Universitetet
i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

BACHELOR'S THESIS

Study program/ Specialization: Bachelor in Engineering / Computer Science	Spring semester 2024 Open/Restricted
Author(s): Fredrik Nilsen Låder, Andreas Solvang Nese	
Faculty supervisor: Alvaro Fernandez Quilez Supervisor(s): Alvaro Fernandez Quilez, Anna Kurbatskaya	
Thesis title: **Robust EEG analysis of Parkinson's disease: A machine learning approach**	
Credits (ECTS): 20	
Keywords: Electroencephalography, Parkinson's disease, Machine Learning, Prediction, Conformal Prediction	Pages: 61 Stavanger 15. mai 2024

Abstract

Parkinson's disease (PD) is one of the most common neurodegenerative disorders worldwide. An electroencephalogram (EEG) has the potential to detect PD. PD is diagnosed by clinically examining the patient's symptoms. As the need for automatic and more reliable procedures in detecting PD increases, different Machine Learning (ML) models have been developed for this purpose. These models have been trained indistinctively with subjects ON and OFF medication, in spite of the known effect that medications such as levodopa can have on the EEG signal.

In this thesis, we aim to explore whether there is a difference when using subjects' ON or OFF medication when training and testing the model in order to assess whether the subjects' medication status affects the model's performance. We also aim to try to improve an already trained ML model by applying a conformal prediction algorithm, making the model more "confident" in the answer it gives. We also evaluated the effect of conformal predictions for different sub-populations based on genders, severity stage and different centres.

The results of the tests, which are used to determine whether there is a difference between using subjects' ON or OFF medication when training and testing the model, show that training on ON medication recordings increases the model's performance compared to training the model on subjects OFF medication. In addition, we also observed that severe subjects ON medication have a big influence in the learning of the algorithm. Another test we used to analyse the effect of medication was to try and predict whether a subject was either ON or OFF medication. The results of the tests indicate that the EEG signals do not contain enough information about the medi-

cation status to distinguish between ON/OFF medication. By applying a conformal prediction we were able to improve the performance of our overall model, with an increase in balanced accuracy from 75.92% ± 6.024 to 79.29% ± 9.601. We also saw different increases within each specified subpopulation, concluding that the use of a conformal prediction can enhance a model's performance.

Acknowledgements

We would like to thank our supervisors Álvaro fernández Quílez and Anna Kurbatskaya for all help provided during the thesis. We are grateful for the quick response to emails, help in creating the models and code and for guidance provided in our meetings.

Disclosure

In this thesis, we have used ChatGPT and Grammarly as tools for rewriting, spell-checking and plagiarism checkers. In spite of using them, we have double-checked their outputs and have not used anything provided by them in its entirety.

Contents

Acronyms

AI Artificial Intelligence

AUC Area under the ROC curve

CT Computer Tomography

CV Cross-Validation

DTree Decision Tree

EC Eyes Closed

EEG Electroencephalography

EO Eyes Open

FN False Negative

FP False Positive

ICA Independent Component Analysis

KNN k-Nearest Neighbor

LR Logistic Regression

ML Machine Learning

MRI Magnetic Resonance Imaging

PD Parkinson's Disease

PSD Power Spectral Density

SVM Support Vector Machine

TN True Negative

TP True Positive

Chapter 1

Introduction

1.1 Motivation

Currently, there is no definitive test for Parkinson's Disease (PD). Electroencephalography (EEG) is a non-invasive technique that could be of use to the detection of PD. Despite its relevance, analyzing EEG recordings requires an expert [1]. Even in the event of a skilful professional reading EEG, the lecture is prone to errors and subjectivity. Machine Learning (ML) could help in the automatic analysis.

1.2 Objectives

Objective 1 - Effect of medication

Several Artificial Intelligence (AI) methods have currently been developed for the detection of PD using EEG, but there is no standardized approach for dealing with medication. It is unclear if the medication can affect the performance of the AIs methods. We know that medication can affect the brain waves represented in the EEG [2], potentially affecting the patterns the algorithm is picking up. We want to verify the effect of the medication

and assess if there is a difference when using data from subjects who have taken medication and from those without medication.

Objective 2 - Predict PD/non-PD using Conformal Prediction

The current developed tools for detecting PD do not have a notion of "confidence". This means that the algorithm will give an output even though it is uncertain in the answer. We want to avoid that and indicate that if the confidence is low, we want to skip that prediction.

1.3 Related Works

Several articles have covered how EEG can be used to detect PD. The works of Kurbatskaya et al. [3] focus on how sub-group populations stemming from different genders almost never are taken into consideration. In their work, they have performed an analysis of the detection ability for the genders using a trained ML model based on Power Spectral Density (PSD) features of EEG. They have found significant differences in detecting ability between males (80.5% accuracy) vs. females (63.7% accuracy) and points towards higher activity for some parietal and frontal EEG channels for males which can explain the difference in detecting ability between the genders.

Kurbatskaya et al. [4] look at using ML models to predict PD/non-PD while also evaluating the effect of harmonization, given the multi-center nature of the datasets. Their validation results show, on average, an improvement in PD detection ability (69.6% vs. 75.5% accuracy) when harmonizing the features and performing feature selection. The final results showed an average global accuracy of 72.2%

The mentioned works are similar to our work for the preprocessing and feature extraction parts, but none of them have the same problems/objectives as our project.

Chapter 2

Background

2.1 Clinical

2.1.1 Parkinson's Disease

PD is a brain disorder that affects the nervous system and the parts of the body controlled by the nerves and worsens over time and is one of the most common degenerative disorders [5]. When nerve cells weaken or die, people can begin to notice problems with movement, tremors, stiffness in limbs or impaired balance. As the disease progresses, these symptoms get more prominent, and the person may find it difficult to walk, sleep or do other simple tasks [6].

Symptomps

Four primary symptoms are characteristic of PD.
Tremor is a shaking that often starts in the hands or feet of the patient [7]. It is most noticeable when the patient is resting and disappears or improves during sleep [6].
Rigidity is muscle stiffness, affecting nearly all patients with PD.The muscles are constantly tense and contracted [8]. If another person tries to move

the limbs, the limbs can move only in certain short and jerky ways.
Bradykinesia is the slowing of automatic movements [9], thus making simple routines, such as getting dressed in the morning, more difficult and tedious.
Postural instability is impaired balance and a change in posturing that can increase the risk of falling [10].
Other, less prominent symptoms of PD is among other things *depression, sleep problems, muscles cramps* and emotional changes [6].

Diagnostics

There is currently no specific test available to diagnose patients with PD. The diagnosis is made via different methods. A clinical examination where the medical and neurological history of the patient is assessed. Blood tests, Computer Tomography (CT) and Magnetic Resonance Imaging (MRI) scans are taken to try to rule out other types of diseases [11] in combination with a physical examination to look at potential symptoms, such as tremors and rigidity [6].

Medication

Although a cure for PD has not yet been developed, there are numerous different medications that can help manage the symptoms and improve quality of life. Dopamine, a neurotransmitter produced in the brain, plays a crucial role in neural communication. When dopamine levels become too low, symptoms of PD appear [12], [13]. This happens because many of the brain cells that produce dopamine have died or are dying. Dopamine as a medication does not work because dopamine can not cross the blood-brain barrier. Doctors prescribe other medications that can act in a similar way. The majority of drug treatments work by doing either one or more of the following:

- Elavating the dopamine levels within the brain.[13]
- Functioning as a surrogate for dopamine by activating the regions of the brain where dopamine operates.[13]

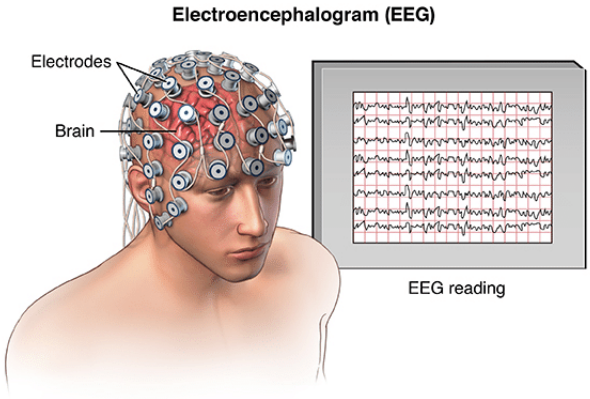
- Preventing the breakdown of dopamine by blocking the action of enzymes or other factors.[13]

Levodopa is one of the most common drugs used by PD patients. Levodopa is a chemical building block that exists naturally in your body, which the body transforms into dopamine in the brain [12], [13]. By taking Levodopa, you are essentially enhancing the supply, which means the nerve cells can produce more dopamine.

2.1.2 Electroencephalography (EEG)

EEG is a non-invasive procedure to record electrical impulses from the brain. An EEG can be taken both when a subject is awake and sleeping. In the awake state, the subject can have an EEG with eyes closed and eyes open. During an EEG, the electrodes are placed on the scalp with some conductive gel to ensure good conductivity between the scalp and electrode. The electrodes are placed with 10-20 standards and given names corresponding to their position on the scalp.[14] One node, called *Reference electrode*, is placed on the neck of the subject. This electrode acts as the 0-value node to ensure that the activity recorded in the EEG is only brain activity and not electrical impulses from other parts of the body [6]. The objective of the reference node is to subtract its signals from all other electrodes. This ensures that the signal that is recorded is only from brain activity. An EEG can be used to help diagnose a number of different diseases and monitor different conditions affecting the brain [15]. A showcasing the overview of how EEG are performed in Figure 2.1

Figure 2.1: Overview of an EEG recording. [16]



2.2 EEG in PD

As mentioned in section 2.1.1, there is no specific test to diagnose a patient with PD. A diagnosis is made by a *neurologist*, a doctor trained in nervous system conditions, based on the patient's symptoms, medical history, and a neurological and physical exam. The need for automated procedures to enhance the accuracy of PD diagnosis is increasing, and EEG is beginning to be recognized as one of the key diagnostic tools for PD[17]. EEG provides a unique opportunity to non-invasive explore and study the temporal patterns of brain activity and cognitive functions.[18]

2.2.1 Preprocessing Techniques

Preprocessing is the process of transforming *raw EEG data* into a more suitable format for further analysis. The signals detected from the scalp may not accurately mirror those originating directly from the brain. Raw EEG data is susceptible to various forms of noise, which can obscure weaker EEG signals, known as the true signal [19]. Additionally, artefacts like eye blinking or muscle movement can further contaminate the data, distorting the overall picture. The primary aim is to separate the relevant neural signals from random activity commonly encountered during the recording of EEG, thereby isolating the true signal.

The PREP Pipeline

The goal of the PREP pipeline is to execute preprocessing procedures aimed at standardizing the data into a format suitable for various applications while retaining the maximum amount of the signal possible.

[20]The first thing the PREP pipeline does is line-noise removal. Since high-pass filtering destroys EEG recordings in some contexts, PREP performs high-pass filtering, calculates a line-noise signal, and subtracts it from the unfiltered signal, returning the non-filtered signal.
Then PREP does robust referencing of the signal relative to an estimate of the average reference. The concept of robust referencing is computing bad channels after data has been referenced to a reference signal that is as similar as possible to the true average of the signal, assuming that none of the channels were bad. The algorithm progresses through two phases: first, estimating the true signal mean, and then utilizing the signal referenced by this mean to identify the *real* bad channels and interpolate.

ICLabel

[19]EEG data commonly have numerous artefacts due to various signal sources such as eye blinks, heartbeats, line noise, channel noise, and muscle artefacts. Independent Component Analysis (ICA) is a signal processing technique used to aid in separating true brain signals from noise sources. ICA performs blind-source separation by decomposing the observed noisy EEG signals into sources that are maximally independent, and is done manually and requires a human to label each component that comes out from an ICA decomposition. This process is prone to human error and becomes challenging to scale up when handling high-dimensional EEG recordings.
ICLabel introduces a statistical model that uses neural networks and a crowdsourced training dataset that automatically labels the ICA components.

Autoreject

PREP only offers the detection and interpolation of bad channels, while we still need something for the detection of bad segments. To achieve this, an algorithm called Autoreject comes into play. Autorecject uses Cross-Validation (CV) and an evaluation metric that can determine the most appropriate peak-to-peak threshold, which identifies bad EEG segments. This is done together with Bayesian optimization, setting a threshold for each sensor and marking segments as bad if a majority of the sensors have a high amplitude artefact, thus mimicking how a human expert would mark bad segments during a visual examination of a subject. [18]

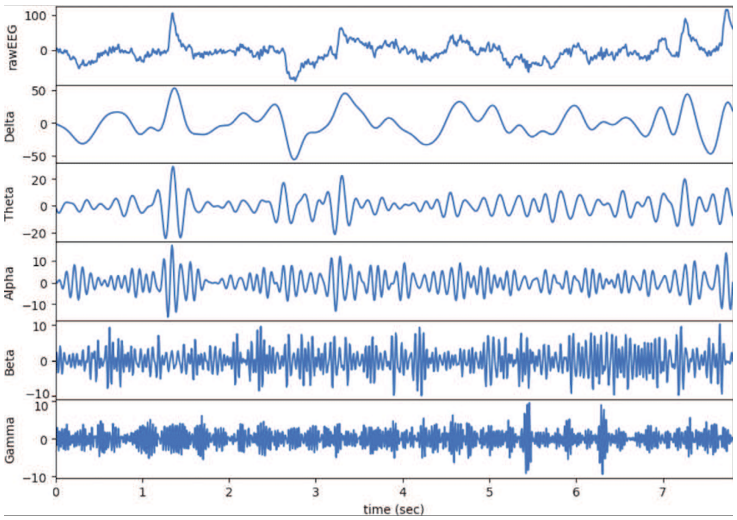
2.2.2 Feature Extraction

After preprocessing, the next step is feature extraction which is a crucial phase of biomedical signal analysis, especially with EEG signals, where dealing with large datasets spanning multiple hours and channels has become common. A fundamental objective of feature extraction is dimensional reduction and data compaction. In essence, this would result in the data being represented in a smaller subset of features [21], which reduces hardware and software resources needed and minimizes computational time.

Frequency Spectrum

EEG waveforms can be characterized based on various attributes, including amplitude, location, frequency, morphology, synchrony, etc. [22] However, frequency-based classification is the most common approach, where EEG waves are named according to their frequency range using Greek numerals [22]. The most frequently studied waveforms include δ spanning from 0.5-4 Hz, θ spanning from 4-8 Hz, α spanning from 8-13 Hz, β spanning from 14-30 Hz, and γ spanning from 30-80 Hz [23]. An example of the different frequency bands in Figure 2.2.

Figure 2.2: Raw EEG signal and corresponding frequency bands.[24]



Power Spectral Density (PSD)

PSD represents the spectral power of the signal at different frequencies and is used under spectral analysis, which is a powerful tool for analysing EEG signals [25]. The PSD can be calculated using the multitaper method, which is a technique known for generating precise, high-resolution spectral estimates without the need for averaging across frequency or time. The way that the multitaper method works is by averaging together multiple independent spectra estimated from a single segment of data [25].

2.3 Machine Learning

2.3.1 Machine Learning Fundamentals

ML has often been defined as the quote from Tom M. Mitchell: *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."* [26]

ML is a subsection of AI and uses computer algorithms that improve through experience and by the use of data. ML enables computers to learn from input data and then make predictions on new data [27].

ML is imitating the way humans learn to process information, i.e. the input, in order to accomplish a goal, i.e. the output [28]. This could be, for example, pattern recognition, where a learner might want to differentiate between dogs and cats. While it is easy for humans to tell the difference between dogs and cats, it is not straightforward for computers. Instead of hard coding the differences into the model, it is instead programmed to find these differences through experience [27].

2.3.2 Supervised Learning

Supervised learning is a learning approach that uses labelled datasets[29]. Training data is paired with its known classification labels[27]. This makes a model understand the similarities and differences within the objects, while still maintaining the ground differences between them. Supervised learning can be separated into two types of problems:

Classification problems

Classification algorithms try to accurately assign test data into the correct categories [28]. A category can, for example, be "spam" or "not spam" for an algorithm prediction of whether an incoming mail is spam or not.

Regression problems

Regression uses algorithms to try to find the relationship between dependent and independent variables. Regression models are useful for predicting numerical values from different data points. [28]

2.3.3 Machine learning Classifiers

In this subsection, we describe the different ML classifiers we have used:

Logistic Regression (LR) is a classifier that models the probability of discrete outputs given the input variables. This type of algorithm is mostly used for classification problems, and useful for predictive analytics. Since the outcome is a probability, the dependent variable is between 0 and 1. A logit transformation is applied to that odds, i.e. the probability of success divided by the probability of failure. [30]

k-Nearest Neighbor (KNN) algorithm uses "distance" or proximity to make classifications about the groupings of individual data points. KNN algorithms are usually used in classification problems. A class label is assigned on the basis of a majority vote, meaning the label that is most often represented around a given data point. For multiple classes, a plurality vote is used.[31]

Support Vector Machine (SVM) is useful for both classification and regression problems and is known to optimize the expected solution. The main objective of SVM is to separate the classes in the training set with a surface that maximises the distance between them, creating a line, called decision boundary, called a hyperplane. SVM aims to optimize the generalization ability of a model.[32]

Decision Tree (DTree) is a non-parametric supervised learning algorithm that is useful for both classification and regression problems. DTree has, as the name suggests, a tree structure consisting of root nodes, branches, internal nodes and leaf nodes. A DTree starts at the top with the root node. The branches feed into internal nodes, called decision nodes. These nodes evaluate the information, which, in the end, is denoted by the leaf nodes.

The leaf nodes show all the possible outcomes of the dataset.[33]

2.3.4 *k*-fold Cross-validation

k-fold CV is a technique used to evaluate the performance of a model on unseen data. The purpose of CV is to prevent *overfitting*, which occurs when a model is performing well on trained data and poor on new data. If a model's performance is only evaluated once, there is no way to know if the results are good or just a lucky run. With CV, the samples are split and evaluated many times. The dataset is split into k subsets, called folds, and the model trains on *k-1* folds, as the last fold is used as the validation fold. After k iterations each fold has been used as the validation set exactly once, and the final result is the average metrics for all the folds [34]. An example of a CV with k = 5 is shown in the Figure 2.3

Figure 2.3: Illustration of CV with k=5

2.3.5 Confusion Matrix

Confusion Matrix is a way to measure the performance of a ML classification. As the figure en Figure 2.4 shows, a confusion matrix is a table of four

different combinations of predicted and actual values. It is a useful metric as it shows the predictions of the model in a more human-readable [35].

- **True Negative (TN)** is a correctly predicted negative result.
- **True Positive (TP)** is a correctly predicted positive result.
- **False Negative (FN)** is an incorrectly predicted negative result.
- **False Positive (FP)** is an incorrectly predicted positive result.

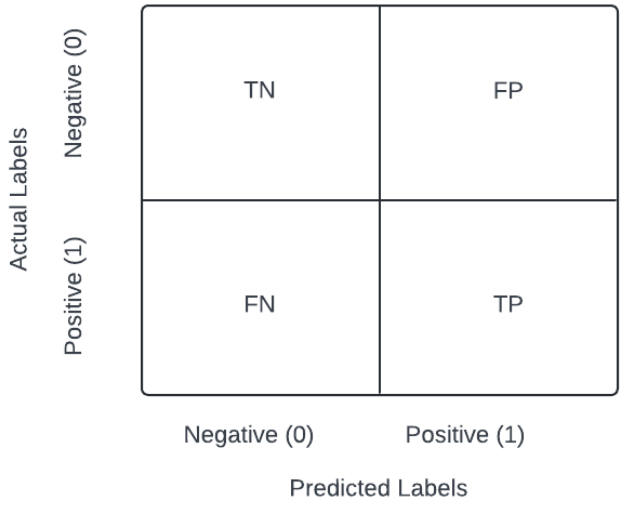


Figure 2.4: Example of a Confusion Matrix

2.3.6 Evaluation Matrices

To understand how the models are performing, it is necessary to have different evaluation metrics. **Accuracy** is used to calculate the accuracy i.e. how often the model is predicting the right outcome. The downside of using accuracy as the main evaluation metric is that it treats all classes equally, and if the dataset used is imbalanced, the accuracy might be misleading. **Balanced Accuracy** counter this imbalance and is more useful when the dataset is imbalanced. **Precision** is a metric showing how many of the total positive predictions are true, while **Recall** measures how often the model is identifying TP from actual true positives. **Area under the ROC curve (AUC)** shows the model's ability to distinguish between classes. **Efficiency** is used to calculate how many sure predictions it has from the total test set. **Validity** is used as a metric to show how many correct predictions it has from the total test set.

2.3.7 Bootstrapping

Bootstrapping is a method used in ML to estimate how a machine performs when making predictions on data not included in the training[36]. This is done by resampling the data with replacement repeatedly, and calculating the metrics for each sample [37].

Using bootstrapping to estimate the skill of an ML model can be summarized as follows: first, choose the sample size n and the number of repetitions r. Then pick n samples with repetition from the dataset from which the model has not been trained, and make predictions on those samples and compute the metrics. Repeat the sampling and prediction r times. Finally, calculate the average of all the metrics. A visual of this can be seen in Figure 2.5

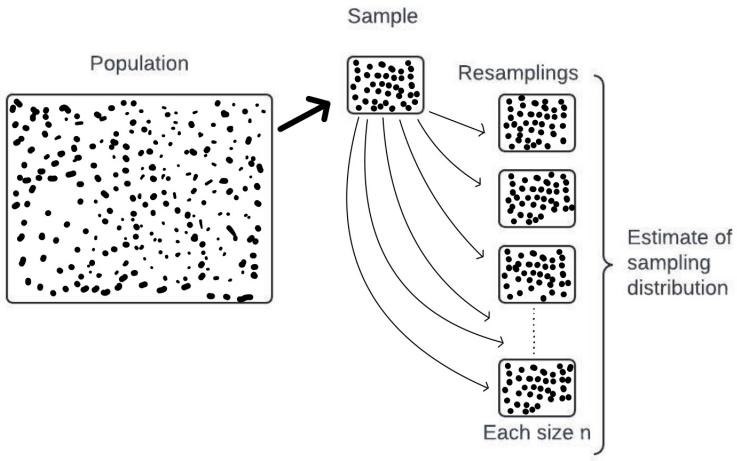


Figure 2.5: Illustration of bootstrapping

2.3.8 Conformal Predictions

Conformal prediction frameworks are used to assess uncertainty in predictions made by prediction algorithms. They convert the predictions made by the chosen algorithm into a prediction set with finite-sample coverage properties. [38] The sets carry explicit, non-asymptotic guarantees without having to make any assumptions about the distribution or the model. Conformal predictions can be used with the trained model to produce sets that contain the ground truth with a specified probability. [39]

Chapter 3

Data and methods

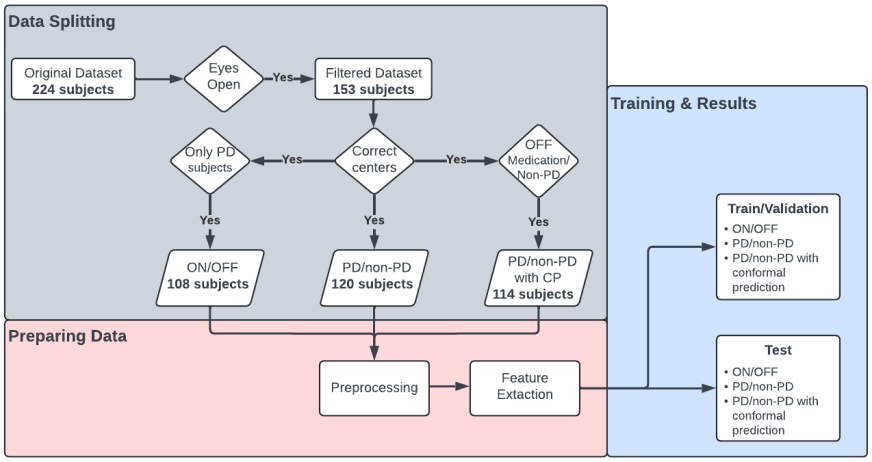


Figure 3.1: Technical approach to the project.

3.1 Datasets

In this project, we use EEG recordings from four cross-sectional studies. The recordings were provided by four different centres from two different countries: San Diego, New Mexico, and Iowa in the United States of Amer-

ica, and Turku in Finland. In total, that amounts to 155 subjects, 78 non-PD subjects and 77 PD subjects. Some more relevant details for each given dataset are depicted Table 3.1. Having datasets from multiple different locations results in a more diverse dataset in terms of demographics such as age, gender and ethnicity. This diversity can help ensure that findings are more generalised across populations.

Table 3.1: Detailes for each included datasets.

Center	Subjects	Diognostic Criterion	Characteristics included
San Diego	16 Non-PD 15 PD	not stated	age, gender, UPDRS, medication status, PD Duration
New Mexico	25 Non-PD 25 PD	not stated	age, gender, UPDRS, medication status, PD Duration
Iowa	14 Non-PD 14 PD	UK Brain Bank	age, gender, UPDRS, medication status, PD Duration
Turku	20 Non-PD 20 PD	UK Brain Bank and MDS	age, gender, UPDRS, medication status, PD Duration

The EEG recordings were acquired with both Eyes Open (EO) and Eyes Closed (EC). The EO recordings have an increase in arousal, compared to EC [40]. In this project, only EO is used. Together with the datasets, a file with clinical info about the subjects was included. In this file, data about their gender, age, and severity was stated.

San Diego Dataset

The San Diego dataset was collected in 2013 from the Scripps Clinic in La Jolla, California. The PD subjects were enlisted from the Scripps Clinic, and the non-PD subjects were drawn from either the local community or spouses of the PD subjects. The study was approved by the University of California, San Diego. The clinical diagnosis of PD was made by a specialist in movement disorders at the PD and movement disorders department of the institution mentioned above. There were collected in total 31 subjects,

16 non-PD and 15 PD subjects. The EEG recordings for the PD group were collected both during the ON and OFF phases of levodopa treatment [41].

New Mexico Dataset

The New Mexico dataset was collected circa 2015 in the Cognitive Rhythms and Computation Lab at the University of New Mexico. There were in total 50 subjects, 25 non-PD and 25 PD. The EEG recordings were collected during both the ON and OFF phases of levodopa treatment [42].

Iowa Dataset

The Iowa dataset was collected from 2017 to 2019. The PD and non-PD subjects were recruited at the University of Iowa, Narayanan Lab. The clinical diagnosis of PD was conducted following the United Kingdom Brain Bank criteria, but information about cognitive diagnosis status was not available. There were, in total, 28 subjects, 14 non-PD and 14 PD. The PD group were examined clinically, and the EEG recordings were collected during the ON phase of levodopa treatment [42].

Turku Dataset

The Turku dataset was collected in 2018. All subjects were recruited at the University of Turku, and Turku University Hospital, Turku, Finland. The PD group was diagnosed either using the United Kingdom Brain Bank criteria or the Movement Disorder's Society (MDS) criteria. There were in total 40 subjects, 20 non-PD and 20 PD. 13 PD subjects were assessed in the OFF phase of levodopa treatment, and the remaining subjects of the PD group were in the ON phase [41].

All the different splits in the experiments are done with the same random seed to ensure a consistent split when performing the different experiments.

3.1.1 Predict ON/OFF medication

For this objective, the subjects from San Diego, New Mexico, Iowa, and Turku were selected. Specifically, only PD subjects were included. In total, there 108 subjects were used from the various centres.

3.1.2 Predict PD/non-PD

For this objective, subjects from San Diego and New Mexico were selected. PD subjects, both the ON and OFF phases of levodopa treatment and non-PD subjects were included. In total, 120 subjects were used from the two centres.

3.1.3 Conformal Predictions - Predict PD/Non-PD

For this objective, subjects from San Diego, New Mexico, and Turku were selected. Specifically, non-PD and PD subjects were only subjects in the OFF phase of levodopa treatment were included. This is to simulate an early detection scenario where the subjects are not under medication. In total, there 114 subjects were used from the various centres.

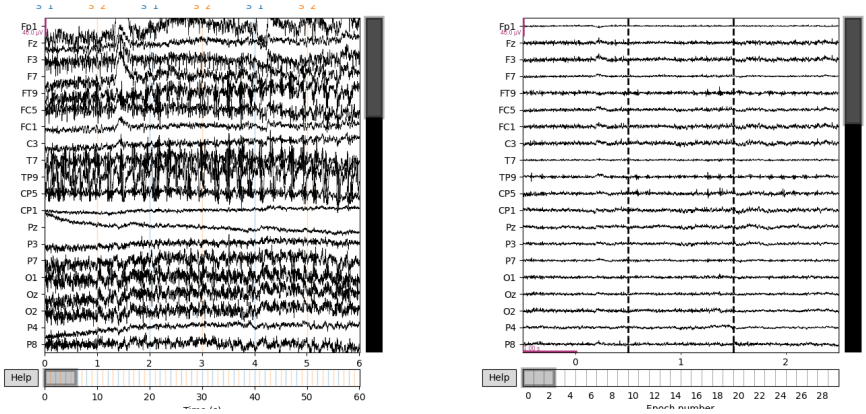
3.2 Data

3.2.1 Preprocessing

In the preprocessing stage, we have transformed the raw EEG data into a more suitable format via various stages, including filtering, line-noise removal, referencing, separating true brain signals from noise sources, and detection of bad EEG segments. Firstly, the PREP pipeline was used on the EEG recordings where high-pass filtering, line-noise removal and referencing were done [20]. Then Autoreject was used, where bad EEG segments were detected [18]. Then ICLabel was used for separating true brain signals from the noise signals [19]. Each stage explained more in subsection 2.2.1

Below there are two examples of the EEG recordings before and after preprocessing was used. Figure 3.2 is a PD subject from New Mexico. Figure 3.3 is a non-PD subject from San Diego. Each of the figures is a six-second section of the EEG signal; the recordings after preprocessing are measured in epochs, where three epochs equals six seconds.

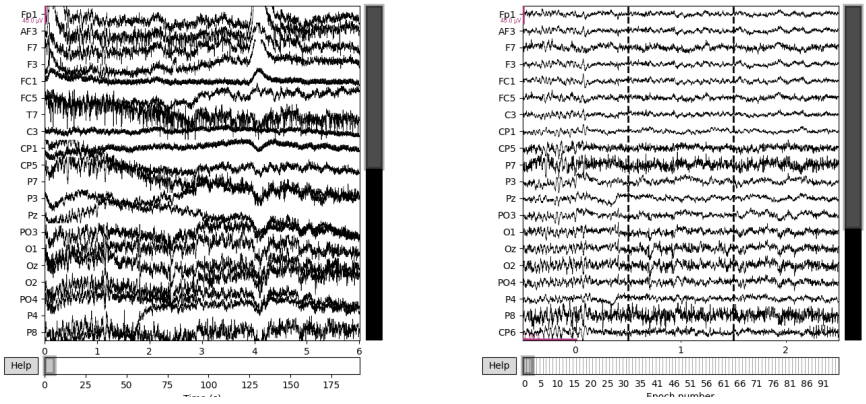
Figure 3.2: PD subject from New Mexico dataset.



Before preprocessing

After preprocessing

Figure 3.3: Non-PD subject from San Diego dataset.



Before preprocessing

After preprocessing

3.2.2 Feature Extraction

In the feature extraction stage, we process EEG data and extract spectral features from different epochs. First, we apply a bandpass filter on the data, limiting the frequency range to between 1-30 Hz and then downsampling the data to 500 Hz. This helps to reduce the computational power needed and to only focus on the bands that we are using.
The frequency spectrum was divided into four bands recommended for clinical research: δ spanning from 1-4 Hz, θ spanning from 4-8 Hz, α spanning from 8-13 Hz, and β spanning from 14-30 Hz [23] [4]. The θ band is divided into two sub-bands, slow-θ spanning from 4-5.5 Hz and pre-α spanning from 5.5-8 Hz [43]. The PSD values for each epoch are computed using the multitaper method [25]. After obtaining the PSD for the specified channel, the median PSD vector across epochs per channel is calculated. Finally, the bandpower is calculated within specific frequency bands with the median PSD vector.

In total, it produced 145 features (5 features $\times$ 29 channels) per subject. We utilise two epoch parameters in the feature extraction process: 'all' and 'min'. The key difference between these parameters lies in the selection of epochs. When the epoch parameter 'all' is used, it uses all the available epochs for the feature extraction. When the epoch parameter 'min' is used, it uses only the first 7 epochs for the feature extraction.

3.2.3 Training, testing and validation

The datasets were split differently for each task. The next section explains the different splits in more detail. For the experiment Effect of Medication - Predict ON/OFF (3.3.1), each subject used had 145 features for each of the epoch parameters 'all' and 'min'. For the experiments Effect of Medication - Predict PD/non-PD (3.3.2) and Conformal prediction - Predict PD/non-PD (3.3.3), each subject used had 145 features for only the epoch parameter 'all'. We used a training set size of 70% and a test set size of 30%. The training set is later used in a k-fold cross-validation system with k = 5, where 80% of the set is used for training and the remaining 20% is used for validation, which is done for each fold. The test set is then used for bootstrap testing on the model that has the best performance in training and validation results.

3.3 Methods

Python was used as the programming language. The open-source library scikit-learn [44] was used to build and train the models. To run the conformal predictions crepes [45] was used. Preprocessing, feature extraction, training and testing were all done using our Macbook Pro's.

3.3.1 Effect of Medication - Predict ON/OFF

Data Splitting

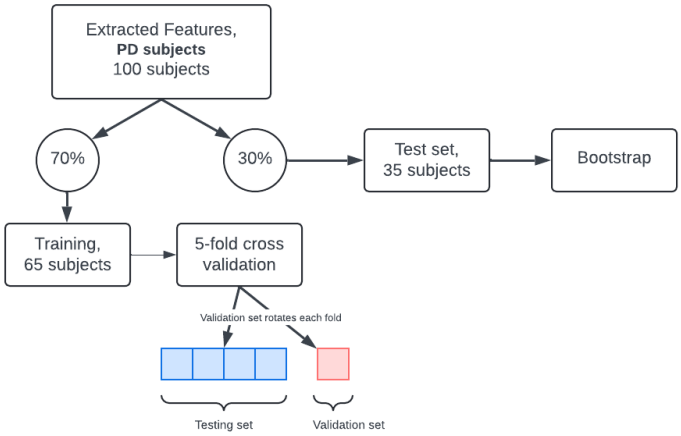


Figure 3.4: Overview of how the splitting is done. The example shown is the dataset split with severe subjects included. Bootstrap is used with the size of the test set and with 100 repetitions.

With Severe Subjects

After preprocessing and feature extraction, there were a total of 100 subjects. Each subject had 145 features for each of the epoch parameters 'all' and 'min' from the feature extraction. The subjects were split into training and testing sets on the subject level to avoid cross-contamination. Subjects that had data for both ON medication and OFF medication were included in the same train, validation or test set to avoid cross-contamination. The training set consists of 70% of the data, resulting in a total of 65 subjects. The testing set consists of the remaining 30% of the data, resulting in a

total of 35 subjects. An overview of the split in Figure 3.4. More detailed information about the split in Table 3.2.

Table 3.2: Detailed information about the training/testing sets

Set	Total	ON Medication	OFF Medication	Mild	Moderate	Severe
Training	65	34	31	38	23	4
Testing	35	17	18	17	18	0

Without Severe Subjects

The splitting was carried in the same manner as above, just without the severe subjects. After preprocessing and feature extraction, there were, in total, 96 subjects without severe subjects. Each subject had 145 features for each of the epoch parameters 'all' and 'min' from the feature extraction. The training set consists of 70% of the data, resulting in a total of 64 subjects. The testing set consists of the remaining 30% of the data, resulting in a total of 32 subjects. More detailed information about the split in Table 3.3.

Table 3.3: Detailed information about the training/testing sets

Set	Total	ON Medication	OFF Medication	Mild	Moderate
Training	64	34	30	40	24
Testing	32	17	15	15	17

3.3.2 Effect of Medication - Predict PD/non-PD

Data Splitting

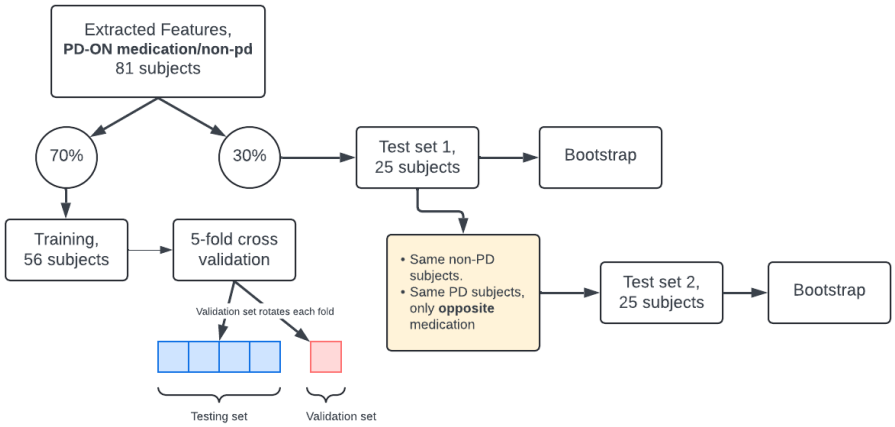


Figure 3.5: Overview of how the splitting is done. The example shown is the dataset split where training is done with subjects ON medication, with severe subjects included. Bootstrap is used with the size of the test set and with 100 repetitions.

With Severe Subjects

After preprocessing and feature extraction, there were in total 106 recordings. Each subject had 145 features for the epoch parameter 'all' from the feature extraction. To avoid cross-contamination, the subjects were split into training and testing sets on the subject level. There were two different types of splits. One where we trained on PD subjects ON medication and non-PD subjects and one where we trained on PD subjects OFF medication and non-PD subjects. On each of the different splits, we did two different test sets. One test set with PD subjects ON medication and non-PD subjects, and one test set with PD subjects OFF medication and non-PD subjects. In the different test sets, the non-PD subjects are the same in both sets. The PD subjects are also the same subjects, just the opposite medication status. Overall, we use the same training and testing sets in both experiments, just with the opposite medication status. An overview of the split in Figure 3.5. The dataset with 81 subjects in the figure, are PD subjects ON medication and non-PD subjects, in a total of 81 subjects out of the total 106 recordings.

The dataset that was split into training and testing sets consists of PD subjects with one medication status and non-PD subjects, in a total of 81

subjects. The training set consists of 70% of the dataset, resulting in a total of 56 subjects. The testing set consists of the remaining 30%, resulting in a total of 25 subjects. Then the non-PD subjects from the original test set and the same PD subjects with opposite medication status are added to an additional test set. More details about the split when training on subjects ON medication in Table 3.4, and when training on subjects OFF medication in Table 3.5.

Table 3.4: Training/testing split when training ON medication

Set	Total	Non-PD	PD	Mild	Moderate	Severe
Train: ON+non-PD	56	28	28	13	14	1
Test: ON+non-PD	25	13	12	5	7	0
Test: OFF+non-PD	25	13	12	5	7	0

Table 3.5: Training/testing split when training OFF medication

Set	Total	Non-PD	PD	Mild	Moderate	Severe
Train: OFF+non-PD	56	28	28	13	14	1
Test: OFF+non-PD	25	13	12	5	7	0
Test: ON+non-PD	25	13	12	5	7	0

Without Severe Subjects

The splitting was carried in the same manner as above, just without the severe subjects. There were a total of 106 subjects after preprocessing and feature extraction. Each subject had 145 features for the epoch parameter 'all' from the feature extraction. The dataset that was split into training test sets consists of PD subjects with one medication status and non-PD subjects, in a total of 80 subjects. The training set consists of 70% of the dataset, resulting in a total of 56 subjects. The testing set consists of the remaining 30%, resulting in a total of 24 subjects. Then the non-PD subjects from the original test set and the same PD subjects with opposite medication status are added to an additional test set. More details about the split when training on subjects ON medication in Table 3.6, and when training on subjects OFF medication in Table 3.7.

Table 3.6: Training/testing split when training ON medication

Set	Total	Non-PD	PD	Mild	Moderate
Train: ON+non-PD	56	28	28	14	14
Test: ON+non-PD	24	13	11	4	7
Test: OFF+non-PD	24	13	11	4	7

Table 3.7: Training/testing split when training OFF medication

Set	Total	Non-PD	PD	Mild	Moderate
Train: OFF+non-PD	56	28	28	14	14
Test: OFF+non-PD	24	13	11	4	7
Test: ON+non-PD	24	13	11	4	7

3.3.3 Conformal prediction - Predict PD/non-PD

Conformal Prediction

An α value is set to indicate how sure the model's prediction has to be to give a result; if not, the result is not reported since it is assumed that the classifier is unsure. Hence, the clinician can not rely on the result. We used two different α values: $\alpha = 0.65$ and $\alpha = 0.85$.
The α values were chosen to have around the same confidence as the consensus for doctors when diagnosis PD. The consensus between doctors when diagnosing PD seems to be from 63% on the low end up towards 83% on the high end [46]. The α values try to catch both the "best" and "worst" consensus.

Data Splitting

There were a total of 113 subjects after preprocessing and feature extraction. To avoid cross-contamination between training and testing, the data gets split at the subject level. The data was split 70% train and 30% test, amounting to 79 subjects in training and 34 subjects in testing. An overview of this splitting can be seen in Figure 3.6. More detailed information about the split in Table 3.8.

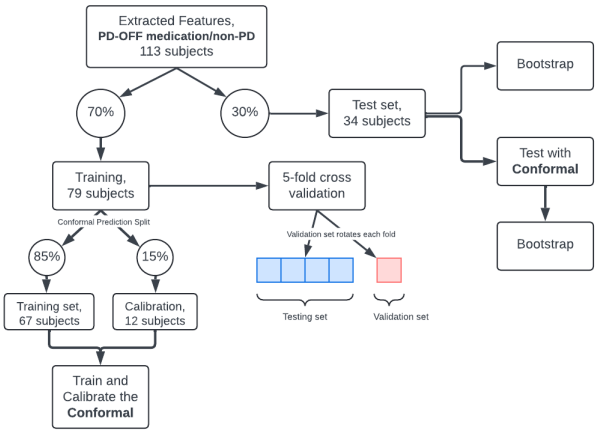


Figure 3.6: Overview of how the splitting is done. Bootstrap is used with the size of the test set and with 100 repetitions.

Table 3.8: Detailed information about the Training and Testing sets for the baseline

Set	Total	Non-PD	PD	Mild	Moderate	Severe
Training	79	42	37	20	14	3
Testing	34	18	16	8	8	0

To make the conformal prediction set, the original training set was split 85%, amounting to 67 subjects for the training of the model and the remaining 15%, totalling 12 subjects, was used as a calibration set. More detailed information about the training, calibration and testing sets are presented in Table 3.9

Table 3.9: Detailed information about the training, calibration and Testing sets.

Set	Total	Non-PD	PD	Mild	Moderate	Severe
PropTraining	67	36	31	18	11	2
Calibration	12	6	6	2	3	1
Testing	34	18	16	8	8	0

Training, testing and validation using conformal prediction

After the model is trained, a conformal prediction is done on the test set. The conformal prediction gives each subject in the test set a value of either [0 1], [1 0], [0 0] or [1 1]. The values [0 1] and [1 0] mean that it is sure that it is prediction is correct. These subjects are then added to the final test set. Bootstrapping is then done on this final test set.

3.3.4 Evaluation

Lastly, bootstrapping was used on the classifier that performed the best in training and validation results. To get an accurate estimate of the distribution we used the testing set size as the sample size, with 100 iterations with repetition when choosing the samples.

3.3.5 Evaluation Metrics

We used different types of evaluation metrics, but we considered **AUC** and **balanced accuracy** as the most important metrics to perform the comparisons and discuss. We want to capture the accuracy of our models, and since the datasets we have are imbalanced, the balanced accuracy accounts for that imbalance, and the AUC measures the accuracy of the diagnostics tests. To evaluate how the conformal prediction performs, we have used Efficiency and Validity.

Chapter 4

Results

4.1 Experimental Results

For the confusion matrices shown, the label 0.0 equals non-PD and label 1.0 equals PD. This counts for every experiment except Section 4.1.1, where label 0.0 equals OFF medication and 1.0 equals ON medication. The test results are reported as Mean ± Standard Deviation. All the experiments are done with the same random seed to ensure a consistent split when performing the different experiments.

4.1.1 Effect of Medication - Predict ON/OFF Medication

In table Table 4.1, we can see the test results from predicting if a subject is either ON or OFF medication. To the left in the table are the test results when including severe-stage subjects. The best model from the training and validation results was DTree for epochs 'all' where 107 features were chosen, which was used on the test set. To the right in the table are the test results when not including severe-stage subjects. From the training and validation results, the model that performed best was KNN for epochs 'min' where 8 features were chosen, which was used on the test set.

From Table 4.1, we can see that, including the severe-stage subjects, the model performed the best when looking at the balanced accuracy (51.76% ± 5.244) compared to the balanced accuracy (47.95% ± 7.749) when not including the severe-stage subjects. The AUC (0.52 ± 0.052) when using the severe-stage subjects was also higher than the AUC (0.48 ± 0.077) when not using the severe-stage subjects. In Figure 4.1, we can see the confusion matrices for the results of the two tests. To the left are the results for when we included the severe-stage subjects, and to the right is when we did not include the severe-stage subjects.

Evaluation Metric	**Result**
Accuracy	50.46% ± 5.162
Balanced Accuracy	51.76% ± 5.244
Macro Recall	51.76% ± 5.244
Micro Recall	50.46% ± 5.162
Macro Specificity	51.76% ± 5.244
Micro Specificity	50.46% ± 5.162
Macro Precision	52.01% ± 5.800
Micro Precision	50.46% ± 5.162
Macro F1	49.76% ± 5.221
Micro F1	50.46% ± 5.162
AUC	0.52 ± 0.052

Evaluation Metric	**Result**
Accuracy	47.25% ± 7.734
Balanced Accuracy	47.95% ± 7.749
Macro Recall	47.95% ± 7.749
Micro Recall	47.25% ± 7.734
Macro Specificity	47.95% ± 7.749
Micro Specificity	47.25% ± 7.734
Macro Precision	47.75% ± 8.649
Micro Precision	47.25% ± 7.734
Macro F1	46.50% ± 7.964
Micro F1	47.25% ± 7.734
AUC	0.48 ± 0.077

Table 4.1: Test results for predicting ON/OFF medication. On the left, we have a table for the test results with severe subjects, for the model DTree. On the right, we have a table for the test results without severe subjects, for the model KNN.

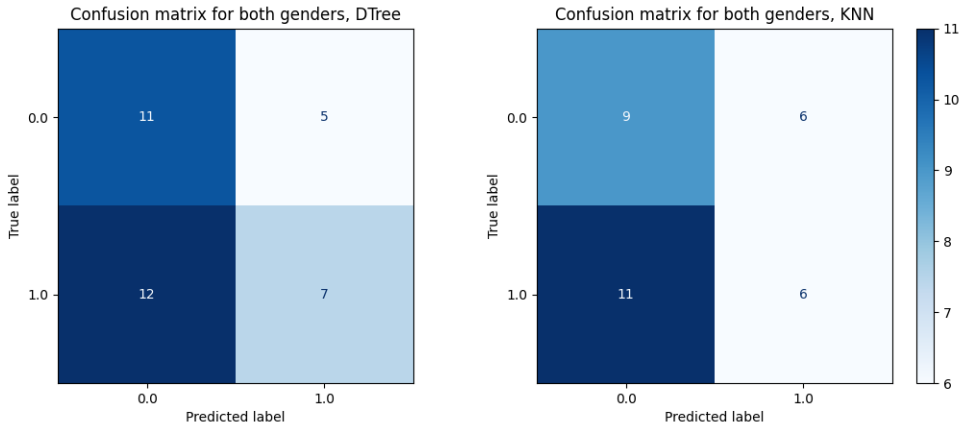


Figure 4.1: Confusion Matrices for predicting ON/OFF medication. On the left, we have a matrix for the test with severe subjects, for the model DTree. On the right, we have a matrix for the test without the severe subjects, for the model KNN.

4.1.2 Effect of Medication - Predict PD/non-PD

With Severe Subjects - Test for ON Medication

In Table 4.2, we can see the test results for when predicting PD or non-PD, with the test set including PD subjects ON medication and non-PD subjects. To the left in the table are the test results for when training the model with subjects ON medication, and to the right are the test results for when training the model with subjects OFF medication. From the training and validation results, the model that performed the best was SVM for epochs 'all' for both of the training runs, which was then used on the test set. The number of features chosen for SVM 'all' when training ON medication was 52 features; when training OFF medication, the number of features chosen was 8 features.

From Table 4.2, we can see that training the model with subjects ON medication performs the best with a balanced accuracy of 80.63%±6.803 compared to training with subjects OFF medication that had a balanced accuracy of 56.34%±8.275. The AUC (0.81±0.068) for training ON medication also performed better than the AUC (0.56±0.083) when training OFF med-

ication. In Figure 4.2, we can see the test results of the predictions of the model. To the left is when training ON medication and to the left is when training OFF medication.

Evaluation Metric	Result
Accuracy	81.12% ± 6.647
Balanced Accuracy	80.63% ± 6.803
Macro Recall	80.63% ± 6.803
Micro Recall	81.12% ± 6.647
Macro Specificity	80.63% ± 6.803
Micro Specificity	81.12% ± 6.647
Macro Precision	83.60% ± 6.384
Micro Precision	81.12% ± 6.647
Macro F1	80.47% ± 7.075
Micro F1	81.12% ± 6.647
AUC	0.81 ± 0.068

Evaluation Metric	Result
Accuracy	56.60% ± 8.263
Balanced Accuracy	56.34% ± 8.275
Macro Recall	56.34% ± 8.275
Micro Recall	56.60% ± 8.263
Macro Specificity	56.34% ± 8.275
Micro Specificity	56.60% ± 8.263
Macro Precision	56.74% ± 8.786
Micro Precision	56.60% ± 8.263
Macro F1	55.88% ± 8.403
Micro F1	56.60% ± 8.263
AUC	0.56 ± 0.083

Table 4.2: Test results for ON medication predicting PD/non-PD with severe stage subjects included. On the left, we have a table for the test results where the model was trained on subjects ON medication, for the model SVM. On the right, we have a table for the test results where the model was trained on subjects OFF medication, for the model SVM.

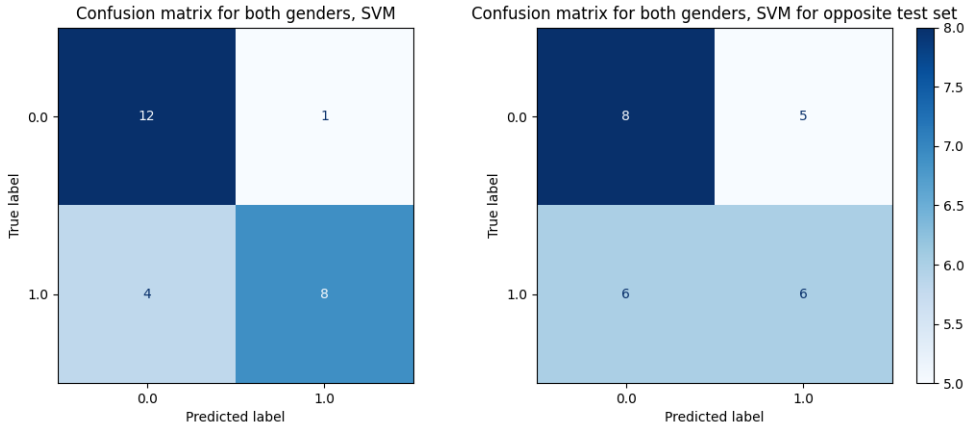


Figure 4.2: Confusion Matrices for ON medication predicting PD/non-PD with severe stage subjects included. On the left, we have a confusion matrix for the test where the model was trained on subjects ON medication, for the model SVM. On the right, we have a confusion matrix for the test where the model was trained on subjects OFF medication, for the model SVM.

With Severe Subjects - Test for OFF Medication

Table 4.3 shows the test results of predicting PD/non-PD, with the test set including PD subjects OFF medication and non-PD subjects. To the left in the table are the test results for when training the model with subjects ON medication, and to the right are the test results for when training the model with subjects OFF medication. From training and validation results, the best model was SVM for epochs 'all' for both of the scenarios, which then was used on the test sets. The number of features chosen for SVM 'all' when training ON medication was 52 features; when training OFF medication, the number of features chosen was 8 features.

From Table 4.3, we can see that training the model with subjects ON medication performs better with a balanced accuracy of 75.12% ± 7.343 and an AUC of 0.75 ± 0.073 in contrast to when training with subjects OFF medication which had a balanced accuracy of 56.19% ± 9.658 and an AUC of 0.56 ± 0.097. Figure 4.3 shows confusion matrices of the model's predictions on the test set. To the left in the figure are the test results for when training the model with subjects ON medication and to the right in the figure are the test results for when training the model with subjects OFF medication.

Evaluation Metric	**Result**
Accuracy	75.80% ± 7.141
Balanced Accuracy	75.12% ± 7.343
Macro Recall	75.12% ± 7.343
Micro Recall	75.80% ± 7.141
Macro Specificity	75.12% ± 7.343
Micro Specificity	75.80% ± 7.141
Macro Precision	79.47% ± 6.981
Micro Precision	75.80% ± 7.141
Macro F1	74.46% ± 7.994
Micro F1	75.80% ± 7.141
AUC	0.75 ± 0.073

Evaluation Metric	**Result**
Accuracy	56.40% ± 9.724
Balanced Accuracy	56.19% ± 9.658
Macro Recall	56.19% ± 9.658
Micro Recall	56.40% ± 9.724
Macro Specificity	56.19% ± 9.658
Micro Specificity	56.40% ± 9.724
Macro Precision	56.47% ± 10.072
Micro Precision	56.40% ± 9.724
Macro F1	55.87% ± 9.784
Micro F1	56.40% ± 9.724
AUC	0.56 ± 0.097

Table 4.3: Test results for OFF medication predicting PD/non-PD with severe stage subjects included. On the left, we have a table for the test results where the model was trained on subjects ON medication, for the model SVM. On the right, we have a table for the test results where the model was trained on subjects OFF medication, for the model SVM.

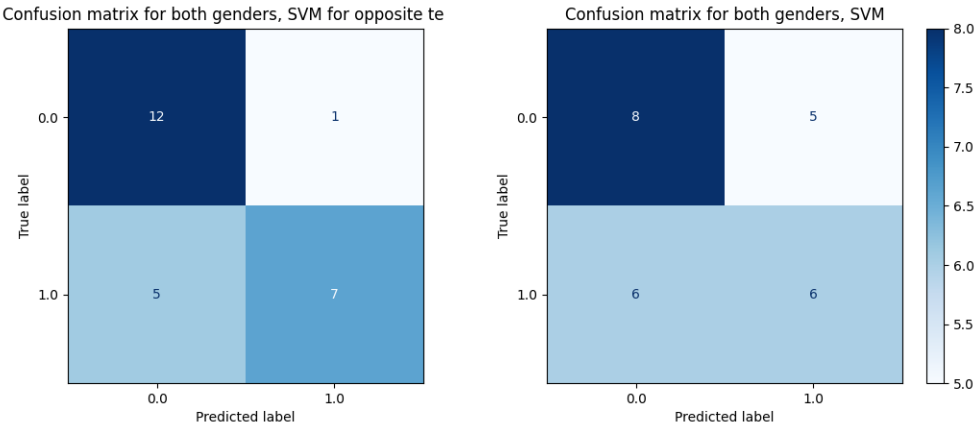


Figure 4.3: Confusion Matrices for OFF medication predicting PD/non-PD with severe stage subjects included. On the left, we have a confusion matrix for the test where the model was trained on subjects ON medication, for the model SVM. On the right, we have a confusion matrix for the test where the model was trained on subjects OFF medication, for the model SVM.

Without Severe Subjects - Test for ON Medication

Table 4.4 shows the test results for predicting PD/non-PD without severe-stage subjects, and with the test set including PD subjects ON medication and non-PD subjects. To the left in the table are the test results for training ON medication and to the right are the test results for training OFF medication. The model that performed the best in the training and validation results was KNN for epochs 'all' for both scenarios. The number of features chosen for KNN 'all' when training ON medication was 5 features; when training OFF medication, the number of features chosen was 48 features.

From Table 4.3 we see that training the model with subjects ON medication performs best with a balanced accuracy of 67.19% ± 8.059 and an AUC of 0.67 ± 0.081 while training the model on subjects OFF medication has a balanced accuracy of 51.10% ± 7.221 and an AUC of 0.51 ± 0.072. In Figure 4.4, we can see the confusion matrices of the predictions the model made on the test set. The left figure shows the predictions of when the model was trained on subjects ON medication, and the right figure shows the predictions of when the model was trained on subjects OFF medication.

Evaluation Metric	**Result**
Accuracy	66.79% ± 7.959
Balanced Accuracy	67.19% ± 8.059
Macro Recall	67.19% ± 8.059
Micro Recall	66.79% ± 7.959
Macro Specificity	67.19% ± 8.059
Micro Specificity	66.79% ± 7.959
Macro Precision	67.69% ± 8.301
Micro Precision	66.79% ± 7.959
Macro F1	66.56% ± 8.012
Micro F1	66.79% ± 7.959
AUC	0.67 ± 0.081

Evaluation Metric	**Result**
Accuracy	50.08% ± 7.120
Balanced Accuracy	51.10% ± 7.221
Macro Recall	51.10% ± 7.221
Micro Recall	50.08% ± 7.120
Macro Specificity	51.10% ± 7.221
Micro Specificity	50.08% ± 7.120
Macro Precision	51.31% ± 7.860
Micro Precision	50.08% ± 7.120
Macro F1	49.44% ± 7.213
Micro F1	50.08% ± 7.120
AUC	0.51 ± 0.072

Table 4.4: Test results for ON medication predicting PD/non-PD without severe stage subjects included. On the left, we have a table for the test results where the model was trained on subjects ON medication, for the model KNN. On the right, we have a table for the test results where the model was trained on subjects OFF medication, for the model KNN.

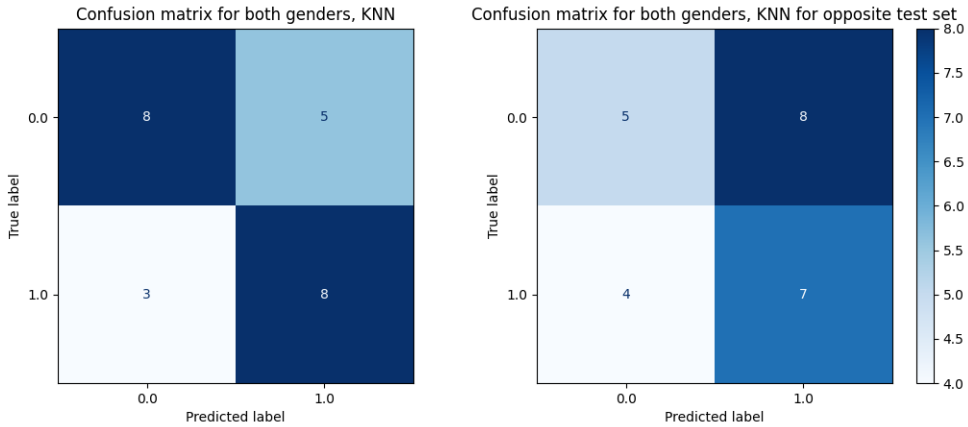


Figure 4.4: Confusion Matrices for ON medication predicting PD/non-PD without severe stage subjects included. On the left, we have a confusion matrix for the test where the model was trained on subjects ON medication, for the model KNN. On the right, we have a confusion matrix for the test where the model was trained on subjects OFF medication, for the model KNN.

Without Severe Subjects - Test for OFF Medication

In Table 4.5, we can see the test results for predicting PD/non-PD without the severe-stage subjects, and with the test set including PD subjects OFF medication and non-PD subjects. The left table shows the test results for training the model on subjects ON medication, and the right table shows the test results for training the model on subjects OFF medication. The model that performed best in the training and validation results was KNN for epochs 'all' for both scenarios. The number of features chosen for KNN 'all' when training ON medication was 5 features; when training OFF medication, the number of features chosen was 48 features.

Table 4.5 shows that training the model with subjects ON medication performs the best with a balanced accuracy of 68.29% $\pm$ 8.347 and an AUC of 0.68 $\pm$ 0.083. When training with subjects OFF medication, the model got a balanced accuracy of 54.83% $\pm$ 6.781 and an AUC of 0.55 $\pm$ 0.068. Figure 4.5 shows a confusion matrix of the predictions the model made on the test set.

Evaluation Metric	Result
Accuracy	67.79% ± 8.309
Balanced Accuracy	68.29% ± 8.347
Macro Recall	68.29% ± 8.347
Micro Recall	67.79% ± 8.309
Macro Specificity	68.29% ± 8.347
Micro Specificity	67.79% ± 8.309
Macro Precision	68.84% ± 6.683
Micro Precision	67.79% ± 8.309
Macro F1	67.59% ± 8.349
Micro F1	67.79% ± 8.309
AUC	0.68 ± 0.083

Evaluation Metric	Result
Accuracy	53.42% ± 6.492
Balanced Accuracy	54.83% ± 6.781
Macro Recall	54.83% ± 6.781
Micro Recall	53.42% ± 6.492
Macro Specificity	54.83% ± 7.781
Micro Specificity	53.42% ± 6.492
Macro Precision	55.97% ± 8.250
Micro Precision	53.42% ± 6.492
Macro F1	52.41% ± 6.428
Micro F1	53.42% ± 6.492
AUC	0.55 ± 0.068

Table 4.5: Test results for OFF medication predicting PD/non-PD without severe stage subjects included. On the left, we have a table for the test results where the model was trained on subjects ON medication, for the model KNN. On the right, we have a table for the test results where the model was trained on subjects OFF medication, for the model KNN

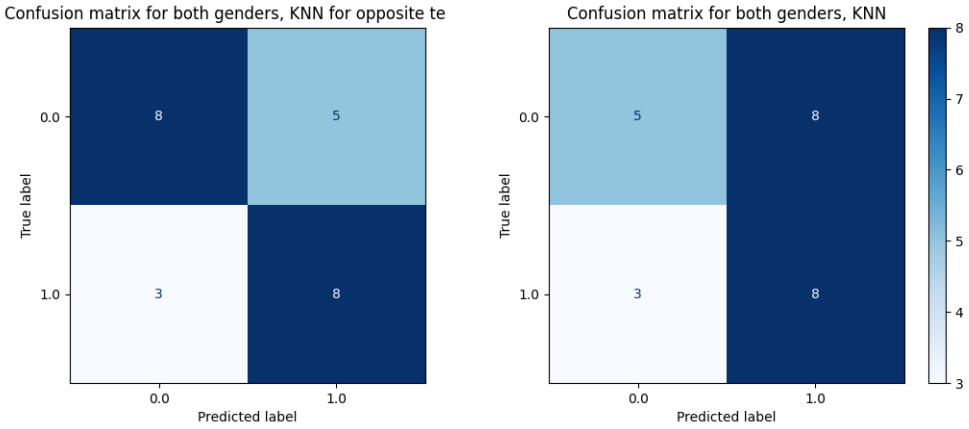


Figure 4.5: Confusion Matrices for OFF medication predicting PD/non-PD without severe stage subjects included. On the left, we have a confusion matrix for the test where the model was trained on subjects ON medication, for the model KNN. On the right, we have a confusion matrix for the test where the model was trained on subjects OFF medication, for the model KNN.

4.1.3 Effect of Severe Stage - Predict PD/non-PD

Train ON Test ON - With and Without Severe Subjects

Table 4.6 shows the results for predicting PD/non-PD when using subjects ON medication for training and testing, and to see if there is any difference between including severe-stage subjects and not including severe-stage subjects. The left table shows the test results when including severe-stage subjects. The best model from the training and validation results was SVM for epochs 'all' where 52 features were chosen, which is used for the test. The right table shows the test results when the severe-stage subjects are not included. The best model from training and validation results was KNN for epochs 'all' where 5 features were chosen, which was used for the test.

In Table 4.6, we can see that the test results when including the severe-stage subjects perform the best, with a balanced accuracy of 80.63% ± 6.803 and an AUC of 0.81 ± 0.068, compared to when the severe-stage subjects were not included which had a balanced accuracy of 67.19% ± 8.059 and AUC of 0.67 ± 0.081. Figure 4.6 shows the confusion matrix for the predictions of the model in the test. The left figure is for when we include the severe-stage subjects, and the right figure is for when we did not include the severe-stage subjects.

Evaluation Metric	**Result**
Accuracy	81.12% ± 6.647
Balanced Accuracy	80.63% ± 6.803
Macro Recall	80.63% ± 6.803
Micro Recall	81.12% ± 6.647
Macro Specificity	80.63% ± 6.803
Micro Specificity	81.12% ± 6.647
Macro Precision	83.60% ± 6.384
Micro Precision	81.12% ± 6.647
Macro F1	80.47% ± 7.075
Micro F1	81.12% ± 6.647
AUC	0.81 ± 0.068

Evaluation Metric	**Result**
Accuracy	66.79% ± 7.959
Balanced Accuracy	67.19% ± 8.059
Macro Recall	67.19% ± 8.059
Micro Recall	66.79% ± 7.959
Macro Specificity	67.19% ± 8.059
Micro Specificity	66.79% ± 7.959
Macro Precision	67.69% ± 8.301
Micro Precision	66.79% ± 7.959
Macro F1	66.56% ± 8.012
Micro F1	66.79% ± 7.959
AUC	0.67 ± 0.081

Table 4.6: Test results for predicting PD/non-PD. On the left, we have the test results with severe-stage subjects included, for the model SVM. On the right, we have the test results without severe-stage subjects included, for the model KNN.

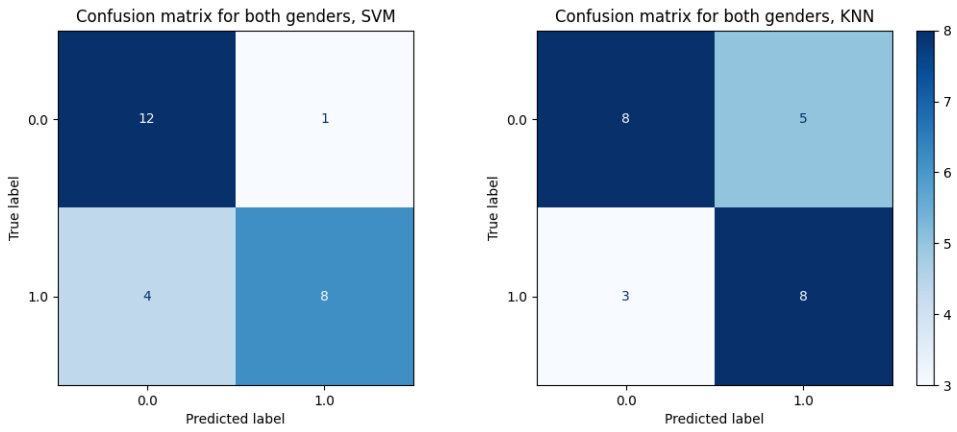


Figure 4.6: Confusion matrices for predicting PD/non-PD. On the left, we have the test results with severe-stage subjects included, for the model SVM. On the right, we have the test results without severe-stage subjects included, for the model KNN.

Train OFF Test OFF - With and Without Severe Subjects

Table 4.7 shows the test results for predicting PD/non-PD when using subjects OFF medication for training and testing, and to see if there is any difference between including severe-stage subjects and not including severe-stage subjects. The left table shows the test results when including the severe-stage subjects, The best model from training and validation results was SVM for epochs 'all' where 8 features were chosen, which was used for testing. The right table shows the test results when the severe-stage subjects were not included. The best model from training and validation results was KNN for epochs 'all' where 48 features were chosen, which was used for testing.

In Table 4.7, we can see that the test result, when including the severe-stage subjects, performs the best, with a balanced accuracy of 56.19% $\pm$ 9.658 and an AUC of 0.56 $\pm$ 0.097. When not including the severe-stage subjects, the model gets a balanced accuracy of 54.83% $\pm$ 6.781 and an AUC of 0.55 $\pm$ 0.068. In Figure 4.7, we can see the confusion matrices displaying the predictions the model made on the test set.

Evaluation Metric	Result
Accuracy	56.40% ± 9.724
Balanced Accuracy	56.19% ± 9.658
Macro Recall	56.19% ± 9.658
Micro Recall	56.40% ± 9.724
Macro Specificity	56.19% ± 9.658
Micro Specificity	56.40% ± 9.724
Macro Precision	56.47% ± 10.072
Micro Precision	56.40% ± 9.724
Macro F1	55.87% ± 9.784
Micro F1	56.40% ± 9.724
AUC	0.56 ± 0.097

Evaluation Metric	Result
Accuracy	53.42% ± 6.492
Balanced Accuracy	54.83% ± 6.781
Macro Recall	54.83% ± 6.781
Micro Recall	53.42% ± 6.492
Macro Specificity	54.83% ± 7.781
Micro Specificity	53.42% ± 6.492
Macro Precision	55.97% ± 8.250
Micro Precision	53.42% ± 6.492
Macro F1	52.41% ± 6.428
Micro F1	53.42% ± 6.492
AUC	0.55 ± 0.068

Table 4.7: Test results for predicting PD/non-PD. On the left, we have the test results with severe-stage subjects included, for the model SVM. On the right, we have the test results without severe-stage subjects included, for the model KNN.

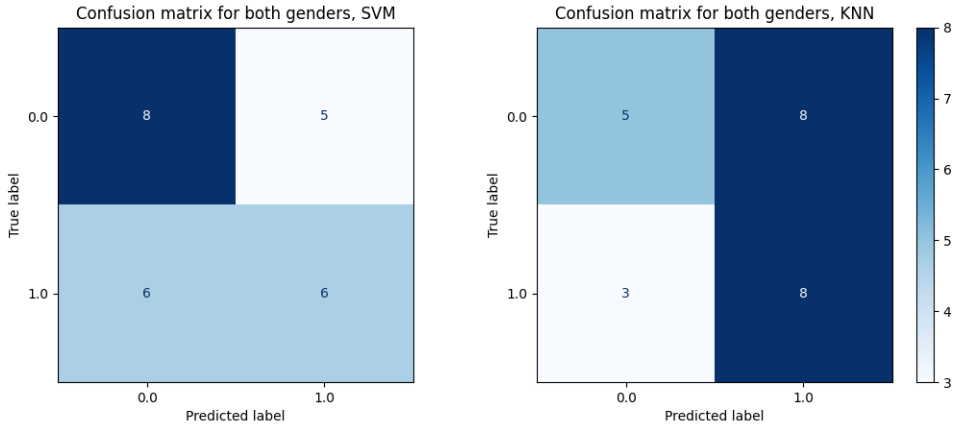


Figure 4.7: Confusion matrices for predicting PD/non-PD. On the left, we have the confusion matrix where severe-stage subjects are included, for the model SVM. On the right, we have the confusion matrix where severe-stage subjects were not included, for the model KNN.

4.1.4 Conformal Predictions - Predict PD/Non-PD

In addition to having used two different α-values, we have also run the model without making any conformal predictions. This has given us a baseline from which to evaluate the conformal predictions. The best classifier is LR

and the number of chosen features is 113.

The Table 4.8 shows the baseline together with conformal predictions for both different α-values. The baseline run is overall very good, with balanced accuracy (75.92% ± 6.024) and AUC (0.76 ± 0.060) being better than $\alpha = 0.65$.
We got the highest balanced accuracy for $\alpha = 0.85$ (79.29% ± 9.601). $\alpha = 0.65$ has both higher efficiency (64.71%) and validity (50%) then $\alpha = 0.85$ (29.41%) and (23.53%).

Figure 4.8 shows the confusion matrices for the baseline, as well as both the different α-values. The matrix in the top left represents the baseline model. The matrix for the model with α=0.65 is in the top right, and the matrix on the last row represents the model with α=0.85.

Evaluation Metric	Baseline	α=**0.65**	α=**0.85**
Accuracy	76.32% ± 6.062	77.23% ± 7.315	80.10% ±7.680
Balanced Accuracy	75.92% ± 6.024	73.77% ± 8.141	79.29% ± 9.601
Macro Recall	75.92% ± 6.024	73.77% ± 8.141	79.29% ± 9.601
Micro Recall	76.32% ± 6.062	77.23% ± 7.315	80.10% ± 7.680
Macro Specificity	75.92% ± 6.024	73.77% ± 8.141	79.23% ± 9.601
Micro Specificity	76.32% ± 6.062	77.23% ± 7.315	80.10% ± 7.680
Macro Precision	77.01% ± 6.440	79.69% ± 8.991	79.41% ± 8.176
Micro Precision	76.32% ± 6.062	77.23% ±7.315	80.10% ± 7.680
Macro F1	75.95% ± 6.102	74.24% ± 8.833	78.85% ± 8.964
Micro F1	76.32% ± 6.062	77.23% ± 7.315	80.10% ± 7.680
AUC	0.76 ± 0.060	0.74 ± 0.080	0.79 ± 0.096
Efficiency	-	64.71%	29.41%
Validity	-	50%	23.53%

Table 4.8: Test results for Predicting PD/Non-PD. On the left, we have the baseline run; in the middle, we have $\alpha = 0.65$ and on the right we have$\alpha = 0.85$. LR has been used, with 113 features.

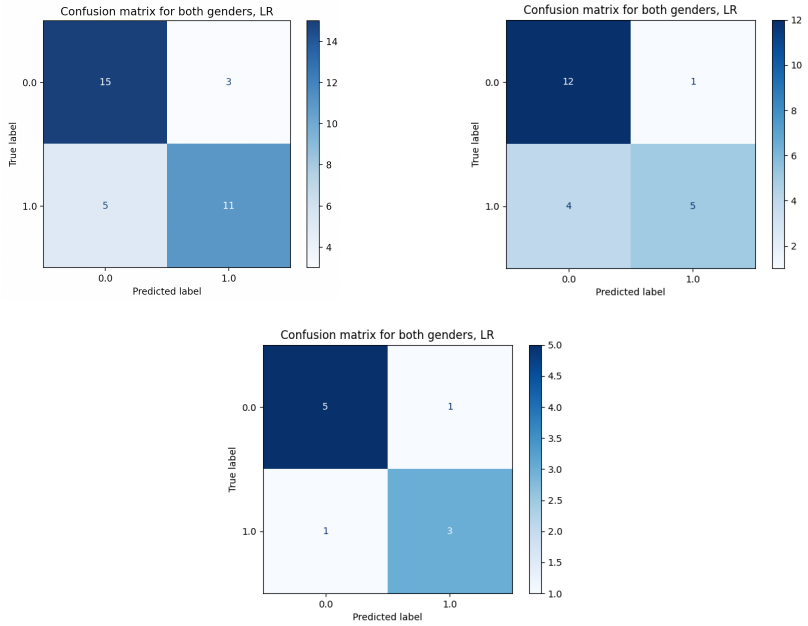


Figure 4.8: Confusion matrices for prediction of PD/Non-PD. In the top left we have the matrix for the baseline. In the top right, we have the matrix for $\alpha = 0.65$ and on the bottom, it is the matrix for $\alpha = 0.85$

4.1.5 Conformal Predictions - Predict PD/Non-PD for different sub-groups

Gender Differences

Table 4.9 shows how the models have performed on males. We can see that the best performance is for $\alpha = 0.85$ (82.67% $\pm$ 12.893). The highest AUC score is also highest for $\alpha = 0.85$ (0.83 $\pm$ 0.13). The best efficiency and validity we got was from $\alpha = 0.65$, with efficiency being 72.22% and validity being 55.56%

Table 4.10 shows the performance for females. Here the highest balanced accuracy was for $\alpha = 0.85$ (83.33%). The lowest score was 66.36% $\pm$ 18.696, coming from $\alpha = 0.65$.

Evaluation Metric	Baseline	$\alpha = 0.65$	$\alpha = 0.85$
Accuracy	71.94% ± 8.366	77.23% ± 8.491	82.67% ± 12.893
Balanced Accuracy	71.94% ± 8.366	78.86% ± 7.884	82.67% ± 12.893
Macro Recall	71.94% ± 8.366	78.86% ± 7.884	82.67% ± 12.893
Micro Recall	71.94% ± 8.366	77.23% ± 8.491	82.67% ± 12.893
Macro Specificity	71.94% ± 8.366	78.86% ± 7.884	82.67% ± 12.893
Micro Specificity	71.94% ± 8.366	77.23% ± 8.491	82.67% ± 12.893
Macro Precision	-	83.99% ± 4.201	87.03% ± 12.999
Micro Precision	-	77.23% ± 8.491	82.67% ± 12.893
Macro F1	70.86% ± 9.108	76.25% ± 9.654	80.97% ± 15.454
Micro F1	71.94% ± 8.366	77.23% ± 8.491	82.67% ± 12.893
AUC	0.72 ± 0.084	0.79 ± 0.079	0.83 ± 0.13
Efficiency	-	72.22%	33.33%
Validity	-	55.56%	27.78%

Table 4.9: Test results for Predicting PD/Non-PD for males. On the left, we have the baseline run; in the middle, we have $\alpha = = 0.65$, and on the right, we have $\alpha = 0.85$. LR has been used, with 113 features.

In Figure 4.9 we can see all confusion matrices for both male and female. The confusion matrices for males are on the left, and on the right, it is for females. The first row represents the baseline models. In the second row are the models with $\alpha = 0.65$, and the last row has the matrices for the models with $\alpha = 0.85$.

Evaluation Metric	Baseline	$\alpha = 0.65$	$\alpha = 0.85$
Accuracy	81.38% ± 8.794	77.67% ± 12.018	75.00% ± 0.000
Balanced Accuracy	81.87% ± 8.652	66.36% ± 18.696	83.33% ± 0.000
Macro Recall	81.87% ± 8.652	66.36% ± 18.696	83.33% ± 0.000
Micro Recall	81.38% ± 8.794	77.67% ± 12.018	75.00% ± 0.000
Macro Specificity	81.87% ± 8.652	66.36% ± 18.696	83.33% ± 0.000
Micro Specificity	81.38% ± 8.794	77.67% ± 12.018	75.00% ± 0.000
Macro Precision	82.61% ± 8.676	65.96% ± 22.047	75.00% ± 0.000
Micro Precision	81.38% ± 8.794	77.67% ± 12.018	75.00% ± 0.000
Macro F1	81.14% ± 8.887	64.86% ± 19.263	73.33% ± 0.000
Micro F1	81.38% ± 8.794	77.67% ± 12.018	75.00% ± 0.000
AUC	0.82 ± 0.087	0.66 ± 0.19	0.83 ± 0.000
Efficiency	-	56.25%	25%
Validity	-	43.75%	18.75%

Table 4.10: Test results for Predicting PD/Non-PD for females. On the left, we have the baseline run; in the middle, we have α= 0.65 and on the right we have α= 0.85. The classifier used is LR for all runs

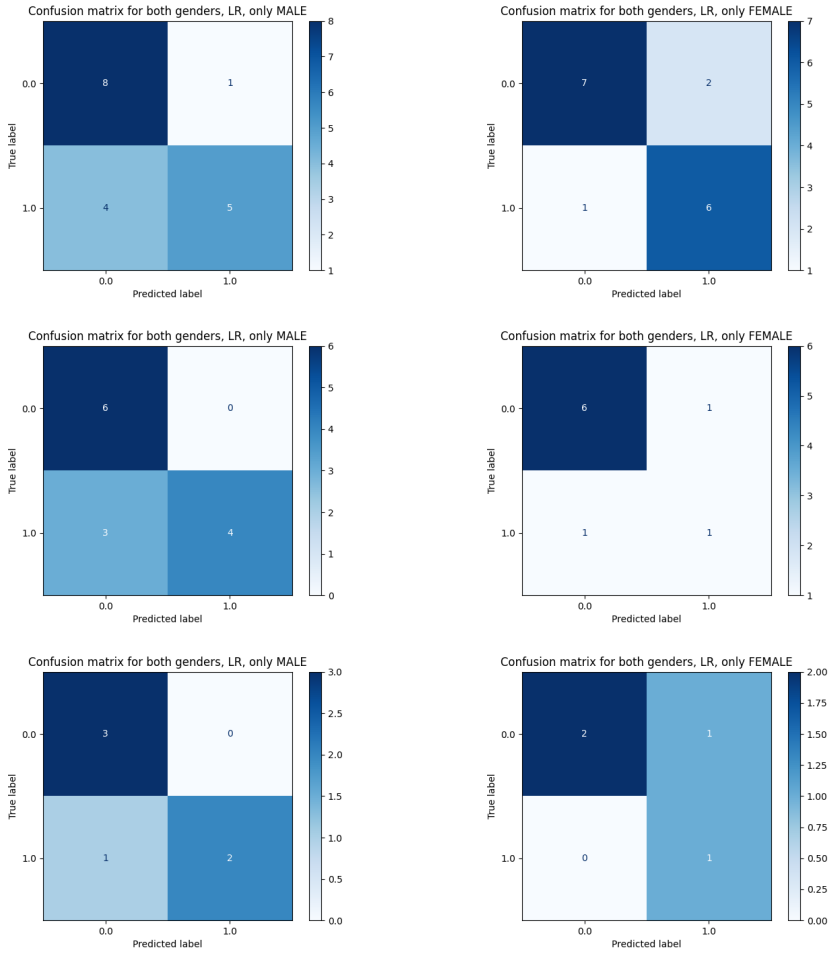


Figure 4.9: Confusion Matrices for predicting PD/Non-PD without conformal. On the left, we have a matrix for the test with only males. On the right, we have a matrix for the test for only females. The matrices in the top row are from the baseline. The matrices in the middle are run with $\alpha = 0.65$ and the two matrices in the last row it is with $\alpha = 0.85$.

Severity Difference

The Table 4.11 shows the test results for subjects with mild PD. Here we can see that balanced accuracy for the baseline is (79.81% ± 6.444). for $\alpha = 0.65$ the balanced accuracy is 80.74$ ± 11.223 and for $\alpha = 0.85$, the balanced accuracy is 91.67%. We can also see that the AUC for the baseline is (0.80 ± 0.064), for $\alpha = 0.65$, the AUC is 0.80 ± 0.11 and for $\alpha = 0.85$, the AUC is 0.92.

Evaluation Metric	Baseline	$\alpha = 0.65$	$\alpha = 0.85$
Accuracy	81.27% ± 5.925	88.50% ± 6.418	89.50% ± 0.000
Balanced Accuracy	79.81% ± 6.444	80.74% ± 11.334	91.67% ± 0.000
Macro Recall	79.81% ± 6.444	80.74% ± 11.334	91.67% ± 0.000
Micro Recall	81.27% ± 5.925	88.50% ± 6.418	87.50% ± 0.000
Macro Specificity	79.81% ± 6.444	80.74% ± 11.334	91.67% ± 0.000
Micro Specificity	81.27% ± 5.925	88.50% ± 6.418	87.50% ± 0.000
Macro Precision	78.56% ± 6.647	83.29% ± 11.865	83.33% ± 0.000
Micro Precision	81.27% ± 5.925	88.50% ± 6.418	87.50% ± 0.000
Macro F1	78.73% ± 6.533	80.73% ± 10.602	85.45% ± 0.000
Micro F1	81.27% ± 5.925	88.50% ± 6.418	87.50% ± 0.000
AUC	0.80 ± 0.064	0.80 ± 0.11	0.92 ± 0.000
Efficiency	-	61.54%	30.77%
Validity	-	53.85%	26.92%

Table 4.11: Test results for Predicting PD/Non-PD for subjects with either "severity" = mild or non-PD. On the left, we have the baseline run; in the middle, we have α= 0.65 and on the right we have α= 0.85. LR has been used, with 113 features.

The Table 4.12 shows the test results for subjects with moderate PD and non-PD. We can see that the AUC for the baseline run is 0.73 ± 0.081. The efficiency for $\alpha = 0.65$ is 73.08%. For $\alpha = 0.85$, the balanced accuracy is 70.92% ± 16.604.

Figure 4.10 shows the confusion matrices for both mild+non-PD and moderate+non-PD. In the top row, the confusion matrices from the baseline run are shown. In the middle row, it is the models with α= 0.65 and the last row shows the confusion matrices when α= 0.85

Evaluation Metric	Baseline	$\alpha = 0.65$	$\alpha = 0.85$
Accuracy	77.54% ± 6.588	78.11% ± 7.282	77.12% ± 8.302
Balanced Accuracy	73.22% ± 8.074	70.09% ± 9.558	70.92% ± 16.604
Macro Recall	73.22% ± 8.074	70.09% ± 9.558	70.92% ± 16.604
Micro Recall	77.54% ± 6.588	78.11% ± 7.282	77.12% ± 8.302
Macro Specificity	73.22% ± 8.074	70.09% ± 9.558	70.92% ± 16.604
Micro Specificity	77.54% ± 6.544	78.11% ± 7.282	77.12% ± 8.302
Macro Precision	74.00% ± 7.816	77.39% ± 12.238	67.36% ± 15.235
Micro Precision	77.54% ± 6.588	78.11% ± 7.282	77.12% ± 8.302
Macro F1	73.21% ± 7.966	71.03% ± 10.594	68.45% ± 15.136
Micro F1	77.54% ± 6.588	78.11% ± 7.282	77.12% ± 8.302
AUC	0.73 ± 0.081	0.70 ± 0.096	0.71 ± 0.166
Efficiency	-	73.08%	30.77%
Validity	-	57.69%	23.08%

Table 4.12: Test results for Predicting PD/Non-PD for subjects with either "severity" = moderate or non-PD. On the left, we have the baseline run; in the middle, we have α= 0.65 and on the right we have α= 0.85. The classifier used is LR for all runs

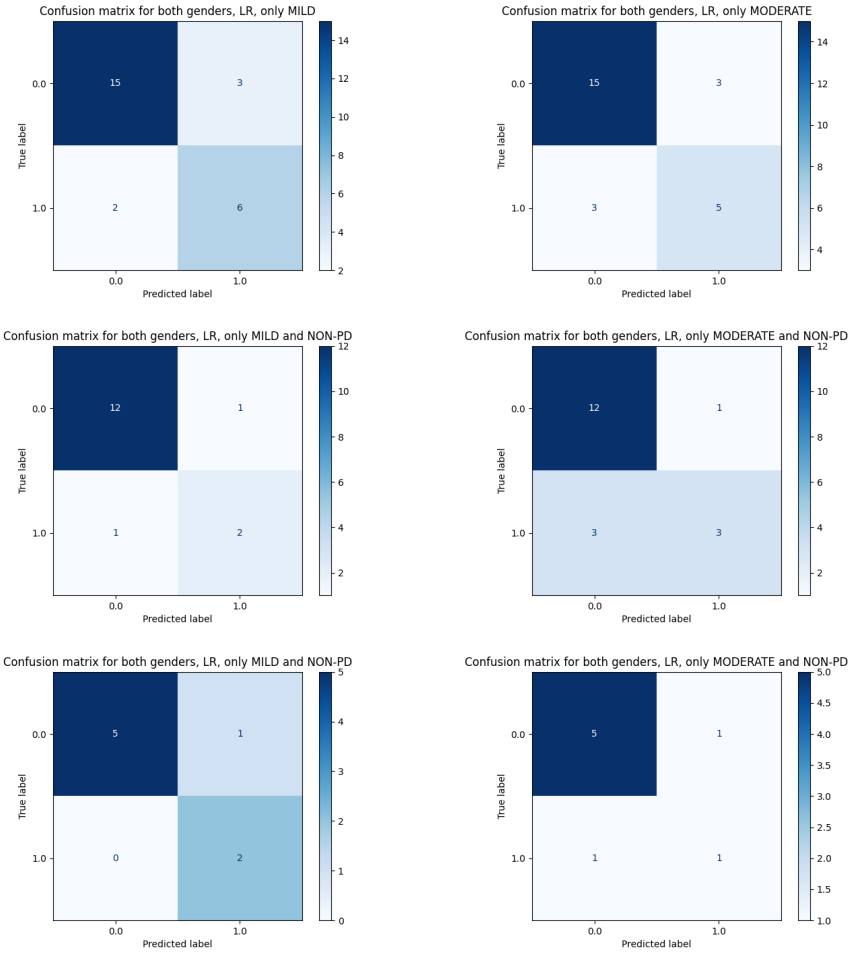


Figure 4.10: Confusion Matrices for predicting PD/Non-PD for "severity"= mild and "severity"= moderate together with non-PD. On the left it is the matrices for mild/non-PD, and on the right are the matrices for moderate/non-PD. The top matrices are for the baseline. The matrices in the middle have α-values of 0.65 and the two on the bottom have α-values of 0.85.

Centre Differences

and Table 4.15 shows the results from each of the different centres. We can see that the balanced accuracy is highest for New Mexico and α=0.85 (88.12% ± 9.417), with the lowest being Turku (50%).

In Table 4.13 we can see that the balanced accuracy for the baseline is 65.00 ± 7.616 and 50.00% for $\alpha = 0.65$. The AUC is 0.65 ± 0.076 in the baseline, and for $\alpha = 0.65$ it is 50.00.

Evaluation Metric	Baseline	$\alpha = 0.65$	$\alpha = 0.85$
Accuracy	66.67% ± 8.462	75.00% ± 0.000	-
Balanced Accuracy	65.00% ± 7.616	50.00% ± 0.000	-
Macro Recall	65.00% ± 7.616	50.00% ± 0.000	-
Micro Recall	66.67% ± 8.462	75.00% ± 0.000	-
Macro Specificity	65.00% ± 7.616	50.00% ± 0.000	-
Micro Specificity	66.67% ± 8.462	75.00% ± 0.000	-
Macro Precision	68.64% ± 11.479	37.50% ± 0.000	-
Micro Precision	66.67% ± 8.462	75.00% ± 0.000	-
Macro F1	64.95% ± 7.660	42.86% ± 0.000	-
Micro F1	66.67% ± 8.462	75.00% ± 0.000	-
AUC	0.65 ± 0.076	50.00 ± 0.000	-
Efficiency	-	44.44%	0.11%
Validity	-	33.33%	0%

Table 4.13: Test results for Predicting PD/Non-PD for subjects from Turku. On the left, we have the baseline run; in the middle, we have α= 0.65 and on the right we have α= 0.85. LR has been used, with 113 features.

The Table 4.14 shows the test results for all subjects from New Mexico. The balanced accuracy for the baseline was 88.00% ± 7.109. The model performs best when α=0.85 (88.12% ± 9.417), and performs the poorest for α=0.65 (83.41% ± 8.858). The same pattern can be seen for the AUC, with the best performance from α=0.85 (0.88 ± 0.942) and the lowest score from α=0.65 (0.81 ± 0.886)

The Figure 4.11 shows the baseline confusion matrices for each centre. The matrix in the top right shows results from Turku, and the matrix on the right shows results from New Mexico. The matrix at the bottom shows

Evaluation Metric	Baseline	$\alpha = 0.65$	$\alpha = 0.85$
Accuracy	88.00% ± 7.109	84.69% ± 8.177	86.43% ± 10.762
Balanced Accuracy	88.00% ± 7.109	83.42% ± 8.858	88.12% ± 9.417
Macro Recall	88.00% ± 7.109	83.42% ± 8.858	88.12% ± 9.417
Micro Recall	88.00% ± 7.109	84.69% ± 8.177	86.43% ± 10.762
Macro Specificity	88.00% ± 7.109	83.42% ± 8.858	88.12% ± 9.417
Micro Specificity	88.00% ± 7.109	84.69% ± 8.177	86.43% ± 10.762
Macro Precision	90.83% ± 4.514	89.47% ± 4.613	89.35% ± 7.388
Micro Precision	88.00% ± 7.109	84.69% ± 8.177	86.43% ± 10.762
Macro F1	87.60% ± 7.724	83.30% ± 9.739	86.21% ± 11.225
Micro F1	88.00% ± 7.109	84.69% ± 8.177	86.43% ± 10.762
AUC	0.88 ± 0.071	0.81 ± 0.886	0.88 ± 0.942
Efficiency	-	81.25%	43.75%
Validity	-	68.75%	37.50%

Table 4.14: Test results for Predicting PD/Non-PD for subjects from New Mexico. On the left, we have the baseline run; in the middle, we have α= 0.65 and on the right we have α= 0.85. LR has been used, with 113 features.

results from San Diego.

The Figure 4.12 shows the confusion matrices for each centre with α=0.65. The matrix in the top right shows results from Turku, and the matrix on the right shows results from New Mexico. The matrix at the bottom shows results from San Diego.

Evaluation Metric	Baseline	$\alpha = 0.65$	$\alpha = 0.85$
Accuracy	68.00% ± 12.793	57.80% ± 21.706	-
Balanced Accuracy	68.75% ± 12.701	56.67% ± 22.174	-
Macro Recall	68.75% ± 12.701	56.67% ± 22.174	-
Micro Recall	68.00% ± 12.793	57.80% ± 21.706	-
Macro Specificity	68.75% ± 12.701	56.67% ± 22.174	-
Micro Specificity	68.00% ± 12.793	57.80% ± 21.706	-
Macro Precision	70.18% ± 13.520	55.46% ± 27.311	-
Micro Precision	68.00% ± 12.793	57.80% ± 21.706	-
Macro F1	67.38% ± 13.045	-	-
Micro F1	68.00% ± 12.793	57.80% ± 21.706	-
AUC	0.69% ± 0.127	0.567 ± 0.221	-
Efficiency	-	55.56%	-
Validity	-	33.33%	-

Table 4.15: Test results for Predicting PD/Non-PD for subjects from San Diego. On the left, we have the baseline run; in the middle, we have α= 0.65 and on the right we have α= 0.85. LR has been used, with 113 features.

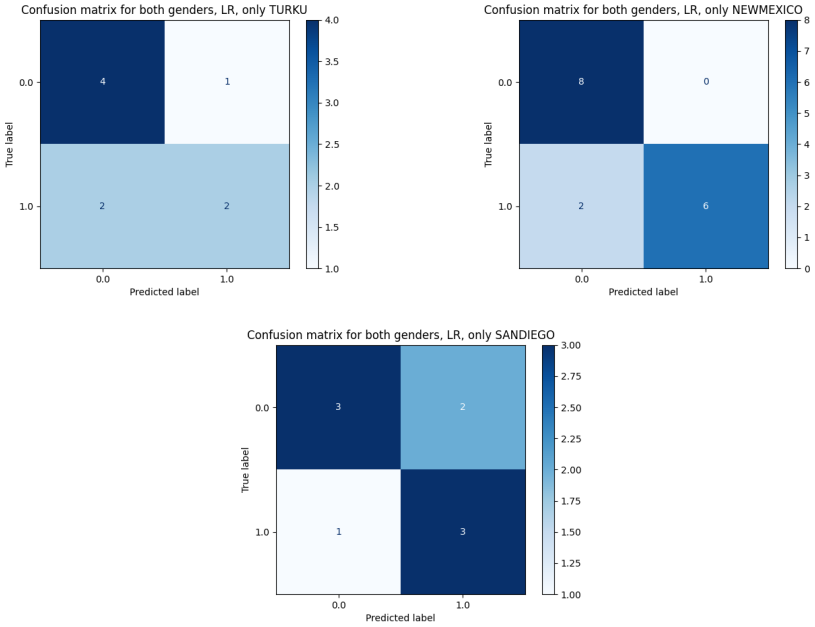


Figure 4.11: Confusion matrixes for predicting PD/non-PD without using conformal predictions sorted by the different centres. Results for Turku are in the top left, with the results for New Mexico in the top right. Results for San Diego are in the bottom centre. LR has been used, with 113 features..

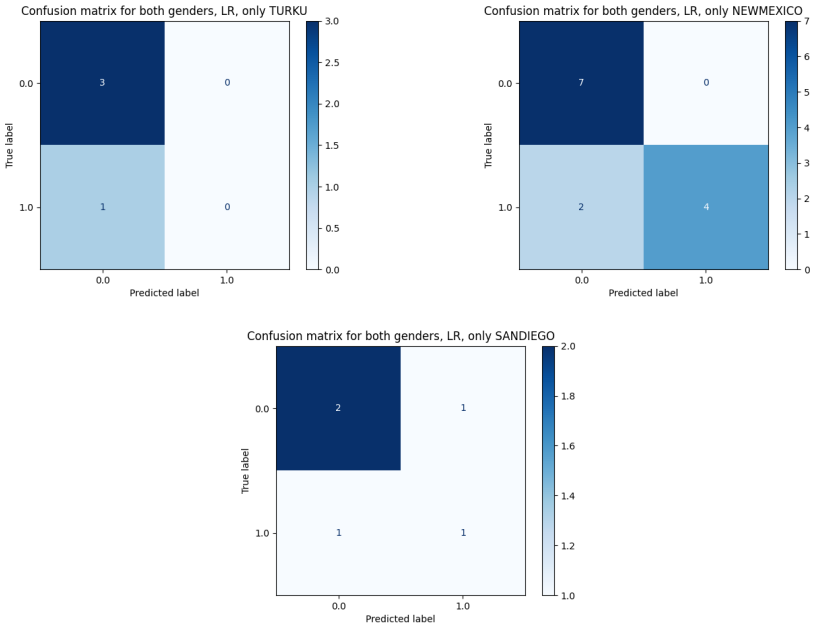


Figure 4.12: Confusion matrixes for predicting PD/non-PD with conformal predictions ($\alpha = 0.65$)sorted by the different centres. Results for Turku are in the top left, with the results for New Mexico in the top right. Results for San Diego are in the bottom centre. LR has been used, with 113 features.LR has been used, with 113 features..

Chapter 5

Discussion and limitations

5.1 Discussion

5.1.1 Effect of Medication

When predicting whether a subject is ON or OFF the medication phase of levodopa without severe-stage subjects, the results indicate that the EEG signals do not contain enough information about their medication status in the mild and moderate subjects. However, when including severe-stage subjects, we saw a slight improvement in the results, improving the balanced accuracy to above 50%, indicating that there might be some information about the medication status in the severe subjects compared to mild/moderate subjects. Overall, the test results indicate that the EEG signals don't contain enough information about the medication status for the model to distinguish between ON/OFF medication, but it seems that severe subjects have some little "traces" of hidden information in the EEG about the medication status.

When predicting whether a subject is PD or non-PD while using a test set of subjects ON medication and non-PD, the results when training with subjects ON medication perform the best. When looking at the results when training with subjects OFF medication, we can see a significant decrease

in the performance of the model. After including the severe-stage subjects, we saw an improvement in the results for both training with subjects ON medication and training with subjects OFF medication. While using a test set of subjects OFF medication and non-PD subjects, the results had a similar outcome. The model performed the best when it was trained on subjects ON medication compared to when it was trained on subjects OFF medication. We can also see an increase in performance when including severe-stage subjects for both training with subjects ON medication and training with subjects OFF medication.
Altogether, the model performs the best when it is trained on subjects ON medication, and including severe subjects in the training also increases performance compared to when not. In addition, the model seems more robust when being trained on subjects ON medication, in that the model is still performing quite well when being used to test on subjects OFF medication.

5.1.2 Effect of Severe Stage

We also made some comparisons to determine the effect of training with or without the severe stage subjects. Most notably, when using subjects ON medication and non-PD subjects for both training and testing, the model achieved far better test results when including the severe stage subjects in the training set. When using subjects OFF medication and non-PD subjects for both training and testing, we observed a performance increase when including severe stage subjects in the training set, though not as significant as the other comparison. In essence, including severe stage subjects in the training set would improve the model and results. And also, the fact that results increase more when using subjects ON medication, shows that the medication status is indeed playing a role in the classification, particularly in the severe subjects.

5.1.3 Conformal Predictions for PD detection

When using a conformal prediction set with $\alpha = 0.85$ we saw an overall improvement of our model. Every metric improved overall, with balanced accuracy improving with 3.4%pt and AUC improved with 0.03.

Using a conformal prediction set with $\alpha = 0.65$ we can see a decrease in the overall performance. However, let's look at the confusion matrix. We can see that the model is performing well in predicting non-PD, with most of the wrong predictions coming from predicting PD, indicating the model is picking up some patterns for the non-PD subjects easier than for PD subjects.

In evaluating the baseline gender models, we saw significant differences in PD detection in males (71.94% ± 8.366) vs females (81.87% ± 8.652). We also see a big difference when applying a conformal prediction set with $\alpha = 0.65$, however, males (78.86% ± 7.884) have much higher accuracy than females (66.36% ± 18.696).
This is similar to the results of Kurbatskaya et al. [3]. Their results got an accuracy of 80.5% for males and 63.7% for females, and they point towards a higher activity for some parietal and frontal EEG channels and frequency sub-bands for male subjects to explain these differences in PD detection ability between the two gender subgroups.
In our work, using a conformal prediction set with α=0.85, these differences become negligible, with the differences becoming less than 1%pt. This indicates that the difference between the genders does not necessarily come from the EEG being unfair, but rather that the classifier was unsure about its prediction.

We did not receive any evaluation metrics for Turku and San Diego when α was set to 0.85. The reason for this is that only a few subjects received a prediction from the conformal prediction set, which is insufficient for calculating any metrics. This also affected the confusion matrices, and that is why they are not displayed.

Chapter 6

Conclusions and future work

6.1 Conclusion

This bachelor's thesis explored how a subject's medication status affected the performance of the model. We also explored if we could improve the models by using conformal prediction, which would give them a notion of "confidence" in their predictions. The results of this thesis can aid in further improvements in detecting PD using EEG signals.

The results show that the goals we set for this thesis were met. We have explored how a subject's medication status can affect the performance of the models, where the results show that the model performs best when it has been trained on subjects ON medication. We have also explored whether a PD subject's EEG signals contain information about their medication status. The test results indicate that the EEG signals don't contain enough information about the medication status to differentiate between ON/OFF medication. We have also explored the effect of training with or without severe stage subjects, where the results show that including severe stage subjects in the training set increases the performance of the model. When implementing a conformal prediction set with a high *alpha* value, we found the performance of the model increased at the expense of dropping a lot of predictions, especially for PD

6.2 Future Works

This thesis has achieved optimistic results in using EEG signals in detecting PD, but there is still room for improvement. Some directions for future work can be:

- Using a bigger dataset when using conformal predictions.
- Looking more into fairness vs. uncertainty for the different genders.
- Explore deeper how the "severe" and other stages affect the model.

Bibliography

[1] *Where is EEG performed?* Epilepsy Foundation. URL: https://www.epilepsy.com/diagnosis/eeg/where-its-performed (visited on 05/14/2024).

[2] Andrew M. Miller et al. "Effect of levodopa on electroencephalographic biomarkers of the parkinsonian state". In: *Journal of Neurophysiology* 122.1 (2019). _eprint: https://doi.org/10.1152/jn.00141.2019, pp. 290–299. DOI: 10.1152/jn.00141.2019. URL: https://doi.org/10.1152/jn.00141.2019.

[3] Anna Kurbatskaya et al. *Assessing gender fairness in EEG-based machine learning detection of Parkinson's disease: A multi-center study.* Mar. 11, 2023. arXiv: 2303.06376[cs,eess]. URL: http://arxiv.org/abs/2303.06376.

[4] Anna Kurbatskaya et al. *Machine Learning-Based Detection of Parkinson's Disease From Resting-State EEG: A Multi-Center Study.* Mar. 2, 2023. arXiv: 2303.01389[cs,eess]. URL: http://arxiv.org/abs/2303.01389 (visited on 05/08/2024).

[5] George DeMaagd and Ashok Philip. "Parkinson's Disease and Its Management: Part 1: Disease Entity, Risk Factors, Pathophysiology, Clinical Presentation, and Diagnosis". In: *P & T: A Peer-Reviewed Journal for Formulary Management* 40.8 (Aug. 2015), pp. 504–532. ISSN: 1052-1372.

[6] NHI. "Parkinson's Disease: Hope Through Research". In: *National Institute of neurological and disorders and stroke* (Mar. 2020). URL: https://www.ninds.nih.gov/sites/default/files/migrate-documents/parkinsons_htr_english_20-ns-139_508c.pdf.

[7] Joji Fujikawa et al. "Diagnosis and Treatment of Tremor in Parkinson's Disease Using Mechanical Devices". In: *Life (Basel, Switzerland)* 13.1 (Dec. 27, 2022), p. 78. ISSN: 2075-1729. DOI: 10.3390/life13010078.

[8] María Del Rosario Ferreira-Sánchez, Marcos Moreno-Verdú, and Roberto Cano-de-la-Cuerda. "Quantitative Measurement of Rigidity in Parkinson´s Disease: A Systematic Review". In: *Sensors (Basel, Switzerland)* 20.3 (Feb. 6, 2020), p. 880. ISSN: 1424-8220. DOI: 10.3390/s20030880.

[9] Matteo Bologna et al. "Evolving concepts on bradykinesia". In: *Brain: A Journal of Neurology* 143.3 (Mar. 1, 2020), pp. 727–750. ISSN: 1460-2156. DOI: 10.1093/brain/awz344.

[10] Bhavana Palakurthi and Sindhu Preetham Burugupally. "Postural Instability in Parkinson's Disease: A Review". In: *Brain Sciences* 9.9 (Sept. 18, 2019), p. 239. ISSN: 2076-3425. DOI: 10.3390/brainsci9090239.

[11] Philippe Rizek, Niraj Kumar, and Mandar S. Jog. "An update on the diagnosis and treatment of Parkinson disease". In: *CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne* 188.16 (Nov. 1, 2016), pp. 1157–1165. ISSN: 1488-2329. DOI: 10.1503/cmaj.151179.

[12] *Levodopa (co-beneldopa and co-careldopa)*. Parkinson's UK. URL: https://www.parkinsons.org.uk/information-and-support/levodopa-co-beneldopa-and-co-careldopa (visited on 05/06/2024).

[13] "Drug Treatments For Parkinson's". In: (July 2019). Ed. by Parkinson's UK. URL: https://www.parkinsons.org.uk/sites/default/files/2019-09/Drug%20treatment%20for%20Parkinson%27s%20booklet%20PRINT_LOWRES.pdf.

[14] J. Satheesh Kumar and P. Bhuvaneswari. "Analysis of Electroencephalography (EEG) Signals and Its Categorization–A Study". In: *Procedia Engineering* 38 (2012), pp. 2525–2536. ISSN: 1877-7058. DOI: https://doi.org/10.1016/j.proeng.2012.06.298. URL: https://www.sciencedirect.com/science/article/pii/S1877705812022114.

[15] NHS. *Electroencephalogram (EEG)*. nhs.uk. Section: conditions. Oct. 18, 2017. URL: https://www.nhs.uk/conditions/electroencephalogram/ (visited on 03/28/2024).

[16] Mahtab Roohi-Azizi et al. "Changes of the brain's bioelectrical activity in cognition, consciousness, and some mental disorders". In: *Medical Journal of the Islamic Republic of Iran* 31 (Dec. 2017), pp. 307–312. DOI: 10.14196/mjiri.31.53.

[17] Majid Aljalal et al. "Parkinson's Disease Detection from Resting-State EEG Signals Using Common Spatial Pattern, Entropy, and Machine Learning Techniques". In: *Diagnostics* 12.5 (Apr. 20, 2022), p. 1033. ISSN: 2075-4418. DOI: 10.3390/diagnostics12051033. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9139946/ (visited on 05/03/2024).

[18] Mainak Jas et al. "Autoreject: Automated artifact rejection for MEG and EEG data". In: *NeuroImage* 159 (2017), pp. 417–429. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2017.06.030. URL: https://www.sciencedirect.com/science/article/pii/S1053811917305013.

[19] Adam Li et al. "MNE-ICALabel: Automatically annotating ICA components with ICLabel in Python". In: *Journal of Open Source Software* 7.76 (Aug. 26, 2022), p. 4484. ISSN: 2475-9066. DOI: 10.21105/joss.04484. URL: https://joss.theoj.org/papers/10.21105/joss.04484 (visited on 05/03/2024).

[20] Nima Bigdely-Shamlo et al. "The PREP pipeline: standardized preprocessing for large-scale EEG analysis". In: *Frontiers in Neuroinformatics* 9 (June 18, 2015). Publisher: Frontiers. ISSN: 1662-5196. DOI: 10.3389/fninf.2015.00016. URL: https://www.frontiersin.org/articles/10.3389/fninf.2015.00016 (visited on 05/03/2024).

[21] Anupreet Kaur Singh and Sridhar Krishnan. "Trends in EEG signal feature extraction applications". In: *Frontiers in Artificial Intelligence* 5 (2023). ISSN: 2624-8212. DOI: 10.3389/frai.2022.1072801. URL: https://www.frontiersin.org/articles/10.3389/frai.2022.1072801.

[22] Chetan S. Nayak and Arayamparambil C. Anilkumar. "EEG Normal Waveforms". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024. URL: http://www.ncbi.nlm.nih.gov/books/NBK539805/ (visited on 05/09/2024).

[23] Claudio Babiloni et al. "International Federation of Clinical Neurophysiology (IFCN) – EEG research workgroup: Recommendations on frequency and topographic analysis of resting state EEG rhythms.

Part 1: Applications in clinical research studies". In: *Clinical Neurophysiology* 131.1 (2020), pp. 285–307. ISSN: 1388-2457. DOI: https://doi.org/10.1016/j.clinph.2019.06.234. URL: https://www.sciencedirect.com/science/article/pii/S1388245719311642.

[24] Nikesh Bajaj. "Wavelets for EEG Analysis". In: Nov. 2020. ISBN: 978-1-83881-947-7. DOI: 10.5772/intechopen.94398.

[25] Michael J. Prerau et al. "Sleep Neurophysiological Dynamics Through the Lens of Multitaper Spectral Analysis". In: *Physiology (Bethesda, Md.)* 32.1 (Jan. 2017), pp. 60–92. ISSN: 1548-9221. DOI: 10.1152/physiol.00062.2015.

[26] Tom M. Mitchell. *Machine Learning*. McGraw-Hill series in computer science. New York: McGraw-Hill, 1997. 414 pp. ISBN: 978-0-07-042807-2.

[27] Issam El Naqa, Ruijiang Li, and Martin J. Murphy, eds. *Machine learning in radiation oncology: theory and applications*. Cham: Springer, 2015. 336 pp. ISBN: 978-3-319-18304-6.

[28] Yinhai Wang, Zhiyong Cui, and Ruimin Ke. *Machine learning for transportation research and applications*. OCLC: on1321786550. Amsterdam, Netherlands: Elsevier, 2023. 239 pp. ISBN: 978-0-323-96126-4.

[29] Qiong Liu and Ying Wu. "Supervised Learning". In: (Jan. 2012). DOI: 10.1007/978-1-4419-1428-6_451.

[30] *LR | IBM*. Mar. 18, 2024. URL: https://www.ibm.com/topics/logistic-regression (visited on 05/03/2024).

[31] ibm. *KNN. What is the k-nearest neighbors algorithm? | IBM*. Apr. 22, 2024. URL: https://www.ibm.com/topics/knn (visited on 05/03/2024).

[32] Jair Cervantes et al. "A comprehensive survey on support vector machine classification: Applications, challenges and trends". In: *Neurocomputing* 408 (2020), pp. 189–215. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.10.118. URL: https://www.sciencedirect.com/science/article/pii/S0925231220307153.

[33] *Decision Tree | IBM*. Mar. 26, 2024. URL: https://www.ibm.com/topics/decision-trees (visited on 05/03/2024).

[34] Jason Brownlee. *K-Fold Cross-Validation*. MachineLearningMastery.com. May 22, 2018. URL: https://machinelearningmastery.com/k-fold-cross-validation/ (visited on 04/24/2024).

[35] *Confusion Matrix - an overview | ScienceDirect Topics*. URL: https://www.sciencedirect.com/topics/engineering/confusion-matrix (visited on 05/03/2024).

[36] Jason Brownlee. *Bootstrap Method*. MachineLearningMastery.com. May 24, 2018. URL: https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/ (visited on 04/20/2024).

[37] Roger W. Johnson. "An Introduction to the Bootstrap". In: *Teaching Statistics* 23.2 (June 2001), pp. 49–54. ISSN: 0141-982X, 1467-9639. DOI: 10.1111/1467-9639.00050. URL: https://onlinelibrary.wiley.com/doi/10.1111/1467-9639.00050.

[38] Ryan Tibshirani. *Conformal Prediction*. UC Berkeley, 2023. URL: https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/conformal.pdf.

[39] Anastasios N. Angelopoulos and Stephen Bates. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification". In: *CoRR* abs/2107.07511 (2021). arXiv: 2107.07511. URL: https://arxiv.org/abs/2107.07511.

[40] Robert J. Barry et al. "EEG differences between eyes-closed and eyes-open resting conditions". In: *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 118.12 (Dec. 2007), pp. 2765–2773. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2007.07.028. URL: https://pubmed.ncbi.nlm.nih.gov/17911042/.

[41] Alberto Jaramillo-Jimenez et al. "Spectral features of resting-state EEG in Parkinson's Disease: A multicenter study using functional data analysis". In: *Clinical Neurophysiology* 151 (July 2023), pp. 28–40. ISSN: 13882457. DOI: 10.1016/j.clinph.2023.03.363. URL: https://linkinghub.elsevier.com/retrieve/pii/S1388245723005989 (visited on 05/15/2024).

[42] Md Fahim Anjum et al. "Linear predictive coding distinguishes spectral EEG features of Parkinson's disease". In: *Parkinsonism & Related Disorders* 79 (Oct. 2020), pp. 79–85. ISSN: 13538020. DOI: 10.1016/j.parkreldis.2020.08.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S1353802020306672 (visited on 05/15/2024).

[43] Laura Bonanni et al. "EEG comparisons in early Alzheimer's disease, dementia with Lewy bodies and Parkinson's disease with dementia patients with a 2-year follow-up". In: *Brain* 131.3 (Jan. 2008). _eprint: https://academic.oup.com/brain/article-pdf/131/3/690/13734947/awm322.pdf,

pp. 690–705. ISSN: 0006-8950. DOI: 10.1093/brain/awm322. URL: https://doi.org/10.1093/brain/awm322.

[44] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. ISSN: 1533-7928. URL: http://jmlr.org/papers/v12/pedregosa11a.html (visited on 05/03/2024).

[45] Henrik Boström. "crepes: a Python Package for Generating Conformal Regressors and Predictive Systems". In: *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*. Conformal and Probabilistic Prediction with Applications. ISSN: 2640-3498. PMLR, Aug. 30, 2022, pp. 24–41. URL: https://proceedings.mlr.press/v179/bostrom22a.html (visited on 04/20/2024).

[46] Elan D. Louis et al. "Reliability of patient completion of the historical section of the unified Parkinson's disease rating scale". In: *Movement Disorders* 11.2 (Mar. 1996), pp. 185–192. ISSN: 0885-3185, 1531-8257. DOI: 10.1002/mds.870110212. URL: https://movementdisorders.onlinelibrary.wiley.com/doi/10.1002/mds.870110212.