THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Feature Selection for Raman Spectra of Alzheimer's Disease Through Recursive Feature Elimination with Random Forest

ISABELLE BIASE
SPRING 2021

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Computer Science
with honors in Computer Science

Reviewed and approved* by the following:

Shengxi Huang
Assistant Professor of Electrical Engineering and Biomedical Engineering
Thesis Supervisor

John Hannan
Associate Professor of Computer Science and Engineering
Honors Adviser

* Electronic approvals are on file.

Abstract

The goal of this thesis is to apply the Random Forest-Recursive Feature Elimination (RF-RFE) algorithm to the classification of Raman spectra related to Alzheimer's disease (AD). In recent research, machine learning methods have demonstrated success in classifying Raman spectra on mouse brain slices with AD and without AD, and important Raman signature bands have been identified by leveraging the feature importance maps of the machine learning models. However, features of a Raman spectrum are necessarily correlated since Raman signature bands span neighboring positions in the spectrum, and it has been shown that a high number of correlated features can worsen classifier performance. In this thesis, we build on this recent research by applying a feature selection algorithm called Recursive Feature Elimination in order to reduce the presence of irrelevant and correlated data points. We compare the abilities of Random Forest-Recursive Feature Elimination and Random Forest alone to classify Raman spectra on mouse brain slices with AD and without AD, and we conclude that RF-RFE performs just as well, despite using significantly fewer features. Removing the variables not needed by the model to distinguish between AD and non-AD spectra yields a better focusing on the important spectral differences and might therefore help with AD biomarker identification.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Professor Shengxi Huang, my thesis supervisor, for her guidance and encouragement throughout the thesis process. I would also like to thank Ziyang Wang for his assistance and advice. Finally, I would like to thank Professor John Hannan, my honors advisor, for his constant support throughout my four years as a Schreyer scholar.

Chapter 1

Introduction

Alzheimer's disease (AD), a progressive neurological disorder, is the most common cause of dementia [1]. In 2018, approximately 5.7 million Americans had Alzheimer's disease, and this figure is projected to reach 13.8 million by mid-century as the population ages [2]. However, AD pathology is not fully understood, and biomarkers of the disease are not fully identified [1]. With the growing need for early diagnosis and treatment of AD, biomarker identification has become a critical area of research.

In previous research, Raman spectroscopy and machine learning were combined to develop a rapid pre-screening approach for AD biomarkers [3]. In this thesis, we build off of that research by applying the Random Forest-Recursive Feature Elimination algorithm to the classification of Raman spectra on mouse brain slices with AD and without AD. By removing correlated and irrelevant variables through Recursive Feature Elimination, our goal is to improve upon the performance of the Random Forest alone and obtain a better feature importance map, which might guide biomarker discovery in future research.

This thesis is organized as follows. In Chapter 2, we provide relevant background information about Raman spectroscopy, the Random Forest algorithm, and Recursive Feature Elimination. In Chapter 3, we present the details of our experiment, discussing how prior research led to our use of a subset of the original dataset as well as how the machine learning classification was implemented. Next, in Chapter 4, we evaluate the RF and RF-RFE classifiers

and compare their performances. Finally, in Chapter 5, we summarize our results and present ideas for future research in this area.

Chapter 2

Background

Raman Spectroscopy

Raman spectroscopy is a spectroscopic technique for the analysis of molecular structure. It is based on a phenomenon called Raman scattering, first observed by Indian scientist C.V. Raman in 1928. When monochromatic radiation, usually from a laser, interacts with molecules, most of the scattered radiation is at the same frequency of the laser source; this is referred to as Rayleigh scattering. A small amount of scattered light, however, is at a new frequency; this is referred to as Raman scattering. The new frequencies depend on the chemical structure of the molecule and together constitute a Raman spectrum. Raman spectra are uniquely characteristic of a molecule and thus serve as an identifying fingerprint [4].

In recent years, there has been a surge in the literature involving the use of machine learning for Raman spectroscopy analysis, especially in biomedical applications such as cancer diagnosis, viral and bacterial infections, and neurodegenerative and autoimmune disorders [5]. Regarding Alzheimer's disease specifically, previous research, which we build off of in this thesis, has combined Raman spectroscopy and machine learning to develop a rapid pre-screening approach for AD biomarkers.

Random Forest Algorithm

Ensemble learning methods are algorithms that build a set of classifiers and combine their outputs by taking a vote of their predictions. The main motivation behind the ensemble

approach is that a group of "weak" learners can be combined to form a "strong" learner [6]. The Random Forest algorithm, proposed by L. Breiman in 2001, is one such ensemble method that combines many decision tree classifiers [7]. Aggregating individual decision trees, each having low bias but high variance, achieves a bias-variance tradeoff and, consequently, better performance [7].

Random Forest, a supervised learning method, can be used for both classification and regression. It has been shown to perform well on high-dimensional data, and it returns measures of variable importance with respect to the prediction [2]. However, when the number of variables is much larger than the number of observations, there are often irrelevant and correlated features present, which can affect Random Forest's ability to identify the strongest predictors. One proposed solution to this problem is the use of the Random Forest-Recursive Feature Elimination (RF-RFE) algorithm [8].

Recursive Feature Elimination

In machine learning, feature selection is the process of choosing a subset of relevant variables from the input data to use to train the model. In a dataset with many variables, some may be highly correlated with others and some may not provide any useful information to the classifier. These irrelevant and correlated variables serve as noise to the model, and total information content can be obtained without them. The goal of feature selection is to determine a subset of features that efficiently describes the input data. Feature selection helps in reducing computation time, improving classifier performance, and understanding the data. Indeed, a

model with a smaller number of variables is more interpretable, and the risk of overfitting is reduced [9].

There are several types of feature selection methods: filter, wrapper, and embedded [9]. Recursive Feature Elimination (RFE) is a wrapper method, meaning it identifies an optimal set of variables among all possible subsets based on the estimator used in the learning algorithm. Because it is not computationally possible to evaluate all possible subsets of variables, wrapper methods typically employ greedy strategies such as forward or backward algorithms [8]. In the case of RFE, a backward algorithm is used.

The general RFE algorithm is an iterative procedure, where at each step of the backward strategy, the classifier is trained, the ranking criterion is computed for all remaining features, and the feature with the smallest ranking criterion is removed. The classifier is first trained on the set of all features, and features are recursively eliminated based on their rank until all features have been ranked (or until the desired number of features to select is reached) [10]. In the case of Random Forest-Recursive Feature Elimination, the feature importance provided by the RF classifier serves as the ranking criterion. An outline of the RF-RFE algorithm is presented below in Figure 1 [10].

Algorithm 1: Random Forest-Recursive Feature Elimination

Input : Training data $\mathbf{X_0} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]^T$,
Class labels: $\mathbf{y} = [y_1, y_2, ..., y_n]^T$
Output: Final feature ranking **r**

Initialize subset of surviving features s = [1, 2, ..., n] ▷ starting with all features
Initialize feature ranking list r = [ ]
Repeat until s = [ ]
    Restrict training examples to surviving feature indices
        $\mathbf{X} = \mathbf{X_0}(:, s)$
    Train the Random Forest classifier
    Compute the feature importance score of each feature based on classifier
    Find the feature with the smallest feature importance score
        f = *argmin*(feature importances)
    Update feature ranking list
        r = [**s**(f), **r**]
    Eliminate the feature with the smallest feature importance score
        s = **s**(1 : f-1, f+1 : length(**s**))

Figure 1: Random Forest-Recursive Feature Elimination pseudocode

Recursive Feature Elimination was introduced by Guyon et al. in 2002 to improve classification performance of Support Vector Machines (SVM) [10]. This strategy has since been applied to Random Forest and has been proven valuable in the presence of correlated features [8, 11]. Because features of a Raman spectrum are necessarily correlated (since Raman signature bands span neighboring positions in the spectrum), RF-RFE was chosen as an appropriate algorithm for our investigation.

Chapter 3

Method

Dataset

In previous research, brain slices were harvested from four types of mice: AD age 4 months, AD age 14 months, without AD age 4 months, and without AD age 14 months. For each brain slice, Raman spectra were collected from 3 brain regions: cortex, hippocampus, and thalamus. The brain slices were immersed in neuroprotectant solution sealed between silicon substrate and a fused quartz cover slide. For some of the measurements, the brain slice was placed in direct contact with monolayer graphene, which was transferred onto the quartz cover slide. Complete details regarding the original dataset can be seen in Tables 1 (graphene-enhanced) and 2 (without graphene). In total, there were 727 samples [3].

ID	Sample	Label	Age	Gender	Region ID	Region
1-81	Sample 1	Non-AD	14 months	Female	1-27	Hippo
					28-54	Thalamus
					55-81	Cortex
82-162	Sample 2	AD	14 months	Male	82-108	Hippo
					109-135	Thalamus
					136-162	Cortex
163-243	Sample 3	Non-AD	4 months	Female	163-189	Hippo
					190-216	Thalamus

					217-243	Cortex
244-324	Sample 4	AD			244-270	Hippo
					271-297	Thalamus
					298-324	Cortex
325-351	Sample 1	Non-AD	14 months	Female	325-333	Hippo
					334-342	Thalamus
					343-351	Cortex
352-378	Sample 2	AD		Male	352-360	Hippo
					361-369	Thalamus
					370-378	Cortex
379-403	Sample 3	Non-AD	4 months	Female	379-403	Hippo

Table 1: Graphene dataset including label, age, gender, and brain region.

ID	Sample	Label	Age	Gender	Region ID	Region
1-81	Sample 1	Non-AD	14 months	Female	1-27	Hippo
					28-54	Thalamus
					55-81	Cortex
82-162	Sample 2	AD	14 months	Male	82-108	Hippo
					109-135	Thalamus
					136-162	Cortex
163-243	Sample 3	Non-AD	4 months	Female	163-189	Hippo
					190-216	Thalamus

					217-243	Cortex
244-324	Sample 4	AD			244-270	Hippo
					271-297	Thalamus
					298-324	Cortex

Table 2. No-graphene dataset including label, age, gender, and brain region.

After preprocessing the raw Raman spectra, which involved implementing Savitzky-Golary filter for spectral smoothing and asymmetric least squares smoothing for baseline correction, and calculating signal-to-noise ratio for spectra with and without the graphene substrate, the research found that graphene-enhanced spectra have much higher signal-to-noise ratio than those without graphene and used the graphene-enhanced samples for further investigation [3].

Next in the research, different machine learning algorithms were applied to the graphene-enhanced Raman spectra, including Linear SVM, Random Forest, XGBoost, and CatBoost. Among the three brain regions, the classification accuracy for every classifier was best on the cortex region, and it was concluded that the cortex is an informative brain region with AD-relevant biomarkers easily captured by Raman spectroscopy. Therefore, building off these findings, we will compare RF and RF-RFE on graphene-enhanced Raman spectra from the cortex region in this thesis. There are 126 samples in this reduced dataset [3].

Machine Learning Classification

Machine learning classification experiments were implemented using scikit learn, a Python machine learning library. We used stratified k-fold cross validation with k=5. Stratified

k-fold cross validation is a type of cross validation in which the class distribution across all folds is kept as similar as possible to the class distribution of the entire dataset, as opposed to being random [12]. Because we have relatively few training data points, this technique was used to preserve the same percentage of samples for each class and improve robustness. For each fold, RF-RFE and RF classifiers were trained and evaluated. In addition, this stratified 5-fold cross validation was repeated five times, meaning each classifier was trained 25 times in total, to verify the results.

The Random Forest-Recursive Feature Elimination classifier was implemented using the RFECV method in scikit learn's feature selection package in combination with scikit learn's built-in Random Forest Classifier. To perform RFE, the number of features to select must be specified in advance, but in most applications, including ours, the optimal number of features is not known ahead of time. RFECV uses cross validation to automatically tune the number of features selected, choosing the number of features that gives the highest cross validation score.

As RFE is performed, the algorithm assigns each feature an integer rank, such that all selected (i.e., estimated best) features are assigned rank 1 and the first feature eliminated has the highest rank. Once we achieved a final model trained on the optimal number of features, we evaluated its performance using test data. This final model also provides feature importance scores for the selected features.

For comparison, we also implemented the Random Forest Classifier without Recursive Feature Elimination. This means it was trained on all 1831 features. Across all 25 models trained, we averaged the F1-scores to assess whether RF-RFE improved upon RF alone. Additionally, we averaged the feature importance scores all 25 iterations for comparison.

Chapter 4

Evaluation

Random Forest-Recursive Feature Elimination

The RF-RFE classifiers achieved an average F1-score of 0.9625, and the average optimal number of features selected was 890 wavenumbers. Among all 25 RF-RFE models trained, the smallest optimal number of features selected was 770, and the largest optimal number of features selected was 1607. A confusion matrix summarizing the performance of each of the 25 RF-RFE classifiers can be found in Appendix A.

Additionally, as RFECV was performed, we plotted the number of features vs. the cross-validation score to visualize how the accuracy changed as features were eliminated. All of the graphs exhibit similar behavior, with the cross-validation score increasing rapidly until approximately 850 features are selected, then plateauing. This pattern, which can be seen in Figures 2-6, suggests that the classifier can perform just as well with only about half of the features. This behavior is also reflected in the fact that the average optimal number of features selected across all 25 RF-RFE models was 890, as previously mentioned.

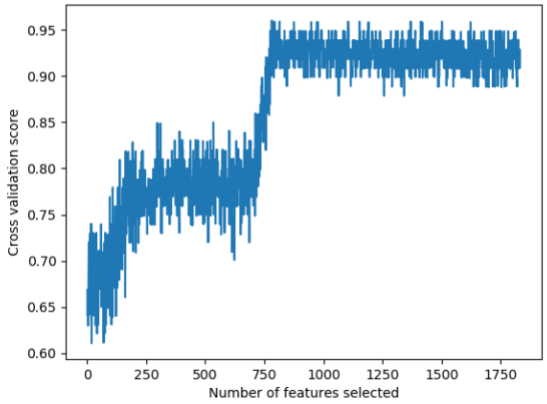


Figure 2: Number of features selected vs. cross validation score for fold 0 of the first repetition of the experiment.

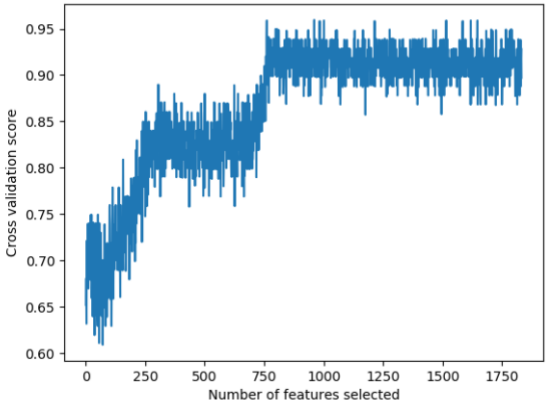


Figure 3: Number of features selected vs. cross validation score for fold 1 of the first repetition of the experiment.

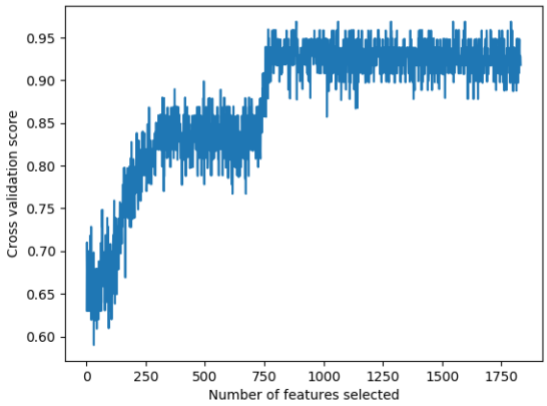


Figure 4: Number of features selected vs. cross validation score for fold 2 of the first repetition of the experiment.

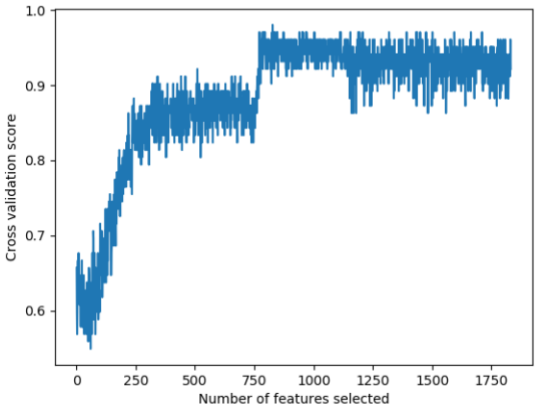


Figure 5: Number of features selected vs. cross validation score for fold 3 of the first repetition of the experiment.

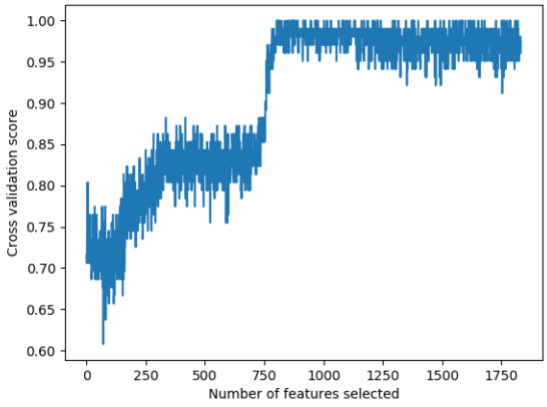


Figure 6: Number of features selected vs. cross validation score for fold 4 of the first repetition of the experiment.

Random Forest

In comparison, the Random Forest classifiers, trained on all 1831 features, achieved an average F1-score of 0.9430. A confusion matrix summarizing the performance of each of the 25 RF classifiers can be found in Appendix B.

RF-RFE vs. RF Comparison

Our results show that feature selection using Recursive Feature Elimination is beneficial for our application. The RF classifiers trained on all features achieved an average F1-score of 0.9430, whereas the RF-RFE classifiers trained on an optimal subset of features achieved a slightly higher average F1-score of 0.9625. Using Recursive Feature Elimination reduced our original set of features from 1831 wavenumbers to 890 wavenumbers on average, meaning that Random Forest could perform slightly better with only about half of the original features. In addition to improved performance, a Random Forest model that uses half the number of variables also benefits from reduced computation time and increased interpretability.

Feature Importance

Each of the trained models provides a feature importance map giving an importance score for each variable. We averaged these maps across all trained models of each type. The average feature importance maps for RF-RFE and RF are shown below in Figures 7 and 8, respectively. A higher score indicates a more important variable.

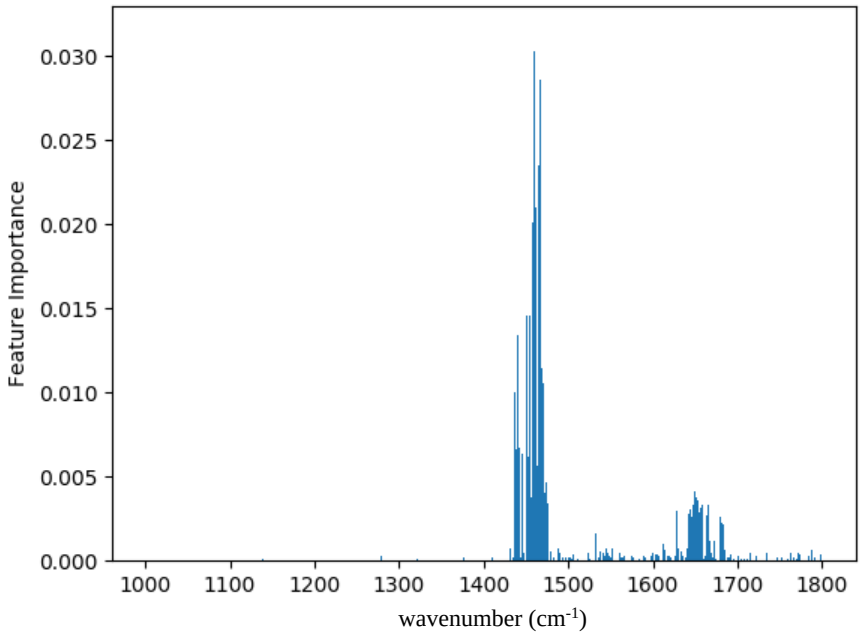


Figure 7: Average feature importance map for RF-RFE classifiers

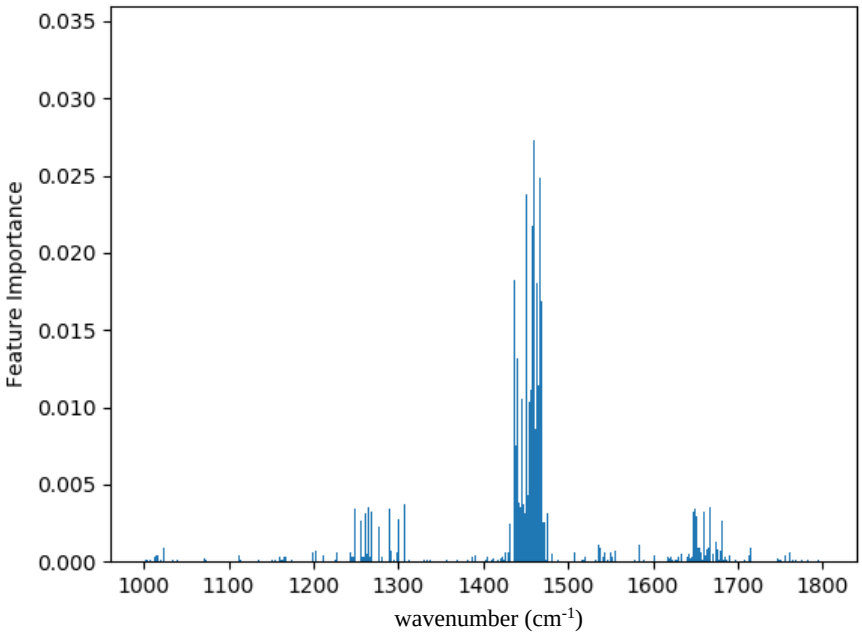


Figure 8: Average feature importance map for RF classifiers

We notice that the feature importance maps for both classifiers have similar shapes. For both RF-RFE and RF, the most important features occur in a peak around 1450 cm^{-1}. In the RF-RFE map, however, the peak around 1650 cm^{-1} is denser, indicating that these wavenumbers were consistently selected through RFE and found to be important in the model's prediction. Additionally, there are small peaks in the 1250-1300 cm^{-1} range of the RF feature importance map that do not appear in the RF-RFE map. In fact, there are hardly any features in the range 1000-1400 cm^{-1} with importance scores above zero in the RF-RFE map, which means that these features were consistently eliminated through Recursive Feature Elimination and may be correlated or irrelevant in predicting AD.

Chapter 5

Conclusion

The application of machine learning to Raman spectroscopy analysis has proven to be an important area of biomedical research. One specific application developed in recent research is the rapid pre-screening for biomarkers of Alzheimer's disease, and in this thesis, we build upon this research, investigating how feature selection through Recursive Feature Elimination improves the performance of a Random Forest classifier on this dataset. In Chapter 1, we discuss the growing importance of AD biomarker identification. In Chapter 2, we provide relevant background information about Raman spectroscopy, Random Forest classifiers, and the Recursive Feature Elimination algorithm. We recall that the Random Forest algorithm is known to perform well on high-dimensional data, but that irrelevant and correlated features can negatively impact its performance, which motivates our investigation of the Random Forest-Recursive Feature Elimination algorithm. In Chapter 3, we discuss the details of our experiment, including the motivations behind the dataset used. Next, in Chapter 4, we evaluate and compare the performances of the RF and RF-RFE classifiers. We conclude that our application benefits from the use of RFE, with the RF-RFE classifiers having a slightly higher average F1-score than the RF classifiers, despite using only about half of the original features. The resulting RF-RFE model better focuses on important spectral differences, which may help with AD biomarker identification.

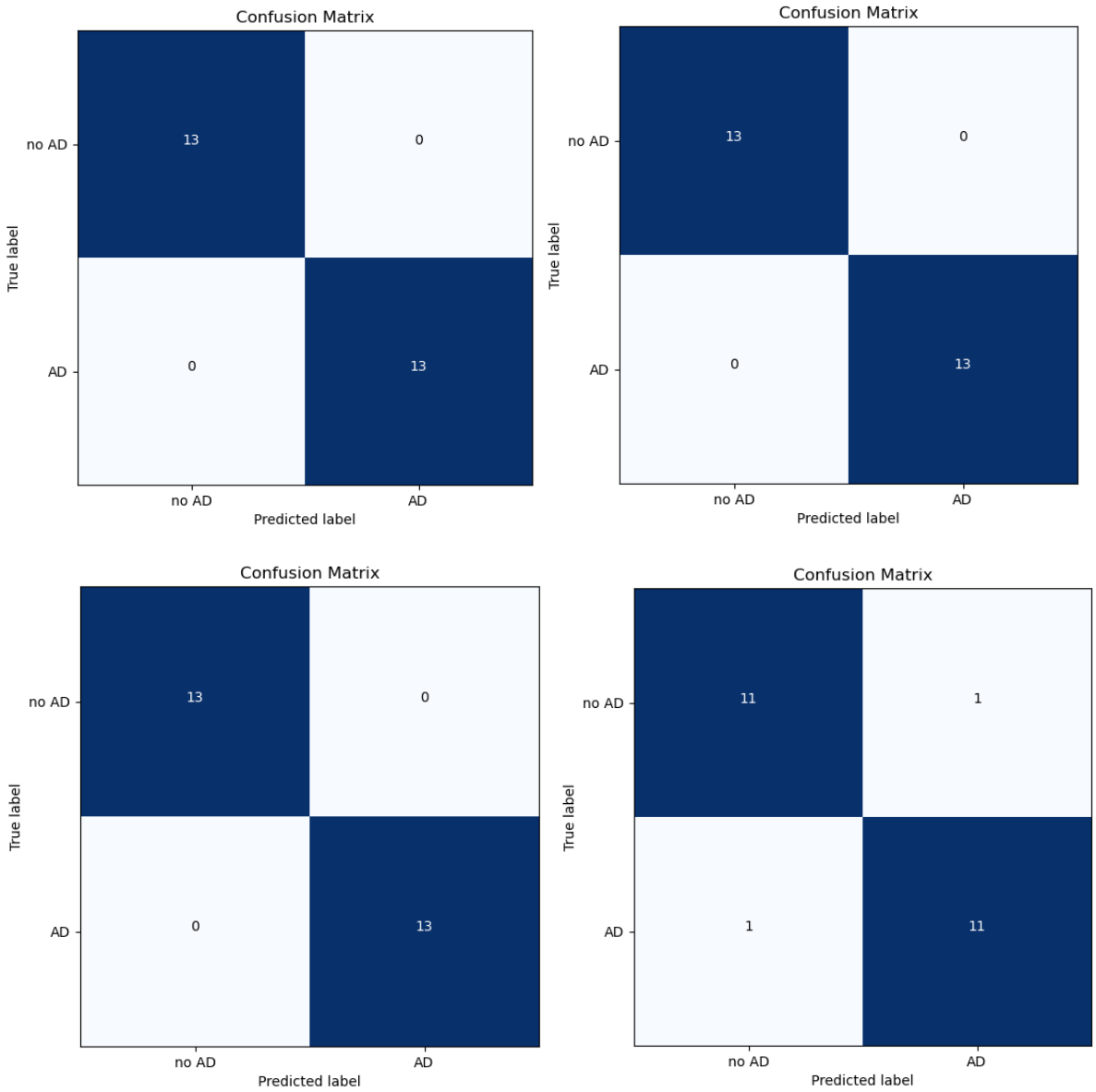
Future Steps

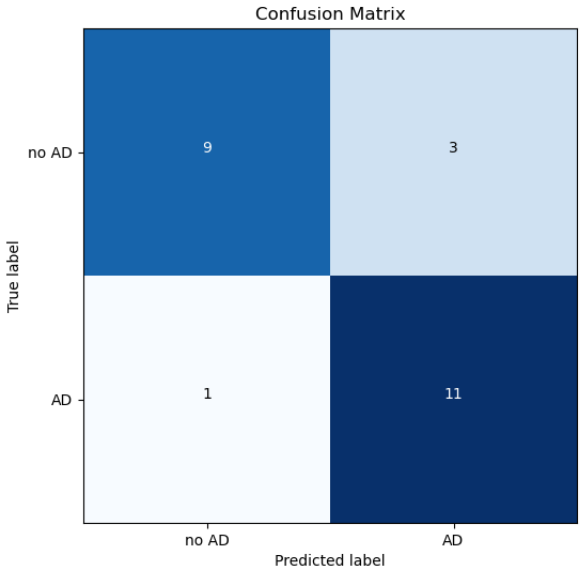
While this thesis validates the benefits of the Random Forest-Recursive Feature Elimination algorithm on the classification of Raman spectra from mouse brain slices with AD and without AD, there are additional future steps that could be taken to better understand our spectral feature importance map and its relationship with potential AD biomarkers. Prior research has developed two metrics to measure this relationship: a Pearson cross-correlation based algorithm, and a matching score based on spectral overlap between important feature ranges and peaks in the Raman spectra of biomarkers [3]. These two metrics could be calculated using the average feature importance map we obtained from the RF-RFE classifiers and the Raman spectra of common brain components. In order to complete these calculations, we would need to obtain Raman spectra of molecular components in the range 1000-1800 cm^{-1}, but this step would help us further interpret our results and screen for potential AD biomarkers.

Appendix A

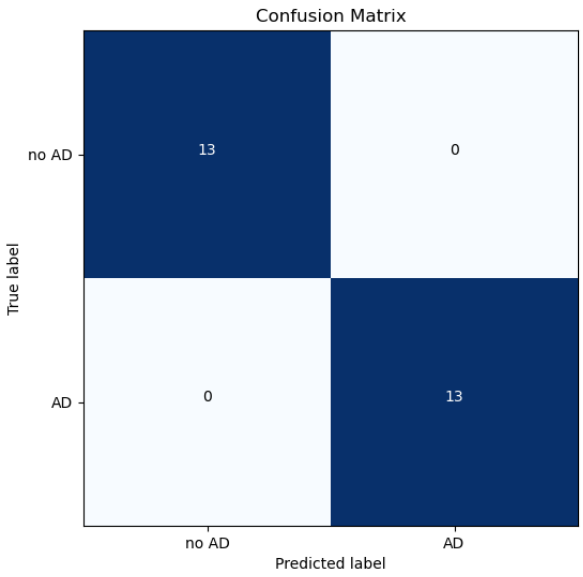
RF-RFE Confusion Matrices

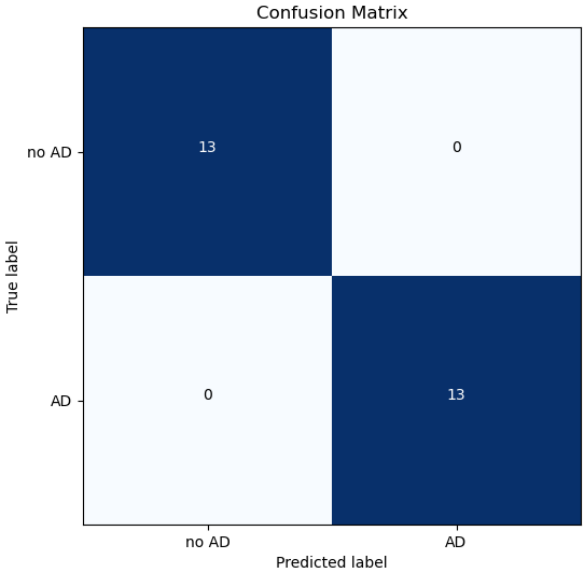
Iteration 1: Confusion matrix for each of the 5 folds

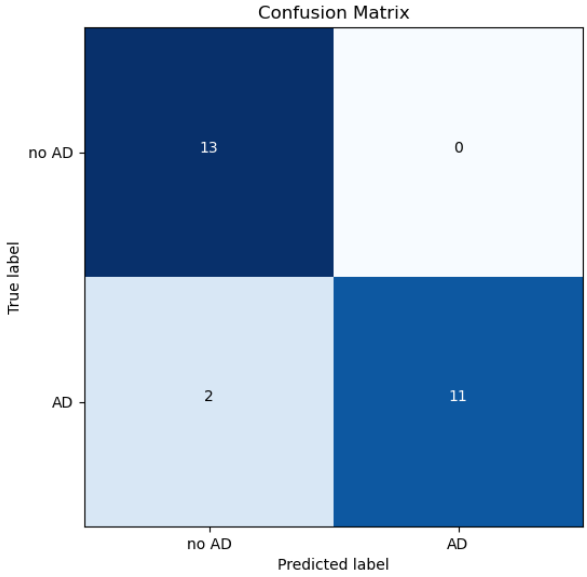


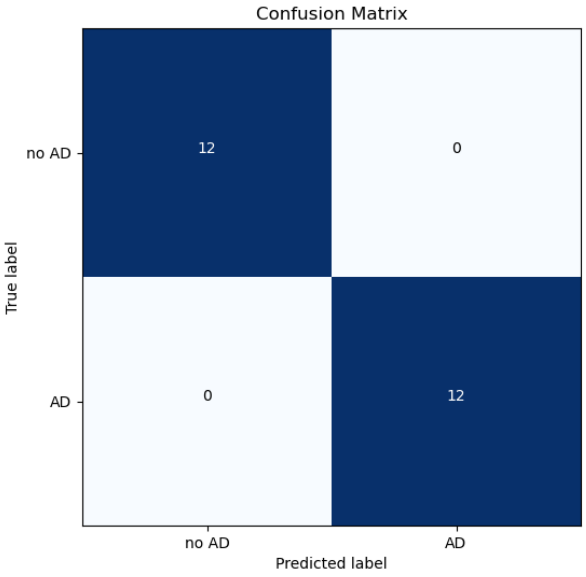


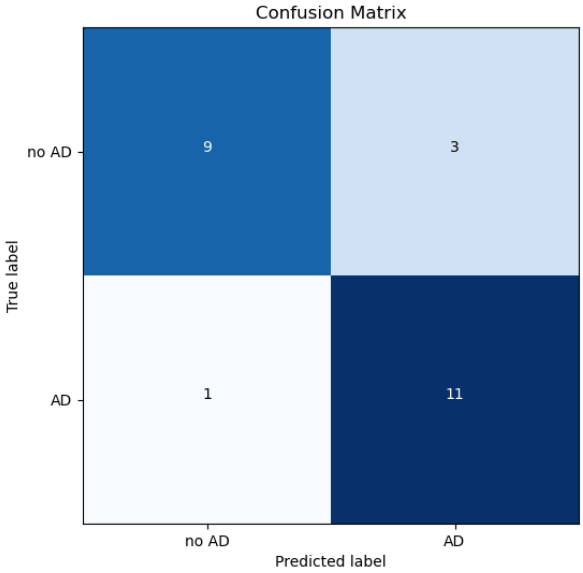
Iteration 2: Confusion matrix for each of the 5 folds



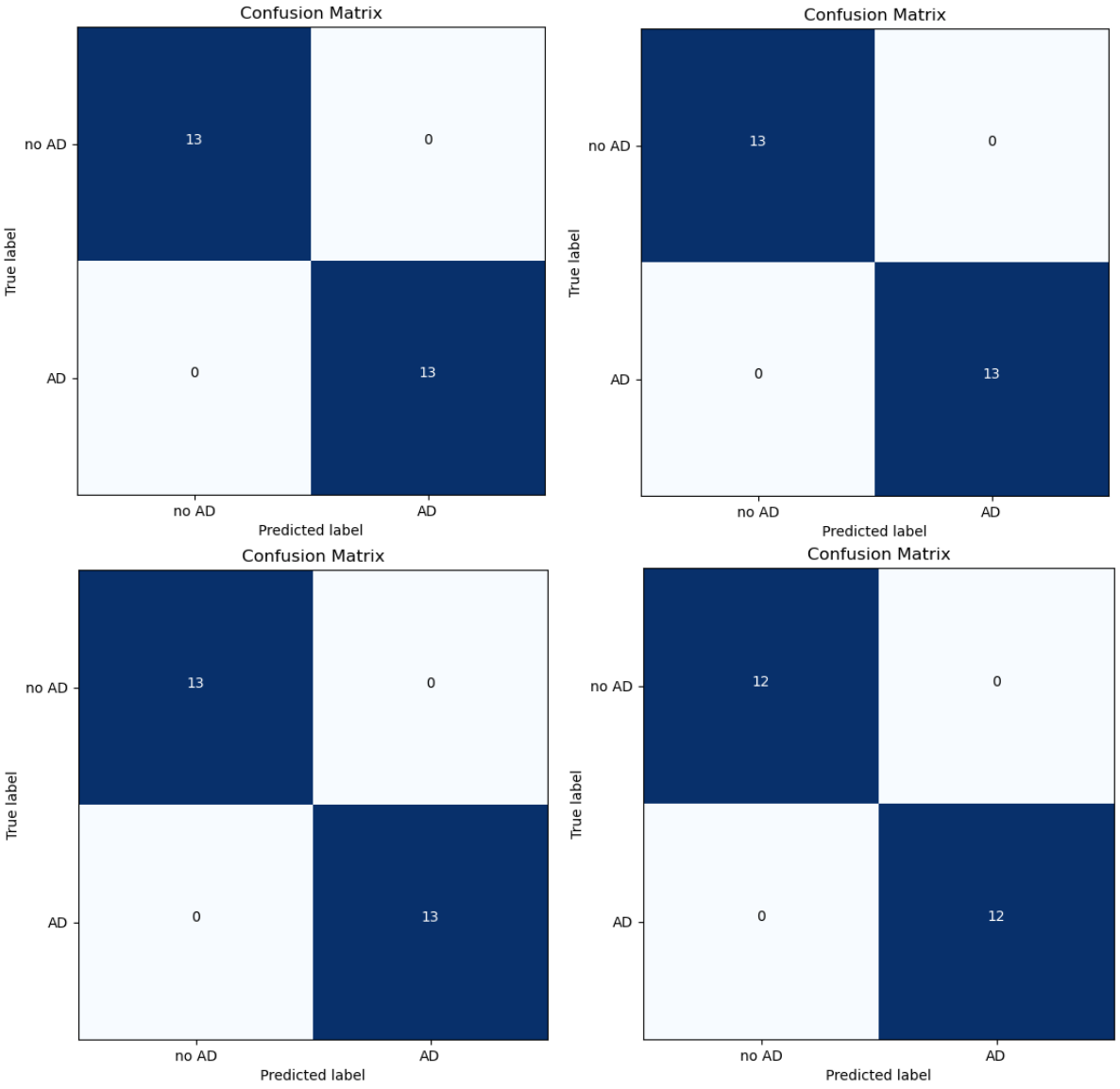


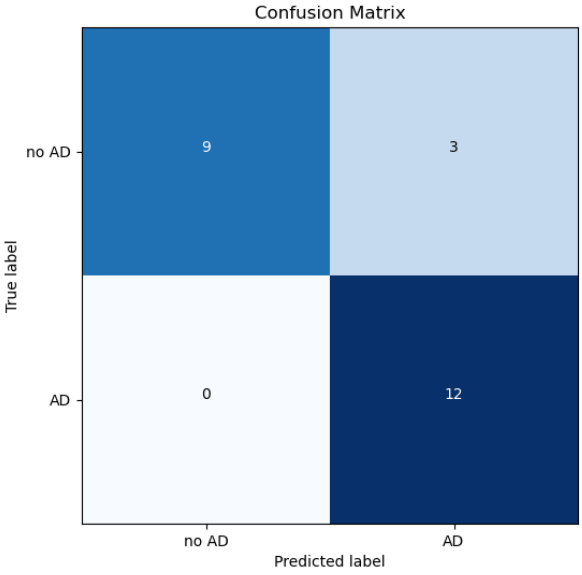

Confusion Matrix
no AD
AD
True label
13
0
2
11
no AD
AD
Predicted label


Confusion Matrix
no AD
AD
True label
12
0
0
12
no AD
AD
Predicted label

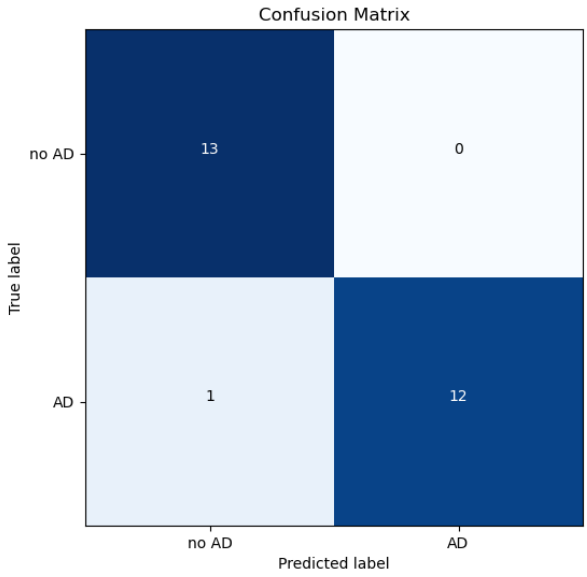

Confusion Matrix
no AD
AD
True label
9
3
1
11
no AD
AD
Predicted label

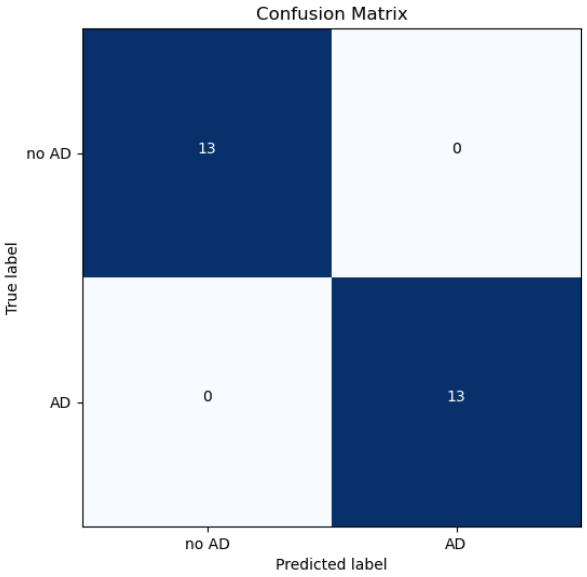
Iteration 3: Confusion matrix for each of the 5 folds

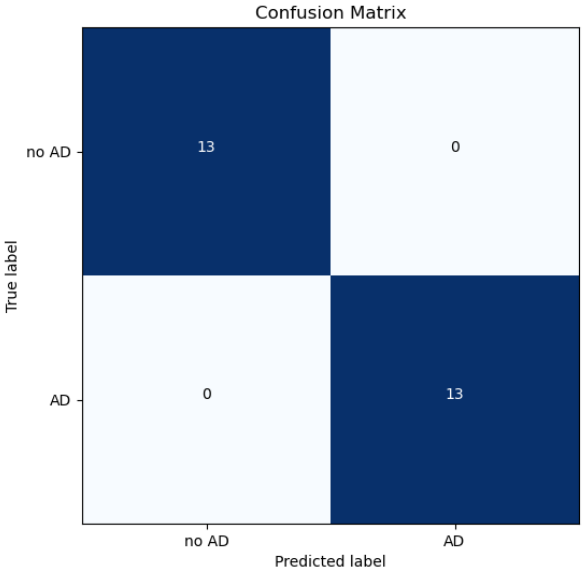


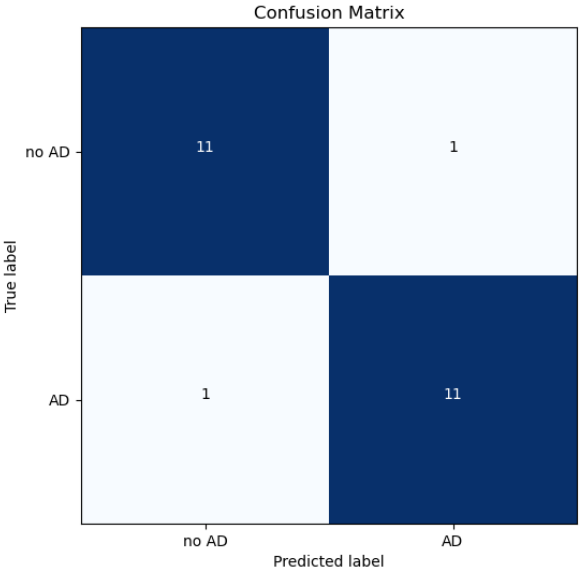


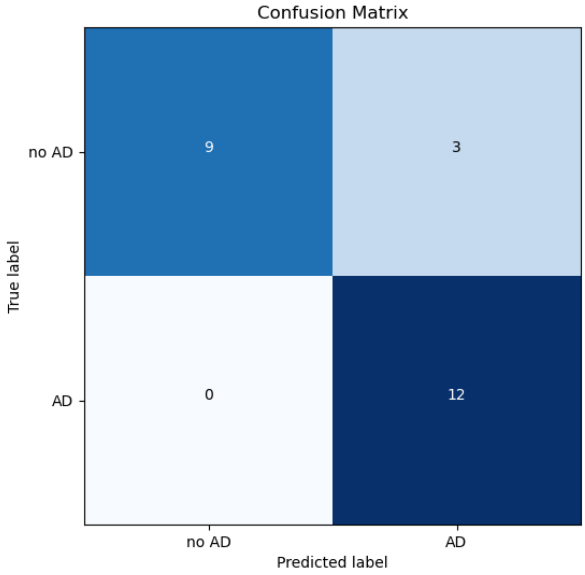
Iteration 4: Confusion matrix for each of the 5 folds



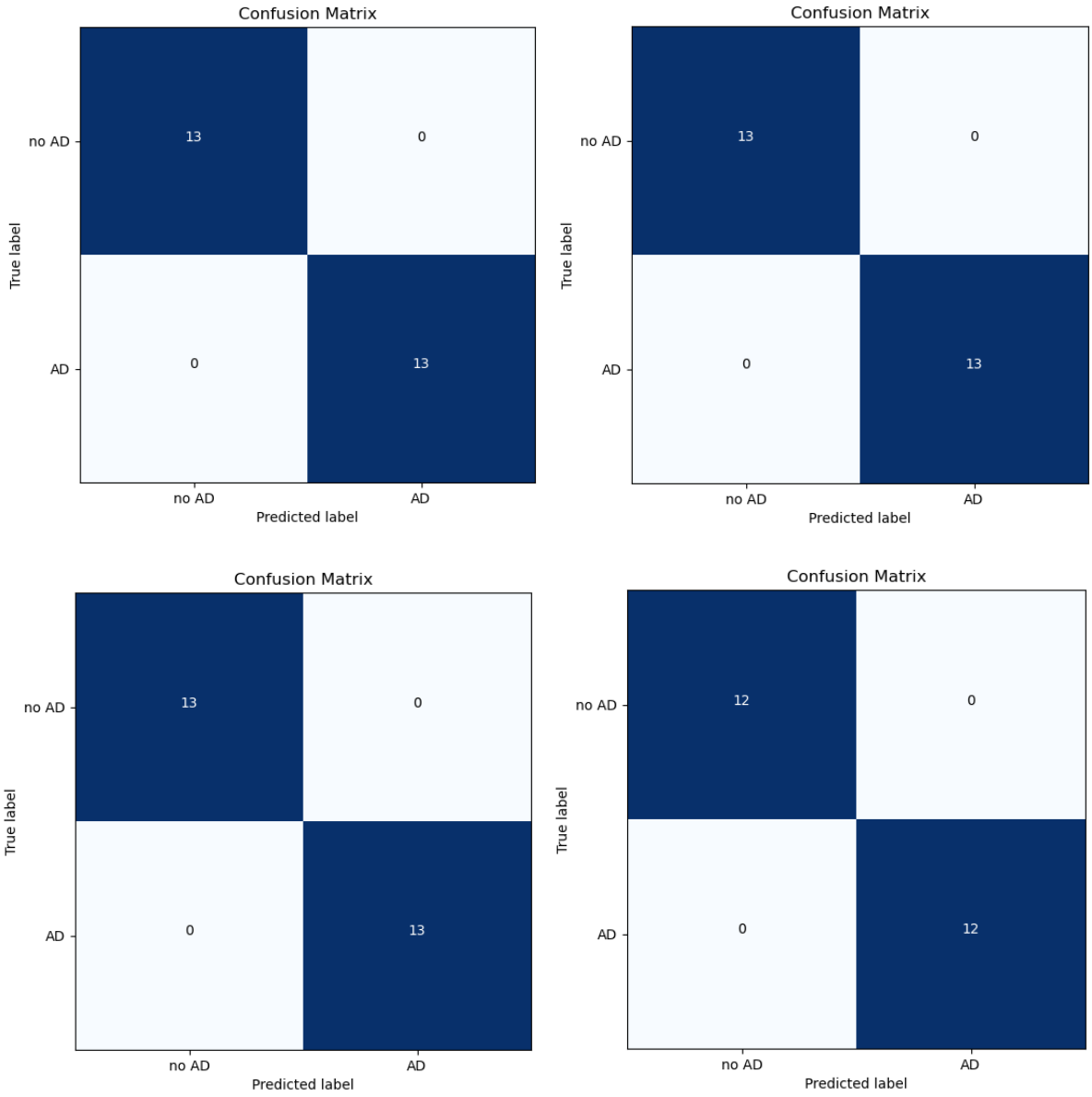


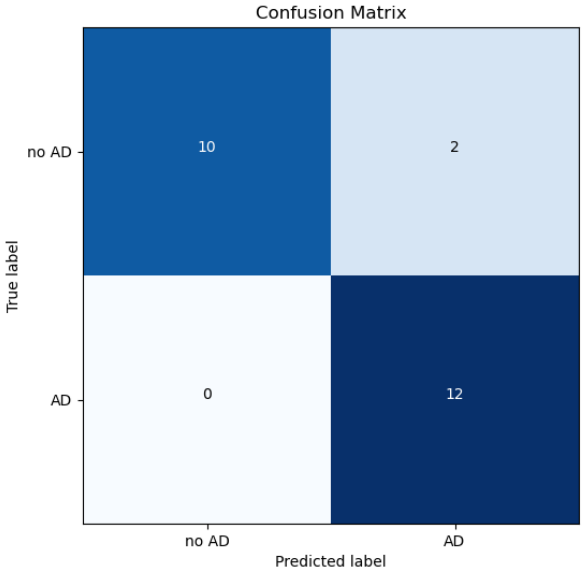

Confusion Matrix
no AD
AD
True label
13
0
0
13
no AD
AD
Predicted label


Confusion Matrix
no AD
AD
True label
11
1
1
11
no AD
AD
Predicted label


Confusion Matrix
no AD
AD
True label
9
3
0
12
no AD
AD
Predicted label

Iteration 5: Confusion matrix for each of the 5 folds

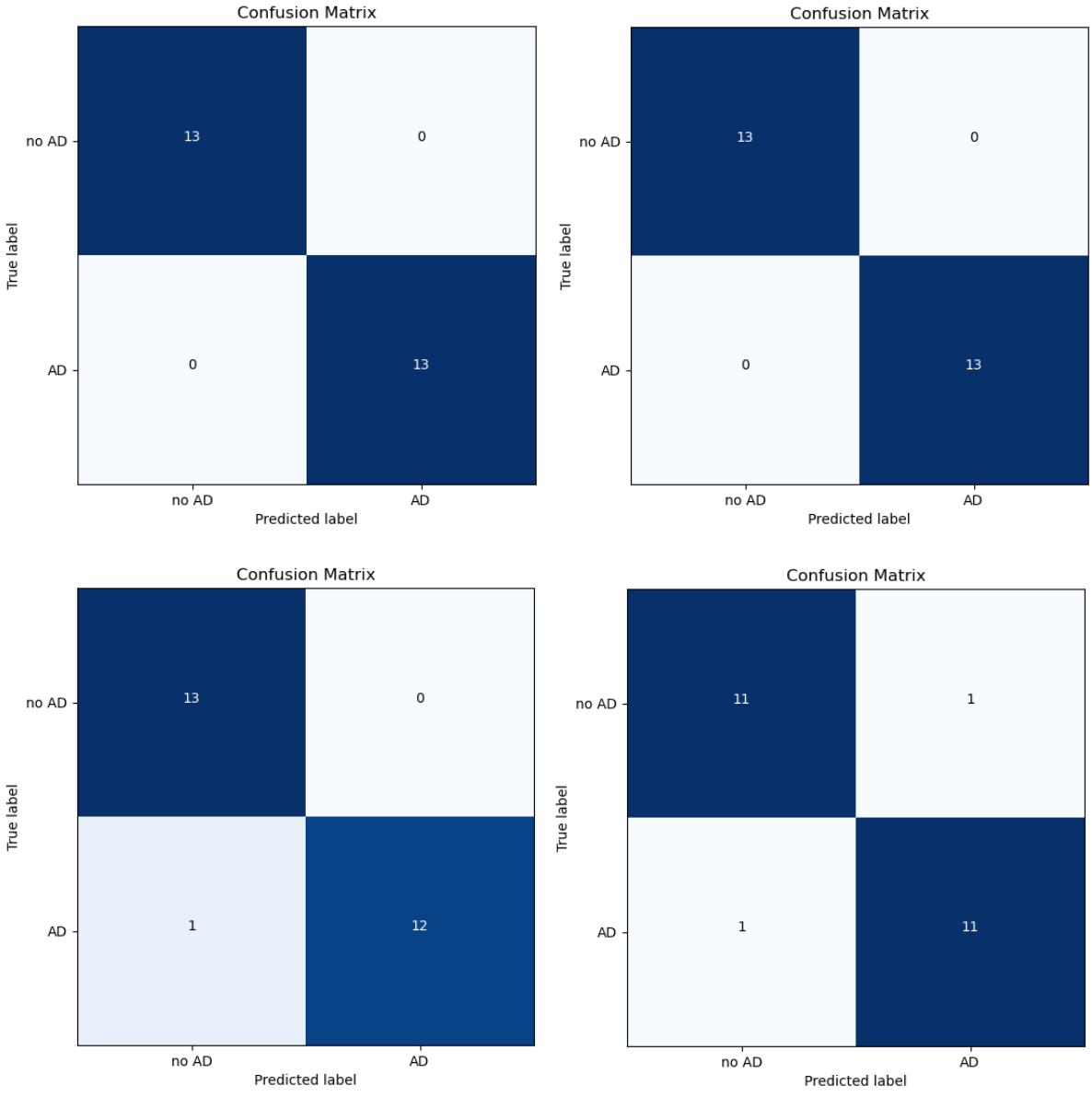


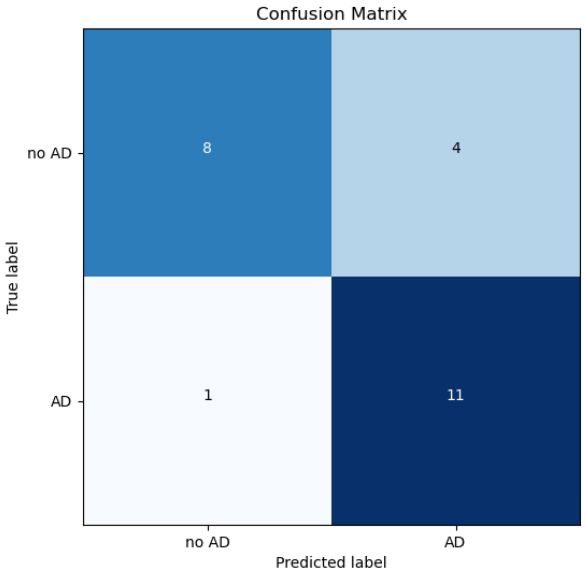

Confusion Matrix
10
2
0
12
no AD
AD
True label
no AD
AD
Predicted label

Appendix B

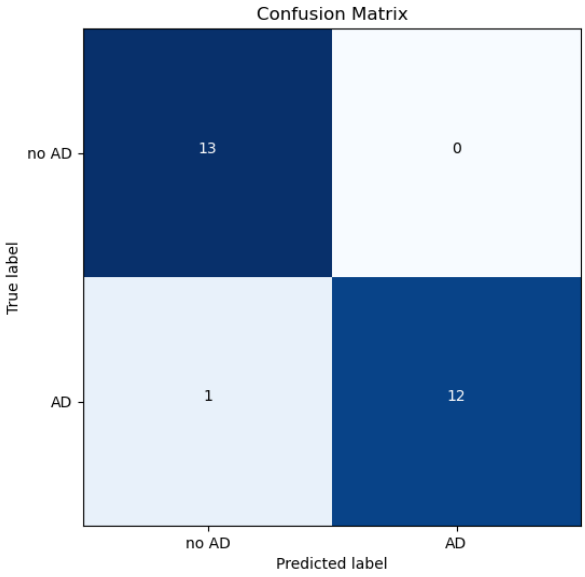
RF Confusion Matrices

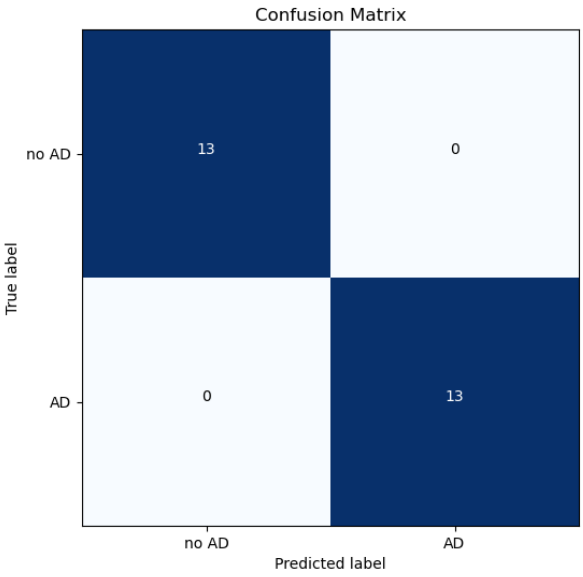
Iteration 1: Confusion matrix for each of the 5 folds

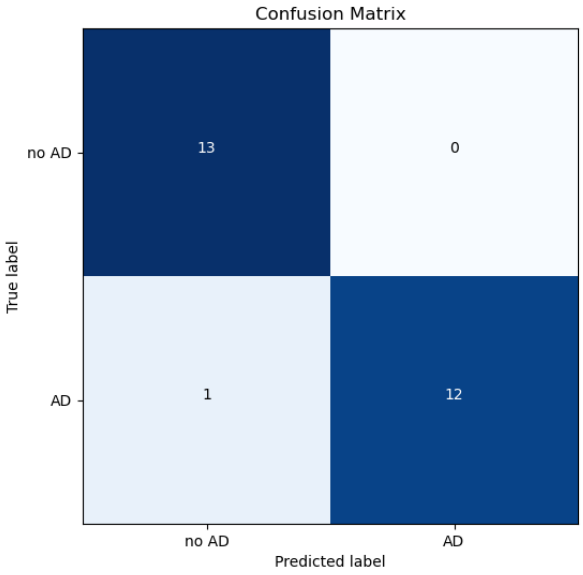


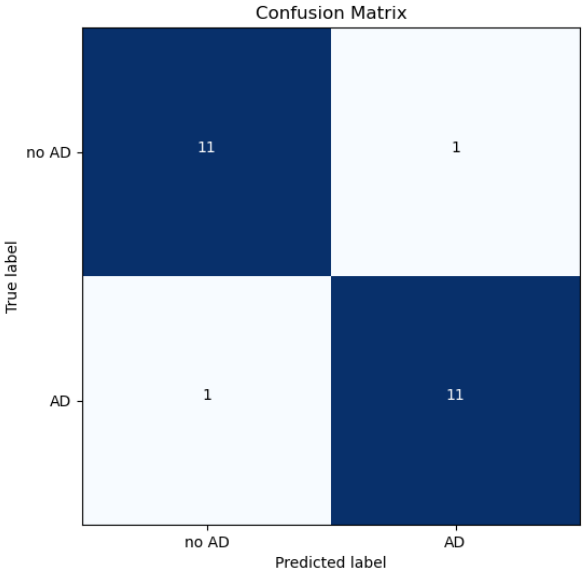


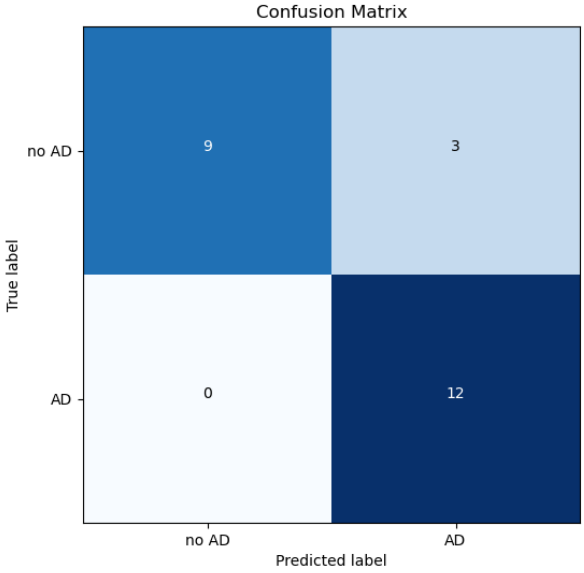
Iteration 2: Confusion matrix for each of the 5 folds



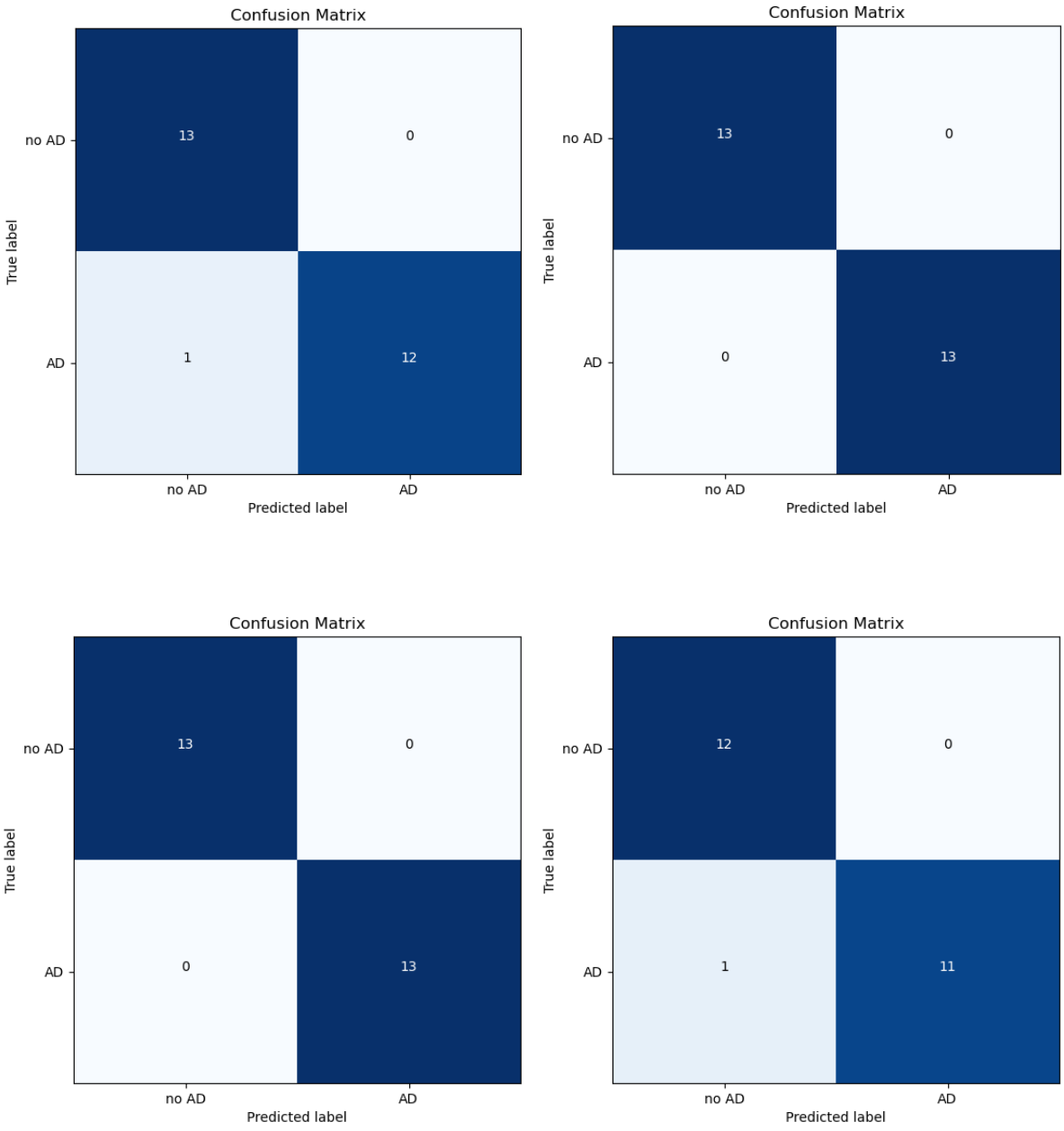


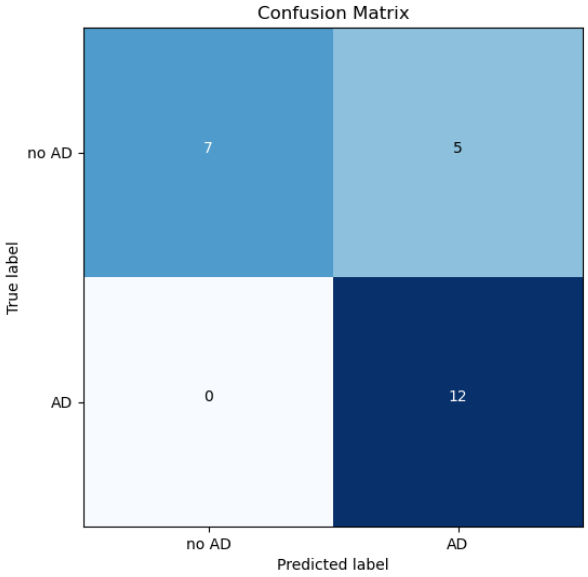

Confusion Matrix
no AD
AD
True label
13
0
1
12
no AD
AD
Predicted label


Confusion Matrix
no AD
AD
True label
11
1
1
11
no AD
AD
Predicted label

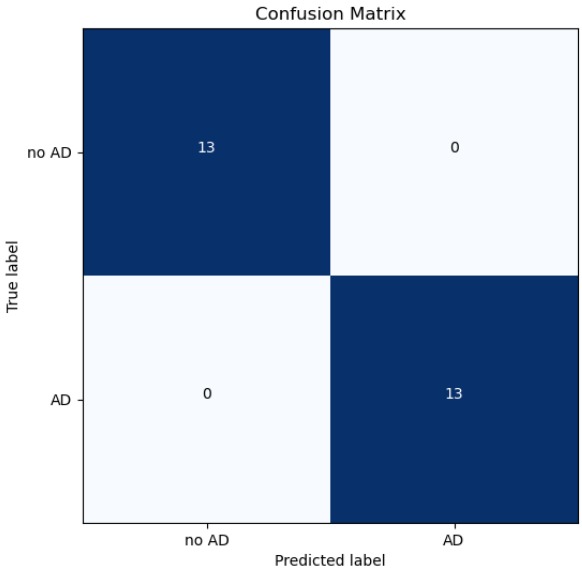

Confusion Matrix
no AD
AD
True label
9
3
0
12
no AD
AD
Predicted label

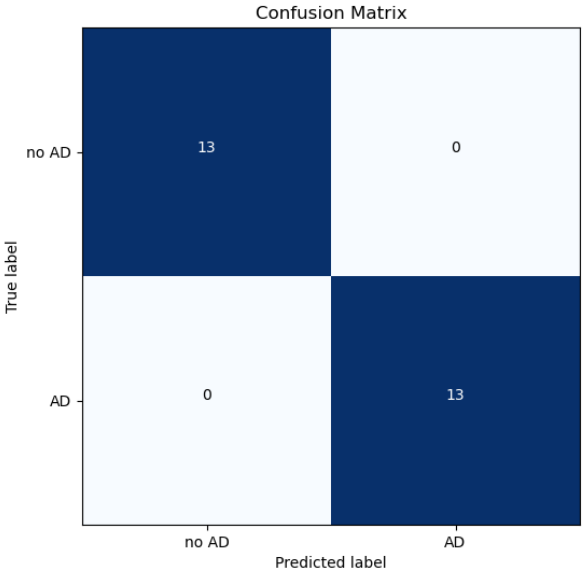
Iteration 3: Confusion matrix for each of the 5 folds

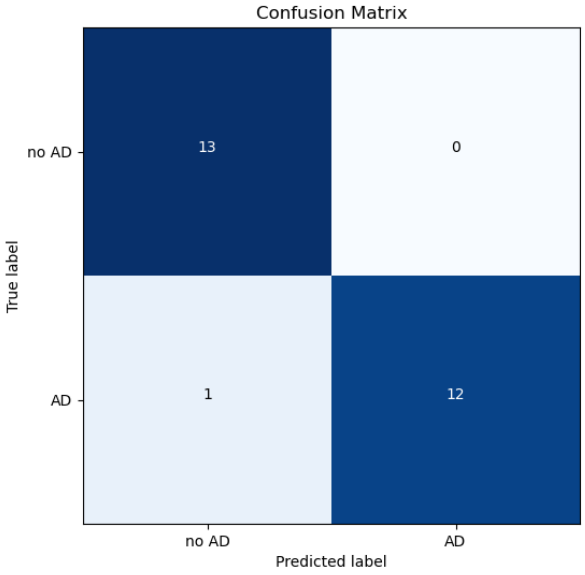


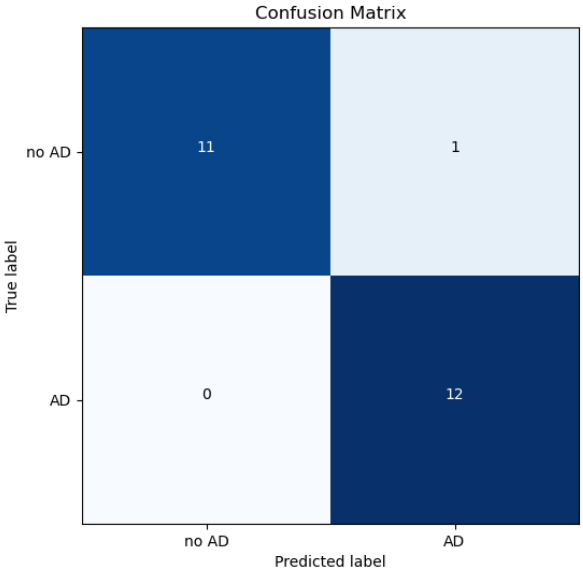


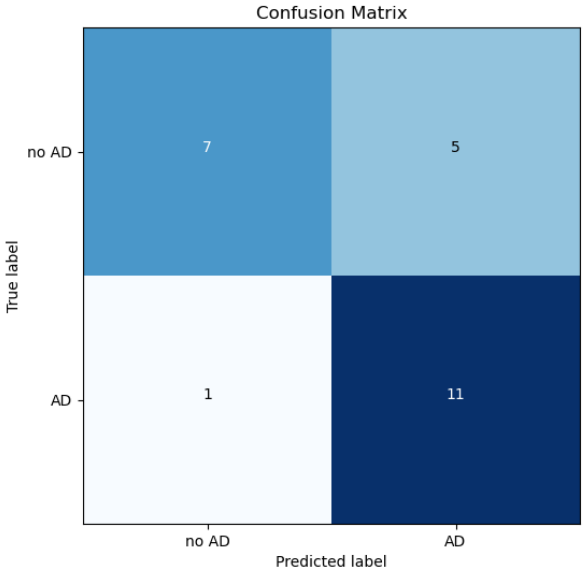
Iteration 4: Confusion matrix for each of the 5 folds



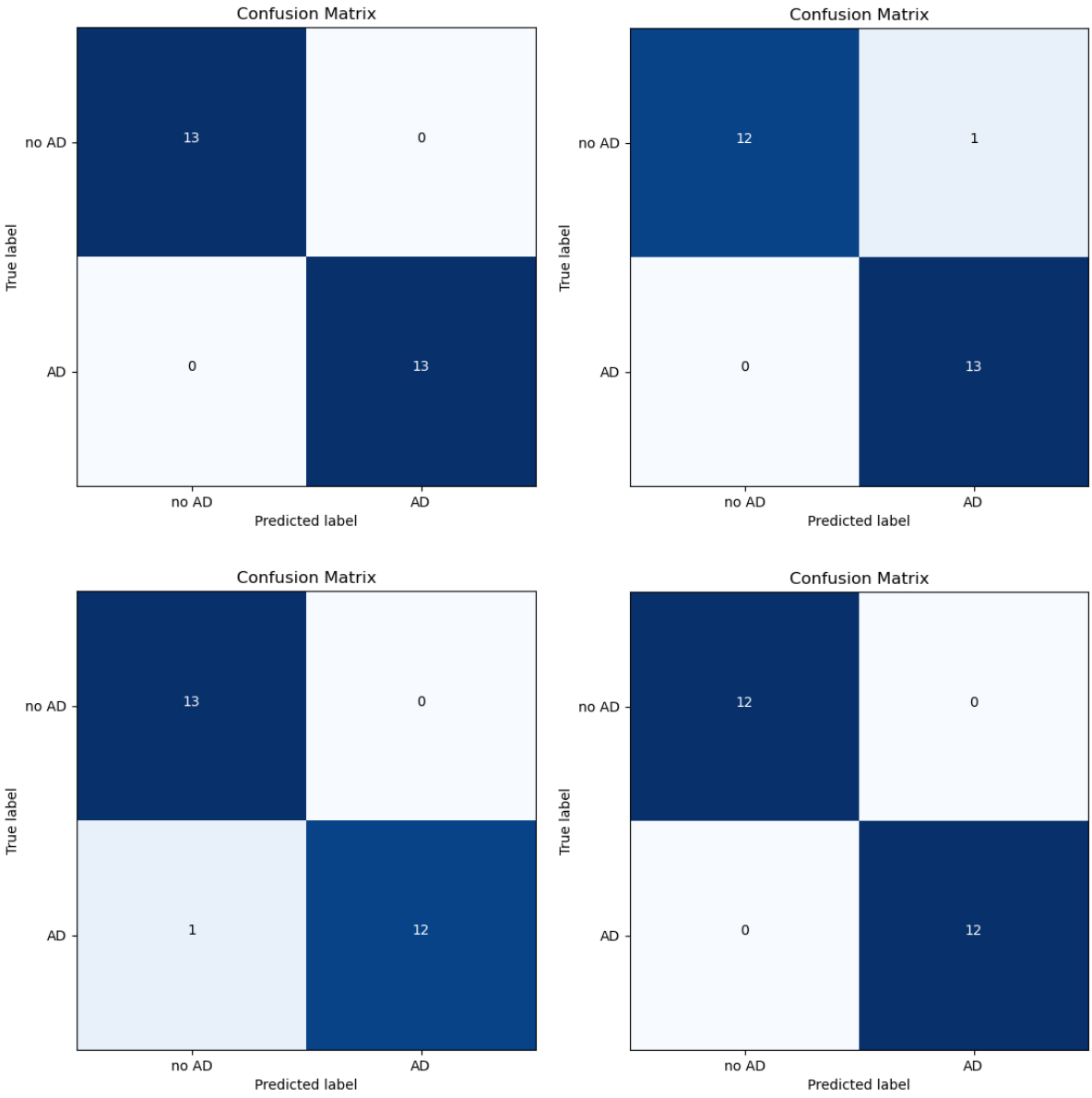


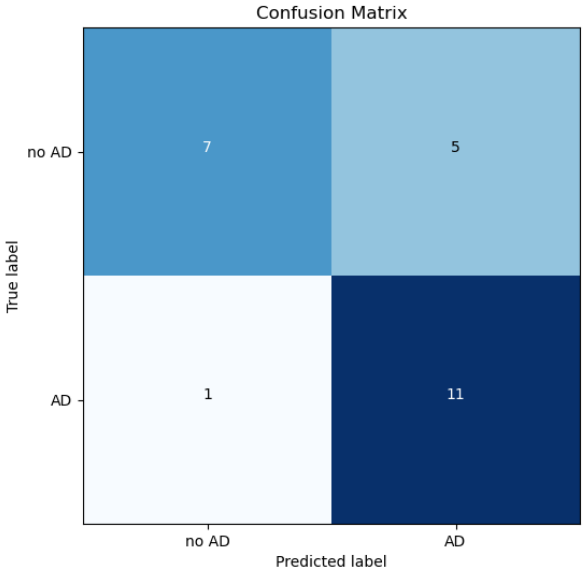

Confusion Matrix
no AD
AD
True label
13
0
1
12
no AD
AD
Predicted label


Confusion Matrix
no AD
AD
True label
11
1
0
12
no AD
AD
Predicted label


Confusion Matrix
no AD
AD
True label
7
5
1
11
no AD
AD
Predicted label

Iteration 5: Confusion matrix for each of the 5 folds




Confusion Matrix
no AD
AD
True label
7
5
1
11
no AD
AD
Predicted label

BIBLIOGRAPHY

[1] Burns, A., & Iliffe, S. (2009). Alzheimer's disease. *BMJ, 338*(Feb05 1). doi:10.1136/bmj.b158

[2] 2018 Alzheimer's disease facts and figures. (2018). *Alzheimer's & Dementia, 14*(3), 367-429. doi:10.1016/j.jalz.2018.02.001

[3] Wang, Z., Ye, J., Ding, L., Granzier-Nakajima, T., Sharma, S., Biase, I., . . . Huang, S. (2021). Rapid Biomarker Pre-screening of Alzheimer's Disease by Machine Learning using Graphene-enhanced Raman Spectroscopy on Mice Brain Slices. Manuscript submitted for publication.

[4] Long, D. A. (1977). *Raman spectroscopy*. New York, NY: McGraw-Hill.

[5] Ralbovsky, N. M., & Lednev, I. K. (2020). Towards development of a NOVEL universal medical Diagnostic method: Raman spectroscopy and machine learning. *Chemical Society Reviews, 49*(20), 7428-7453. doi:10.1039/d0cs01019g

[6] Brown, G. (2017). Ensemble learning. *Encyclopedia of Machine Learning and Data Mining,* 393-402. doi:10.1007/978-1-4899-7687-1_252

[7] Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST, 25*(2), 197-227. doi:10.1007/s11749-016-0481-7

[8] Gregorutti, B., Michel, B., & Saint-Pierre, P. (2016). Correlation and variable importance in random forests. *Statistics and Computing, 27*(3), 659-678. doi:10.1007/s11222-016-9646-1

[9] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16-28. doi:10.1016/j.compeleceng.2013.11.024

[10] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning, 46*(1/3), 389-422. doi:10.1023/a:1012487302797

[11] Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics, 19*(S1). doi:10.1186/s12863-018-0633-8

[12] Stratified cross validation. (2017). *Encyclopedia of Machine Learning and Data Mining,* 1191-1191. doi:10.1007/978-1-4899-7687-1_788

ACADEMIC VITA

Isabelle Biase

EDUCATION

The Pennsylvania State, Schreyer Honors College **University Park, PA**
College of Engineering *Class of May 2021*
Bachelor of Science in Computer Science
Relevant Coursework: Object-Oriented Programming, Web-Based Applications, Computer Organization & Design, Data Structures & Algorithms, Systems Programming, Programming Language Concepts, Operating Systems, Applications Programming, Deep Learning

PROFESSIONAL EXPERIENCE

Facebook **Menlo Park, CA**
Software Engineer *December 2021*
Software Engineer Intern *June 2020 – August 2020*

- Created a feature on Facebook Stories to autoplay video previews of stories content in the thumbnail, resulting in a 6% increase in video story views
- Integrated product & performance logging and set up A/B tests to evaluate the impact of this feature, culminating in a full public launch
- Collaborated with various cross-functional partners to drive product, design, and engineering work

Microsoft **Redmond, WA**
Explore Intern *May 2019 – August 2019*

- Contributed to an interactive dashboard feature that provides aggregate visualizations of metadata from a set of documents
- Explored PM role by participating in two customer interviews and adding to mockups based on learnings
- Translated wireframes into user interface components in React and TypeScript that interacted with backend API and utilized React Context for state management

Lockheed Martin **Rockville, MD**
Data Analytics Intern *June 2018 – Aug 2018*

- Created a Python script to convert XML English Wikipedia dump file to a CSV file to be ingested and used as an ontology in Brainspace natural language processing software
- Built a web data connector for Tableau using JavaScript to facilitate synergy between applications in data analytics suite

PERSONAL PROJECTS

THON Augmented Reality iOS Application **University Park, PA**
Developer *Nov 2018 – Feb 2019*

- Designed an immersive experience for the world's largest student run philanthropy to reach pediatric cancer patients at Hershey Medical Center who were unable to attend the dance marathon
- Integrated augmented reality features into iOS application using Apple's ARKit2 and SceneKit

Nittany AI Associates **University Park, PA**
Associate Developer *Aug 2019 – Dec 2019*

- Developed proof-of-concept to improve Penn State's Giving site, which included defining a database model and improving search results
- Built web application implementing Model View Control architecture using Django

Nittany AI Challenge **University Park, PA**

Participant *Feb 2018 – Sep 2018*

- Developed an artificial intelligence-based web application with two teammates to help college students find relevant skills and experiences to pursue in order to obtain their dream job
- Programmed the front-end of a web application to create a consumable user interface
- Accepted into Happy Valley LaunchBox FastTrack Accelerator, a 15-week startup accelerator program, to continue development of startup

LEADERSHIP & INVOLVEMENT

Penn State IFC/Panhellenic Dance Marathon **University Park, PA**

Technology Captain | THON Information Network Developer (THINK) *Apr 2020 – Present*

- Maintained and developed the THINK website to support THON volunteers
- Completed various projects based on requests from other Captains, involving data model changes, adding new forms and views, etc.

Women in Engineering Program **University Park, PA**

Facilitated Study Group Leader | Programming and Computation I *Aug 2020 – Dec 2020*

- Met weekly with small group of first- and second-year engineering women to review concepts, complete practice problems, study for exams, and help students succeed
- Prepared weekly review slides and study materials for exams

Pennsylvania State University **University Park, PA**

Learning Assistant | Programming and Computation I *May 2019 – Dec 2020*

Learning Assistant | Introduction to Systems Programming *Aug 2019 – Dec 2019*

- Held regularly scheduled office hours to answer students' questions about lab assignments and course material
- Assisted with course administration, including exam proctoring, grading, and review sessions

Presidential Leadership Academy **University Park, PA**

Member *Aug 2018 - Present*

- Selected to join a competitive three-year leadership program that accepts thirty rising sophomores
- Improved leadership, discourse, and critical thinking skills through a class with Penn State President Eric Barron and meetings with esteemed leaders in trips around the country

TECHNICAL SKILLS

Python, Django, Swift, Java, Objective C, React

HONORS & AWARDS

Chris Mader Memorial Scholarship, Wolgemuth Scholarship in Engineering, President's Freshman Award, Penn State Provost Award, Selembo Trustee Scholarship, Academic Excellence Scholarship, Elks Pennsylvania State Scholarship, National Merit Scholarship Program Finalist, National AP Scholar Award