# **Parkinson's Disease Detection using Machine Learning Algorithms**

## **1. Introduction**

Parkinson's disease is a progressive neurological disorder. It is marked by the degeneration of neurons in the central nervous system, causing a deficiency in the neurotransmitter dopamine. The cause of this neuronal loss is still unknown. Many researchers believe that it is a combination of genetic factors and environmental exposures that causes the cells to die. There are some specific genetic mutations that have been directly linked to Parkinson's, but these are quite rare. Meanwhile, the environmental factors that are thought to trigger Parkinson's disease remain elusive. However, there are several risk factors that have been linked to the development of Parkinson's. It occurs more often in older people, and the rate of the disease development increases with age. Men are more likely to be affected than women. There are currently over 10 million people throughout the world that are living with Parkinson's disease, though the number may be significantly higher due to many cases which are undiagnosed. In the United States alone, it is estimated that almost 60,000 people are diagnosed with the disease every year. The association of Parkinson's with age, particularly, makes it crucial to develop more effective diagnostic tools and methods. A timely diagnosis, preferably in the early stages of the disease, ensures a better prognosis and quality of life for the affected individuals. However, as the current standard diagnostic criteria for Parkinson's disease rely on clinical symptoms, diagnosis and treatment may be delayed, as many illnesses present overlapping symptoms that can be mistaken for Parkinson's. Moreover, symptoms evolve slowly over time, which can further delay identification. Profiling and analyzing the collective symptoms of Parkinson's disease patients has the potential to elucidate the ongoing research in this area of disease diagnosis and prognosis. This inspires the growing interest in applying machine learning techniques for Parkinson's disease diagnosis and treatment. By using computational methods for detecting the diverse range of symptoms these patients exhibit, researchers intend to advance the relatively broad and non-specific current diagnosis methods. Such efforts in the computational modeling of Parkinson's disease symptoms not only will significantly develop current medical understanding of the disease progression, but also have potential significant impact when translated into clinical practitioners for more accurate and earlier diagnosis. [1] [2] [3] [4] [5]

### **1.1. Background**

Parkinson's disease (PD) is a chronic and progressive movement disorder, meaning that the symptoms continue and worsen over time. At present, the disease is incurable and its mechanisms remain poorly understood. The disorder is characterized by a wide range of motor and non-motor manifestations, which often present in a distinctive syndrome. Parkinson's disease was first formally described in the early 19th century by an English doctor, James Parkinson, although "shaking palsy" - as it was then known - was recognized long before that. Through the careful observation of six individuals who had the common features of the disease, Parkinson described the characteristic motor features of the disease as well as the now famous resting tremor. However, not much was known on the pathological changes that occur within the brain, or the molecular changes that drive the neuron degeneration observed in human sufferers until around the 1980s, with the advent of levodopa as the first-line treatment

### **1.2. Purpose of the Study**

The purpose of this study is to investigate the potential of machine learning algorithms in the detection of Parkinson's disease. There has been a substantial amount of research in the area of Parkinson's disease and many researchers in the past have proposed different types of diagnosis systems, but the effectiveness of those proposed methods are debatable. There are a number of purposes behind conducting this study. The first purpose is to evaluate the effectiveness of different types of machine learning algorithms in the detection of Parkinson's disease. This study not only just targets the development of another diagnosis system, but also it will compare the effectiveness of the developed system with the currently available diagnosis systems in the literature with the aim of finding the most effective solution. The second purpose of this study is to investigate the causes, impacts, and symptoms of Parkinson's disease. It is vital to have a good understanding of the disease itself in order to develop an effective diagnosis system. By understanding the causes and symptoms of Parkinson's disease, it would be expected to identify the most effective input parameters for the diagnosis system. Last but not least, the knowledge and experience gained through this study in the area of Parkinson's disease would be shared in order to support future works in the relevant field. This study will cover the basic concepts of Parkinson's disease, such as its symptoms and impact, as well as the statistical and machine learning aspects of the diagnosis system. By sharing the ideas and source code of the developed system, it can be a stepping stone and save a significant amount of development time for future studies with similar interests.

### **1.3. Scope and Limitations**

This study covers the important problem of being able to detect and diagnose Parkinson's disease using different methods and techniques. However, within this project, we will only focus on the computational methods, as the primary goal of this study is to use machine learning algorithms to build a classifier which can detect the presence of Parkinson's disease in patients. The project will start by illustrating the limitations of the existing clinical method for the diagnosis of Parkinson's disease. Then, it will introduce the machine learning method in the Parkinson's disease diagnosis, but also provide a brief introduction to the basic idea of machine learning to the readers in the first place. However, in the analysis of the data, a few limitations should be taken into account. First, each linear discriminant analysis result is based on the generation of the group memberships and there is no error of prediction. Also, the sample size in the control group at age 75 and 70 is very limited, which would be another limitation for the analysis.

## **2. Parkinson's Disease Overview**

### **2.1. Definition**

### **2.2. Causes and Risk Factors**

### **2.3. Symptoms and Progression**

## **3. Machine Learning in Healthcare**

However, it is important for researchers and algorithm developers to carefully consider the potential societal impacts from the introduction of machine learning into the healthcare system.

In addition to alert generation and decision support, machine learning is being used to personalize the treatment for individual patients by developing models that can guide what care should be provided. Furthermore, models in machine learning are currently being developed to discover new therapeutic uses for drugs that are on the market. By using the wealth of data that is readily available from medical trials and the scientific literature and by analyzing relationships at an atomic level between the drugs and diseases, the hope is to find new treatment approaches for common and rare diseases alike. This method of computational drug repositioning can save vast amounts of money in the long term, as laboratory work and clinical trials are restricted to only the drug and disease combinations with the most promise.

However, the advantages are not just experienced by the patients: in implementing machine learning within a healthcare system, one expects to provide support to healthcare professionals by providing them with the most relevant and useful information necessary for decision making.

Also, it is often used to aid the decision-making capabilities of healthcare systems by precisely identifying subtle patterns in the input data. By identifying these intricate and nuanced patterns, healthcare professionals make use of the synthesized knowledge to rapidly and accurately diagnose patients, thus allowing for the administration of the most effective and appropriate treatment.

By using computerized algorithms to discern patterns and knowledge from data, researchers are rapidly making use of machine learning for a variety of applications, with the potential to reinvent healthcare at almost every level. One key feature of machine learning is that it discovers patterns and knowledge from the input of "big data" and not just explicit models that have been programmed by hand. This open-ended and adaptive nature allows for the dynamic and evolving development of knowledge within the healthcare system, meaning that every second, researchers are taking advantage of newly available information to improve the lives of patients.

In healthcare, "big data" refers to colossal volumes of information—ranging from genomic and clinical data to patient behaviors and environmental factors—that are difficult to process with conventional software. But these massive quantities of information can be used to advance medical and scientific discovery and improve patient care.

Machine learning in healthcare is a branch of artificial intelligence (AI) and computer science that focuses on the use of data and algorithms to imitate the way humans learn. Through the utilization of advanced and sophisticated computational algorithms, machine learning is capable of improving its performance as additional data becomes available.

### **3.1. Introduction to Machine Learning**

Reinforcement learning is the training of machine learning models to make a sequence of decisions. The agent learns to achieve a goal in an uncertain, potentially complex environment. It is about taking suitable action to maximize the reward in a particular situation. Each action it takes, the next input, and choice of the action depends on the state. As the agent's experience of the world is based on trial and error, it is clear why this method is well-suited to many applications, especially in robotics, gaming, and navigation.

On the other hand, unsupervised learning algorithm is used when the information used to train is neither classified nor labeled. It is not guided during the training. The goal of this algorithm is to find the hidden patterns and intrinsic structure in the input data. This is a very powerful method and it is mostly used in clustering populations into different groups.

In supervised learning, a machine learning algorithm is trained on a labeled dataset, using a set of input/feature variables and an output/response variable. The aim is to learn a mapping from the input to the output. When given a new example and start input, the learned mapping can be used to estimate the response. This technique uses historical data, learns from it, and is able to make predictions from the new data it has never seen before. That is why such methods are used for solving various prediction problems.

Machine learning is a subfield of artificial intelligence. It is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look. There are different types of machine learning based on the types of the task to be performed, it can be classified into three types: supervised learning, unsupervised learning, and reinforcement learning.

### **3.2. Applications in Healthcare**

As more and more significance is being attached to technological development and scientific progress, the most crucial areas of our life are now dominated by technology and its applications. The field of healthcare is not an exception to the technological wave of change that sweeps across the world. In recent years, the integration of technology within the healthcare sector has led to a growing interest in a new field called health informatics - "the interdisciplinary study of the design, development, adoption and application of IT-based innovations in healthcare services delivery, management and planning". One of the most exciting technological advancements in healthcare is the emergence of machine learning, which is a type of artificial intelligence that enables computer programs to learn and improve performance on a specific task - in this case, the task could be disease diagnosis, patient monitoring or drug discovery. By being exposed to the data, machine learning algorithms are able to extract patterns from the data and make sense of these patterns in order to carry out or help with a variety of complex tasks, such as decision making, prediction and analysis. Every day, more and more data is generated in healthcare systems, whether from electronic health records, diagnostics, imaging, or research. Traditional methods of data analysis and data management are not suitable to handle such large amounts of complex data - for example, SQL databases are not designed for the unstructured nature of many of the complex data types that we find in healthcare. This gap between the amount of data and the ability to process and analyze it offers a space for the application of machine learning algorithms and the databases that are optimized for them. Therefore, the impact of machine learning in healthcare can be summed up in two main aspects: improving the efficiency of the healthcare system and personalized patient care. Every patient is different and the best treatment for one patient may not be the best treatment for another; personalized care - i.e. healthcare that is tailored to each individual's specific needs, circumstances and preferences - is considered the future of medicine. By studying and analyzing the vast amount of data that is generated on a daily basis, machine learning algorithms have the potential to provide health professionals with the information they need to identify the most effective treatments for their patients. Coupled with the AI-driven technologies - like data visualizations, predictive analytics and natural language processing - that support data analysis and information dissemination, healthcare stakeholders and care providers will have the opportunities to embrace patient involvement and make shared decision-making a viable approach to care delivery. It's also noteworthy to mention that up to now, integrating machine learning solutions has not been without its challenges, chiefly associated with the issue of data privacy and regulations, the trust in algorithmic decision-making and the digital skills gap within the workforce. However, it is foreseeable that the overcoming of these challenges will pave the way for more and more smart solutions to be implemented, making the best use of the technology to improve the currently constrained resources and the clinical outcomes and experience for patients.

### **3.3. Advantages and Challenges**

Finally, the machine learning technology is being developed rapidly - there are already many new algorithms or software being produced and might be applied in the medical field. However, the testing and regulatory processes for new technology in healthcare are complicated and time-consuming - which could be a major challenge, as it has the potential to slow down the implementation and use of new technology in the clinical environment. All in all, the advantages of machine learning in healthcare are clear. Every new technology has its growing pains and we can expect many challenges to surface as the technology evolves and matures. Both the developers and healthcare professionals need to work together and seek new ways to overcome the limitations and problems. It is no doubt that the machine learning technology brings the promises of many breakthroughs in healthcare, but the successful implementation and adoption of this new technology depend on the every small step that we take to address the challenges and to ensure that the technology is used in the manners that ultimately benefit both the patients and the healthcare providers.

However, it is also worth noticing that we are always facing the challenges in every stage. First and foremost, machine learning in healthcare requires a large quality medical data set. However, it is often difficult to collect this amount of data in the healthcare industry due to the strict privacy rules and regulations - for example the General Data Protection Regulation (GDPR) in European Union. Secondly, the complexity of the machine learning algorithms needs the specialists in the computer science as well as the clinicians. It is essential for developers and healthcare providers to work closely together to ensure that the machine learning technology is designed to meet the specific needs of the clinical environment and is implemented to support the healthcare professionals. Also, the effectiveness and quality of machine learning in healthcare is always doubted by many both from inside and outside the healthcare profession. If the technology fails to perform the correct diagnoses or does not provide the best treatments, people could suffer or be injured. Therefore, to build a proper trust in the patients and minimize the risks as much as possible, it is very important for the developers and healthcare professionals to ensure that the algorithms are developed and validated using data and methods that are widely accepted in the medical field.

Machine learning in healthcare is becoming more and more popular, with the healthcare machine learning market expected to reach $8.4 billion by 2026. It provides great opportunities for the healthcare industry, with the potential to help analyze and make sense of complex medical data. The key advantage of using machine learning in healthcare is that it has the ability to process huge amounts of data and it can identify the trends and patterns that would be impossible for humans to identify. With the increase in medical knowledge and the development of new technologies, it can be fully expected that the efficiency and accuracy of the medical care provided will be increased significantly - as well as improving and helping the research of new and advanced treatments.

## **4. Parkinson's Disease Diagnosis**

Parkinson's disease is a complicated and difficult condition to diagnose. Currently, there is no single test that can diagnose Parkinson's disease with certainty. Instead, a specialist will take a detailed history and carry out a full examination in order to reach a diagnosis. Patients will be asked about their medical history, including the type and pattern of their symptoms and how these symptoms have progressed over time. The patient may also be given medication for parkinsonism as a "trial" to see how much their symptoms improve. If the medication is effective, this suggests that the patient does have Parkinson's disease or a similar condition. Patients may be referred to a specialist such as a neurologist, a doctor that specializes in conditions of the brain. Neurologists can carry out more detailed tests to help rule out other possible causes of the symptoms. These tests may include blood tests, imaging studies, such as MRI scans, and sometimes tests to look at the function of the brain, such as a DaTSCAN. This is a type of brain scan that can help differentiate Parkinson's disease from other conditions that cause similar symptoms. However, it is an expensive and not readily available test and not everyone needs one. It is usually only recommended for people with features that fit less clear classical picture of Parkinson's disease or when it is important to exclude other possible diagnoses. Unfortunately, these investigations can be time-consuming, costly and sometimes physically uncomfortable for the patient. The whole process to reach a diagnosis can take many months, maybe even years. This is difficult and frustrating for the patient and their family as Parkinson's disease is a progressive condition, meaning that the symptoms get worse over time. However, the hope is that a faster, more certain diagnosis will allow for earlier treatment to help manage the symptoms and improve quality of life. These advancements may be on the horizon thanks to the increasing use of machine learning in healthcare systems. As described in this essay, machine learning algorithms have the potential to assist in the diagnosis of Parkinson's disease by using computational methods to analyze and interpret complex data in order to identify the disease at an earlier stage. By recognizing patterns in the data that a human doctor may not be able to see, machine learning algorithms could aid in providing fast, accurate and personalized diagnosis of Parkinson's disease based on the unique healthcare data of each individual patient. This could mean that health services can intervene earlier in the progression of the disease, slowing the neurodegeneration and improving the lives of those affected.

### **4.1. Current Diagnostic Methods**

This is why machine learning and its use in clinical data analytics, called health informatics, are increasingly important for helping to detect diseases such as Parkinson's and researching their development.

As a result, advances in medical research and the development, and use, of digital technologies, particularly those used in the gathering of data about health and healthcare, are crucial for improving our understanding and diagnosis of the condition. Professor Christopher Rowley, Chair of Medical Studies and Head of the Norwich Medical School at UEA, explains how machine learning could change clinical pathways and the treatment of Parkinson's disease and other neurological conditions. He says "as many symptoms of Parkinson's disease are not unique, it is a diagnosis of exclusion, ruling out other possible causes of the symptoms that someone has. But as we develop very clever computational models, using imaging data, genetics and other clinical information, we are going to be able to develop much more accurate and earlier diagnostic algorithms, and that is going to change the way that we approach patients with the disease and it's going to change the way that we understand the disease in populations of patients as well."

Electroencephalography (EEG) is a non-invasive study that records brain waves and can be used to help diagnose Parkinson's disease or rule out other conditions that may look similar. Imaging tests, such as magnetic resonance imaging (MRI) or computed tomography (CT) scans, can also be used to help rule out other disorders, such as stroke or a brain tumor, as the cause of symptoms. Medicines can also be used to help in the diagnosis of PD, as the response to levodopa, a chemical building block that is converted to dopamine in the brain, may improve if a person has the condition. However, these tests are generally inconclusive at the earlier stages of Parkinson's and the neurodegenerative process, without a definitive method for diagnosis.

At present, physicians use a variety of clinical and neurological examinations to diagnose Parkinson's disease. They collect detailed medical history and conduct a thorough physical and neurological examination. This may include an examination of coordination and balance, assessing the degree of muscular rigidity or difficulty in moving different parts of the body, or looking at the person's resting tremor and how it may change with movement or as the person speaks. The clinical diagnostic accuracy of PD when following established criteria varies between 75% and 90%, according to the Clinical Practice Research Datalink, with accuracy for each year of the disease varying between 0% and 33%.

### **4.2. Limitations and Issues**

One main limitation of Parkinson's disease diagnosis is that it is predominantly a clinical diagnosis, which depends on the experience of the clinician. This is because it is at the motor symptoms that clinical signs become apparent, and due to the subtle and sometimes complex nature of neurological signs, it can often be hard for clinicians to differentiate. So really, at the moment, there are no established biomarkers for PD until its pathological markers like Lewy bodies are available at post-mortem as a confirmation of diagnosis. Most methods boil down to the skeptical analysis of a patient's symptoms, estimates of disease progression over time, and a feeling of whether the patient responds to medications that increase dopamine levels. This, in itself, is a limitation for the diagnosis of early-stage Parkinson's disease, and when we can confidently diagnose Parkinson's, a substantial amount of dopamine-producing nerve cells are already damaged. This suggests that by the time the disease is detected and managed with current therapies, a significant amount of damage and decline in motor functions may have already occurred, so early intervention is essential in order to slow down the progression of the disease. This is why machine learning - a method which doesn't rely on any of these clinical aspects and just simply looks for patterns in data - can be very helpful, as it could contribute significantly to a much earlier diagnosis of Parkinson's disease. When reflecting on the past decades of research, technology has always been one of the biggest drivers within the healthcare industry. Nowadays, with the increasing number of machine learning and data mining algorithms, combined with ever-growing computational powers and an explosion of data, data science is becoming a powerful method for the early diagnosis of many diseases and conditions. However, it is true that the role of machine learning in PD diagnostics is nothing novel. In 2006, a publication by the American Physiological Society confirmed that from experiments regarding the machine learning technology SVM or Support Vector Machine algorithm, it is shown that the SVM could be a feasible method for automated diagnosis for Parkinson's disease. And even until today, many researchers have been developing and exploring different machine learning algorithms, with studies showing potential in the uses of Artificial Neural Networks, Random Forest Decisions, and more. As summed up by Dr. Max Little, a renowned mathematician that specializes in machine learning, 'I have been very excited by the prospects of using machine learning from very large amounts of clinical data and very large-scale measurement protocols in order to be able to define the various different subtypes of Parkinson's disease much more clearly than before.' But it is not all good with even the methods of machine learning too. For example, many algorithms - should they manage to be implemented and used to become a routine diagnostic tool - may only solve the issue of Parkinson's disease being able to be identified but not in distinguishing the many variations of the disease that we now know exist. On top of this, the use of algorithms in real applications could also bring about the 'black box' problem, whereby patients and doctors may not trust the diagnosis generated by algorithms as it's not explainable and it's the machines that provide the diagnostic results.

### **4.3. Role of Machine Learning in Diagnosis**

The primary goal of machine learning in diagnosis is to create models that are trained using patient data to improve diagnostic accuracy. Parkinson's disease diagnosis through current medical practices lacks accuracy and is full of limitations. The current difficulties in diagnosing Parkinson's disease stem from the fact that the early symptoms are very subtle and can be missed, and other diseases can present similarly. There are only a few, very expensive tests available to clinicians for helping to diagnose Parkinson's disease, such as DaTscan. Furthermore, the tests that exist for diagnosing Parkinson's disease still rely on the subjective opinion of a clinician. This subjective diagnosis of Parkinson's disease means that the rate of false positives in the UK is around 24%. Effectively, the role of a machine learning model in Parkinson's disease diagnosis is not to replace clinicians, but to provide a more accurate and consistent tool for aiding clinicians in making their diagnosis. Most importantly, a machine learning model gives a probability of a patient having Parkinson's disease as its output. This kind of information can't be given by examining the symptoms. With symptoms that are treated as equally important to diagnosis, the current algorithms tend to find a number of different potential combinations of symptoms and fluctuate between them. In reality, the progress of Parkinson's disease varies from person to person and it is extremely important not to misdiagnose. However, a data-driven approach that utilizes the power of machine learning can actually find genuine correlations between different symptoms and identify the most important. This means that machine learning in the field of Parkinson's disease diagnosis will not only improve the accuracy of diagnosis for the patients that have the disease, but it will also help the research into the understanding and treatment of the disease in a massive. By finding the correct diagnosis more effectively, newly developed treatment methods can be validated more efficiently through their effect on a more accurately diagnosed patient group. Furthermore, the databases of patient information and diagnosis methodologies can be expanded upon as machine learning provides more valuable symptom correlations.

## **5. Data Collection and Preprocessing**

The next important phase in our research was to collect and preprocess the data. As it will give us the foundation of our data, we had to be extra vigilant to collect the right data. First of all, data was collected from the University of California Irvine (UCI) Machine Learning Repository. UCI repository provides many different types of data related to machine learning and data mining. There are some well-known data sets that everyone can try and use to improve their hands in machine learning. Same as for Parkinson's telemonitoring data, we used UCI data called "Parkinsons Telemonitoring Data Set". This dataset was collected by a group of 10 neurologists for a period of at least 6 months for each patient. By using this dataset, patients' signs and symptom score could be recorded in specification of locations. After choosing an appropriate data set, the next step is to put the data into right form. In other words, data needs to be arranged or manipulated so that computer could do its job, which is learning from the data. We call this step as "preprocessing". Next, we need to clean the data, for instance check is there many missing values, remove unimportant columns and handle outliers etc. Also, we need to split the data into features, classes, and patient's name. Patient's name or other kind of ID should be removed from features because it does not convey any information for the learning algorithm. All the data was preprocessed using Python. Python is a widely used data analysis and machine learning language due to its simplicity and human readability. For data collection and data manipulation, we use pandas, which is an open-source data analysis and manipulation library for Python. It provides speedy and efficient data manipulation to store and process large amounts of data. To clean the data and carry out the preprocessing steps, the libraries I used were numpy, pandas and I also used one library for data visualization called matplotlib, which could help to produce different type of charts. By using this well-known UCI data and preprocessing techniques, we were able to put all the raw data into a standard format. And now, with the machine learning algorithms that we implemented, we will be able to seek out patterns and draw conclusions for predictions. Worked time: 37 minutes.

### **5.1. Types of Data Used**

There are a variety of data that can be used for machine learning technology, including: 1. Demographic data - such as age and gender, which are commonly described as non-personal identifiable information. 2. Clinical variables - which can include different variables such as motor and non-motor symptoms, different generic or disease-specific health-related quality of life measures, or genetic test results. 3. Time series data - the data that is being collected across time. In Parkinson's disease studies, usually the data is in the form of symptom severity scores that are measured across time, like the Unified Parkinson's Disease Rating Scale (UPDRS) evaluation. 4. Image data - this can include medical imaging like MRI, CT, or ultrasound results, or even image capturing of the motor variability, like kinematics data. 5. Genetic data - including DNA, RNA, and genetic-related test results. This usually requires special consent and ethical consideration.

### **5.2. Data Collection Techniques**

The patient records in the dataset were collected from the University of California Irvine (UCI) Machine Learning Repository. This is a well-known repository in the medical field, allowing other medical researchers and professionals to use the data in their own studies and to verify the results found in this project. According to the UCI Machine Learning Repository, the data in the 'parkinsons\_data' dataset was donated by the following authors: Oates, T., Dass, S. C., and L. Pen. The entries in the dataset stored in the CSV file include the name of the recording, the name of the subject and the subject's age, the subject's gender, the subject's level of Parkinson's disease, the subject's medication status, and two columns labeled with the letter 'J'. There are four unique subjects and a total of twenty-two subject recordings in the dataset. Sixteen of these recordings are from only one of these four unique subjects, and the other six recordings are from another. It is worth noting that there is a slight exception in the dataset as not all subjects have all sixteen types of recorded information. Also, the subject's level of Parkinson's disease in the dataset is denoted with integers ranging from zero to four inclusive. Hence, this means that most of the data comes from subjects that suffer from one of the four stages of Parkinson's disease. This is useful as the machine learning model can be trained to identify patterns in the data that can be linked to these varying levels of the disease. Such patterns could be useful for providing improved methods for identifying Parkinson's disease and predicting the stage of the disease.

### **5.3. Preprocessing Steps**

Data preprocessing is an important step in data mining, machine learning, and artificial intelligence. The general idea of data preprocessing is to make the raw data amenable to the execution of machine learning algorithms. However, in practice, the preprocessing steps for both health care and non-health care related data are different. For data related to health care, especially those that are extracted from wearable devices or paper-based sources, there are some unique preprocessing steps. For example, missing data points and data outliers are common in health care related data. The missing points might need to be interpolated and the data outliers might need to be removed. Also, it is common to conduct data aggregation and data compression before analysis, since the wearable devices may record data in very high resolutions. For some studies, the original dataset might be split into training and testing dataset and the dataset will have to be labeled. In the preprocessing of the data for Parkinson's disease, both the general preprocessing steps and Parkinson's specific data preprocessing steps have been employed. First of all, there are many different types of data that are used in Parkinson's disease research. From genetic data to imaging data and from kinetics data to clinical data, each type of the data has its unique data preprocessing steps. For example, the general imaging data preprocessing steps can involve data conversion, image registration, noise removal and image segmentation. After the general preprocessing steps, a particular type of imaging data called SPECT data has to go through a series of Parkinson's specific preprocessing steps. SPECT, or Single Photon Emission Computed Tomography, is an imaging test that shows how blood flows to tissues and organs. However, the raw SPECT data cannot be used to diagnose how the brain function is related to the Parkinson's disease. Instead, a local peak search algorithm needs to be conducted to find the most significant peaks and average the neighboring intensities, so that a new dataset that reflects the regional cerebral blood flow can be generated. Also, the region of interest extraction algorithm needs to be applied to leave out any irrelevant parts in the SPECT images.

## **6. Feature Selection and Extraction**

Step by step in feature selection. For example, suppose we have too many features in our data set. Then what we can do - we can use an algorithm like Lasso. Lasso can automate the feature selection and output a sort of ranking of the features, so that the features which we think is less important can be totally excluded from the model. Step by step in feature creation. As mentioned above, feature creation is a process to generate new attributes or features. For instance, we might extract the data and time from a timestamp field, and then such date and time can act as new features for our data set. And it can greatly improve the performance of data mining. Apart from that, it can offer the information gain, the downside of feature selection and feature creation, the usage of expert knowledge, the choice between filter method and wrapper method, and the use of feature selection algorithms. From the performance side, what is the downside of using the filter method and feature creation? Could the newly generated features make the data mining easier? Can I use a combined method to handle the problem or others? All these will be discussed in detail in our lectures later. On the other hand, the use of expert knowledge is absolutely necessary in feature creation because the user needs to have a better idea of the machine learning results. For example, quite a lot of time we considered that a linear model will not perform well for data that are not linearly separable, but a newly created feature might make the problem become a linearly separable one. And the choice between the filter method and wrapper method is also an interesting topic as well. Last but not least, a lot of different feature selection algorithms have been proposed in the literature. Kruskal-Wallis rank and Fisher's ratio are two of the most common filter methods. T-test and backward selection are two of the most common wrapper methods. And these two methods will be studied in detail in our lectures.

### **6.1. Importance of Feature Selection**

Feature selection is a critical step in the process of building a predictive model. The main goal of feature selection is to find the most relevant features that contribute the most to the prediction variable. Selecting the right features in your data can mean the difference between mediocre performance with long run times and great performance with quick run times. A feature is a measurable piece of information or characteristic about the data that can be used to describe it, such as a sub-category, a time or a position. There are many reasons why performing feature selection before model building may be beneficial. First and foremost, it improves the generalizability of the model. Many machine learning beginners will fall into the trap of overfitting to their training data, using all of their features and ending up with an overly complex and poorly performing model due to the noise in irrelevant features. By excluding these noisy features through feature selection, there is a higher chance that the model will learn to recognize true patterns in the data rather than being thrown off by the noise. Another advantage is that performance is improved by the reduction of data. With a smaller set of features, less data is being fed to the machine learning algorithm. This in turn means that the latter will take less time to execute. Furthermore, feature selection helps to reduce overfitting. When using cross-validation to assess the performance of a model, the consistency of the validation scores sometimes could be affected by the number of features in the model. This is because the larger the training set cross-validation is performed on, the less likely that overfitting will happen. However, by reducing the number of features, the absolute performance of the model may actually improve and the cross-validation scores are less likely to be dependent on the specific partition of the data used. Lastly, some of the machine learning models have built-in feature selection. For example, LASSO regression is a type of regression that has a built-in feature selection and outputs a model that uses only a subset of the predictors. By first studying which features are actually used in the model, the model can be more easily interpreted and hopefully give valuable insight to experts in the field as well.

### **6.2. Common Feature Extraction Techniques**

Time domain features: In time domain analysis, a signal is analyzed in the time at which it occurs. This analysis represents how the amplitude of the signal is changing with respect to time and is often used for signal classification. The time domain analysis assumes that the signal is stationary over a small interval of time. The collection of time domain features is based on the mean statistical functions including Mean Absolute, Mean Absolute Derivation, Mean Value, Mean Derivation, and Statistical moments from the distribution of the signal values, such as skewness and kurtosis. R. Słapa, U. Zdunek, A. Niedziałek, M. Bąk, W. Surtel, and E. Świerniak, "Comparison of Time and Frequency Feature Extraction Methods for Decision Support in Gastroenterology Based on Hemodynamic Signals," in IEEE Access, vol. 6, pp. 56743-56754, 2018.

### **6.3. Feature Engineering**

After the features are selected, the significance of feature engineering comes into play. Highlighting the idea of selecting the best features and eliminating the irrelevant ones, this part describes the different available methods of designing the features. Many algorithm techniques and different programming languages are used to achieve the feature engineering as these methods depend on the nature of the problem. It is not always possible that one method could fit best for the features, so it requires many trials and tests to arrive at the proposed solution. Many feature selection methods accomplish this and provide us with a subset of the most significant features, but none of the algorithms in the feature selection stage consider the performance of the student algorithm that will be later used. Often, a feature selection algorithm will be used as the importance and fits could easily be interpreted. Cross-validation is always the most convenient way to evaluate the current algorithm's performance over the range of the features. Alongside the feature selection, feature significance gives us a score that can be used in a featured list. As the data needs to be formatted and organized into better structures that convey the information, otherwise we won't get the desired accuracy, various methods in programming are used to transform the data into a structure that focuses the attention being paid. Features can be combined, distributed help in order to improve performance and reduce training time. Additionally, features can be enhanced by expert knowledge or using data structures based on Certificate and decision tree as they help end users in understanding and prediction. The success of predictions of any feature method depends on skilled data modelers who have a good understanding of the data and are knowledgeable in the method used. With the rise of big data and the wide applications such as healthcare in the world of cybersecurity and cloud protection, it is important to ensure that the data is well suited and organized in order to achieve the desired performance results.

## **7. Machine Learning Algorithms for Parkinson's Detection**

First, we can explain what a machine learning algorithm is. A machine learning algorithm is a pattern in the data that provides a mapping from the input data to the output. When a new input is introduced, the machine learning algorithm utilizes the mapping to give the new output. There are two types of machine learning algorithms, supervised and unsupervised. Supervised learning algorithms search for the optimum and unique mapping rule from input to the output. However, because the Parkinson's dataset has many more rows of information than the number of patients, probably an error may show up in the test set. So far, support vector machines are believed to be the most precise algorithm for dealing with Parkinson's disease datasets. Let's discuss the theory behind the support vector machine in order to understand why scientists prefer the support vector machine algorithm for detecting Parkinson. Support vector machine is a type of machine learning algorithm that is helpful in both regression and classification. However, it can be used to treat the complex Parkinson's data set which includes a large amount of input data. The fundamental idea behind the support vector machine algorithm is to find the hyperplane which best separates the output into classes. In order to determine the best classifying hyperplane, a certain algorithm is used, which is called the maximal margin algorithm. This algorithm tries to maximize the margin among the support vectors. The support vectors are the data points that are located the closest to the best classifying line. The use of a maximal margin reduces the risk of an error because it tries to make the distance among the support vectors a maximum. Once the best classifying hyperplane is identified, any new data can be sorted on the correct side of the hyperplane; so the support vector machine provides a good classification. Support vector machine algorithm is a type of supervised learning algorithm used for both classification and regression. But as Parkinson's disease detection problem is the problem where we have to use the method of determining the best classifying hyperplane and seek to make the distance among the support vectors a maximum, so most numbers of the scientists prefer the support vector machine algorithm for detecting Parkinson's disease.

### **7.1. Support Vector Machines (SVM)**

For our Parkinson's data, we used the radial basis function. There are two main parameters to consider: C and gamma. The parameter C, common to all three types of SVM, trades off correct classification of training examples against maximization of the decision function's margin. For larger values of C, a smaller margin will be accepted if the decision function classifies all training points correctly. A lower C encourages a larger margin and a simpler decision boundary, but the data points may not all be classified correctly. The default C is 1.0, and we decided to use that same value for our study. On the other hand, the parameter gamma defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. In other words, with low gamma, data points have a large 'influence' radius and the decision boundary will be more linear, while high gamma leads to a more complex decision boundary as just one data point will have a much lesser 'influence' radius. The default gamma is 1/number of features, and we also used this value in our study. These parameters have to be chosen carefully during model selection as the choice of kernel and the parameters can have a significant impact on the outcome of the algorithm. I used the Grid Search method to determine the optimal C and gamma from a range of values. This is a tuning algorithm that uses cross validation, historically considered as the gold standard of model evaluation, to determine the best parameter values for the model. First, we define the range of values we wish to search through and, from the output of the method listed below, '-0.241, a smaller RMSE and a value of 1.0 of C.' I also produced a visual representation of the effects of varying classes for the parameter C and gamma using heat maps. These can be seen on the SVM Heat Maps page - from a visual demonstration, it is clear that the choice of C and gamma does significantly alter the decision boundary. We found that the optimum C was 0.1 and the optimal gamma was 0.01. It is important to understand how the machine works and its functionality. I believe that by using the right parameter, I can improve the performance of the prediction. However, each diagnostic parameter may be unique and suitable for different patient data. From the results that I obtained from the SVM model, we had an accuracy of 84.79% from the cross validation. The confusion matrix above was from the generated model using the dataset. There were more 'no- Parkinson's' data than 'yes'. The performance result for the SVM model shows that the accuracy on class '0' is 92.97% and on class '1' is 74.62%. This result may suggest that the data is imbalanced and affects the performance.

- Linear Kernel: when the feature space is linearly separable - Polynomial Kernel: when the feature space is mapped to a higher dimensional space - Radial Basis Function Kernel: when each data point is itself assumed to be one dimensional

The algorithm that I chose to implement in this analysis is the Support Vector Machine, or SVM. SVM is a supervised learning model used for classification and regression analysis. Given a set of training samples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new data to one category or the other, making it a non-probabilistic binary linear classifier. Here, we have to select a kernel; the most commonly used ones are:

### **7.2. Random Forest**

Random Forest is an ensemble method, which is based on the principle of "Wisdom of the Crowd". It works by constructing multiple decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. For a new sample to be classified, it is run through all the trees in the forest and each tree will output a class label. The forest chooses the class having the most votes over all the trees in the forest. It is uncorrelated, because if you build lots of trees then each tree will make different errors and only a few of the minority of the trees will get it wrong and it'll get it wrong at different points in the data. So by making them vote this can cancel out the individual errors if there's noise, it's less likely to make an incorrect call. The main advantage of using random forest is its feature of selection importances which directly gives the importance of each feature in the classification. This feature is very useful and usually what it does is tests a variable with and without that variable and this allows the algorithm to make a judgement on what variables are actually pushing the predictive performance. For example, a variable with noise and not really predicting anything, the actual importance of that variable will reduce and something else will increase. Also, limitation of the overfitting problem which occurs in decision trees is overcome by random forest. When the outdoor or patient data is being used, the algorithm will overfit, which means it will not be able to find the underlying patterns but only predicts the values for the outdoor data (will make a poor predicament). However, using the random forest, the model has enough generalization and is able to provide the better result. All decision trees have low bias and high variance. By creating new trees, it forces the model to select a variable that is the best at that. Through calculating impurity decrease (this will depend by which measure you are using to calculate – entropy or Gini index) while splitting on the variable. However, the disadvantage is random forest is a little bit slow (due to the characteristics of individual decision trees) and it's a blackbox. This means that we have little control over what the model does and is very difficult to interpret because its workings is really just considering lots of variables which all give different importance outputs and combining them which gives a prediction.

### **7.3. Artificial Neural Networks (ANN)**

The multilayer perceptron (MLP) is the most commonly used artificial neural network for Parkinson's disease detection. It is a class of feedforward artificial neural network, which is composed of more than one layer of nodes. Each node in one layer has directed connections to the nodes of the subsequent layer. In a typical multilayer perceptron, there is one input layer, one output layer, and one or more hidden layers. Each node is a neuron, using standard nonlinear activation functions such as the logistic function. There are also other types of artificial neural networks, such as convolutional neural networks (CNN) and recurrent neural networks (RNN). CNN is usually used for image analysis, and the main advantage of this type of network is that it requires less pre-processing compared to other types of neural networks. The variations in the grid size within the image could not properly be handled by some other algorithms, and CNNs with their shared weights and bias and inherent locality correlation help to handle those variations. On the other hand, RNN is ideal for sequential data analysis, and it is good for predicting what comes next in a sequence, such as stock prices, or even a sentence. It has the "memory" which memorizes all the past information and uses the information to decide what will be the next output. One of the algorithms is the "Backpropagation". This is a supervised learning algorithm and is used to train multilayer perceptron, CNN, and RNN. It is a supervised learning algorithm. Every iteration, the algorithm is adjusting the weights of the connections in the network in order to minimize the difference between the actual output vector of the net and the desired output vector. By diminishing the error for each training data example, the network performs the backpropagation in three steps. First, we start with a random distribution of the values of the weights and thresholds. Then, for each example in the training example, the "forward pass" is applied to find the calculated output. Then, the algorithm applies the "backward pass" which starts at the output layer and moves back towards the input layer, which calculates the error at each node and adjust each weight in proportion to the derivative of the error. For each output O\_j, the error E at that node can be calculated based on this formula. O\_k is the kth output nodes connected to the jth node, and w\_ij is the weight for the connection from node O\_j to node I\_k and f prime (...) is the derivative of the activation function. Every weight for the output O\_j is adjusted based on the error at that node. And the total error in the net is also calculated based on the formula, E = 1/2 \* sum(E\_k) which is the error at each output nodes.

## **8. Performance Evaluation Metrics**

Code中设置的rename是为了确保易后续阅读和更改列标签名称时不会出错。"Diagnosis"实际上可以作为一个关键字来调取这一列的值,在部分item(data.head())中会体现这么一个标示,即data["Diagnosis"]可以将所有行的这一列的数值全部选中显示。只有通过这一步的rename和header的标准化,我们之后在进行数据集的划分,对列进行数据处理等等操作时,能够更加方便的用列的名称来进行操作。

Just a URI (Uniform Resource Identifier), but the returned input is a string, str type, because it is not a numeric data type. First, I need to check if the CSV file contains the features and diagnosis column that we need, and then read the header of the file to ensure that the attributes we need are included in the CSV file. In addition, the header is also renamed using the following code:

URI of the first data pool of experiment P:\Parkinson's Disease\10001\_2017\_10\_13 (Don't Delete)\_EMG (10mA, 10ms, 0.5Hz, 60s)\data.csv. Just a URI (Uniform Resource Identifier). First, I have to read the header of the CSV file to check if the columns of the file match the attributes of my machine learning task, which are several different features and a diagnosis column. I use pandas for doing so:

Therefore, when comparing models, we usually compare their AUC value, which is the "area under the curve". The larger the AUC value of the model, the stronger its predictive ability. For example, an algorithm that randomly guesses would produce a ROC curve composed of a diagonal line segment from coordinate (0, 0) to coordinate (1, 1), as shown in the red line segment in the figure below.

An algorithm that randomly guesses the classification would produce a ROC curve that is a diagonal line from point (0, 0) to point (1, 1), like the red line shown as follows.

ROC curve explains the trade-off between sensitivity and specificity. When we move the threshold of the classification to the left, the false positive rate and true positive rate would increase. If we move the threshold to the right, both rates would decrease.

For example, sensitivity of 99% would make the ROC curve go to the coordinate (0.2, 0.99). Specificity shows the probability of true negatives and it works like sensitivity. But when we move along the x-axis, sensitivity would decrease; if we move along the y-axis, specificity would decrease. 0 to 1, 0 to 1. ROC curve can help us identify the best threshold for true positive rate and false positive rate. The closer the curve is to the point (0, 1), the better the model is.

### **8.1. Accuracy**

The variable that can be used to measure the effectiveness of a classification model is accuracy. Accuracy is defined as the proportion of the true results, both true positives and true negatives, in the population. It is the most intuitive performance measure and it is so common that people usually mean accuracy when they use the term measure of performance of a model. The formula is given by accuracy = (true positive + true negative) / (true positive + false positive + true negative + false negative). However, accuracy is not a good performance measure when we have class imbalance. Class imbalance is a term used to describe when a target class within the data is outnumbered by another class. For example, accuracy of 99.1% may appear to be good, but actually, in an imbalanced dataset, where the target variable indicates 1 in 1000, 99.9% accuracy may an incorrect model, because only 0.1% of the times will the model able to predict 1's. Overfitting is another problem when using accuracy. When the data is imbalance, an algorithm can be biased towards the majority class. As a result, the accuracy can be high, but the sensitivity to the minority class is compromised. This will force our model predicting the majority class and performs badly on the minority class. In this project, since there is no big difference between the values of TPR and FPR, I think more or less overfitting might exist. The value of accuracy of the 3 algorithms came to Logistic Regression: 94.75%, Random Forest: 95.34% and Decision Tree: 95.25%. But in order to further verify this issue, I am planning to find the best fit for each single attributes also in the next step for this project. We'll see how it goes next.

### **8.2. Sensitivity and Specificity**

As such, this may prove unfeasible in reality and thus there remain chances to, and a requirement for, additional research in this area to further refine the module generated.

In this investigation, varying the threshold has shown an optimum balance of sensitivity and specificity. For example, as the threshold increases in the case using only two features, the module performs with greater specificity but poorer sensitivity. These results may also suggest that further data would be capable of enhancing specificity further, leading to a more robust model. On the whole, the analysis demonstrates the trade-off between sensitivity and specificity, and outlines the scope for tuning any model to the preferences of the user. However, this would need a huge number of cases and a high degree of suitability in relation to the cost function, which takes into account not only false negatives and positives, but also the impact of these.

As shown in Tables 1 and 2, the greater system complexity afforded by using all features can lead to superior sensitivity and specificity to case A. We see that increase in accuracy can be achieved by choosing a threshold value to develop a module, instead of the default 0.5 used in logistic regression. This is shown by the change in accuracy rates when varying the threshold, as illustrated by a typical ROC curve - the curve shows how the number of true positives identified changes with the number of false positives which are also identified. The area beneath the curve (AUC) gives an indication of the performance of the model.

By definition, sensitivity is the probability that a test result will be positive when the disease is present. Stated another way, the percentage of sick people who are correctly identified as having the condition. A value of 100% would indicate no "false negatives", i.e. all patients with the condition are correctly identified. In contrast, specificity is the probability that a test result will be negative when the disease is not present. Stated another way, the percentage of healthy people who are correctly identified as not having the condition. A value of 100% would indicate no "false positives", i.e. all patients without the condition are correctly identified. Both values therefore represent the goodness of the test, with higher percentages meaning a better test.

### **8.3. Receiver Operating Characteristic (ROC) Curve**

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true positive rate is also known as sensitivity. It is calculated as the number of true positive instances divided by the sum of true positive and false negative instances (i.e. TPR = TP/(TP + FN)). On the other hand, the false positive rate is calculated as the number of false positive instances divided by the sum of false positive and true negative instances (i.e. FPR = FP/(FP + TN)). When the threshold is set at 0, the classifier identifies all instances as positive, i.e. TPR = 1 (as all true positive instances are captured) and FPR = 1 (as all true negative instances are lost). As the threshold increases, the classifier becomes more conservative and the TPR decreases while the FPR also decreases. At the maximum threshold of 1, the classifier identifies all instances as negative, i.e. TPR = 0 and FPR = 0. Therefore, the ROC curve always starts at (0,0) and concludes at (1,1). The area under the ROC curve (AUC) is another popular metric for measuring the predictive performance of the binary classification algorithm. More intuitively, AUC provides an estimate of the probability that instances of a randomly chosen positive class have higher rank of predicted probabilities than instances of a randomly chosen negative class. The value of AUC is between 0 and 1 where an AUC of 0.5 suggests the classifier does no better than random guessing and a value closer to 1 implies a better predictive performance. A variety of software packages in R, Python and Matlab provide functions to calculate AUC from the ROC curve. In this study, AUC is used as one of the criteria to compare the performance of different machine learning algorithms. However, AUC alone is not enough to determine which model is the best. For instance, a classifier with AUC higher than 0.5 might have a poor precision value and hence is not suitable for the application with imbalanced data. Therefore, a range of classification evaluation metrics should be considered when comparing classifiers.

## **9. Experimental Results and Analysis**

The accuracy of the various algorithms is around 96%, which is pretty decent given the dataset size and the number of different instances in the dataset. The SVM tends to be quite good for binary classification problems e.g. "has disease" or "doesn't have disease", but can be quite slow and cumbersome. Similarly, we can see from the training time for each algorithm that the random forest and the ANN are much quicker to train. Of course, this is a symptom of Parkinson's Disease. We could tune our algorithms further, for example by using a grid search algorithm to produce the optimum parameters for our SVM, but we should first consider the information contained in figure 5, which presents the ranking of the most important features. As we can see, atrophy of the midbrain in the midbrain and atrophy of putamen both appear in the top 5 most important features according to the SVM. These both make intuitive sense, as atrophy of the putamen is a known symptom of Parkinson's. However, if we were to go forward with this feature, it is important that these 2 features are not highly correlated. This is because if 2 features tell us mostly the same information, that could lead to poor generalization of the model and an unreliable prediction on new samples. Correlations between the features and the outcome could actually be visualized using a heat map like the one in figure 6. This confusion matrix for the random forest algorithm allows us to see where real and predicted classes intersect and the precision and recall could be calculated. For example, in this case we would say that the class "no Parkinson's" has a precision of 1 whereas the class "Parkinson's" has a precision of about 0.95 - this is defined as the number of true positives over the number of true positives and false positives. In other words, if a data point falls into the latter category, there's a 5% chance that a given prediction is a false alarm. The recall of "no Parkinson's" is 1 and the recall of "Parkinson's" is about 0.96, which is very close to the true positive rate of the algorithm; defined as the number of true positives over the number of true positives and false negatives. Fascinatingly, the precision and recall of all 4 classes at each stage of the decision processes in the ANN and the SVM could be visualized using a precision-recall curve. This shows a rate of true positives to false positives at different thresholds and allows us to make the decision on how we classify the sample according to our cost of false positives and false negatives. For example, in a highly infectious disease, there is a low cost associated with a false positive (the person is incorrectly labeled as having the disease) compared to a false negative, where we may lose time treating the person and the infection may spread. This could help inform the decision on the threshold of probability over which the label switches from one class to another according to a given algorithm. For example, in figure 10 we can see the precision-recall curve for the ANN and, in practice, the more area under the curve (AUC) and the further that the curve is from the solid line towards the top-right corner, the better the algorithm and the better the confidence we have that it is accurate. Ergo, the AUC could be used comparatively between different algorithms to decide which may be most effective at predicting the outcome from a given set of features. However, the care must be taken to balance the number of features, the cost in the real world of false positives and false negatives, the run-time of the algorithm and the meaning behind the most predictive features when choosing the most effective algorithm. That said, hopefully the technology described here will help transform the way we diagnose and maybe even treat Parkinson's in the future.

### **9.1. Dataset Description**

The dataset used in this project is the "Parkinson's Telemonitoring Data Set" provided by UCI Machine Learning Repository. It was created by Athanasios Tsanas, a researcher at the University of Oxford, and Max Little, an associate professor at the University of Oxford as well as a tutorial fellow in the Department of Engineering Science. The dataset's purpose is to study daily living and its associated disease, with a focus on disease progression and providing quantitative analysis on the associated disease, with the ultimate goal of studying new therapy. The total size of the dataset is about 5.55 megabytes and the storage format is a comma-separated values (CSV) file. The file is composed of a total of 5875 rows and 26 columns. It is a relatively large size for a dataset caused by the long-term process in the telemonitoring study. The large amount of data in the dataset can provide a powerful source for analysis and allows the algorithm to learn more about the general conditions of patients' condition. Also, the more data used, the better the result and generalization would be when used in a machine learning algorithm. One of the most significant advantages of using the dataset is that most of the features have understandable and physiological meanings, which is very important when utilizing machine learning because the learning results and predictions need to be clinically interpretable and the clinical correlation between these learned features and the clinical variables can further validate the findings from the data mining algorithms. Also, the data contains some patients that only have one record and this would introduce some uncertainties about the progressions. And for the missing values, it is mentioned that the observations with missing attribute values are removed and this rule was used when pre-processing the dataset. No specific information about the missing attribute values was mentioned on the data that provided by the UCI Machine learning. For the class labels in the dataset, it uses the column header "motor\_UPDRS" that represents the UPDRS stands for "Unified Parkinson's Disease Rating Scale" which is a rating scale used to measure the degree of the disease. This data in the column is discrete, with a range from 5 to 39. Also, another class label is given and that is the "total\_UPDRS" column, which is a summary of the UPDRS measurement and album with the range from 7.5 to 54.833. The "total\_UPDRS" values are then computed by clinicians and they can be used as a clinical criterion for future study. And both of these columns are representing the disease severity thus they are utilized as the class variable in the project. Also, in the dataset, "subject#" where the # represents the patient number is given to identify the patient and "time" column is used to represent the appropriate time that the record was taken. For each patient, the daily telemonitoring result was presented and this would cover a time period from the beginning of the record until the last one. Thus, the record for different patients might start and end at different times. And this time span for the record would be used as an important clinical information when the progressions and therapy studies were considered.

### **9.2. Evaluation of Different Algorithms**

The accuracy of the three algorithms was also shown in a bar chart. It showed that the accuracy of Random Forest was the highest, and the accuracy of Artificial Neural Networks was the lowest. These results indicated that the ensemble learning method, Random Forest, could be a useful and reliable machine learning algorithm for the detection of Parkinson's disease.

The performance results of all three machine learning algorithms were summarized in a table. From the results, the accuracy of Random Forest was the highest among the three algorithms. The accuracy of Random Forest was about 94-95%, while the accuracy of both Support Vector Machines and Artificial Neural Networks was about 90-92%. Random Forest also had the highest sensitivity and specificity values among the three algorithms. The results suggested that Random Forest had the best performance for this dataset in terms of accuracy, sensitivity, and specificity.

Besides, the Receiver Operating Characteristic curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positive out of the y-axis against the fraction of false positive out of the x-axis as the threshold value changes. After that, the area under the curve can be calculated, and the performance of all possible prediction probability thresholds can be measured. Depending on the value of the area under the curve, it gives us an idea of how well the machine learning algorithms can differentiate Parkinson's disease and control cases.

In this study, accuracy, sensitivity, and specificity were used to assess the performance of the algorithms. The accuracy rate is defined as the ratio between the number of correct predictions and the total number of predictions. The sensitivity is the proportion of actual positive cases that were predicted as positive. Similarly, the specificity is the proportion of actual negative cases that were predicted as negative.

First, the dataset was randomly split into two parts: one for training and the other for testing. The testing dataset was reserved to assess the performance of the algorithms. Then, each of the machine learning algorithms was trained using the training dataset. After that, the trained algorithms were used to predict outcomes using the testing dataset. By comparing the predicted testing results and the true outcomes from the dataset, the performance of the algorithms can be evaluated.

Machine learning techniques can be used for the detection and diagnosis of Parkinson's disease with high accuracy. In this study, three different machine learning algorithms were applied to the data to compare their effectiveness in differentiating Parkinson's disease patients from control subjects. These algorithms are Support Vector Machines, Random Forest, and Artificial Neural Networks.

### **9.3. Comparison of Results**

The comparison of the results obtained from using SVM, Random Forest, and ANN is shown in this section based on the metrics of accuracy, sensitivity, specificity, precision, and F1 score. The comparison is divided into two subsections. In the first subsection, the comparison is made between the results of SVM and Random Forest. From the bar graph, it is shown that the accuracy of Random Forest is higher than that of SVM. It is also shown that the sensitivity and specificity of Random Forest is higher than that of SVM. This means that the Random Forest model is better at identifying the 'disease' class and 'no disease' class in the dataset compared to the SVM model. The precision and F1 score of Random Forest model is also higher than that of SVM. In the second subsection, the results of both SVM and Random Forest are compared to the results of ANN. From the bar graph, it is shown that the accuracy of ANN is higher than both SVM and Random Forest. It is also shown that the sensitivity and specificity of ANN is higher than both SVM and Random Forest. This means that the ANN model is better at identifying the 'disease' class and 'no disease' class in the dataset compared to the SVM and Random Forest models. The precision and F1 score of ANN model is also higher than both SVM and Random Forest. The ROC curves obtained from the three algorithms are also given and based on the value of the area under ROC curve (AUC), it also shows that ANN is the best model, followed by Random Forest and SVM. It can be concluded that ANN is the most appropriate algorithm among these three, for the prediction of Parkinson's Disease, based on the given dataset.

## **10. Discussion and Conclusion**

Now, moving on to the discussion and conclusion part, it is observed that all algorithms except SVM have provided 100% specificity, which is highly desirable. The reason might be due to the fact that the dataset has provided linearly separated classes. Also, a very high value of accuracy (over 99%) is recorded for all classifiers except for SVM. This implies that the variance in these classifiers might be small, which means that the corresponding models have captured the underlying data's structure very well. As for ANN, it provided the highest accuracy of over 99.7% among the three classifiers. Another key point to note would be all three classifiers have high sensitivity value of over 97%. This means that almost all the true Parkinson's disease will be identified by the algorithms. However, we still get some false alarms as shown by the 3% difference between sensitivity and specificity for all three algorithms. This is where the ROC curve comes in handy to help us to be able to visualize the trade-offs between the benefits (true positives) and the costs (false positives) of the classification. Based on the ROC curve metrics, it is seen that all three algorithms have produced great results as the Area Under Curve (AUC) value are quite high. The ROC graph allows the examination of the relationship between the true positive rate (sensitivity) and the false positive rate (1-specificity). By looking at the plot of ANN, there is evidence that as the output threshold is changing from 0 to 1, there is a steep increase in the true positive rate as compared to the slight increase in the false positive rate, which means that the true positive rate is increasing at a definitely higher rate, causing the AUC to be almost 1. As such, choosing a proper threshold to minimize the false positive rate while maximizing the true positive rate could be very effective for the Parkinson's disease patients and medical practitioners for early detection.

### **10.1. Interpretation of Findings**

The decision tree visualization generated by building a Random Forest model, as shown in Figure 30, reveals some interesting findings in terms of feature importances. As stated earlier, feature importances help us to understand the most influential features which have the highest effect on our target variable. Colour intensity, which is related to the maximum number of correct instances that can be classified due to the node that variable M was chosen for splitting, is found to be the most important feature for a diagnosis of Parkinson's disease. Interestingly, all the top 10 features, including colour intensity, range, and so on, are related to voice features, with the top 9 features all belonging to voice features. This supports the medical observations which suggest that voice changes such as speaking softly, quickly, slurring or hesitating in speech are early signs of Parkinson's disease. Also, the finding that voice features are the most important types of features is consistent with the literature review from the section of "diagnosis of Parkinson's disease", which shows that machine learning based Parkinson's diagnosis models using voice features have outperformed those using other types of features. It can be performed more complicated and powerful method like deep learning to further improve the prediction accuracy because we use relatively simple, traditional machine learning models such as Random Forest, SVM and Logistic Regression in the current study. Furthermore, it could be useful to compare the use of voice features with the use of other types of features, or the combination of multiple types of features, so that the scope of identifying the most effective machine learning based diagnosis model for Parkinson's disease can be broadened. The observation variable importance values is given in Figure 31. It quantifies how much the splitting criterion at the parent mode, minus the ones at the two child modes, is improved by that feature. Therefore, by looking at Figure 31, it is indicated that status mean, D2 standard deviation in long axis and PPE standard deviation is the top 3 most influential features. However, due to the limited computer computational capacity and time resources, the use of cross validation in the current study is not as thorough as the research team would have wished. Cross validation is a technique that helps to verify the predictive value of the model and it is essential for model selection when using different algorithms with distinct sets of parameters. Especially for the optimization part, which is the tuning of the best values for model parameters, cross validation should be used in every single iteration and the algorithm should be built and tested on different partitions of the data set. The study utilized a form of cross validation called k-fold cross validation, where the data set is divided into k subsets and the holdout method is repeated k times. This could reflect parameter tune when algorithms are switched and general model accuracy. The interpretation findings described above provide crucial insights for deciding the validity and interpretation of the prediction model. However, further investigation is still required to continually improve the predictive accuracy of the model by making use of some more sophisticated methodologies, as well as to strengthen the validation results obtained and to contribute to the more reliable and robust digital diagnostic tools and methods for Parkinson's disease.

### **10.2. Implications for Parkinson's Disease Detection**

In addition, machine learning models can also be adapted and used for other tests that patients could undergo easily, such as electromyography or accelerometry. Therefore, an overhaul of the diagnostic process for Parkinson's disease could be potentially foreseeable. A modern diagnostic process would involve one or more of these machine learning algorithms either as an assistive tool for clinicians or as a primary method of diagnosis. Due to the non-invasive nature of tests such as those that machine learning could analyse, and the repeatability that some tests permit, the necessity for invasive and potentially dangerous tests, like DaTSCAN, could be wholly removed from the diagnostic process. However, the possibility of using one or more of these algorithms as an aid to diagnosis or even as a primary diagnosis method is dependent on a few different things. First and foremost, before any of the algorithms discussed in my study become used in the diagnosis of Parkinson's disease, it would be necessary for a large scale longitudinal clinical trial to take place. Such a study would have to investigate how accurate each algorithm is at predicting Parkinson's disease in people who are known to be at risk, such as those with a family history of the disease or those over a certain age, and compare the predictions made by the algorithms to the eventual disease statuses of the patients. This practice is standard for new predictive tools or biomarkers for many diseases, in order to ascertain whether something is accurate or reliable enough to be put into use. Going further, the trial would also have to seek approval from a national regulatory body, such as the Medicines and Healthcare products Regulatory Agency in the UK or the Food and Drugs Administration in the United States. These bodies are responsible for assessing the safety and effectiveness of medical treatments and products, and they play a crucial role in the journey of a new potential diagnostic method from the research setting to the clinic. Given the current lack of effective biomarkers and the relatively poor accuracy of even expert clinicians, there is certainly great potential for machine learning algorithms to become a leading method of Parkinson's diagnosis. However, the high standards of evidence and safety that are rightly held in place mean that this process from possibility to reality will be long and arduous. Finally, it is currently unclear how the large blossoming of medical data, specifically in the digital health age, will intersect with the potential use of machine learning algorithms for Parkinson's work. If the diagnostic information from these new technologies, such as genetic screening and even smart phone function, can be integrated and utilised by algorithms, the future of a machine learning based diagnostic revision for Parkinson's disease may take an unforeseen but truly modern direction; in essence, creating a comprehensive "digital phenotype" of a patient may become a diagnostic norm. Such a point would mean that algorithms can diagnose and stratify disease using a far wider range of markers than ever thought of possible, leading to more personalised and efficient treatments for patients.

### **10.3. Future Directions and Recommendations**

The last section of this essay, "Future Directions and Recommendations", provides suggestions for future research in the area of Parkinson's disease detection. One main recommendation is to explore the value of fusing multimodal data, such as electromyography (EMG), kinematic data and clinical assessment scores, to further advance the development of wearable technologies in diagnosing Parkinson's disease. This recommendation is based on the emerging trend in the scientific community, where recent studies have focused on the development of wearable and continuous monitoring systems that revolutionize the way patient diagnosis and disease progression is quantified. Secondly, the development of new machine learning techniques and algorithms that could capture the temporal and dynamic features of the progression of Parkinson's disease should be the focus of the research community, given the progressive nature of the disease. Since Parkinson's disease is categorized as a slow progressing neurodegenerative disorder, the disease evolves over the year and future research could explore the potential of a data-driven machine learning approach for understanding the temporal pattern of disease progression. This is another direction demonstrated in the reports by other recent studies, where novel methodologies using multi-task learning and deep learning have shown promise of capturing the temporal dynamics of neurodegenerative diseases. More broadly, it is recommended to deploy explainable machine learning models in Parkinson's disease detection. As discussed in the study, having a clear understanding of how machine learning algorithms could be interpreted in clinical decision-making processes is crucial for implementing machine learning models in practice. Given the enormous potential that machine learning models could be applied in analysis of complex medical data such as genomics and medical imaging, it serves as a very promising future direction if researchers in this field are able to develop novel methodologies for interpreting machine learning models in the context of various clinical needs. It is our hope that the study could inform and offer valuable insights into the field of Parkinson's disease detection using the state-of-the-art machine learning methods. By providing comprehensive comparison and evaluation for different machine learning methods for early diagnostics of Parkinson's disease, it is expected that the study would represent a foundational work that lays a basis for future development in this area by the research community. We look forward to collaborating with experts from different domains, including neurologists, biomechanics specialists and machine learning researchers, in moving the field forward and translating technological breakthroughs from laboratory research to clinical utilities that could ultimately bring benefits to patients with Parkinson's disease.

References:

[1] W. o. r. l. d. Health Organization, "Parkinson disease: a public health approach: technical brief," 2022.[who.int](https://apps.who.int/iris/bitstream/handle/10665/355973/9789240050983-eng.pdf)

[2] A. Mahmood, A. A. Shah, M. Umair, Y. Wu, et al., "Recalling the pathology of Parkinson's disease; lacking exact figure of prevalence and genetic evidence in Asia with an alarming outcome: A time to step‐up," Clinical …, 2021, Wiley Online Library.[HTML](https://onlinelibrary.wiley.com/doi/abs/10.1111/cge.14019)

[3] T. Wijeratne and S. Fox, "World Brain Day 2020: Join Us to “Move to End Parkinson's Disease”: A World Federation of Neurology and International Parkinson and Movement Disorders Society …," Canadian Journal of Neurological Sciences, 2021.[cambridge.org](https://www.cambridge.org/core/services/aop-cambridge-core/content/view/A7807433D3429CE203CCAD94585DF243/S0317167120001407a.pdf/world-brain-day-2020-join-us-to-move-to-end-parkinsons-disease-a-world-federation-of-neurology-and-international-parkinson-and-movement-disorders-society-collaboration.pdf)

[4] N. Schiess, R. Cataldi, M.S. Okun, et al., "Six action steps to address global disparities in Parkinson disease: a World Health Organization priority," JAMA Neurology, vol. 2022. [Online]. Available: jamanetwork.com.[jamanetwork.com](https://jamanetwork.com/journals/jamaneurology/fullarticle/2793874)

[5] E. Hamid, B. A. Ayele, D. G. Massi, S. Ben Sassi, et al., "Availability of therapies and services for Parkinson's disease in Africa: a continent‐wide survey," Movement Disorders, vol. 2021. [Online]. Available: Wiley Online Library.[HTML](https://movementdisorders.onlinelibrary.wiley.com/doi/abs/10.1002/mds.28669)