2022

Classification of Electropherograms Using Machine Learning for Parkinson's Disease

Soroush Dehghan
University of Denver

Classification of Electropherograms Using Machine Learning for Parkinson's Disease

Abstract

Parkinson's disease (PD) is a neurodegenerative movement disorder that progresses gradually over time. The onset of symptoms in people who are suffering from PD can vary from case to case, and it depends on the progression of the disease in each patient. The PD symptoms gradually develop and exacerbate the patient's movements throughout time. An early diagnosis of PD could improve the outcomes of treatments and could potentially delay the progression of this disorder and that makes discovering a new diagnostic method valuable. In this study, I investigate the feasibility of using a machine learning (ML) approach to classify PD patients from a healthy group. A set of plasma samples were collected from both PD patients and healthy people. Then the data were processed in a custom-designed capillary zone electrophoresis (CZE) system. CZE allows us to study metabolomics, which is the chemical processes and comprehensive analysis of small molecules in regard to metabolism within an organism such as a cell or body fluids like plasma. Metabolic profiling can demonstrate changes in the composition and therefore it can potentially reflect an underlying condition or may provide valuable information for disease diagnosis. After preprocessing the generated electropherograms or output data from the CZE system, I developed and applied various machine learning algorithms to distinguish the PD samples from the healthy samples based on the biomarkers extracted using the CZE system. Our experimental results demonstrate that there are clearly different features in two groups of samples. Therefore, it was possible to reach the classification accuracy of 94% in a very small set of samples.

Document Type

Thesis

Degree Name

M.S.

Department

Computer Engineering

First Advisor

Mohammad H. Mahoor

Second Advisor

Daniel Paredes

Third Advisor

Yun-bo Yi

Keywords

Capillary electrophoresis, Machine learning, Parkinson's disease

Subject Categories

Artificial Intelligence and Robotics | Computer Engineering | Electrical and Computer Engineering | Engineering

Classification of Electropherograms Using Machine Learning for Parkinson's Disease

A Thesis

Presented to

the Faculty of the Daniel Felix Ritchie School of Engineering and Computer Science

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Soroush Dehghan

March 2022

Advisors: Dr. Mohammad H. Mahoor and Dr. Daniel Paredes

Author: Soroush Dehghan
Title: Classification of Electropherograms Using Machine Learning for Parkinson's Disease
Advisors: Dr. Mohammad H. Mahoor and Dr. Daniel Paredes
Degree Date: March 2022

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative movement disorder that progresses gradually over time. The onset of symptoms in people who are suffering from PD can vary from case to case, and it depends on the progression of the disease in each patient. The PD symptoms gradually develop and exacerbate the patient's movements throughout time. An early diagnosis of PD could improve the outcomes of treatments and could potentially delay the progression of this disorder and that makes discovering a new diagnostic method valuable. In this study, I investigate the feasibility of using a machine learning (ML) approach to classify PD patients from a healthy group. A set of plasma samples were collected from both PD patients and healthy people. Then the data were processed in a custom-designed capillary zone electrophoresis (CZE) system. CZE allows us to study metabolomics, which is the chemical processes and comprehensive analysis of small molecules in regard to metabolism within an organism such as a cell or body fluids like plasma. Metabolic profiling can demonstrate changes in the composition and therefore it can potentially reflect an underlying condition or may provide valuable information for disease diagnosis. After preprocessing the generated electropherograms or output data from the CZE system, I developed and applied various machine learning algorithms to distinguish the PD samples from the healthy samples based on the biomarkers extracted

using the CZE system. Our experimental results demonstrate that there are clearly different features in two groups of samples. Therefore, it was possible to reach the classification accuracy of 94% in a very small set of samples.

ACKNOWLEDGMENTS

I would like to express my deep and sincere gratitude to my advisor Dr. Mohammad Mahoor and my co-advisor Dr. Daniel Paredes for giving me the opportunity of doing research on this project and for their invaluable guidance and continued support. My special thanks to Dr. Luis Hernandez for all his constructive support and tremendous help in sample preparation and data acquisition. I would also like to thank my thesis committee members, Dr. Mohammad Matin, and Dr. Yun-Bo Yi.

I am deeply grateful to my parents for their great support and encouragement throughout my research work. I sincerely appreciate my lab mates for giving me determination and motivation during this research. I want to express my gratitude to the University of Denver administrators and faculty for providing a thriving environment during my study at DU.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PD	Parkinson’s disease
NC	Normal Control
DBS	Deep Brain Stimulation
CZE	Capillary Zone Electrophoresis
CE	Capillary Electrophoresis
FITC	Fluorescein Isothiocyanate
CB	Carbonate Buffer
BF	Borate Buffer
SVMs	Support Vector Machines
RF	Random Forest
DTW	Dynamic Time Warping

CHAPTER ONE: INTRODUCTION

1. Parkinson's Disease

Parkinson's disease (PD) is a progressive neurodegenerative disorder that causes movements malfunctions. It is possible people start showing PD symptoms at an early age, but the average age is around 60. In the United States, it is estimated that the prevalence of PD reaches nearly 1 million at present. This number is predicted to reach 1.2 million cases by 2030 [1]. The symptoms develop gradually, and they progress and get worse over time.

In PD, specific neurons start to malfunction, and gradually it will lead to the death of these vital nerve cells [2]. The loss of these neurons that their presence is essential to produce a neurotransmitter called dopamine. Dopamine is responsible for controlling movement and coordination, and the decrease of its amount causes abnormal brain activity such as impaired movement. The progress of PD exacerbates dopamine production in the brain, leaving a person unable to control movement normally [3].

However, the exact cause of PD is unknown; researchers believe genetic mutations and environmental triggers could play a role.

Primary motor signs of Parkinson's disease include the following:

- Tremor of the hands, arms, legs, jaw, and face
- Bradykinesia or slowness of movement

- Rigidity of the limbs and trunk
- Postural instability or impaired balance and coordination.

There is no specific test to diagnose Parkinson's disease. Neurologists will diagnose PD based on the patient’s medical history, evaluations of signs and symptoms, and a neurological and physical examination [4], [5]. Due to the lack of a reliable test, there is always a possibility for misdiagnosis. Moreover, the onset of PD is currently determined when the motor symptoms start to emerge but there is evidence suggesting that when first motor symptoms manifest about 50% of substantia nigra dopaminergic neurons are compromised [6], [7].

Even though there is no cure for PD, there are two main types of treatment to improve the symptoms. Firstly, administrating carbidopa-levodopa can help to produce dopamine in the brain and compensate for the lack of it. Carbidopa will help to prevent the degeneration of levodopa in the bloodstream [8]. This treatment is for the early stages of PD and by the progression of the disease, levodopa loses its effectiveness. The other treatment is deep brain stimulation (DBS). DBS is an invasive method and it requires surgery. A surgeon will put electrodes in a specific area of the brain, called the subthalamic nucleus and it has proven that the stimulation can improve the symptoms significantly [9].

2. Metabolomics

The study of chemical processes and comprehensive analysis of small molecules in regard to metabolism within an organism such as a cell or body fluids like plasma refers to metabolomics. Metabolomics provides unique chemical fingerprints and profiles associated with distinguishing chemical processes [10], [11]. Metabolic profiling can be utilized to demonstrate changes in the composition of the metabolites, and therefore it can potentially reflect an underlying condition or may provide valuable insights about it. Due to recent advances, metabolomics is considered an emerging technology that delivers a powerful tool for precision medicine and provides a direct "functional readout of the physiological state" of an organism [10]. Different analytical techniques have been widely exploited in metabolomics, such as Gas Chromatography (GC), Liquid Chromatography (LC) interfaced with Mass Spectroscopy (MS) as well as Nuclear Magnetic Resonance (NMR) Spectroscopy and Capillary Electrophoresis [12].

3. Capillary Zone Electrophoresis

Capillary Zone Electrophoresis is described as a technique of detecting specific particles, sample ions, such as different molecules, proteins, peptides, and nucleic acids within a larger unknown compound in a narrow bore (25-100 micron) capillary tube filled with an electrolyte solution (buffer) [13]. The buffer solution provides conductivity through the capillary tube allowing the flow of current.

By applying high voltage across the capillary, the generated current which flows through the tube results in the separation of the injected compound. The sample solution (typically 1-20 nL) is placed at the end of the capillary tube away from the detector. The light detector, photomultiplier, is somewhere near the capillary right before it enters the buffer reservoir that is connected to the ground. The capillary tube ends from both sides are dipped into reservoirs containing electrolyte solution and high-voltage electrodes. One electrode from a cable leads to the output of the high-voltage power supply, whereas the other one is connected to the ground cable. To have the best conductivity, the electrodes usually are made of platinum [14].

The sample solution goes through a preparation process. It is combined with a biomarker, fluorescein isothiocyanate (FITC), that can show the volume of each particle. The measurements happen by finding the intensity of the glowing fluorescent, resulting from the laser when it hits the capillary. The generated plot of detector response through time is termed an electropherogram.

The system consists of different parts that are described in the following:

- Capillary: It is a long, very small tube with a length and diameter of about 60 cm and 25 microns respectively. The capillary has a yellow coating which should be removed where the laser hits to be transparent for it.
- Laser: it is a solid-state continuous waveform laser with a 488nm wavelength.
- Photomultiplier: to measure the light intensity resulting from laser hitting the capillary.

- Pump: there is a pump for conditioning the capillary and the injection of the sample.
- Valves: there are two valves that control the injection and conditioning.
- Pressure chamber: this keeps the pressure constant for the different runs of the samples.
- Buffer reservoir: this reservoir containing the buffer provides an environment for the buffer inside the capillary to be connected to the ground and keep the circuit closed for the current.
- Power supply: It is responsible for providing a high voltage of 27000 volts.

The block diagram of the CZE is depicted in Figure 1.1.

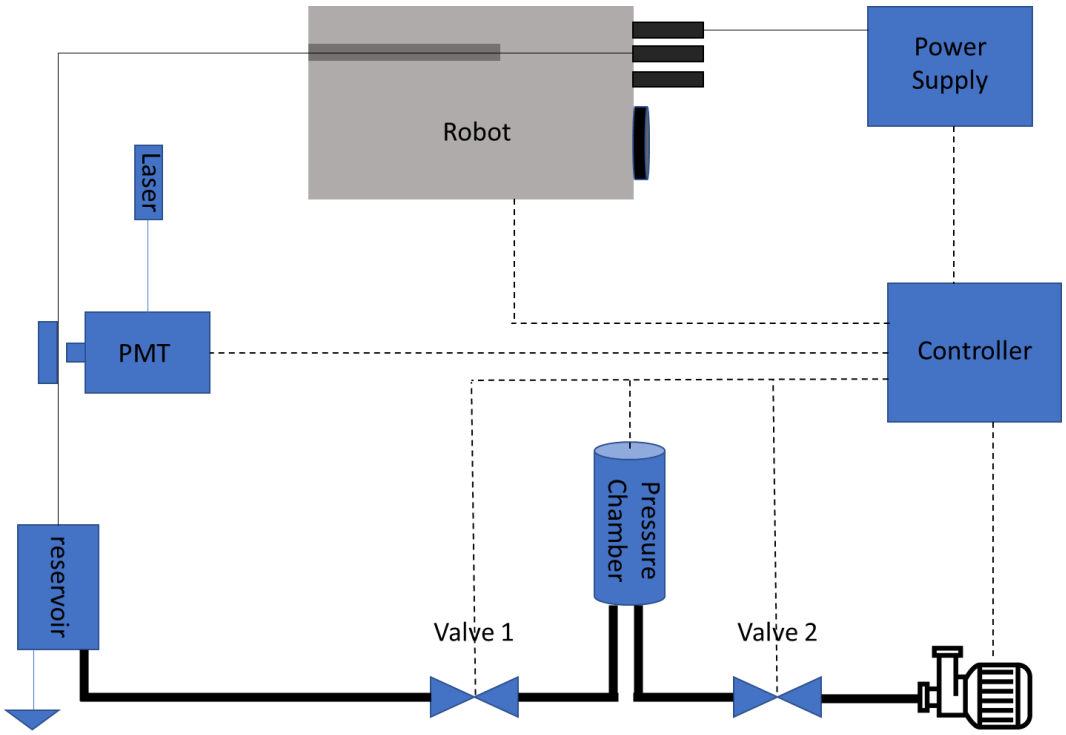

Figure 1.1 CZE block diagram

3.1 Electrophoresis

Electrophoresis refers to the movement of sample ions under the influence of an electric field [14], [15]. The applied voltage causes ions to move toward the appropriate electrode and pass through the detector. The size and number of ionic charges affect the migration rate or mobility resulting in different shapes of time series. For instance, a smaller ion moves faster than a larger one considering both have the same number of charges, or an ion with higher charges moves faster than an ion in the same size with only one charge. The ionic mobility is described as Equation 1.1

$$\mu_E = \frac{q}{6\pi}\eta r \qquad 1.1$$

Where μE is electrophoretic mobility, q number of charges, η solution viscosity, and r radius of the ion.

Therefore, in electrophoresis of solutions of different ions, the smaller particle will appear sooner in the plot of detector response.

Furthermore, the electrophoretic velocity (Equation 1.2) of the ions is related to their mobilities and the magnitude of the applied electric field.

$$v = \mu_E\, E \qquad 1.2$$

Where v is the velocity of the ion, and is E the applied electric field.

3.2 Electro-Osmotic Flow (EOF)

The applied voltage across the capillary causes a flow of solution due to the presence of an electrolyte. This solution flow pushes solute ions along the capillary toward the same direction as the flow generated by the applied voltage. This flow happens due to the ionization of the acidic silanol groups on the inside of the capillary because of being in contact with the buffer solution. At high pH, these groups are separated, resulting in a negatively charged surface followed by cations appearing near the surface to maintain electroneutrality. The applied voltage across the capillary pushes these cations towards the cathode (Figure 1.2). Due to the presence of water molecules solvating the cations, a flow of solution along the capillary occurs due to cations movement (Figure. 1.2). This effect refers to as an "electric pump."

The flow intensity is dependent on the charge on the capillary, the buffer viscosity, and the dielectric constant of the buffer, which is shown in Equation 1.3:

$$\mu_{EOF} = \frac{\varepsilon\zeta}{\eta} \qquad 1.3$$

where μEOF = "EOF mobility," η = viscosity, and ζ = Zeta potential (charge on the capillary surface).

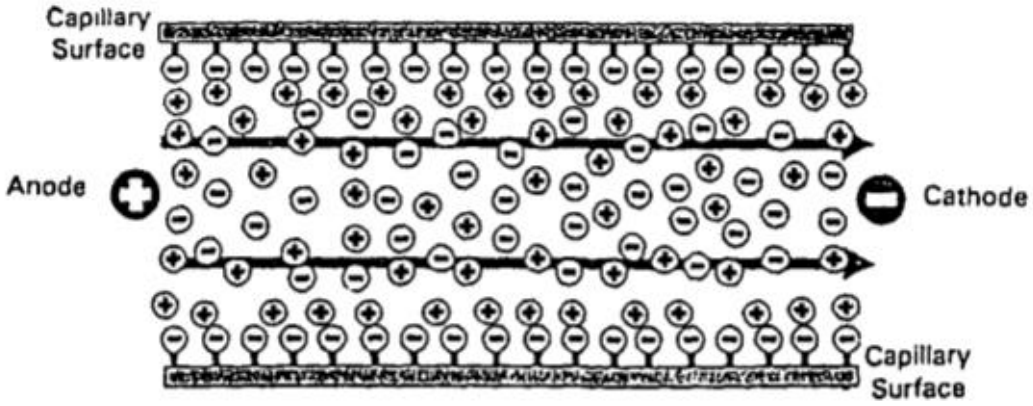

Figure 1.2 Electroendosmosis flow [14]

The ζ potential is highly affected by the ionization of the acidic silanols, resulting in the EOF intensity's dependence on electrolyte pH. Below pH 4, since the ionization is small, the EOF flow rate is neglectable. On the contrary, above pH 9, the silanols are completely ionized, leading to a strong EOF [16]. The relation of pH and EOF is depicted in Figure 1.3. The EOF level reduces with the increase in electrolyte concentration while the ζ potential is diminished.

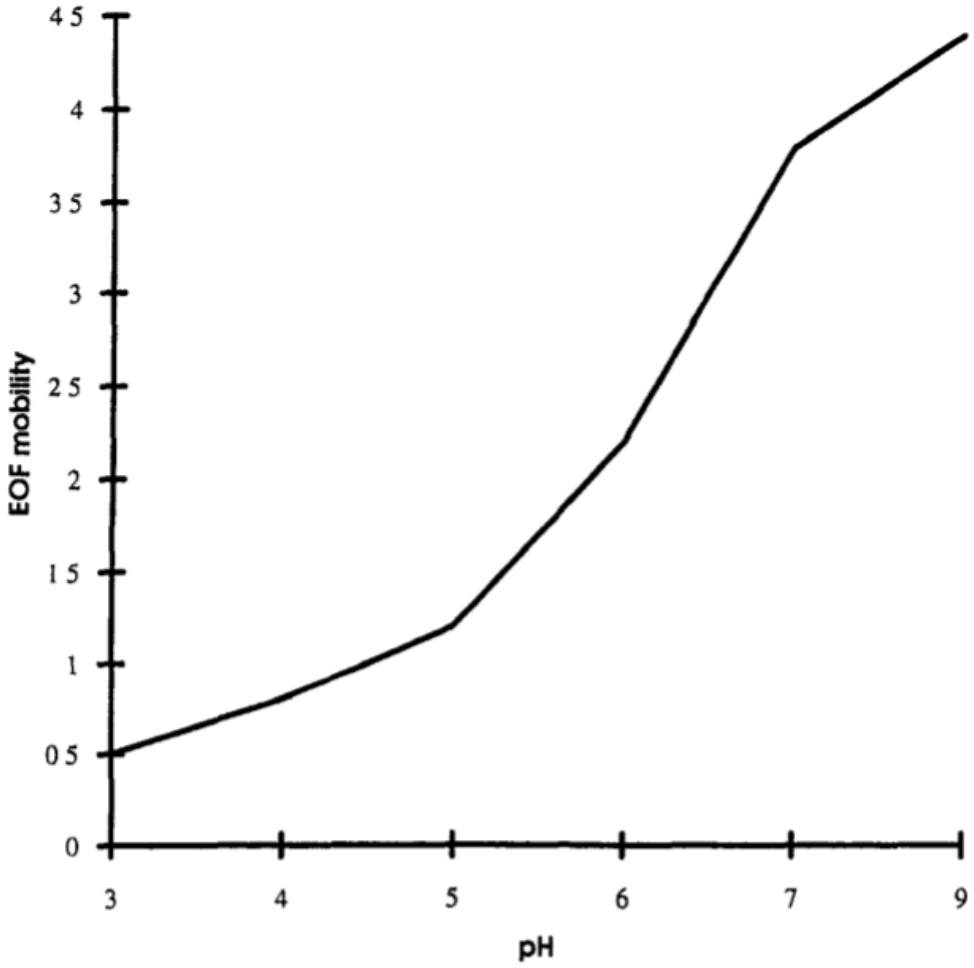

Figure 1.3 Changes of EOF based on pH [14]

The separation occurs due to the presence of EOF, which allows detection of both cations and anions within a single analysis, due to sufficient level of EOF at pH 7 and above, to push anions towards the cathode regardless of their charge. The electropherogram results from a study of a mixture of cations, neutral compounds, and anions. Each peak represents a compound and the corresponding migration time for that peak is when the compound passes through the detector.

Since the smaller anions show more resistance against the EOF, they are detected later than anions with lower mobility. Additionally, anions with higher charges move more strongly against the EOF, and therefore they will be detected later. Thus, pH is considered as the most critical operating parameter since it has significant impacts on the separation of the ionic particle by affecting both solute charged level and the EOF intensity.

Finally, the migration time of a solute is defined with regard to the mobility of the solute and its intensity. Equation 1.4 describes the apparent mobility equation, which is measured from the migration time.

$$\mu_A = \mu_E + \mu_{EOF} = \frac{lL}{tV} \qquad 1.4$$

where l = length along the capillary (cm) to detector, V = Voltage, and L = total length (cm) of the capillary.

4. Capillary Electrophoresis Background

For disease diagnosis, it is crucial to discover the associated biomarkers [17] and proteins and peptides constitute one of the essential groups of biomarkers. The various approaches that have been used to detect these biomarkers are not without difficulties since they require the handling of real samples [18].

We can find these biochemical biomarkers in tissues or biological fluids such as blood and cerebrospinal fluid (CSF). For a diagnosis purpose, the physiological process of the disease affects the selection of the biological fluid. Therefore, the patient's disease state can be defined by the concentration of the biomarkers in the collected sample.

There are different strategies reported for capillary electrophoresis (CE) analysis in diagnosis which is developed by using body fluids. By using CE, not only could we detect biomarkers but also quantize them, and this technique is highly advantageous since it provides high efficiency and resolution with a small amount of sample volume. Also, it is possible to accomplish high sensitivity when it is coupled with laser-induced fluorescence detection (LIF) [19] or mass spectrometry (MS). However, dealing with biomarker analysis by CE is not without difficulties. Firstly, the complexity of the sample matrix and then their low concentration could play an important role in this analysis.

For example, a very common biological fluid is blood containing many proteins in very different concentrations. This could cause problems due to interfering substances that could deteriorate the separation and also because of potential interaction between the targeted substance and the rest of plasma proteins leading to changing the accuracy. Proteins and peptides, additionally, have a tendency to interact with the surface of the silicone capillary affecting the resolution and sensitivity of the system. To keep the system, different functional actions must be taken in addition to the initial CE separation system.

Biological fluids containing biomarkers consist of proteins and compounds that could lead to many potential interferences due to similar behavior during the CE analysis process. Therefore, prior to running the sample, a preparation step must be performed in order to reduce or eliminate the adversarial factors that could potentially interrupt CE analysis. The preparation process not only helps to get rid of the non-interesting compounds but also to magnify the biomarker effects by increasing their concentration of them.

The most common body fluids that are used for protein analysis are plasma, serum, and cerebrospinal fluid. Protein analysis from these fluids is complicated since it requires sample cleanup before other steps. Many methods have been presented for the sample treatment to extract, isolate, filtrate, or concentrate the targeted substance from these body fluids. This study uses liquid-liquid-based extraction techniques to eliminate interfering substances in the plasma, such as centrifugation or precipitation. The steps for sample treatment are thoroughly explained in the next chapter.

5. Literature Review

The applications of artificial intelligence (AI) and machine learning (ML) have been gaining more popularity in biomedical research. Most research is focused on medical diagnosis and detection of different stages in different diseases. In spite of the fact that several studies have been carried out to explore biomarkers associated with PD, a reliable test to diagnose PD remains unknown [20]. Presently, most diagnoses are made after the onset of motor symptoms, and the emergence of symptoms occurs once the damage has progressed drastically. Despite the vast knowledge of the neuropathology of PD, an accurate test for diagnosing PD remains challenging, especially in the PD early stages [21].

Even though many studies are being conducted on developing and validating biomarkers for PD, the success rate remains far from satisfactory [22]. Thus, the discovery of diagnostic biomarkers is necessary to provide enough time to intervene by early diagnosis at the onset of the disorder and affect the efficiency of the course of treatments.

As mentioned, one of the most well-known techniques to detect and identify different chemical compounds is capillary electrophoresis (CE). This analytical tool presents a simple, cost-efficient, and fast separation method that can be exploited in the analysis and detection of different components and particles of larger unknown compounds. The applications of CE have been rapidly growing in a chemometric study, such as diagnosis of diabetes [23] or quantifying the amount of polyamines [24] which can be an important way to understand the progression of PD. Since CE can help measure the amount of a particular component within a given chemical compound, it has been used to quantify the amount of putrescine in red blood cells of PD patients. It is shown that the levels of putrescine concentration were significantly higher for PD patients compared to healthy people [24].

Additionally, it has been demonstrated that the CE can offer a classification method for different chemical compounds [25]. They propose an ML approach to classify the quality of olive oil using CZE with UV detection to generate electropherograms of three different groups of samples. This study shows that using CE makes it possible to reach a high classification accuracy for separating the various quality of olive oil.

Furthermore, in toxicology, some research has been conducted for using pattern recognition in the classification of urine samples. They used different ML algorithms for determining the type of cadmium dosage, acute or chronic. The ML models were trained based on the urine profiling that CE offers [26]. Capillary electrophoresis has been used on urine profiling for cancer patients as well [27]. They have shown that by using a neural

network, it is possible to separate normal people from patients who are suffering from different kinds of cancer.

To the best of our knowledge, there has been very little work on pattern recognition and machine learning for capillary electrophoresis analysis for PD research. As mentioned [9], PD patients could have a higher concentration of putrescine in their blood cells. Furthermore, in another study, they show that spermine levels and also spermine-spermidine ratio were significantly declined in the PD group [28]. Therefore, we can assume there are other molecules and chemical components that their concentration can be affected by PD. This study aims to investigate the application of CE further and by using the whole electropherogram instead of a few peaks in the signal, we want to dig deeper into differences in PD patients and healthy people.

In this study, before data analysis, the plasma samples went through a preparation process. Next, I run all the samples based on a designed protocol and the electropherograms of the plasma samples are acquired by using CZE. Firstly, the corresponding peaks across samples are detected and they are used to do a classification between the PD group and healthy people. Finally, the classification algorithms are developed using whole electropherograms. Several steps are taken to preprocess the data before analysis, namely noise reduction and baseline correction.

6. Thesis Goals and Impacts

As discussed earlier, there is no reliable test to determine whether someone has PD or not based on blood biomarkers and the available methods are physical exam and evaluation of brain imaging. However, there is a high chance of misdiagnosis in the mentioned methods. The main goal of this thesis is to provide a way for early and accurate diagnosis of PD patients. It is shown that electropherogram signals contain vital information about the contents of a compound. In other words, the thesis aims to investigate the correlation of patterns associated with PD and find a biomarker for this disease. This could help classify healthy people and detect anomalies in their samples in the early stage of the disease using machine learning algorithms. By developing a data analysis pipeline, I extracted features from electropherograms and used them to do the classification of the dataset into the two groups of PD and NC.

CHAPTER TWO: DATA ACQUISITION

1. Subjects

For collecting data, the samples should be prepared. The plasma sample was taken from 22 people, with 11 subjects for the PD group and 11 subjects for normal control (NC). The demographic information of each subject is shown in Table 2.1 and Table 2.2.

Table 2.1 Normal control group

Study ID	SEX	AGE
C-1	male	62
C-2	male	60
C-3	female	64
C-4	female	65
C-5	female	62
C-6	female	64
C-7	female	61
C-8	female	61
C-9	female	68
C-11	female	72
C-12	female	64

Table 2.2 PD subjects

Study ID	SEX	AGE	DISEASE DURATION since diagnosis (years)
P-1	female	70	7
P-2	male	61	10
P-3	female	67	8
P-4	male	71	5
p-6	female	61	7
p-7	male	67	8
p-8	male	69	8
p-9	male	60	7
p-10	female	68	8
p-11	female	62	6
p-12	female	71	10

The plasma samples of subjects were kept in the freezer at -80 Celsius degree temperature.

2. Sample Preparation

After thawing samples at room temperature, each plasma sample goes through a preparation process to be ready to be run in the systems. This step is essential to remove any interfering substance from the samples. A brief explanation of this process is given:

- Adding acetonitrile to samples to precipitate the proteins in plasma.
- Centrifuging the samples to separate the plates in plasma.
- Extracting the clear supernatant from centrifuged plasma.
- Adding Fluorescein isothiocyanate (FITC), which is mixed with carbonate buffer and acetone to derivatize the samples with equal volume as the samples.
- Leave the samples for 24 hours to be combined with FITC.

To prepare the dye to be used as biomarkers, the Fluorescein isothiocyanate (FITC) was mixed with acetone with a dilution factor of 1mg per 1ml. After mixing it well, the solution was added to the same amount of carbonate buffer. Next, the dye needs to be filtered before adding to derivatize the samples.

After preparing the samples, it is crucial to dilute them with the correct dilution factor to get the most applicable spectrum of samples. Five microliters of each sample were taken and added to 95 microliters of H2O for these samples. Then 25 microliters of the new solution were taken and mixed with 12.5 microliters of H20. The different dilution factors have impacts on the generated spectrum. The high concentration of FITC without proper dilution will cause saturation of the signals in most peaks, making it hard to distinguish between them.

There are two different buffers that are used, the carbonate buffer and the borate buffer. The carbonate buffer was resulted by combining 200mM sodium carbonate and 200mM sodium bicarbonate. The amount of each solution is 5 ml and 5 ml, respectively. Then 90 ml of H2O was added to make 100 ml of 20 mM carbonate buffer. And 40mM

sodium tetraborate decahydrate and 20mM sodium dodecyl sulfate was used in equal volume to make the borate buffer. The carbonate buffer was used in the dye, and then for running the samples, the capillary was filled with the borate buffer.

3. Data Acquisition

For running the samples, the system must be prepared. Before injecting the sample to the capillary, the capillary was conditioned with 0.1 mM NaOH. Then it was rinsed with water. For these processes, the capillary was washed using positive pressure by applying constant pressure on the syringes containing NaOH and H2O for two minutes. Next, the capillary was filled with a buffer that provides its current conductivity.

The protocol to run a sample is as described:

- Condition the capillary with 0.1mM NaOH for two minutes with positive pressure or 8 minutes with negative pressure.
- Rinse with H2O for two minutes with positive pressure or 8 minutes with negative pressure.
- Filled the capillary with BF for two minutes with positive pressure or 8 minutes with negative pressure.
- Inject the sample for 2 sec using the pressure chamber and the pump.
- Inject the BF for 2 sec using the pressure chamber and the pump.
- Activate the power supply a let it run for about 40 minutes.

The signal acquired by the photomultiplier represents the amount of some specific molecules within the samples. By mixing the plasma samples with FITC, the particles of the plasma get attached to the FITC. After applying high voltage across the capillary, the particles start to separate from each other due to the polarity of molecules. The negative particles go towards the cathode and the negative ones towards the anode. There are some neutral molecules that will be pushed by the Na ion existing inside the buffer since Na is moving towards the anode. This difference between polarities of molecules causes the electropherogram to have three groups of peaks. The first group has a negative charge, the second group is neutral, and the last one is positive.

When these molecules pass the laser, the fluorescent starts to glow, providing a phenomenon that the photomultiplier can catch. The intensity of this light shows the amount of FITC attached to the molecules, which can quantize the amount of the compartments.

The acquired electropherograms from two NC and PD groups samples are depicted in figures 2.1 and 2.2, respectively.

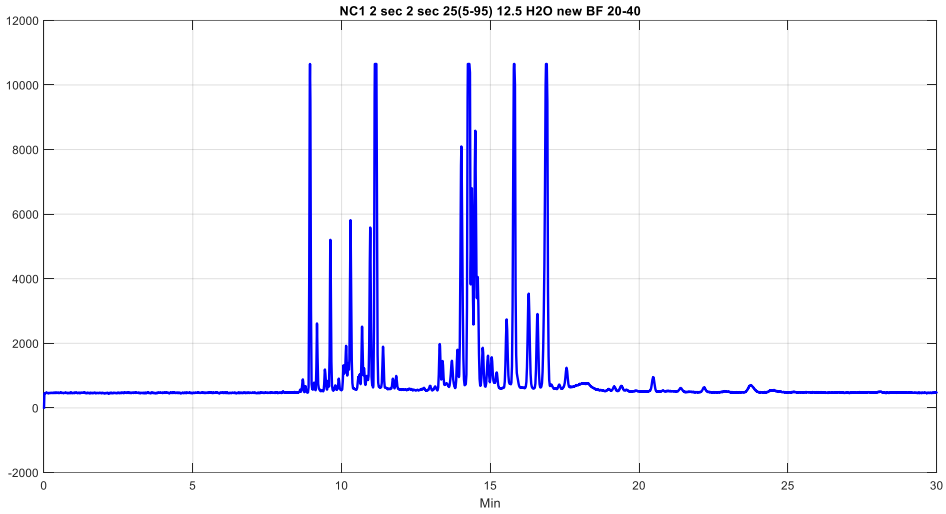

Figure 2.1 Electropherogram of a healthy person

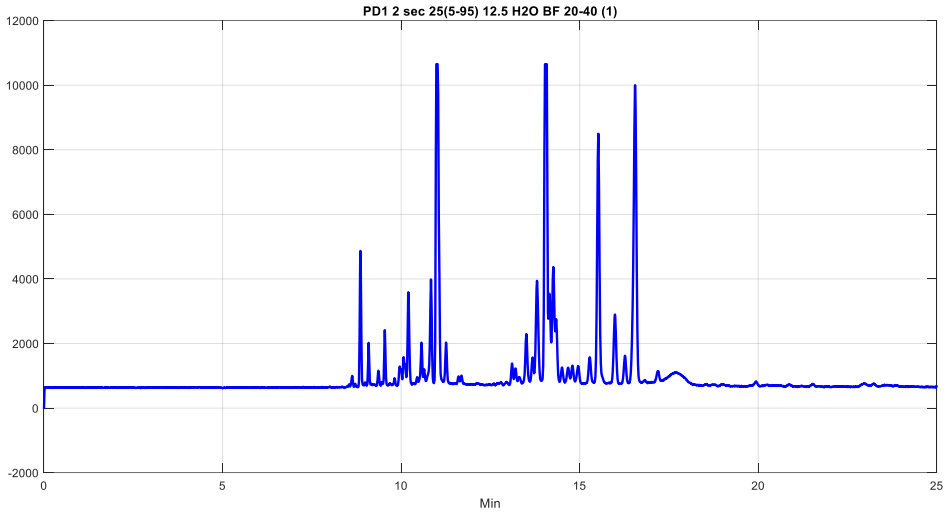

Figure 2.2 Electropherogram of a PD patient

The concentration of plasma samples in the newly prepared solution is essential to give the most detailed information about each sample. For instance, one of the samples of the PD group was rum three times with three different dilution factors. The first sample is prepared by mixing 5 ul of the original solution and 95 ul of H2O. The second one uses 25 ul of the previously described solution with 12.5 ul of H2O. For the final solution, another 25 ul of the same solution is used with the difference in the amount of water added, which is 25 ul. The spectrums of these solutions, the first solution (blue), the second solution (purple), and the last solution (green) are shown in figures 2.3 and 2.4.

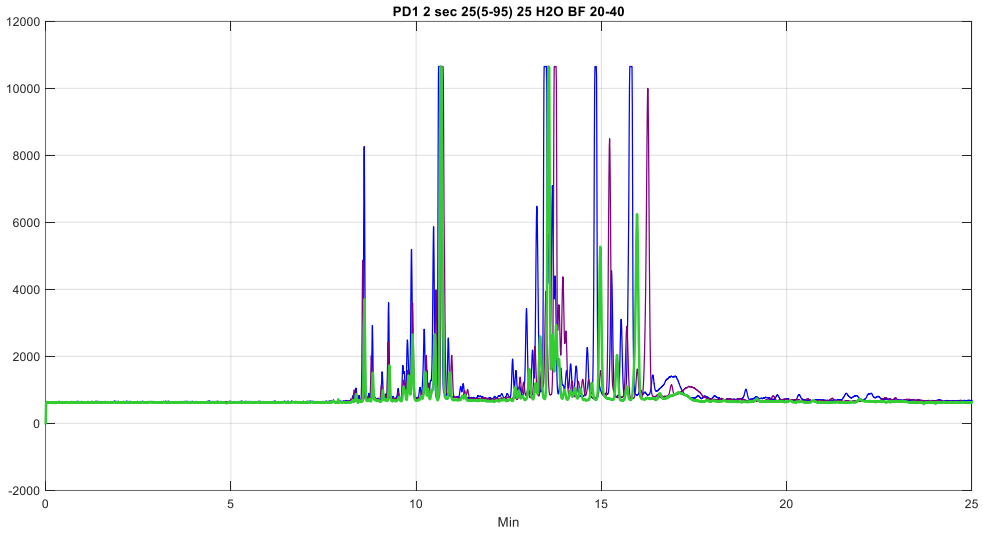

Figure 2.3 Electropherogram of a PD patient with different concentrations of FITC

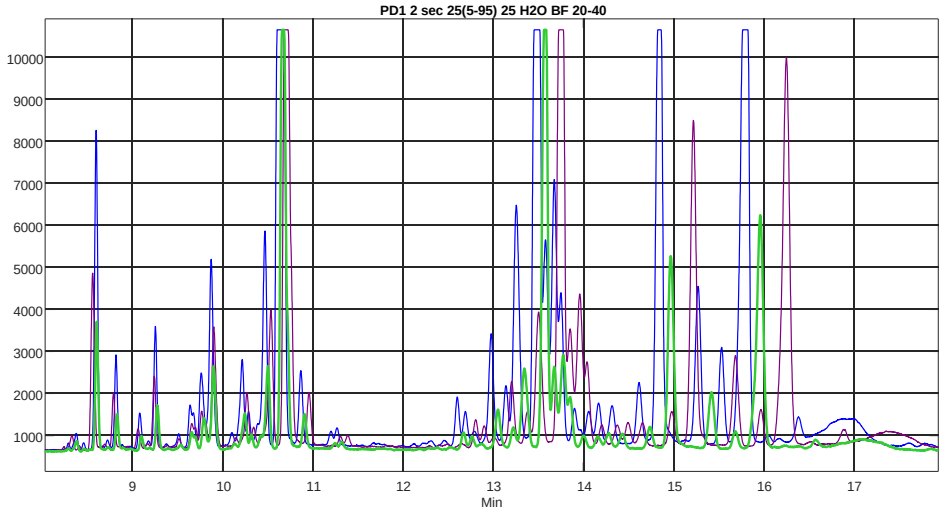

Figure 2.4 Area of interest; Electropherogram of a PD patient with different concentration of FITC

By adding more water to the samples, we can decrease the number of saturated peaks, which is important to show the differences between any other of them. However, more water can cause a decrease in the amplitude of smaller peaks. Therefore, the concentration of the second solution is considered as the optimal diluted solution.

4. Sample Spiking

Each peak represents a different molecule. To find the targeted molecule, first, the plasma sample was run alone. Next, we inject the sample for two seconds along with the injection of a specific molecule right after that. In this way, the concentration of each molecule inside the capillary drops except for the molecule that was injected separately. By comparing the newly acquired electropherogram with the spectrum of the sample, we

can identify the location of the specific molecule among all the peaks. For instance, in Figure 2.5, the spectrum for the PD sample of subject six is shown along with the spectrum of the same sample with an injection of Alanin for 2 seconds.

As can be seen in figure 2.2, at a time of about 7.2 minutes, there is only an increase in the amplitude of the peaks. Thus, this peak represents the arginine.

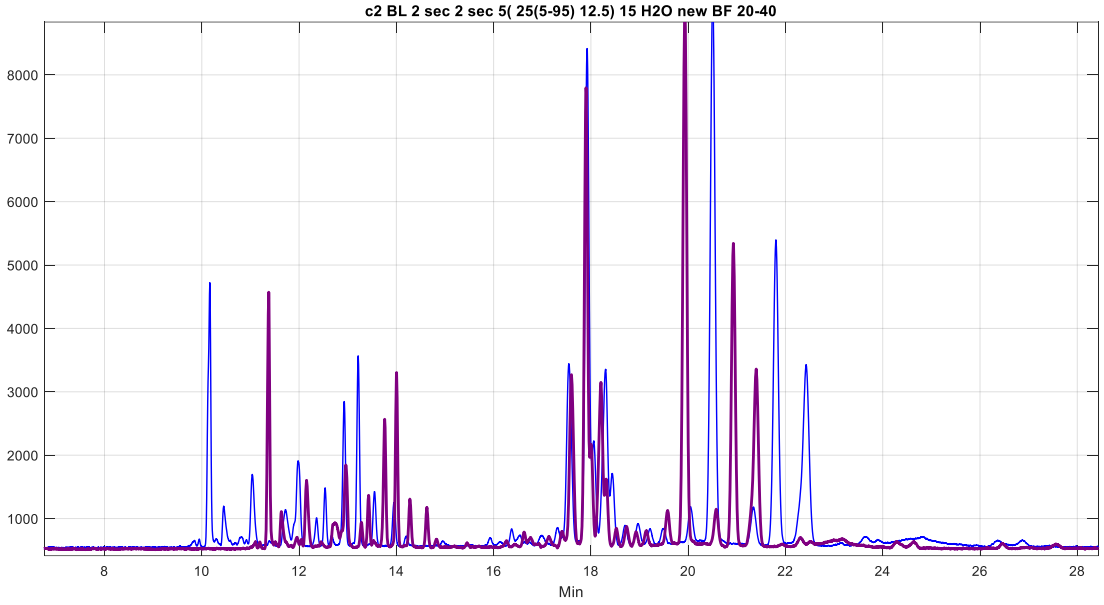

Figure 2.5 Electropherogram of a PD patient compared to the Electropherogram of the same sample with Alanin injection

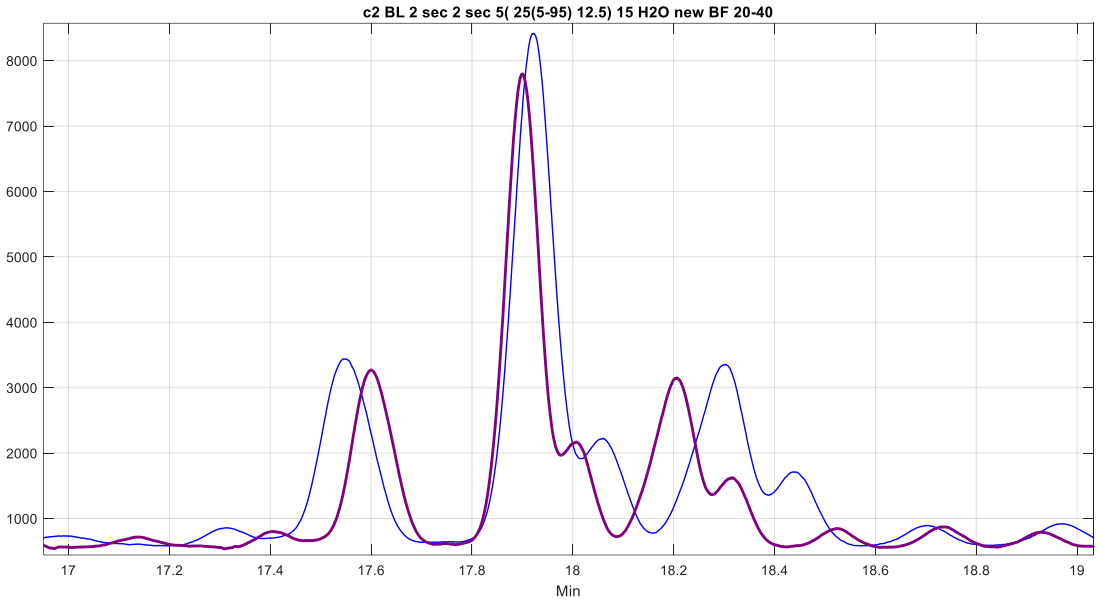

Figure 2.6 Peak associated with Alanin

This relationship is the reason why the electropherogram contains crucial information and can be utilized to diagnose distinctive diseases. Since PD has different symptoms and they can have impacts on the metabolism of the body's organs, the concentration of each molecule in the plasma could change based on the severity of the PD in the patients. In this method, by taking all common peaks into consideration, we can find the pattern associated with PD and use it for early diagnoses.

CHAPTER THREE: DATA PREPROCESSING

1. Denoising

For noise suppression, two different methods are carried out depending on the accuracy of the algorithm to preserve the vital information of the electropherograms. Firstly, noise suppression was accomplished by using wavelet transform Figure 3.1. Prior to further analysis the wavelet decomposition of the signal at level 7 with ‘Symlet4’ was computed.

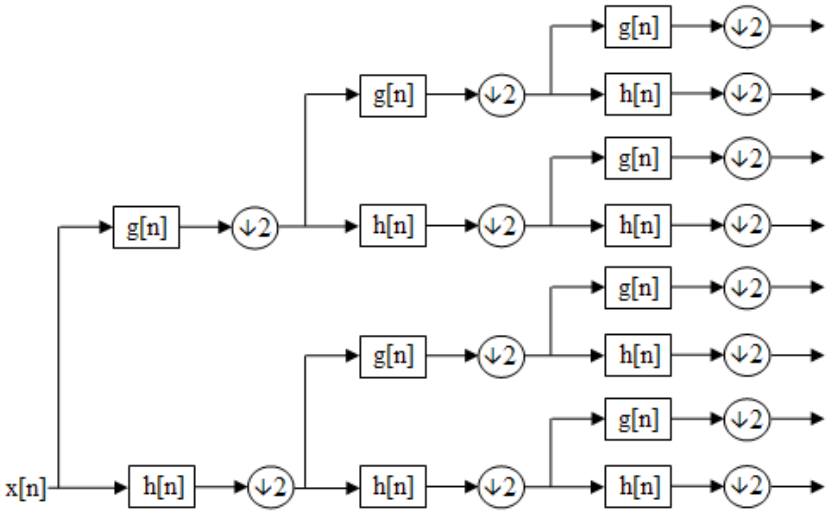

Figure 3.1 Wavelet coefficients in different level

After thresholding the detail coefficients, the denoised signal was reconstructed using the original coefficients of level 7 and the modified detail coefficients of level from 1 to 7.

Secondly, a moving average filter is used to remove the noise in the signals. The procedure is carried out by sweeping a window of a specific length through the signal and taking the average of the data points within the window. It was shown that the moving average performs better over the wavelet to keep the value of peaks in the electropherograms.

2. Baseline Correction

Due to changes in experimental conditions such as possible residue of the dye inside the capillary and erosion of the inner side of the capillary through time, the signal can drift with different experiments. To keep the shape of the electropherogram consistent, it is necessary to remove the bias in the baseline. For the baseline correction, the signal passes a low-pass filter to omit small fluctuations in order to find the prominent peaks of the signals. The valleys of the signals were determined by using the reversed signal and detection of peaks of specific height in the new signal. Then by using a cubic interpolation among the detected valleys, a baseline was drawn. Finally, the fitting curve was subtracted from the original electropherogram to correct the baseline.

3. Peak Detection

Detection of peaks in the electropherogram is important since each peak represents a molecule in the chemical compound. The peak was detected by using the first derivation of the signal and adding some restrictions to obtain the relevant ones, such as the height and minimum distance of two adjacent peaks.

In order to find the corresponding peaks, an algorithm called Dynamic Time Warping (DTW) was used. In DTW, the distance between all two pairs of peaks is calculated and, based on those pairs; these calculations result in a matrix of distances for all peaks. In this matrix, a path that has the shortest distance for all pairs of peaks is the answer to the alignment [29].

With these links that exist between each pair, the path between two corresponding peaks is known, and by assigning each peak to the corresponding point in the other signal the aligned signals are obtained.

4. Feature extraction

For the feature extraction, since the corresponding peaks across all electropherograms are found, by calculating the height and the area under each peak we have a vector of features that can be used to classify the spectrums into two groups. This process is described in more detail in the next chapter for each different dataset.

5. Software Implementation

The code for all of the above parts was implemented on MATLAB and a graphical user interface was developed to provide a user-friendly environment for manipulating the peaks.

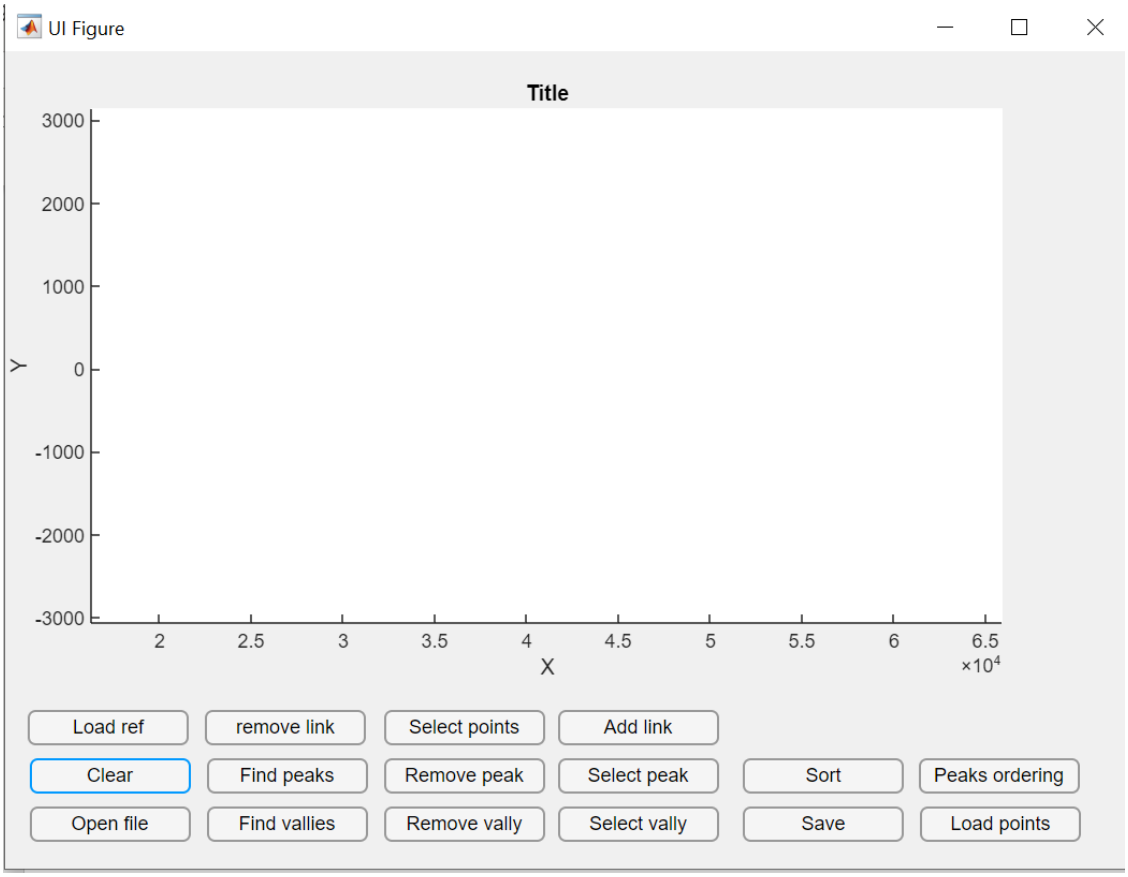

Figure 3.2 Graphical interface of the peak detection pipeline

CHAPTER FOUR: CLASSIFICATION

1. Algorithms for Binary Classification

The goal of this thesis is to classify electropherograms into two different groups. Therefore, a binary classifier is needed. Several methods are evaluated for this classification, such as support vector machine (SVM), random forest, k-nearest neighbors algorithms, and deep learning.

1.1 Support Vector Machine (SVM)

The SVM has been used widely in pattern recognition applications. SVM finds the optimal hyperplane that separates the two training classes [30]. As can be seen in figure 4.1a there are a number of potential hyperplanes. The support vectors refer to the samples of a class that is in the closest distance of the other class, and the margin is defined by the distance between the hyperplane and the support vectors. SVM chooses the hyperplane with the largest margin of separation which can be seen in Figure 4.1b.

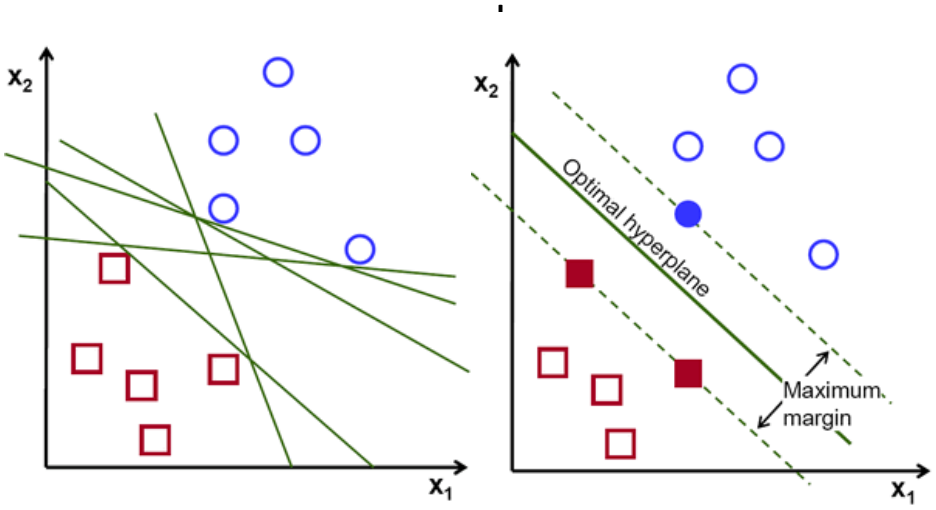

Figure 4.1 SVM chooses the maximum margin various among potential hyperplanes

Given a training set of n points with the following form

$$(x_1, y_1), \ldots, (x_n, y_n)$$

Where the y_i indicates in which classes xi belongs. In this algorithm, the goal is to find the maximum-margin hyperplane that can divide the points xi into different groups. Any hyperplane that can satisfy the following (Equation 4.1) can split the points into different classes.

$$w^T x_i - b = 0 \qquad 4.1$$

where w is the normal vector of the hyperplane.

SVM solves the following optimization problem (Eq 4.2) for the given X, Y vectors.

by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i(w^T x_i - b)\right) + \lambda\|w\|^2 \qquad 4.2$$

1.2 Random Forests

Random decision forests are an ensemble learning method for classification that operates by constructing a large number of decision trees at training time. The classification output of the model is chosen by the output class of most decision trees [31].

The RF functions are based on decision trees, a type of ML algorithm for supervised learning. In decision trees, the data is continuously split into different parts according to a certain parameter. These parameters could be interoperated as features to train the model. For training a random forest model, many decision trees will be created and the result of the majority of models determines the output class of the algorithm.

1.3 K-nearest Neighbors

K-nearest neighbors (KNN) is a type of classification where the algorithm relies on the distance of each set of points from each group [32]. In this classification algorithm, k is a user-defined constant which determines the number of training samples closest to the query point. An unlabeled vector (a query point) is classified by assigning the label of the most frequent class among the group of k training samples nearest to that vector.

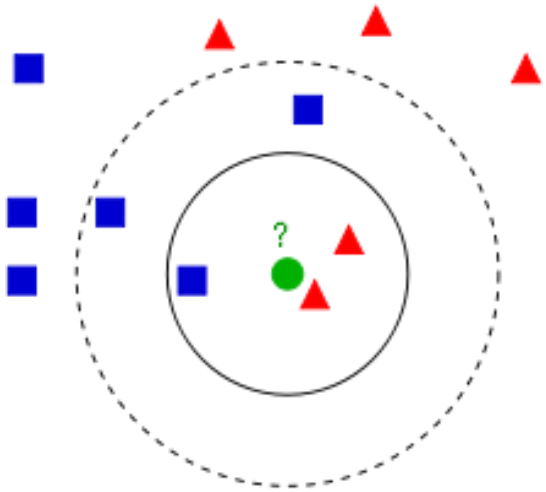

Figure 4.2 KNN classification for two groups of samples

1.4 Deep Learning for the Classification

Most of the previous research on capillary electrophoresis analysis was based on using traditional machine learning to classify data into different classes. The problem with those approaches is the need for aligning signals prior to starting using classification algorithms. In this method, the alignment process was disregarded by training a deep model that is not sensitive to the timing of electropherograms.

This model, first, was developed on a different dataset. The electropherograms of this dataset are acquired by running the eye chamber fluid. First, Pancreatic islets of diabetic subjects were implanted in the eye chamber of mice. Next, the fluid inside the eye chamber was extracted and was derivatized similar to parkinsonian subjects.

The samples belong to two groups. The first group describes the mice that rejected the islets, and the other group belongs to mice that tolerated the implants. A sample of the

data for each group is shown in Figure 4.3. Ceballos et al. performed the classification based on a machine learning approach [33]. The goal of developing this new deep learning model is to be able to classify electropherograms correctly without the need for alignment.

To develop this new model, a convolution neural network (CNN) is connected to the recurrent neural network (RNN), providing features for long short-term memory networks (LSTM) [34], [35]. Since this network is not stable during different runtimes, five similar architectures of the CNN-LSTM network are aggregated. Finally, a fully connected network is used to convert the outputs of these five networks into two classes. The neural network is depicted in Figure 4.4.

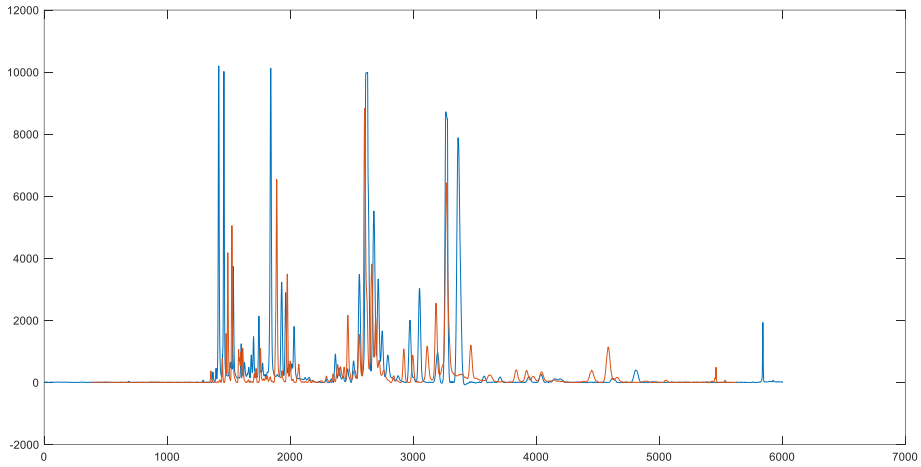

Figure 4.3 The electropherograms of rejected (orange) and tolerant (blue) group

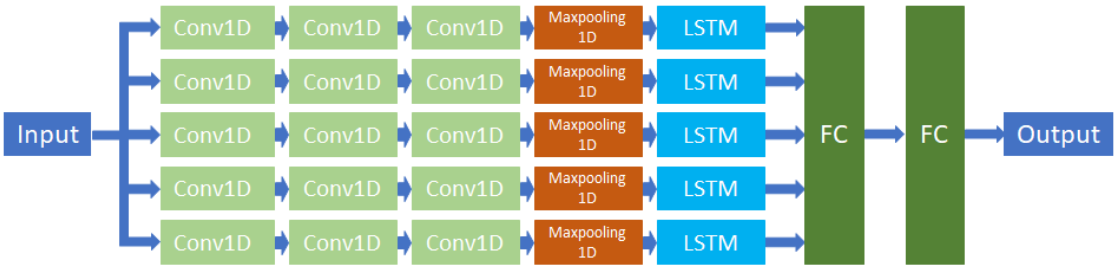

Figure 4.4 Schematic of the deep learning model

Wavelet transform is also used to transfer the input signal to a shorter length in order to be fed to the network. The coefficients of the 5th level of 'Symlet4' are used for wavelet transformation.

For making the DL model insensitive to the timing of the input signal, data augmentation was used. In this method, several electropherogram signals are generated by shifting the signal to the right and left. Using those newly generated signals to train the DL model means there is no need to align the electropherograms before feeding them to the model. The accuracy that was achieved for the original samples from the eye chamber fluids by using this new dl model was 91%.

2. Results

2.1 Using the Integrated Peaks

A total of 30 peaks were selected based on the observation of all electropherograms and the height and the area under the curve of each peak were calculated. Firstly, All the extracted features are used in the classification problem. Furthermore, to improve the result

of machine learning algorithms, based on the mean and the standard deviation of the height and the area under each peak, 19 peaks were handpicked and were used as features for training the ML algorithm. The features are selected based on the statistical hypothesis test (t-test).

The calculation of height and area under each peak results in the first set of features. Therefore, a total number of 49 features are exploited to be used in the machine learning algorithms. The highest accuracy that was reached using all integrated peaks was 68%. This was achieved by using SVM algorithms as the classifier with linear kernel. To improve the results, a feature reduction algorithm, principal components analysis (PCA), is performed on the features, resulting in an accuracy of 72 percent with the same ML algorithms. Furthermore, based on the p-value of features in each group, 19 peaks were selected. This new approach helped to reach the accuracy of 91 percent. For classification, the three methods described earlier are used. As it is shown, the best result belongs to SVM. In each time, the model was trained based on 11-fold cross-validation 100 times. The confusion matrix, accuracy, sensitivity, and specificity for SVM, Random Forest, and KNN are reported respectively.

The c value in SVM is 3, and among the kernels, the linear kernel offered the most promising results. The confusion matrix is shown in Table 4.1.

Table 4.1 Confusion matrix of SVM on selected peaks

	PD	NC
PD (predicted)	465	85
NC (predicted)	13	537

$$\text{Accuracy: } \frac{TP+TN}{\text{TP+FP+TN+FN}} = 0.918 \qquad 4.3$$

$$\text{Sensitivity: } \frac{TP}{\text{TP+FN}} = 0.97 \qquad 4.4$$

$$\text{Specificity: } \frac{TN}{\text{FP+TN}} = 0.86 \qquad 4.5$$

For the random forest, the number of estimators is 100 and entropy was used for the criterion, and the confusion matrix is described in Table 4.2 and the result for KNN classification is depicted in Table 4.3.

Table 4.2 Confusion matrix of RF on selected peaks

	PD	NC
PD (predicted)	369	181
NC (predicted)	121	429
Accuracy: 0.73 Sensitivity: 0.75 Specificity: 0.70		

Table 4.3 Confusion matrix of KNN on selected peaks

	PD	NC
PD (predicted)	300	250
NC (predicted)	102	448
Accuracy: 0.68 Sensitivity: 0.74 Specificity: 0.64		

2.2 Using the Whole Electropherogram

It was realized that due to differences in the speed of reaction between the samples and FITC, the amplitudes of the peaks can vary across different samples. To tackle this problem, a new set of samples are derivatized. However, this time the samples were allowed to sit for 5 days in darkness. The samples went through the same sample treatment, dilution procedure, and the same running protocol.

Before starting the analysis for the newly acquired dataset, the same moving average algorithm to suppress the noise is used. Once the filtering with moving average filter is performed, the baseline correction needs to be carried out. Therefore, after the noise reduction, the valleys of the signal are detected, and they are used to draw a curve based on the cubic interpolation of these points. Next, the fitting curve is subtracted from the signal resulting in baseline correction. Due to changes in experimental uncontrolled conditions such as temperature, or current variation within different samples run-times, the peaks migration time could vary, and these time shifts lead to differences in the occurrence of corresponding peaks across samples. Therefore, it is crucial to perform a signal alignment algorithm for further analysis. Dynamic time warping (DTW) is used to align the signals. DTW is a well-known approach that widely has been used to deal with different retention times in CE analysis [36]. The DTW algorithm performs the alignment by calculating the distance between a pair of points in two given time series. To perform DTW, one sample is needed to be the reference for the alignment of the rest of the signals. Once the signal alignment is performed (Figure 4.5), due to the high resolution of CE and since the whole electropherogram is utilized for the classification, I investigated the effect of

reducing the dimension of data on the performance of the models. Principle component analysis (PCA) is used to extract the most relevant features of the signals in the dataset [37]. Finally, the random forest algorithm and SVM are used to classify the outcome of PCA and the preprocessed signals without dimension reduction into two groups of PD and NC.

Due to the small number of samples, the k-fold cross-validation procedure is used to reach the most generalized result. The value for k is 11, and therefore in each observation, 10% of the samples belong to the validation set. To apply the DTW method on the dataset for alignment, one electropherogram is randomly selected as the reference. For the PCA, I assigned a variance coefficient of 99% and it is applied on noise reduced, baseline corrected, and aligned electropherograms. Once the PCA is carried out, the dataset is transferred to 19 principal components. As can be seen (Figure 4.6), there are considerable differences in the PD and NC groups. This illustration is based on only the first three important principal components of the dataset.

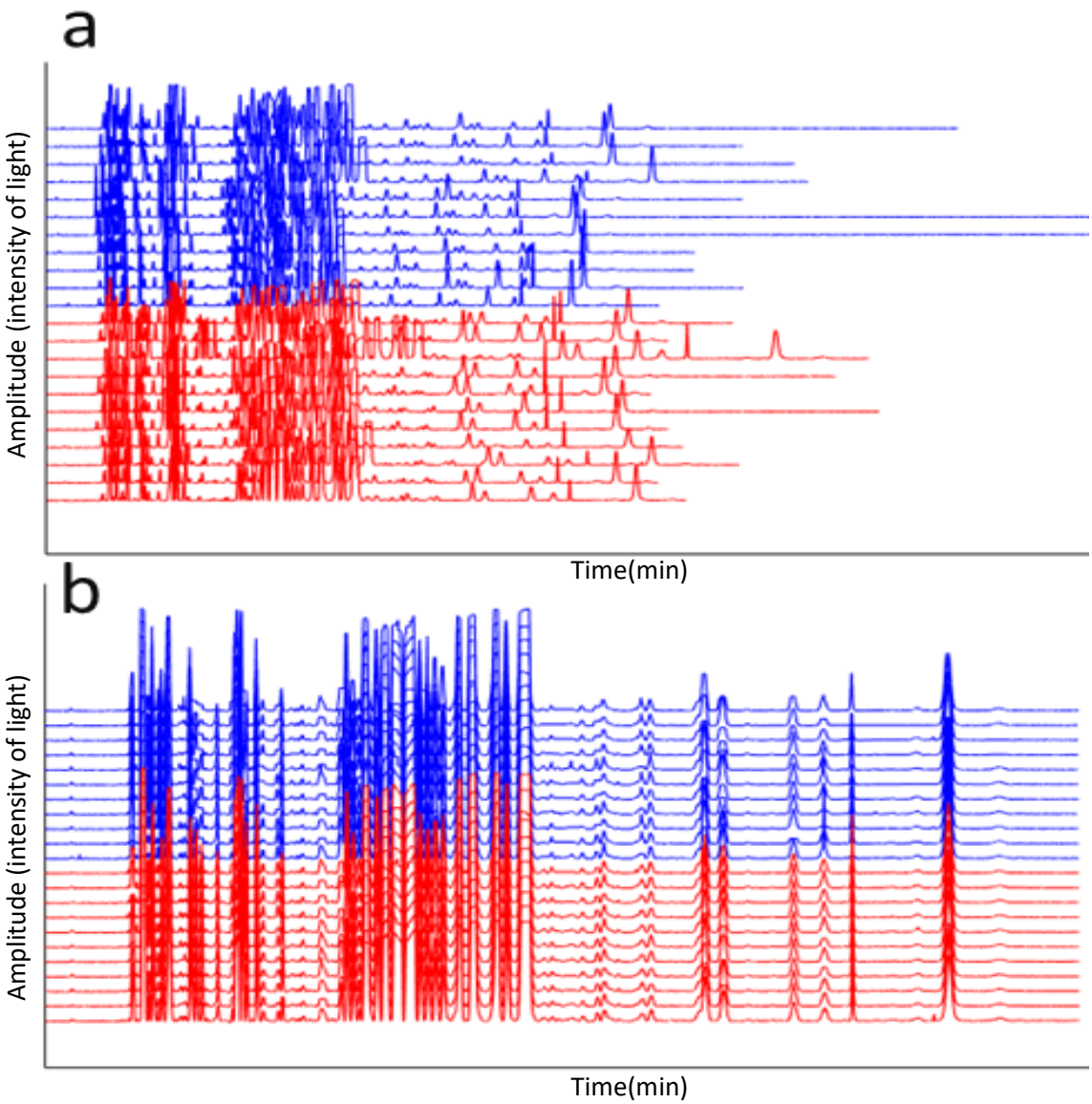

Figure 4.5 a) The samples after noise reduction and baseline correction b) DTW is used to align the electropherograms; PD (red) and NC (blue)

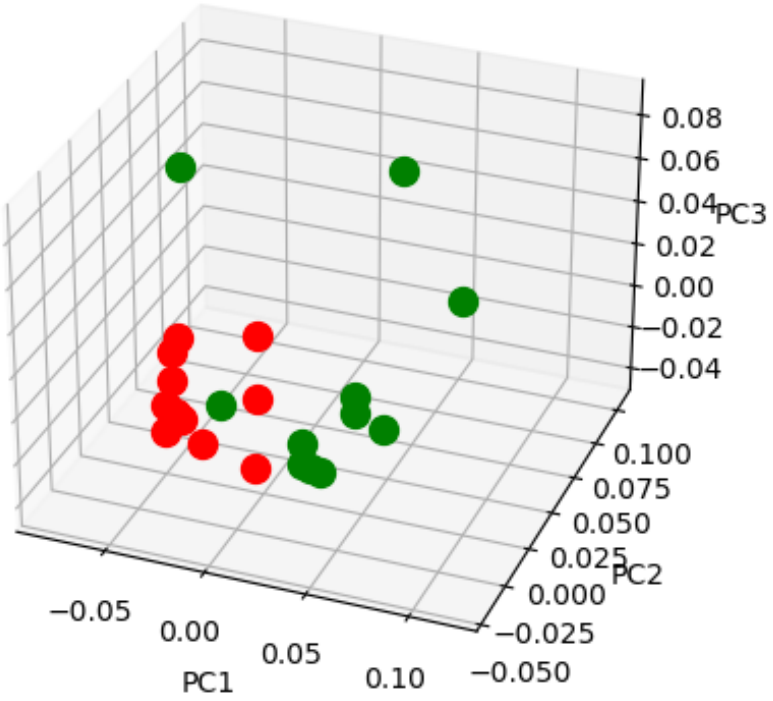

Figure 4.6 Sample of PCA implementation on the dataset based on a random reference for alignment; PD (Red) and NC (green)

The performance of the classification algorithms is evaluated using a loop within the signals. Due to the selection of different references for the DTW method, a loop is implemented, and therefore, I can sweep across the dataset and choose each of the electropherograms as a reference for one time. To consider all features, I studied the performance of different ML classifiers. Firstly, the classification is carried out based on the Random Forest (RF) algorithm. To train the RF model, I used 500 decision trees. Moreover, I deployed the Support Vector Machine (SVM) model on the dataset. For the SVM model, I used a linear kernel with a C of 100. Lastly, within the loop of different electropherogram references, the classification is performed by the RF and SVM models which are deployed to the preprocessed, aligned signals and the PCA version of the dataset.

The highest accuracy of 94 ± 5 % belongs to the RF model on noise-suppressed, baseline-corrected aligned signals. This is the average accuracy of the classification due to different electropherogram references along with the standard deviation of the classification accuracies. The results of SVM improved on the PCA version of signals, compared to the outcome of the same model without dimension reduction. On the other hand, the PCA did not improve the accuracy of the RF model and as the result, the performance of the model dropped to 85 ± 7 %. Since the alignment signal is different, the length of the model input can vary within the loop. The average length of signals is 61034 and the average number of features after downsampling and alignment is 6129. The results of this analysis are thoroughly illustrated in Table 4.4. the values are shown based on the average of different performances with regard to different DTW references and the standard deviation of them. The accuracy for each model with different DTW references is calculated by 11-fold cross validation. For the implementation of the data analysis pipeline, I used Python and Scikit-learn machine learning package.

Table 4.4 Results of different models for electropherogram classification

Algorithm	Accuracy	Validity
RF	94.00 ± 5.19 %	Sensitivity: 94.84 ± 7.20 % Specificity: 94.22 ± 5.73 %
PCA-RF	85.12 ± 7.27 %	Sensitivity: 86.24 ± 8.88 % Specificity: 85.14 ± 8.32 %
SVM	91.32 ± 8.43 %	Sensitivity: 92.35 ± 9.27 % Specificity: 90.89 ± 8.86 %
PCA-SVM	93.13 ± 5.77 %	Sensitivity: 94.93 ± 5.40 % Specificity: 92.05 ± 7.32 %

The deep learning model that was originally developed for a different dataset consisting of electropherograms of eye chamber fluids is used to classify the data. The same approach for data augmentation was used. However, the accuracy achieved was not satisfactory. Several steps were taken to tackle the low accuracy such as different augmentation methods and changing the architecture of the original model, but it did not improve the results.

3. Conclusion and Discussion

This research shows that using electropherogram signals is a reliable method to classify PD patients from healthy people. The best result was achieved using the RF on the whole preprocessed signals for classification of the electropherograms of the PD group from normal control with an average accuracy of 94 percent. As it was shown, by using a wider number of peaks, higher accuracy can be achieved since it will help with the adequate number of features on the model. However, every person has a unique composition of different components in their blood which reflects their distinctive metabolism. Therefore, this would lead to having a unique electropherogram for each person. This uniqueness is the reason why we see the variation of the ML model accuracy with different alignments. There are some molecules that do not exist among all cases that can be the cause of electropherograms misalignment, and it has impacts on the accuracy of the machine learning model. Nevertheless, it has been shown that the highest average classification accuracy across samples is 94% which proves the hypothesis of this study about the potential use of CE as a new diagnostic tool in precision medicine. Moreover, this

classification (PD vs NC) was accomplished without identifying any of the molecules in the electropherogram, and therefore, the overall pattern of the electropherograms is only important.

Moreover, this study indicates that even by using a few selected numbers of peaks, we can distinguish PD samples from the healthy group with an SVM model and reach an accuracy of 91%. This suggests that if we only use targeted peaks that can be identified by spiking, we can reach a reliable accuracy and limit the variation that is caused due to misalignment. It is demonstrated that the RF algorithm does a better classification when we have a larger number of features due to the model characteristic in building a group of decision trees compared to SVM which is more applicable for a fewer number of features.

This study demonstrates the feasibility of using plasma samples as a testing method to diagnose PD. This would help the discovery of PD biomarkers because a chemical element is reflected in a peak in capillary electrophoresis and by identifying this peak in CE of PD patients, we can find the indicator of Parkinson's disease. This is not just for an early and accurate diagnosis since it can also play a role in prognosis, by administrating that element to the patient in order to maintain a healthy state of metabolism.

4. Future work

For future study, an internal standard can be added to the samples while they are being derivatized. This can help the analysis of electropherograms in two ways. First, one of the problems with the current approach is that the injection system of the CZE can withdraw a different number of samples at different times. An internal standard such as fluorescence can be used to scale the electropherogram across all the samples. Since fluorescence does not interact with any of the molecules inside the samples, its intensity in the detector signal must stay the same assuming there is a similar amount of fluorescence in all derivatized samples. Furthermore, an internal standard (IS) can help in aligning the electropherograms together. It can be exploited to break down the electropherogram into two parts helping the alignment procedure to find better the corresponding peaks in two signals.

Another important issue that needs to be considered is the fact that similar diseases that can have common symptoms such as essential tremor (ET) can be misdiagnosed and including samples from ET patients to the future work will be helpful for building a generalized model.

Moreover, identifying the peaks existing in the electropherogram will be very helpful for both classification problems and also determining the biomarkers that are associated with Parkinson's disease. Finding the biological components that affect the PD condition is important since changing the amount of that components to the normal level could potentially affect the prognosis of Parkinson's disease.

Furthermore, to improve the accuracy of the capillary electrophoresis system, we need to provide an environment for the system to maintain the temperature inside and outside the capillary to avoid air bubbles. Also, it is important to build a platform where the laser hits the capillary so there will be no need for the alignment of the capillary with the laser light and the amplitude of the signals remains persistent with different times of running samples.

REFERENCES

[1] C. Marras *et al.*, "Prevalence of Parkinson's disease across North America," *NPJ Parkinson's disease*, vol. 4, no. 1, pp. 1–7, 2018.

[2] J. M. Fearnley and A. J. Lees, "Ageing and Parkinson's disease: substantia nigra regional selectivity," *Brain*, vol. 114, no. 5, pp. 2283–2301, 1991.

[3] C. R. Freed *et al.*, "Transplantation of embryonic dopamine neurons for severe Parkinson's disease," *New England Journal of Medicine*, vol. 344, no. 10, pp. 710–719, 2001.

[4] L. M. Chahine and M. B. Stern, "Diagnostic markers for Parkinson's disease," *Current opinion in neurology*, vol. 24, no. 4, pp. 309–317, 2011.

[5] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of neurology, neurosurgery & psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[6] A. M. Lozano and A. E. Lang, "Pallidotomy for Parkinson's disease," *Neurosurgery clinics of North America*, vol. 9, no. 2, pp. 325–336, 1998.

[7] A. Dagher, "Functional imaging in Parkinson's disease," in *Seminars in Neurology*, 2001, vol. 21, no. 01, pp. 23–32.

[8] S. Ovallath and B. Sulthana, "Levodopa: History and therapeutic applications," *Annals of Indian Academy of Neurology*, vol. 20, no. 3, p. 185, 2017.

[9] A. L. Benabid, “Deep brain stimulation for Parkinson’s disease,” *Current opinion in neurobiology*, vol. 13, no. 6, pp. 696–706, 2003.

[10] U. Roessner and J. Bowne, “What is metabolomics all about?,” *Biotechniques*, vol. 46, no. 5, pp. 363–365, 2009.

[11] C. B. Clish, “Metabolomics: an emerging but powerful tool for precision medicine,” *Molecular Case Studies*, vol. 1, no. 1, p. a000588, 2015.

[12] K. Segers, S. Declerck, D. Mangelings, Y. vander Heyden, and A. van Eeckhaut, “Analytical techniques for metabolomic studies: a review,” *Bioanalysis*, vol. 11, no. 24, pp. 2297–2318, 2019.

[13] W. Beck and H. Engelhardt, “Capillary electrophoresis of organic and inorganic cations with indirect UV detection,” *Chromatographia*, vol. 33, no. 7, pp. 313–316, 1992.

[14] K. D. Altria, “Fundamentals of capillary electrophoresis theory,” in *Capillary Electrophoresis Guidebook*, Springer, 1996, pp. 3–13.

[15] N. Volpi and F. Maccari, *Capillary electrophoresis of biomolecules: methods and protocols*. Springer, 2013.

[16] K. D. Altria and C. F. Simpson, “High voltage capillary zone electrophoresis: Operating parameters effects on electroendosmotic flows and electrophoretic mobilities,” *Chromatographia*, vol. 24, no. 1, pp. 527–532, 1987.

[17] B. D. W. Group *et al.*, “Biomarkers and surrogate endpoints: preferred definitions and conceptual framework,” *Clinical pharmacology & therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[18] K. Mesbah, R. Verpillot, J. B. Falmagne, and M. Taverna, “Contribution of CE to the analysis of protein or peptide biomarkers,” *Capillary Electrophoresis of Biomolecules*, pp. 167–190, 2013.

[19] O. O. Dada, B. J. Huge, and N. J. Dovichi, “Simplified sheath flow cuvette design for ultrasensitive laser induced fluorescence detection in capillary electrophoresis,” *Analyst*, vol. 137, no. 13, pp. 3099–3101, 2012.

[20] D. B. Miller and J. P. O’Callaghan, “Biomarkers of Parkinson’s disease: present and future,” *Metabolism*, vol. 64, no. 3, pp. S40–S46, 2015.

[21] E. Tolosa, A. Garrido, S. W. Scholz, and W. Poewe, “Challenges in the diagnosis of Parkinson’s disease,” *The Lancet Neurology*, vol. 20, no. 5, pp. 385–397, 2021.

[22] K. M. Prakash and E.-K. Tan, “Development of Parkinson’s disease biomarkers,” *Expert review of neurotherapeutics*, vol. 10, no. 12, pp. 1811–1825, 2010.

[23] G. A. Ceballos, L. F. Hernandez, D. Paredes, L. R. Betancourt, and M. H. Abdulreda, “A machine learning approach to predict pancreatic islet grafts rejection versus tolerance,” *PloS one*, vol. 15, no. 11, p. e0241925, 2020.

[24] L. Betancourt *et al.*, “Micellar electrokinetic chromatography with laser induced fluorescence detection shows increase of putrescine in erythrocytes of Parkinson’s disease patients,” *Journal of Chromatography B,* vol. 1081, pp. 51–57, 2018.

[25] N. Arroyo-Manzanares, F. Gabriel, A. Carpio, and L. Arce, “Use of whole electrophoretic profile and chemometric tools for the differentiation of three olive oil qualities,” *Talanta*, vol. 197, pp. 175–180, 2019.

[26] S. Zomer, C. Guillo, R. G. Brereton, and M. Hanna-Brown, “Toxicological classification of urine samples using pattern recognition techniques and capillary electrophoresis,” *Analytical and bioanalytical chemistry*, vol. 378, no. 8, pp. 2008–2020, 2004.

[27] R. Zhao, G. Xu, B. Yue, H. M. Liebich, and Y. Zhang, “Artificial neural network classification based on capillary electrophoresis of urinary nucleosides for the clinical diagnosis of tumors,” *Journal of Chromatography A*, vol. 828, no. 1–2, pp. 489–496, 1998.

[28] S. Saiki *et al.*, “A metabolic profile of polyamines in parkinson disease: A promising biomarker,” *Annals of neurology*, vol. 86, no. 2, pp. 251–263, 2019.

[29] S. Salvador and P. Chan, “Toward accurate dynamic time warping in linear time and space,” *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[30] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[31] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[32] N. S. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[33] G. A. Ceballos, L. F. Hernandez, D. Paredes, L. R. Betancourt, and M. H. Abdulreda, “A machine learning approach to predict pancreatic islet grafts rejection versus tolerance,” *PloS one*, vol. 15, no. 11, p. e0241925, 2020.

[34] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in *2017 international conference on engineering and technology (ICET)*, 2017, pp. 1–6.

[35] A. Sherstinsky, “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network,” *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[36] F. Karabiber, “An automated signal alignment algorithm based on dynamic time warping for capillary electrophoresis data,” *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 21, no. 3, pp. 851–863, 2013.

[37] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1–3, pp. 37–52, 1987.