**EEG Feature Extraction for Diagnosis of Parkinson’s Disease and Schizophrenia using Machine learning techniques**

*Mustafa Mehmood, Hafeez Ullah Amin\*,*

*School of Computer Science*

*Faculty of Science and Engineering*

*The University of Nottingham,*

*43500 Semenyih, Selangor, Malaysia*

*\*Email: hafeezullah.amin@nottingham.edu.my*

**Abstract:** This study aims to present an electroencephalography (EEG)-based machine learning approach for diagnosing patients with Parkinson's disease, and Schizophrenia, for treatments and improving the quality of life. Further, the study evaluated the performances of different machine learning classifiers for adopting the most robust machine learning classifier for diagnosis and modelling the EEG datasets, providing valuable insights into the most effective methods for diagnosis. Two EEG datasets, one for patients with Parkinson's disease, and the second with Schizophrenic patients were used in this study. The raw EEG data signals were first pre-processed in both datasets by using band-pass filtering, average referencing, and independent component analysis (ICA) to remove eyeblink artifacts and signal noise. The pre-processed EEG datasets were then analysed using statistical methods and the power spectral density method for extracting useful information as features. For statistical features, mean, standard deviation, variance, minimum, maximum, max/min indices, mean square, root mean square, skewness, and kurtosis were calculated. Fast Fourier Transforms with the windowing method were adopted to calculate averaged EEG power in different frequency bands. The extracted features were then used to train machine learning models such as support vector machines, k-nearest neighbours, random forests, logistic regression, and Naïve Bayesian classifier. The results showed that with statistical features, the random forests classifier outperformed the other machine learning techniques in both data sets, with an accuracy of 96% for Parkinson's disease and 92% for schizophrenia. The logistic regression also showed acceptable classification results, slightly lower than the random forests. The proposed machine learning-based approach has the potential to efficiently diagnose Parkinson's Disease and schizophrenia in EEG signals.

Keywords:- EEG Signals, Parkinson's Disease, Schizophrenia Disease, Feature Extraction, Statistical Features, Power Spectral Density, Machine learning techniques

**Introduction:**

Mental disorders, such as schizophrenia and Parkinson's disease, are a significant public health concern, with millions of people worldwide living with these conditions. Schizophrenia is a serious mental disorder that is characterized by abnormal interpretations of reality, including hallucinations, delusions, and disorganized thinking and behavior (Owen, 2016). These symptoms can significantly impact daily functioning and may require lifelong treatment. Parkinson's disease, on the other hand, is a neurodegenerative condition that causes uncontrolled movements, such as tremors, stiffness, and difficulty coordinating and balancing (Kalia & Lang, 2015). It is caused by damage to the substantia nigra, a region of the brain that produces dopamine, a chemical messenger that is necessary for proper body movement and speech.

The diagnosis of mental disorders such as schizophrenia and Parkinson's disease is typically made by a neurologist, but any tool that can assist in the diagnosis process is welcome. One promising approach for automating the diagnosis of these disorders is the use of neuroimaging techniques, such as electroencephalograms (EEGs). (Amzica & Lopes da Silva, 2011)EEGs are a functional neuroimaging technique that records brain activity from the head surface at high temporal and spatial resolutions. They offer several advantages over other technologies, such as CT scans and MRIs, including low cost, portability, and the ability to record brain activity for a longer period of time at a faster rate. By analysing EEG data with machine learning techniques, researchers have been able to detect various mental disorders, including autism spectrum disorder, Alzheimer's disease, and major depression. However, the use of EEGs for the diagnosis of Parkinson's disease has not been fully explored.

The objective of this research paper is to use machine learning techniques, including both conventional methods and deep learning models, to diagnose Parkinson's disease and schizophrenia using EEG data. The authors propose using a variety of methods, including support vector machines (SVMs), k-nearest neighbours (K-NN), random forests, logistic regression, and Naive Bayes. By testing the performance of these methods on EEG data, the authors aim to determine the most effective approach for the automated diagnosis of these mental disorders. This research has the potential to improve the accuracy and speed of diagnosis, ultimately leading to better care and outcomes for individuals with schizophrenia and Parkinson's disease.

One of the challenges of using machine learning techniques for the diagnosis of mental disorders is the need for high-quality training data. In order to effectively train a machine learning model to differentiate between EEG data from individuals with schizophrenia or Parkinson's disease and those without these disorders, a large and diverse dataset is necessary.

Overall, the use of machine learning techniques on EEG data has the potential to significantly improve the diagnosis of schizophrenia and Parkinson's disease. By automating the diagnosis process and utilizing the analytical power of machine learning, more people may have access to effective care, leading to better outcomes and a higher quality of life for those living with these mental disorders.

**Literature Review**

Several studies have utilized EEG data collected during mental or muscular tasks, such as the Timed Up and Go Task, to classify PD patients with high accuracy rates. For example,(Ly et al., 2017b) used Bayesian Neural Networks to classify PD patients based on their EEG data during the Timed Up and Go Task, achieving an accuracy rate of 86.2%. Another study used support vector machines (SVMs) and wavelet transforms to classify gait initiation failures (GIFs), a type of freezing of gait experienced by PD patients, with an accuracy rate of 86.3%. (Ly et al., 2017a). (Oh et al., 2020) Used a 13 layered Convolutional Neural Network which didn’t require an explicit feature selection, extraction or classification. They achieved an accuracy of 88.25%.(Zhang et al., 2022) proposed two methods: time–frequency analysis with deep learning, tunable Q-factor wavelet transform with deep residual shrinkage network (TQWT-DRSN) and the wavelet packet transform with deep residual shrinkage network (WPT-DRSN). The accuracies reached 99.92% for 2-class classification tasks. WPT-DRSN has accuracies of 97.81% and 92.59% in 3-class classification and 4-class classification tasks, respectively, which are higher than TQWT-accuracies DRSN's of 95.20% and 90.46%.

Other studies have focused on identifying EEG parameters or characteristics that are effective in detecting PD. For example,(Yuvaraj et al., 2018) (Oh et al., 2020) Yuvaraj et al used higher-order spectrum (HOS) feature extraction to develop an automated diagnosis of PD, and found that bi-spectrum features were relevant for PD detection. Resting-state EEG recordings have also been used for the automated diagnosis of PD, with several machine-learning techniques being applied to achieve high classification accuracy. In addition to these approaches, wavelet transforms have been used to decompose EEG signals into subbands and extract statistical features for PD classification. (Khare et al., 2021)Smith K.K. et al proposed the use of wavelet transforms to extract five features from subbands, which were then classified using various machine learning techniques.

The use of machine learning techniques has been applied for the detection of schizophrenia using EEG signals in the following studies: (Jahmunah et al., 2019)employed digital filtering and non-linear features extraction, followed by dimension reduction using the Student's t-test, and classification using a support vector machine (SVM) classifier. (Dvey-Aharon et al., 2015) used channel selection and time-frequency analysis-based features extraction, followed by k-nearest neighbor (KNN) classification using leave one out cross validation. (Santos-Mayo et al., 2017)employed digital filtering and time-frequency analysis-based features extraction, followed by linear discriminant analysis (LDA) based dimension reduction and multi-layer perceptron (MLP) classifiers. (Vittala et al., 2020) employed multivariate empirical mode decomposition (MEMD) and modes entropy-based features extraction, followed by recursive feature elimination (RFE) based dimension reduction and support vector machine with radial basis function (SVM-RBF) classification. These studies demonstrate the potential of machine learning techniques to improve the accuracy of schizophrenia detection using EEG signals, however further research is needed to determine the optimal combination of methods for this task.

Overall, the use of EEG data and machine learning techniques has shown promise in the diagnosis of PD and SZ. By automating the diagnosis process and utilizing the analytical power of machine learning, more people may have access to effective care, leading to better outcomes and a higher quality of life for those living with these diseases. However, more research is needed to fully explore the potential of EEG and machine learning in the diagnosis.

**Materials and Methods**

Hyperparameter tuning.

Figure Methodology of proposed approach

In this study, two public datasets of raw EEG recordings were obtained for patients with Schizophrenia and Parkinson's disease. The raw signals were pre-processed in MATLAB to remove noisy data, and independent component analysis was applied to extract and remove components with artifacts. The pre-processed data was then normalized to improve the accuracy of the subsequent analysis.

Two independent methods were used for feature extraction: statistical features and Average band power using Welch's periodogram, which averages successive Fourier transforms of small windows. These methods were employed to obtain a comprehensive understanding of the EEG signals and to ensure that the results were reliable and robust.

Conventional machine learning techniques were utilized to train the extracted data. Hyperparameter tuning was carried out for all the machine learning techniques using GridSearchCV. The data was split into train and test sets using cross-validation with n-fold to reduce the risk of overfitting and to obtain accurate results. The detailed explanation of the procedure in explained in the following paragraphs.

**Raw Data-sets Sources**

In this study, two public EEG datasets were used to test the proposed methods. The first dataset (Rockhill et al., 2020)was collected from patients with Parkinson's disease (PD) and healthy participants at the Scripps Clinic in La Jolla, California. The PD dataset included raw EEG data from 15 patients (eight females, mean age 63.2 years) who were taking and not taking dopaminergic medication, as well as 16 healthy participants (nine females, mean age 63.5 years). All PD patients were diagnosed by a movement disorder specialist and were right-handed. The EEG data was collected using a 32-channel BioSemi ActiveTwo system, sampled at 512 Hz. Participants were seated and asked to fixate on a cross presented on a screen, and additional electrodes were placed to monitor eye blinks and movements. The data collection was conducted in a counterbalanced order for patients taking and not taking medication, and the EEG data for this study was only considered for patients who were off medication.

Table PD demographics

|  |  |  |
| --- | --- | --- |
| “US San Diego PD Dataset” | PD | Control |
| Number | 15 | 15 |
| Sex | 8 Females/ 7 Males | 9 Females/ 7 Males |
| Age ( Mean ± SD) | 63.2 ± 8.2 | 63.5 ± 9.6 |
| UPDRS | 39.2 ± 9.7 (OFF) | - |
| Years diagnosed | 4.5 ± 3.5 | - |

*UPDRS United Parkinson’s Disease Rating Scale*

The second dataset was collected from the Institute of Psychiatry and Neurology in Warsaw, Poland (Olejarczyk and Jernajczyk, 2017) (Olejarczyk & Jernajczyk, 2017)and included recorded EEG signals from 14 patients with schizophrenia and 14 normal individuals matched in terms of age and gender. The EEG signals were recorded with the eyes closed for 15 minutes, using standard 10-20 electrodes with a sampling frequency of 250 Hz. The electrodes included Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, and O2.

Table SZ demographics

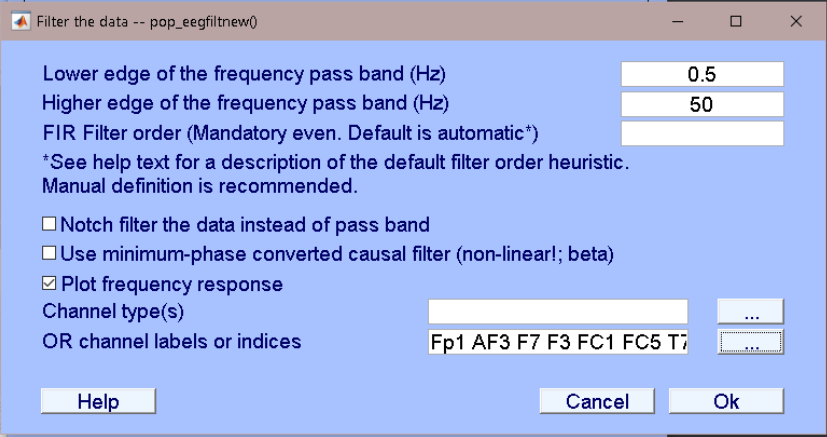
|  |  |  |
| --- | --- | --- |
| “Warsaw SZ Dataset” | SZ | Control |
| Number | 14 | 14 |
| Sex | 7 Females/ 7 Males | 7 Females/ 7 Males |
| Age | 27.9 ± 3.3(M), 28.3 ± 4.1(F) | 26.8 ± 2.9(M), 28.7 ± 3.4(F) |

**Data Pre-processing**

Pre-processing raw EEG data is an important step in the analysis process, as it helps to improve the quality and reliability of the data. Raw EEG data is often contaminated with artifacts such as eye blinks, muscle activity, and electrical noise, which can interfere with the analysis of the underlying brain activity. Pre-processing involves a series of steps to remove or reduce the impact of these artifacts and improve the quality of the data.

To pre-process the raw EEG data in this study, we used the MATLAB application specifically the EEGLAB library . The first step in the pre-processing process was to remove the mean of each electrode and use the average of all electrodes as a reference. Any electrodes that were excessively noisy were excluded from the analysis.

Figure Applying low and high pass filter EEGLAB



A high pass filter at 0.5 Hz was then applied using two-way FIR filters to eliminate low-frequency drift. This type of filter is designed to pass frequencies above a certain threshold while attenuating frequencies below that threshold. In the case of EEG data, a high pass filter is often applied at a frequency of 0.5 Hz or higher to remove low-frequency drift, which can be caused by factors such as movement or changes in body position. Removing low-frequency drift can help to improve the accuracy and reliability of the analysis, as it reduces the impact of these extraneous factors on the data.

After the high pass filter was applied, independent component analysis (ICA) was performed on the data. ICA is a statistical technique used to identify and extract independent sources of information from a dataset. In the context of EEG data analysis, ICA can be used to identify and remove artifacts such as eye blinks, muscle activity, and electrical noise, which can interfere with the analysis of the underlying brain activity. It can also be used to identify patterns of activity that are associated with different mental states or cognitive processes, such as attention or memory.

Figure Run ICA Example using runica Algorithm

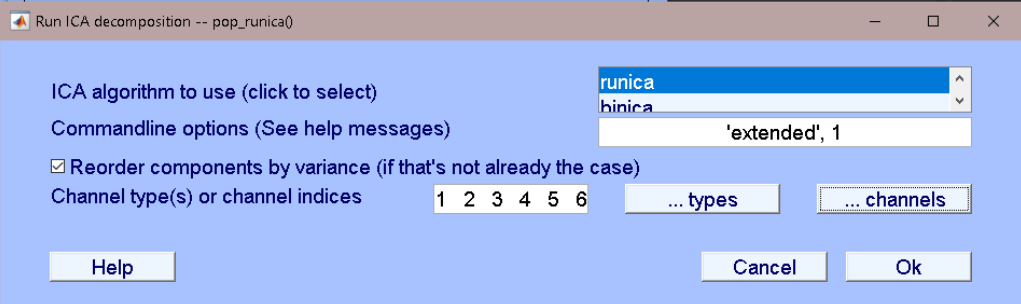


Figure Extracted components Example after ICA

Graphical user interface, text

Description automatically generated

Figure Visual representation for the extracted components from ICA

A picture containing graphical user interface

Description automatically generated

Following ICA, the data was examined for any remaining artifacts, such as eye blinks, muscle activity, electrical noise, and other sources of noise. These artifacts were then removed from the data to further improve the quality and reliability of the analysis. Overall, the pre-processing of the raw EEG data involved several steps to ensure that the data was free of artifacts and other sources of noise that could potentially interfere with the analysis.

**Data Analysis and Feature Extraction**

To diminish the number of features in the EEG datasets, two feature selection methods were utilized in this study. The initial method encompassed the extraction of statistical characteristics, including mean, standard deviation, and skewness, from the raw EEG data. Meanwhile, the second method involved extracting features by averaging the power of EEG bands, enabling an understanding of the frequency content of the EEG signals. The process was conducted through the use of diverse Python libraries specifically created for machine learning, EEG, and data management.

The identification of activity patterns that are relevant to the task at hand is made possible by feature extraction, which is a crucial step in the analysis of EEG data. Eventually, the accuracy and effectiveness of the analysis will increase as a result of the process of reducing the number of characteristics since it places a more intense focus on the most crucial patterns of activity. Moreover, feature extraction can enhance machine learning algorithms' performance by lowering the complexity of the data and allowing the algorithm to concentrate on the most crucial elements.

Features selection and extraction are helpful techniques for studying EEG data since they aid in identifying activity patterns and boost the performance of machine learning algorithms.

**Statistical Features**

Statistical features refer to the characteristics of a dataset that are determined using statistical methods. In the context of EEG data analysis, statistical features can be extracted from raw EEG signals to obtain insights into the signals' distribution and properties. In this study, statistical features were extracted from the EEG signals by segmenting each recorded signal into 5-second frames that did not overlap. Various statistical measures were computed for each frame, including the mean, standard deviation, variance, minimum, maximum, max/min indices, mean square, root mean square, the absolute difference in the signal, skewness, and kurtosis. These features were then utilized for data processing and training purposes.

Table List of statistical features

|  |  |
| --- | --- |
| Name of Statistical feature | Formula/ description |
| Standard error |  |
| Standard deviation |  |
| Sample variance |  |
| Kurtosis |  |
| Skewness |  |
| Max/Min value | Maximum/Minimum signal point value |
| Max/Min index | Index of the Maximum/Minimum signal point |
| Mean | Arithmetic average of distribution |
| Median | Middle value separating the greater and lesser halves of the data. |

An effective method for describing signals in terms of their distribution and qualities when examining EEG data is to calculate statistical features. Statistical characteristics, like the existence of outliers or abnormal numbers, provide insight into the broader patterns and trends in the data. The nature of the EEG signals and how they could change over time can be more thoroughly understood by include statistical aspects in the study.

**Average band power (****Power Spectral Estimate Welch’s periodogram)**   
To extract features from the EEG data using average band power, the following steps were followed:

1. The raw EEG signals were divided into overlapping segments of 2 seconds in length.
2. An estimate of the power spectral density was calculated for each segment using Welch's periodogram, which is a method that improves the accuracy of the classic periodogram by averaging successive Fourier transforms of small windows of the signal.
3. The power spectral density was divided into frequency bands of interest, such as delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30-100 Hz).

Figure EEG Signal Bands

1. The average power within each frequency band was calculated by summing the power values within the band and dividing by the number of values.
2. The average band power values for each segment were then used as features for further analysis and machine learning.

The Welch's method drastically reduces this variance by averaging the periodograms obtained over short segments of the windows. However, this comes at the expense of a lower frequency resolution. In fact, the frequency resolution is defined as follows:

Where is the Sampling frequency of the signal, N the total number of samples and t is the signal duration in seconds.

Average band power features provide insight into the frequency content of the EEG signals and is useful for identifying patterns of activity that are related to different mental states or cognitive processes. They can also be useful for training machine learning algorithms to classify or predict different mental states or cognitive processes. By dividing the power spectral density into frequency bands and calculating the average power within each band, it is possible to summarize the contribution of each frequency band to the overall power of the signal. This can be particularly useful for identifying patterns of activity that are related to specific mental states or cognitive processes, as different frequency bands are often associated with different brain functions.

Welch's periodogram (Welch, 1967)is a useful tool for estimating the power spectral density of a signal because it can accurately capture the time-varying nature of EEG signals. By averaging the periodograms obtained over short segments of the signal, Welch's method can reduce the variance of the estimate and improve the accuracy of the spectral estimate. However, it does come at the cost of lower frequency resolution, which can be a trade-off.

**Dataset Training Testing Split**

For this study, the dataset was partitioned into various subsets based on groupings, utilizing a technique known as K-cross-validation. This approach is used to evaluate the performance of machine learning algorithms, involving the division of the dataset into K unique subsets or folds. One of the K folds is used for testing, while the remaining folds are utilized for training. The algorithm is trained and tested on each of the K folds, and the outcomes are averaged to derive an overall estimate of the algorithm's performance.

It is particularly advantageous for small or imbalanced datasets as it allows for a more reliable estimation of the algorithm's generalization ability. Compared to conventional techniques like the train/test split method, cross-validation leverages more subsets of the data, which can result in a more precise estimate of the algorithm's performance, particularly when the available data is restricted.

**Training: Hyperparameter tuning**

Optimizing hyperparameters is crucial to enhance the performance of machine learning models, and different models have varying hyperparameters. One way to optimize hyperparameters is through grid search, which involves trying all possible combinations of hyperparameters and assessing model performance using cross-validation. However, grid search can be computationally intensive, especially with many hyperparameters or possible values.

In this study, hyperparameter tuning was applied to all machine learning models used. An exhaustive grid search technique was used, but only likely parameter values were considered to reduce computational complexity. This was done to improve model performance and achieve better results in the diagnosis of mental disorders such as Parkinson's disease and Schizophrenia.

Figure 7 Cross Validation and Hyperparameter tuning

Training Set

Hyperparameters

Grid Search

Model

|  |  |
| --- | --- |
| Training Fold | Validation fold |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1st iter | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
| 2nd iter | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
| 3rd iter | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
| 4th iter | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |

. . . . …………………………………………………

………………………………………………………

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 10th iter | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |

*E1 to E10 are loss functions during training and E is cross validation loss. Grid search method is used to find optimum E*

**Machine Learning Methods**

The utilization of machine learning techniques in the diagnosis of mental disorders is advantageous due to their ability to handle vast amounts of data and generate accurate predictions or classifications. Automated analysis of EEG data is made possible through these techniques, ensuring quick and efficient processing. Additionally, machine learning methods are highly flexible, enabling easy adaptation to different types of data and diverse mental disorders. These methods can improve over time, as they have the ability to learn from their mistakes and adapt to new data, thus enhancing their efficiency and accuracy in diagnosing mental disorders.

**Support Vector Machine**

Support vector machines (SVMs) are a powerful type of supervised learning algorithm that can be utilized for classification, regression, and outlier detection. The basic concept behind SVMs is to locate a hyperplane within a high-dimensional space that maximizes the separation of distinct classes.

Regarding classification, SVMs endeavor to identify the hyperplane that provides the maximum separation between two or more classes. This hyperplane is situated at a distance as far as feasible from the nearest training data points of any class. These training points, known as support vectors, are employed to determine the hyperplane. After identifying the hyperplane, new data points can be classified based on their position relative to the hyperplane.

SVMs excel at dealing with high-dimensional and complexly related data. Additionally, they can handle imbalanced datasets, where one class is much more prevalent than the other.

The most significant advantage of SVMs is their capacity to deal with non-linearly separable data using the kernel trick. This method involves mapping the data into a higher-dimensional space where it becomes linearly separable. This allows SVMs to categorize data that cannot be separated using a linear hyperplane in the original space.

**K- Nearest Neighbours**

The K-nearest neighbours (KNN) method is a non-parametric classification strategy that seeks to locate the K data points that are most likely to be related to a particular observation and categorise it in accordance with the majority class of its neighbours. Machine learning, statistical pattern identification, and data mining all frequently employ this technique. KNN is a reasonably simple method to build since, unlike other algorithms, it doesn't need any training data. The method can also handle any number of classes and distance measurements, making it very versatile. It is computationally demanding, though, and the precision of the categorization is dependent on the value of K, which establishes how many neighbours are taken into account. In cases where the decision boundaries are erratic and the input is not uniform, the KNN method is very helpful.

**Random Forest**

Random forest classification is a supervised learning algorithm used for regression and classification tasks. This method relies on constructing multiple decision trees trained on a random subset of the training data and combines their predictions to make a final decision. Each decision tree has an internal node representing a feature or attribute, and its branches represent possible values of that attribute. To classify a new data point, the algorithm traverses the branches of each decision tree based on the features' values until reaching the leaf node representing the final prediction.

Random forest classifier uses many decision trees, with each tree using a different subset of training data and selecting features to split nodes at random. Final predictions are generated by averaging the predictions of all the decision trees, or by choosing the majority class predicted by the decision trees. This algorithm is robust against overfitting, can handle high-dimensional data with many features, and works for regression and classification tasks. However, constructing multiple decision trees can be computationally expensive.

**Logistic Regression**

For categorization issues, a popular machine learning technique is logistic regression. As the dependent variable in this kind of regression analysis is categorical, there are only a few possible outcomes. Finding a correlation between the dependent variable and a group of independent variables, which may be numerical or categorical, is the aim of logistic regression.

In the logistic regression model, the dependent variable's likelihood of falling into a certain category is expressed using a logistic function. The logistic function returns a number between 0 and 1, which represents the likelihood that the dependent variable will fall into the positive category. Several industries, including as banking, marketing, and healthcare, heavily rely on logistic regression.

**Naive Bayes**

Using the Bayes theorem, the Naive Bayes classification approach determines the likelihood that a particular data point will belong to a specific class. This method of guided learning. This method is based on the Bayes theorem, a statistical theory that determines the likelihood of an event based on information about its circumstances.

Estimating the likelihood that a given data point will be placed in a certain class based on its attributes is the fundamental goal of the Naive Bayes classification approach in the classification domain. This method is extensively used by many various industries, including spam filtering, sentiment analysis, and natural language processing, to mention a few.. This is done using Bayes' theorem, which can be written as:

P(C|X) = (P(X|C) \* P(C)) / P(X)

Where P(C|X) is the probability that the data point belongs to class C, given its features X, P(X|C) is the probability of the features X given that the data point belongs to class C, P(C) is the prior probability of class C, and P(X) is the probability of the features X.

Given the class, the Naive Bayes classification method makes the premise that each data point's characteristics is independent of the others. Because it frequently proves to be false in reality, this is known as the "naive" assumption. Naive Bayes classification can still perform well in many situations despite this presumption, though.

A straightforward and effective technique, naive Bayes classification needs very minimal data preparation before use. It may be applied to applications requiring binary and multi-class classification and can manage high-dimensional data. It is, however, sensitive to the notion of feature independence and may not function properly when this assumption is broken.

**Experimental Results and Discussion**

In the first experiment, conventional machine-learning techniques were applied to statistical features of data sets after they were pre-processed by the before mentioned techniques for the diagnosis of Schizophrenia and Parkinson's Disease. The techniques tested included Support Vector Machine (SVM), k-nearest neighbours (KNN), Gaussian Naive Bayes (GNB), Random Forest (RF), and Logistic Regression. The epoch length was set to 5 seconds and all the algorithms were hyperparameter tuned using the GridSearchCV technique in python. Statistical features were extracted from the data sets and used as input for the machine learning models. The results of the experiment are shown in the following tables:

Table Parkinson’s Disease Data Set - Features: Statistical

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Precision | F1-score |
| SVM | 0.94 | 0.90 | 0.95 |
| K-NN | 0.89 | 0.85 | 0.90 |
| GaussianNB | 0.64 | 0.72 | 0.54 |
| Random Forest | 0.96 | 0.92 | 0.95 |
| Logistic Regression | 0.90 | 0.85 | 0.80 |

Table Schizophrenia Data Set - Features: Statistical

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Precision | F1-score |
| SVM | 0.78 | 0.83 | 0.78 |
| K-NN | 0.67 | 0.64 | 0.69 |
| GaussianNB | 0.51 | 0.54 | 0.15 |
| Random Forest | 0.92 | 0.93 | 0.91 |
| Logistic Regression | 0.89 | 0.88 | 0.86 |

Figure ROC Curve for Random Forest SZ detection -Statistical Features

Chart, line chart

Description automatically generated

Figure ROC Curve Random Forest PD Detection Statistical Features

Chart, line chart

Description automatically generated

Figure Statistical features Accuracy comparison best performing classifiers

For the Parkinson's Disease data set, the SVM algorithm had an accuracy of 0.94, a precision of 0.94, and an f1-score of 0.95. The k-nearest neighbours’ algorithm had an accuracy of 0.89, a precision of 0.89, and an f1-score of 0.90. The Gaussian Naive Bayes algorithm had an accuracy of 0.64, a precision of 0.72, and an f1-score of 0.54. The Random Forest algorithm had the highest accuracy of 0.96, a precision of 0.95, and an f1-score of 0.95. The Logistic Regression algorithm had an accuracy of 0.90, a precision of 0.85, and an f1-score of 0.80.

For the Schizophrenia data set, the SVM algorithm had an accuracy of 0.78, a precision of 0.83, and an f1-score of 0.78. The k-nearest neighbour’s algorithm had an accuracy of 0.67, a precision of 0.64, and an f1-score of 0.69. The Gaussian Naive Bayes algorithm had an accuracy of 0.51, a precision of 0.54, and an f1-score of 0.15. The Random Forest algorithm had an accuracy of 0.92, a precision of 0.93, and an f1-score of 0.91. The Logistic Regression algorithm had an accuracy of 0.89, a precision of 0.88, and an f1-score of 0.86.

Overall, it can be seen that the SVM and Random Forest algorithms performed the best in both data sets, with the SVM algorithm having the highest f1-score for the Parkinson's Disease data set and the Random Forest algorithm having the highest accuracy and f1-score for the Schizophrenia data set. The k-nearest neighbours, Gaussian Naive Bayes, and Logistic Regression algorithms had lower performance than the SVM and Random Forest algorithms. It is important to note that the results may vary depending on the specific data set and the chosen machine learning model and hyperparameters.

In the second experiment, conventional machine learning techniques were applied to data sets for the diagnosis of Schizophrenia and Parkinson's Disease using features extracted from the Average Band Powers of the EEG signal. The techniques tested included Support Vector Machine, k-nearest neighbours, Gaussian Naive Bayes, Random Forest, and Logistic Regression. A window size of 2 seconds was used and the algorithms were hyperparameter tuned using the GridSearch technique. The results of the experiment are shown in the following tables:

Table Parkinson’s Disease Data Set - Features: Average Band Power

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Precision | F1-Score |
| SVM | 0.81 | 0.73 | 0.61 |
| K-NN | 0.64 | 0.56 | 0.53 |
| GaussianNB | 0.70 | 0.43 | 0.39 |
| Random Forest | 0.81 | 0.73 | 0.69 |
| Logistic Regression | 0.76 | 0.64 | 0.48 |

Table Schizophrenia Data Set - Features: Average Band Power

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Precision | F1-Score |
| SVM | 0.64 | 0.24 | 0.12 |
| K-NN | 0.67 | 0.63 | 0.58 |
| GaussianNB | 0.62 | 0.08 | 0.06 |
| Random Forest | 0.72 | 0.71 | 0.68 |
| Logistic Regression | 0.68 | 0.60 | 0.36 |

Figure ROC PD- Average Band Power Random Forest

Chart, line chart

Description automatically generated

Figure ROC SZ Average Band Power

Chart, line chart

Description automatically generated

For the Parkinson's Disease data set, the SVM algorithm had an accuracy of 0.81, a precision of 0.73, and an f1-score of 0.61. The k-nearest neighbour’s algorithm had an accuracy of 0.64, a precision of 0.56, and an f1-score of 0.53. The Gaussian Naive Bayes algorithm had an accuracy of 0.70, a precision of 0.43, and an f1-score of 0.39. The Random Forest algorithm had the highest accuracy of 0.79, a precision of 0.73, and an f1-score of 0.69. The Logistic Regression algorithm had an accuracy of 0.76, a precision of 0.64, and an f1-score of 0.51.

For the Schizophrenia data set, the SVM algorithm had an accuracy of 0.64, a precision of 0.24, and an f1-score of 0.12. The k-nearest neighbour’s algorithm had an accuracy of 0.67, a precision of 0.63, and an f1-score of 0.58. The Gaussian Naive Bayes algorithm had an accuracy of 0.62, a precision of 0.08, and an f1-score of 0.06. The Random Forest algorithm had an accuracy of 0.72, a precision of 0.71, and an f1-score of 0.68. The Logistic Regression algorithm had an accuracy of 0.68, a precision of 0.60, and an f1-score of 0.36.

In terms of evaluation metrics, accuracy is the proportion of correct predictions made by the model, while precision is the proportion of correct positive predictions made by the model. The f1-score is a combination of precision and recall, which is the proportion of correct positive predictions made by the model out of all positive predictions made.

**Comparative Analysis with Previous Studies**

Table Comparing our work to previous approaches for SZ detection.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Authors | Techniques | Subjects | Accuracy |
| 1 | (Jahmunah et al., 2019) | Digital filtering, non-linear features extraction, dimension reduction with Student's t-test, SVM classifier | Normal: 14, SZ: 14 | 92.91% |
| 2 | (Dvey-Aharon et al., 2015) | Channel selection, time-frequency analysis-based features extraction, KNN with Leave one out cross validation | Normal: 25, SZ: 25 | 92.70% |
| 3 | (Santos-Mayo et al., 2017) | Digital filtering, time-frequency analysis-based features extraction, LDA based dimension reduction, MLP classifiers | Normal: 31, SZ: 16 | 93.42% |
| 4 | Shim et al. [24], 2016 | "Sensor-level" and "source-level" features extraction , trained SVM using Leave one out cross validation | Normal: 34, SZ: 34 | 88.24% |
| 5 | (Vittala et al., 2020) | MEMD, modes entropy-based features extraction, RFE based dimension reduction, SVM with Radial Basis Function (SVM-RBF) | Normal: 14, SZ: 14 | 93.00% |
| 6 | Present work | Independent Component Analysis, Statistical Features, Averaged EEG power in different frequency Bands, Cross Validation, KNN, Logistic Regression, SVM, Random Forest, Naïve Bayes with Hyper Parameter Tuning. | Normal: 14, SZ: 14 | Statistical features : 92.8% (Random Forest) Average Band Power : 72.3% ( Random Forest) |

Table Comparing our work to previous approaches for Parkinson’s Disease detection.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Authors | Techniques | Subjects | Accuracy |
| 1 | (Oh et al., 2020) | A thirteen-layer CNN architecture, feature extraction, selection, classification not required, Cross Validation | Normal: 20, PD: 20 | 88.25% |
| 2 | (Zhang et al., 2022) | Time–frequency analysis with deep learning, tunable Q-factor wavelet transform with deep residual shrinkage network (TQWT-DRSN) , wavelet packet transform with deep residual shrinkage network (WPT-DRSN) | Normal REM:13, PD REM:12 , REM disorder :11 | Three - Class classification:  97.81% |
| 3 | (Ly et al., 2017b) | Independent component analysis, Bayesian Neural Networks, Timed Up and Go Task experiments. | PD: 6  Normal: n/a | 86.2% |
| 4 | Present work | Independent Component Analysis, Statistical Features, Averaged EEG power in different frequency Bands, Cross Validation, KNN, Logistic Regression, SVM, Random Forest, Naïve Bayes with Hyper Parameter Tuning. | PD: 15  Normal: 16 | Statistical features: 96.1% (Random Forest)  Average Band power: 81.4% (SVM) |

**Conclusion**

According to the findings of the machine learning tests done for this study, average band powers of the EEG signal and characteristics derived using statistical methods can be used to diagnose both schizophrenia and Parkinson's disease more accurately along with conventional approaches. The Random Forest algorithm had the greatest performance on the Parkinson's Disease data set, and the Random Forest and Support Vector Machine algorithms had the best performance on the Schizophrenia data set. Nonetheless, it should be emphasised that the feature set employed affected how well these algorithms performed. Significantly, statistical feature sets outperformed Average Band powers in producing superior outcomes. In order to assess the efficacy of these strategies more accurately, it would be advantageous to carry out studies in the future utilising bigger data sets.

Also, the algorithms' performance was enhanced by hyperparameter tweaking using the GridSearch approach, emphasising the need of carefully choosing the hyperparameters during the machine-learning process.

Overall, the outcomes of these investigations show the promise of machine learning methods for neurological illness diagnosis. To get the most precise and trustworthy diagnoses, it's crucial to keep enhancing and perfecting these methods while also looking into alternative ideas. To further boost the effectiveness of these algorithms, future work can entail adding more features or using more sophisticated machine learning methods, including deep learning.

References

Amzica, F., & Lopes da Silva, F. (2011). Cellular substrates of brain rhythms. *Niedermeyer’s Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, *6*, 33-63.

Dvey-Aharon, Z., Fogelson, N., Peled, A., & Intrator, N. (2015). Schizophrenia detection and classification by advanced analysis of EEG recordings using a single electrode approach. *PLoS One*, *10*(4), e0123033. <https://doi.org/10.1371/journal.pone.0123033>

Jahmunah, V., Lih Oh, S., Rajinikanth, V., Ciaccio, E. J., Hao Cheong, K., Arunkumar, N., & Acharya, U. R. (2019). Automated detection of schizophrenia using nonlinear signal processing methods. *Artif Intell Med*, *100*, 101698. <https://doi.org/10.1016/j.artmed.2019.07.006>

Kalia, L. V., & Lang, A. E. (2015). Parkinson's disease. *The Lancet*, *386*(9996), 896-912.

Khare, S. K., Bajaj, V., & Acharya, U. R. (2021). Detection of Parkinson’s disease using automated tunable Q wavelet transform technique with EEG signals. *Biocybernetics and Biomedical Engineering*, *41*(2), 679-689.

Ly, Q. T., Handojoseno, A. A., Gilat, M., Chai, R., Martens, K. A. E., Georgiades, M., Naik, G. R., Tran, Y., Lewis, S. J., & Nguyen, H. T. (2017a). Detection of gait initiation Failure in Parkinson's disease based on wavelet transform and Support Vector Machine. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),

Ly, Q. T., Handojoseno, A. A., Gilat, M., Chai, R., Martens, K. A. E., Georgiades, M., Naik, G. R., Tran, Y., Lewis, S. J., & Nguyen, H. T. (2017b). Detection of turning freeze in Parkinson's disease based on S-transform decomposition of EEG signals. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),

Oh, S. L., Hagiwara, Y., Raghavendra, U., Yuvaraj, R., Arunkumar, N., Murugappan, M., & Acharya, U. R. (2020). A deep learning approach for Parkinson’s disease diagnosis from EEG signals. *Neural Computing and Applications*, *32*(15), 10927-10933.

Olejarczyk, E., & Jernajczyk, W. (2017). Graph-based analysis of brain connectivity in schizophrenia. *PLoS One*, *12*(11), e0188629.

Owen, M. (2016). Sawa A. *Mortensen pb. Schizophrenia Lancet*, *388*, 86-97.

Rockhill, A. P., Jackson, N., George, J., Aron, A., & Swann, N. C. (2020). uc san diego resting state eeg data from patients with Parkinson’s disease. *Available:*.

Santos-Mayo, L., San-Jose-Revuelta, L. M., & Arribas, J. I. (2017). A Computer-Aided Diagnosis System With EEG Based on the P3b Wave During an Auditory Odd-Ball Task in Schizophrenia. *IEEE Trans Biomed Eng*, *64*(2), 395-407. <https://doi.org/10.1109/TBME.2016.2558824>

Vittala, A., Murphy, N., Maheshwari, A., & Krishnan, V. (2020). Understanding Cortical Dysfunction in Schizophrenia With TMS/EEG. *Front Neurosci*, *14*, 554. <https://doi.org/10.3389/fnins.2020.00554>

Welch, P. (1967). The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, *15*(2), 70-73.

Yuvaraj, R., Rajendra Acharya, U., & Hagiwara, Y. (2018). A novel Parkinson’s Disease Diagnosis Index using higher-order spectra features in EEG signals. *Neural Computing and Applications*, *30*(4), 1225-1235.

Zhang, R., Jia, J., & Zhang, R. (2022). EEG analysis of Parkinson's disease using time–frequency analysis and deep learning. *Biomedical Signal Processing and Control*, *78*, 103883.