LR, RF, XGB, RF, RF, RF, RF, RF, RF, RF, and LR. We observe that random forest is the best method for 16 out of the 25 tasks, indicating its strong performance and robustness in classifying Alzheimer's disease. In both experiments, random forest has the best predictive power among the three methods. Despite that the mean accuracies over the 25 tasks for logistic regression and XGBoost have minimal difference, logistic regression is the best method for 7 tasks while XGBoost is the best for only 2 tasks. Although XGBoost is the most preferred method in developing predictive models due to its remarkable accuracy and many notable advantages [13], the low occurrence of XGBoost suggests that other methods are more suitable for these 25 specific tasks.

CHAPTER 5

Conclusion and Future Work

The goal of this project is to implement three machine learning algorithms, logistic regression, random forest, and extreme gradient boosting, on the handwriting dataset and evaluate model performance using model accuracy, ROC curves, and AUC interpretations. The results from chapter 3 and 4 demonstrate that employing random forest on the 18 features extracted from the 25 tasks, totaling 450 features, yielded better performance than the models fitted solely on the 18 features from each task. The resulting model accuracy is 91.43% with AUC of 0.9441. From the results, we can infer that the feature set of task 17 is effective in distinguishing Alzheimer's patients with the healthy individuals.

All together, the results support the fact that including different handwriting tasks with different levels of difficulty allows for the discrimination of Alzheimer's patients with healthy individuals. Furthermore, these results prove that the three methods all achieved a mean accuracy over 70%, indicating the set of features used in this project is capable of capturing the distinctive handwriting characteristics of Alzheimer's patients. All the models, except for logistic regression applied to all 450 features, have demonstrated good performance in predicting Alzheimer's disease.

Even though the three methods demonstrated good performance, there is still room for improvement. For example, we can reduce bias introduced by randomness by performing multiple runs. We could also perform k-fold cross validation before training the data to yield a more reliable estimate model performance. In addition, utilizing the tuning parameter that minimizes the cross-validated error to build the final model helps prevent overfitting

[14]. Lastly, instead of fitting the models on each task separately, we could apply feature selection techniques to extract the most relevant features in model construction to increase model accuracy and efficiency. Since each feature selection technique has its own strengths and weaknesses, it is common that researchers combine multiple techniques to extract the most relevant features [15].

REFERENCES

[1] C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. Tavera-Romero, "Early-stage alzheimer's disease prediction using machine learning models," *Front Public Health*, vol. 10, Mar. 2022, doi: 10.3389/fpubh.2022.853294.

[2] S. S. Rana, X. Ma, W. Pang, and E. Wolverson, "A multi-modal deep learning approach to the early prediction of mild cognitive impairment conversion to alzheimer's disease," *2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, pp. 9–18, 2020, doi: 10.1109/BDCAT50828.2020.00013.

[3] K. Huang et al., "A multipredictor model to predict the conversion of mild cognitive impairment to alzheimer's disease by using a predictive nomogram," *Neuropsychopharmacol.*, vol. 45, pp. 358–366, Jan. 2020, doi:10.1038/s41386-019-0551-0.

[4] N. D. Cilia, G. D. Gregorio, C. D. Stefano, F. Fontanella, A. Marcelli, and A. Parziale, "Diagnosing alzheimer's disease from online handwriting: A novel dataset and performance benchmarking," *Engineering Applications of Artificial Intelligence*, vol. 111, Mar. 2022, doi: 10.1016/j.engappai.2022.104822.

[5] F. Fontanella, "Darwin," *UCI Machine Learning Repository*, 2022, doi: 10.24432/C55D0K.

[6] G. Ambrish, B. Ganesh, A. Ganesh, C. Srinivas, Dhanraj, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 127–130, June 2022, doi: 10.1016/j.gltp.2022.04.008.

[7] B. Akkaya and N. Çolakoğlu, "Comparison of multi-class classification algorithms on early diagnosis of heart diseases," *y-BIS Conference 2019: Recent Advances in Data Science and Business Analytics*, pp. 162–172, 2019.

[8] K. Hajian-Tilaki, "Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation," *Caspian J Intern Med. 2013 Spring*, vol. 4, no. 2, pp. 627–635, Sep. 2013.

[9] A. A. H. de Hond, E. W. Steyerberg, and B. van Calster, "Interpreting area under the receiver operating characteristic curve," *Lancet Digit Health*, vol. 4, no. 12, pp. e853–e855, Dec. 2022, doi: 10.1016/S2589-7500(22)00188-1.

[10] M. Velazquez, Y. Lee, and for the Alzheimer's Disease Neuroimaging Initiative, "Random forest model for feature-based alzheimer's disease conversion prediction from early mild cognitive impairment subjects," *PLoS ONE*, vol. 16, no. 4, Apr. 2021, doi: 10.1371/journal.pone.0244773.

[11] S. Doki, S. Devella, S. Tallam, S. S. Reddy Gangannagari, P. Sampathkrishna Reddy, and G. P. Reddy, “Heart disease prediction using xgboost,” *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, pp. 1317–1320, 2022, doi: 10.1109/ICICICT54557.2022.9917678.

[12] E. LeDell, M. J. van der Laan, and M. Petersen, “Auc-maximizing ensembles through metalearning,” *Int J Biostat*, vol. 12, no. 1, pp. 203–218, May 2016, doi: 10.1515/ijb-2015-0035.

[13] D. Tarwidi, S. R. Pudjaprasetya, D. Adytia, and M. Apri, “An optimized xgboost-based machine learning method for predicting wave run-up on a sloping beach,” *MethodsX*, vol. 10, Mar. 2023, doi: 10.1016/j.mex.2023.102119.

[14] D. Berrar, “Cross-validation,” *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, pp. 542–545, doi: 10.1016/B978- 0-12- 809633-8.20349- X.

[15] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O’Sullivan, “A review of feature selection methods for machine learning-based disease risk prediction,” *Front Bioinform*, vol. 2, June 2022, doi: 10.3389/fbinf.2022.927312.