VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Santhibastawad Road, Machhe
Belagavi - 590018, Karnataka, India

PROJECT WORK PHASE- 1 (18CSP77) REPORT

ON

“Anomaly detection in Parkinson's disease using unsupervised machine learning Approach”

Submitted in the partial fulfillment of the requirements for the award of the degree of

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

For the Academic Year 2021-2022

Submitted by

Abhay PJ	(1JS18IS001)
Hemanth Kumar A	(1JS19IS406)
Jeevan KV	(1JS18IS039)
Manoj GH	(1JS18IS048)

Under the Guidance of
Dr Malini M Patil
Associate Professor, Department of Information Science and Engineering

2021-2022

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

JSS ACADEMY OF TECHNICAL EDUCATION

JSS Campus, Dr.Vishnuvardhan Road, Bengaluru-560060

JSS MAHAVIDYAPEETHA, MYSURU

JSS ACADEMY OF TECHNICAL EDUCATION

JSS Campus, Dr.Vishnuvardhan Road, Bengaluru-560060

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that Project Work Phase -1(18CSP77) Report entitled “**Anomaly detection in Parkinson's disease using unsupervised machine learning Approach**” is a bonafide work carried out by Abhay PJ [1JS18IS001], Hemanth Kumar A [1JS19IS406], Jeevan KV [1JS18IS039], Manoj GH [1JS18IS048] in partial fulfillment for the award of degree of Bachelor of Engineering in Information Science and Engineering of Visvesvaraya Technological University Belagavi during the year 2021- 2022.

Signature of the Guide

Dr. Malini M Patil
Associate Professor
Dept. of ISE
JSSATE, Bengaluru

Signature of the HOD

Dr. Rekha P M
Associate Professor & Head
Dept. of ISE
JSSATE, Bengaluru

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible. So with gratitude, we acknowledge all those whose guidance and encouragement crowned my effort with success.

First and foremost we would like to express our heartfelt deep sense of gratitude to **His Holiness. Jagadguru Sri Shivarathri Deshikendra Mahaswamiji**, whose divine blessings have motivated and helped us in completing this project work.

Dr. Mrityunjaya V Latte, Principal, JSSATE, Bangalore for providing an opportunity to carry out the Project Work Phase – 1 (18CSP77) as a part of our curriculum in the partial fulfillment of the degree course.

We express our sincere gratitude for our beloved Head of the department, Dr**. Rekha P M**, for her co-operation and encouragement at all the moments of our approach.

It is our pleasant duty to place on record our deepest sense of gratitude to our respected guide **Dr. Malini M Patil, Associate Professor,** for the constant encouragement, valuable and timely guidance in every possible way in completing this work.

We are thankful to the Project Coordinators **Dr. Nagamani N P,** Asst. Professor and **Mrs. Sahana V** Asst. Professor, for their continuous co-operation and support.

We would like to thank all **ISE department teachers** and **non teaching staff** for providing us with their valuable guidance and for being there at all stages of our work.

Abhay PJ	(1JS18IS001)
Hemanth Kumar A	(1JS19IS406)
Jeevan KV	(1JS18IS039)
Manoj GH	(1JS18IS048)

TABLE OF CONTENTS

ABSTRACT:

This Project focuses on anomaly detection techniques on Parkinson’s disease, the insight generated by the algorithm is useful for doctors to diagnose the patient efficiently. It can help the patient with symptoms to diagnose whether he/she is positive or negative of Parkinson’s disease. Parkinson's Outlier analysis is very useful in real life as it is very difficult to identify a specific symptoms as every individual go through different type of symptoms and by going through all undesired symptoms we can get a brief knowledge on the disease.

The Project uses Google collab as the AI/ML platform and Tableau as the visualization tool. The Front end of the project is developed using Flask/Django with AWS/Azure as the service provider. Accuracy of Parkinson’s disease is very less at early stage. The possibility of getting accurate prediction is low therefore we study the undesirable symptoms leading to Parkinson’s disease by this manner we get a bigger picture of Parkinson’s symptoms.

INTRODUCTION:

What is artificial intelligence?

Artificial intelligence leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind.

While a number of definitions of artificial intelligence (AI) have surfaced over the last few decades, John McCarthy offers the following definition in this 2004 paper, " Artificial Intelligence is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable."

Application of AI?

Artificial Intelligence has various applications in today's society. It is becoming essential for today's time because it can solve complex problems with an efficient way in multiple industries, such as Healthcare, entertainment, finance, education, etc. AI is making our daily life more comfortable and fast. Figure 1 illustrates the application of artificial intelligence in real world.

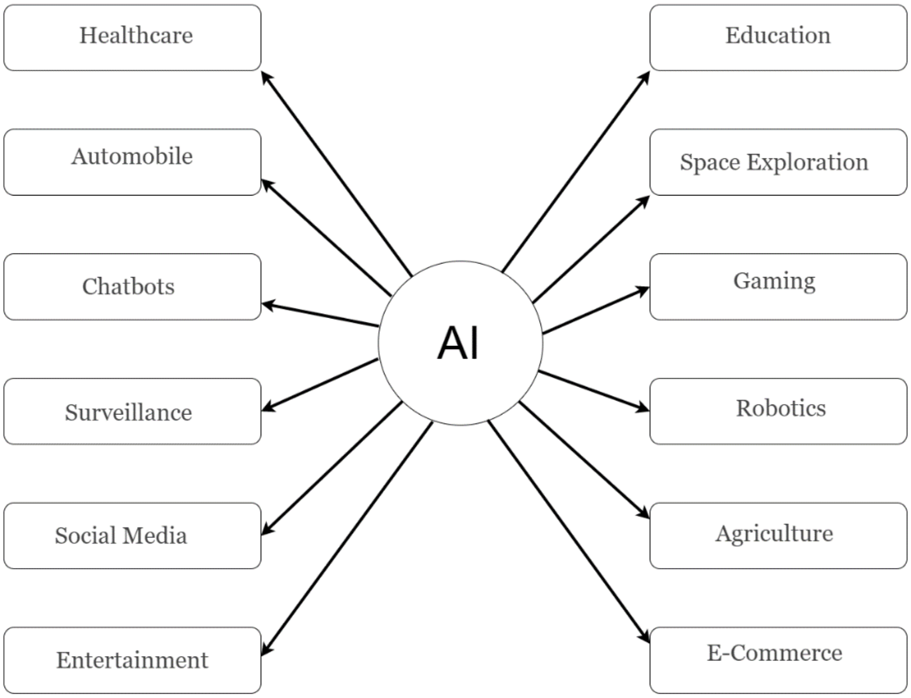

Figure 1. Application of Artificial Intelligence

What is Machine learning?

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

Methods of Machine learning

Machine learning classifiers fall into three primary categories.

- **Supervised machine learning**

Supervised learning, also known as supervised machine learning, is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids over fitting or under fitting.

- **Unsupervised machine learning**

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, image and pattern recognition.

- **Semi-supervised learning**

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of having not enough labeled data (or not being able to afford to label enough data) to train a supervised learning algorithm.

- **Reinforcement Learning:**

Reinforcement learning is a type of machine learning method where an intelligent agent (computer program) interacts with the environment and learns to act within that.

Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.

Data Mining

Data mining is a process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.

Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. . Figure 2 explains about Data mining process and Figure 3 explains Data mining as core in knowledge

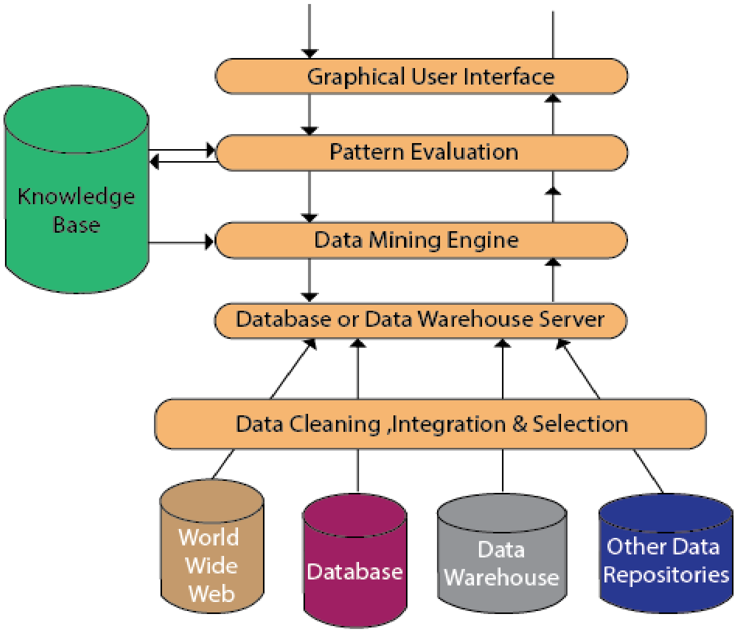

Figure 2. Data mining process

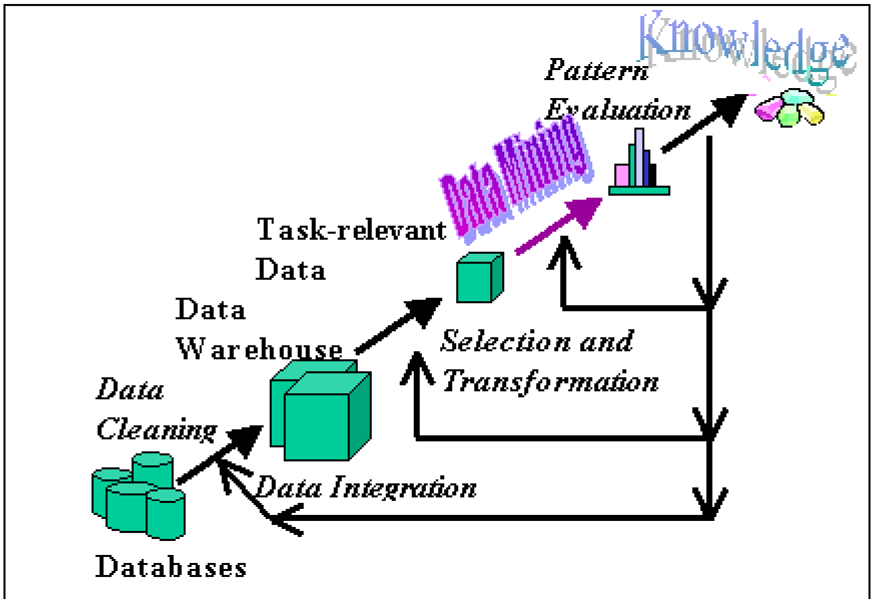

Figure 3. Data mining as core in knowledge

Data mining techniques

Data mining works by using various algorithms and techniques to turn large volumes of data into useful information. Here are some of the most common ones: Figure 4 explains Data mining Technique

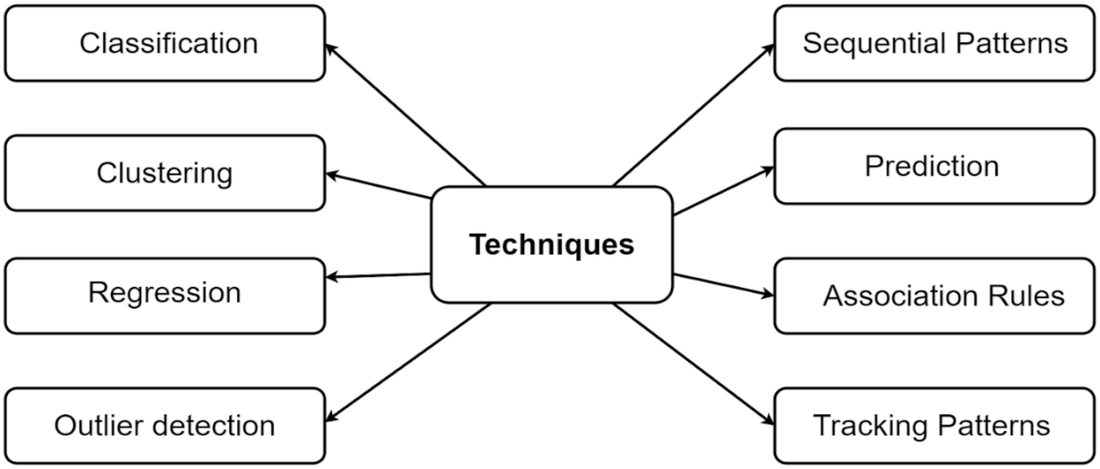

Figure 4. Data mining technique

Outlier Mining

Outliers are extreme values that deviate from other observations on data, they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample.

Outlier detection in high-dimensional datasets is a fundamental and challenging problem across disciplines that has also practical implications, as removing outliers from the training set improves the performance of machine learning algorithms. Figure 5 illustrates Outlier in Scatter Plot

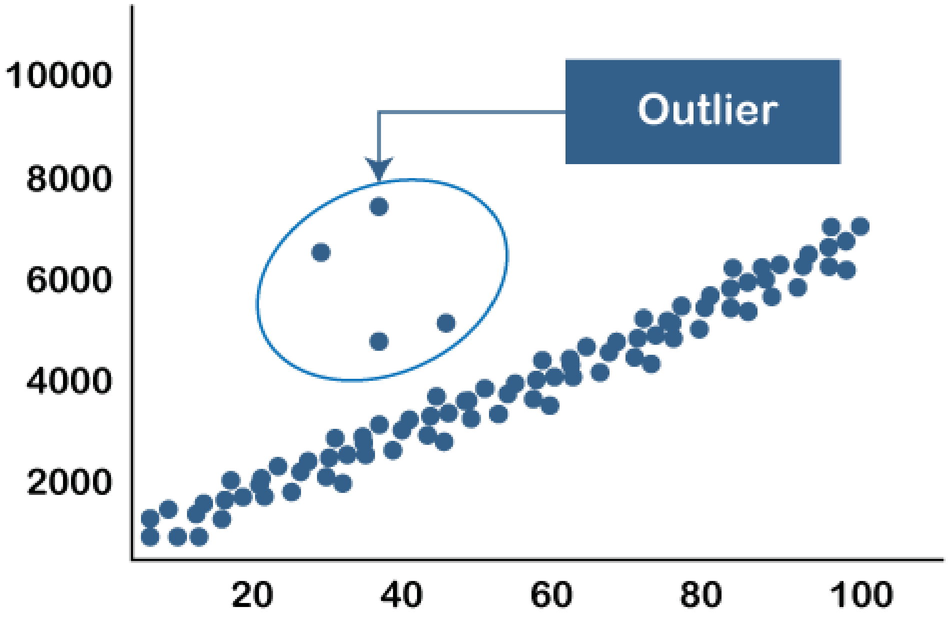

Figure 5. Outlier in Scatter Plot

Types of outliers

Outliers can be of two kinds: univariate and multivariate.

- **Univariate**: This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a unilabiate data can be height.
- **Multivariate**: When the data involves three or more variables, it is categorized under multivariate. Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined. It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).

Parkinson's disease

Parkinson's disease is a brain disorder that leads to shaking, stiffness, and difficulty with walking, balance, and coordination.

Parkinson's symptoms usually begin gradually and get worse over time. As the disease progresses, people may have difficulty walking and talking. They may also have mental and behavioral changes, sleep problems, depression, memory difficulties, and fatigue.

Symptoms of Parkinson’s disease

Parkinson's disease has four main symptoms:

- Tremor (trembling) in hands, arms, legs, jaw, or head
- Stiffness of the limbs and trunk
- Slowness of movement
- Impaired balance and coordination, sometimes leading to falls

Treatment of Parkinson’s disease

Although there is no cure for Parkinson's disease, medicines, surgical treatment, and other therapies can often relieve some symptoms.

Medicines for Parkinson's Disease. Medicines prescribed for Parkinson's include:

- Drugs that increase the level of dopamine in the brain
- Drugs that affect other brain chemicals in the body
- Drugs that help control no motor symptoms

LITERATURE SURVEY:

1. Mei J, Desrosiers C, Frasnelli J. Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature. Front Aging Neurosci. 2021 May 6;13:633752. doi: 10.3389/fnagi.2021.633752. PMID: 34025389; PMCID: PMC8134676.

Diagnosis of Parkinson’s disease (PD) is commonly based on medical observations and assessment of clinical signs, including the characterization of a variety of motor symptoms. However, traditional diagnostic approaches may suffer from subjectivity as they rely on the evaluation of movements that are sometimes subtle to human eyes and therefore difficult to classify, leading to possible misclassification. In the meantime, early non-motor symptoms of PD may be mild and can be caused by many other conditions. Therefore, these symptoms are often overlooked, making diagnosis of PD at an early stage challenging. To address these difficulties and to refine the diagnosis and assessment procedures of PD, machine learning methods have been implemented for the classification of PD and healthy controls or patients with similar clinical presentations (e.g., movement disorders or other Parkinsonian syndromes). To provide a comprehensive overview of data modalities and machine learning methods that have been used in the diagnosis and differential diagnosis of PD

2. Cheng, Zhangyu, Zou, Chengming, Dong, Jianwei, 2019/09/24, 161, 168, 978-1-4503-6843-8, Outlier detection using isolation forest and local outlier factor, 10.1145/3338840.3355641, RACS '19: Proceedings of the Conference on Research in Adaptive and Convergent Systems.

Outlier detection, also named as anomaly detection, is one of the hot issues in the field of data mining. As well-known outlier detection algorithms, Isolation Forest(iForest) and Local Outlier Factor(LOF) have been widely used. However, iForest is only sensitive to global outliers, and is weak in dealing with local outliers. Although LOF performs well in local outlier detection, it has high time complexity. To overcome the weaknesses of iForest and LOF, a two-layer progressive ensemble method for outlier detection is proposed. It can accurately detect outliers in complex datasets with low time complexity. This method first utilizes iForest with low complexity to quickly scan the dataset, prunes the apparently normal data, and generates an outlier candidate set. In order to further improve the pruning accuracy, the outlier coefficient is introduced to design a pruning threshold setting method, which is based on outlier degree of data. Then LOF is applied to further distinguish the outlier candidate set and get more accurate outliers. The proposed ensemble method takes advantage of the two algorithms and concentrates valuable computing resources on the key stage. Finally, a large number of experiments are carried out to verify the ensemble method. The results show that compared with the existing methods, the ensemble method can significantly improve the outlier detection rate and greatly reduce the time complexity.

3. McDonnell MN, Rischbieth B, Schammer TT, Seaforth C, Shaw AJ, Phillips AC. Lee Silverman Voice Treatment (LSVT)-BIG to improve motor function in people with Parkinson's disease: a systematic review and meta-analysis. Clin Rehabil. 2018 May;32(5):607-618. doi: 10.1177/0269215517734385. Epub 2017 Oct 5. PMID: 28980476.

The technique called Lee Silverman Voice Treatment (LSVT)-LOUD has previously been used to improve voice quality in people with Parkinson's disease. The objective of this study was to assess the effectiveness of an alternate intervention, LSVT-BIG (signifying big movements), to improve functional mobility.

4. W. Wang, J. Lee, F. Harrou and Y. Sun, "Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning," in *IEEE Access*, vol. 8, pp. 147635-147646, 2020, doi: 10.1109/ACCESS.2020.3016062.

Accurately detecting Parkinson's disease (PD) at an early stage is certainly indispensable for slowing down its progress and providing patients the possibility of accessing to disease-modifying therapy. Towards this end, the premotor stage in PD should be carefully monitored. An innovative deep-learning technique is introduced to early uncover whether an individual is affected with PD or not based on premotor features. Specifically, to uncover PD at an early stage, several indicators have been considered in this study, including Rapid Eye Movement and olfactory loss, Cerebrospinal fluid data, and dopaminergic imaging markers. A comparison between the proposed deep learning model and twelve machine learning and ensemble learning methods based on relatively small data including 183 healthy individuals and 401 early PD patients shows the superior detection performance of the designed model, which achieves the highest accuracy, 96.45% on average. Besides detecting the PD, we also provide the feature importance on the PD detection process based on the Boosting method.

5. T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins and R. H. Ghomi, "Parkinson's Disease Diagnosis Using Machine Learning and Voice," *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2018, pp. 1-7, doi: 10.1109/SPMB.2018.8615607.

This paper explores the effectiveness of using supervised classification algorithms, such as deep neural networks, to accurately diagnose individuals with the disease. Our peak accuracy of 85% provided by the machine learning models exceed the average clinical diagnosis accuracy of non-experts (73.8%) and average accuracy of movement disorder specialists (79.6% without follow-up, 83.9% after follow-up) with pathological post-mortem examination as ground truth [3]

6. JOUR, Maitin, Ana, García-Tejedor, Alvaro, Romero Muñoz, Juan Pablo, 2020/12/03, 8662,
T1 - Machine Learning Approaches for Detecting Parkinson's Disease from EEG Analysis: A Systematic Review, 10, 0.3390/app10238662, Applied Sciences

The review process was performed following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. All publications previous to May 2020 were included, and their main characteristics and results were assessed and documented. Results: Nine studies were included. Seven used resting state EEG and two motor activation EEG. Subsymbolic models were used in 83.3% of studies. The accuracy for PD classification was 62–99.62%. There was no standard cleaning protocol for the EEG and a great heterogeneity in the characteristics that were extracted from the EEG. However, spectral characteristics predominated. Conclusions: Both the features introduced into the model and its architecture were essential for a good performance in predicting the classification. On the contrary, the cleaning protocol of the EEG, is highly heterogeneous among the different studies and did not influence the results. The use of ML techniques in EEG for neurodegenerative disorders classification is a recent and growing field.

7. Wang, C., Liu, Z., Gao, H., & Fu, Y. (2019). Applying anomaly pattern score for outlier detection. IEEE Access, 7, 16008-16020.

Outlier detection is an important sub-field of data mining and studied intensively by researchers in the past decades. For neighborhood-based outlier detection methods like KNN and LOF, different settings in the number of neighbors (indicated by a parameter k) would greatly affect the model's performance. Thereby, there are some recent studies which focus on identifying the optimal value of k by analyzing the global or local structure of the dataset.

8. Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. Ieee Access, 7, 107964-108000.

Detecting outliers is a significant problem that has been studied in various research and application areas. Researchers continue to design robust schemes to provide solutions to detect outliers efficiently. In this survey, they have presented a comprehensive and organized review of the progress of outlier detection methods from 2000 to 2019. First, they have offered the fundamental concepts of outlier detection and then categorized them into different techniques from diverse outlier detection techniques, such as distance, clustering density ensemble, and learning-based methods.

9. Singh, K., & Upadhyaya, S. (2012). Outlier detection: applications and techniques. International Journal of Computer Science Issues (IJCSI), 9(1), 307.

Outliers is regarded as noisy data in statistics and has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques have been developed specific to certain application domains, while some techniques are more generic. Some application domains are being researched in strict confidentiality such as research on crime and terrorist activities. The techniques and results of such techniques are not readily forthcoming. A number of surveys, research and review articles and books cover outlier detection techniques in machine learning and statistical domains individually in great detail. This paper has made an attempt to bring together various outlier detection techniques, in a structured and generic description.

10. Djenouri, Y., Belhadi, A., Lin, J. C. W., Djenouri, D., & Cano, A. (2019). A survey on urban traffic anomalies detection algorithms. IEEE Access, 7, 12192-12205.

This paper reviews the use of outlier detection approaches in urban traffic analysis. They have divided existing solutions into two main categories: flow outlier detection and trajectory outlier detection. The first category groups solutions that detect flow outliers and includes statistical, similarity and pattern mining approaches. The second category contains solutions where the trajectory outliers are derived, including off- line processing for trajectory outliers and online processing for sub-trajectory outliers. Solutions in each of these categories are described, illustrated, and discussed, and open perspectives and research trends are drawn.

11. Park, C. H. (2019). Outlier and anomaly pattern detection on data streams. The Journal of Supercomputing, 75(9), 6118-6128.

A data stream is a sequence of data generated continuously over time. A data stream is too big to be saved in memory, and its underlying data distribution may change over time. Outlier detection aims to find data instances which significantly deviate from the underlying data distribution. While most outlier detection methods work in batch mode where all the data samples are available at once, the necessity for efficient outlier and anomaly pattern detection methods in a data stream has increased. Outlier detection is performed at an individual instance level, and anomalous pattern detection involves detecting a point in time where the behavior of the data becomes unusual and differs from normal behavior.

12. JOUR, Singh, Karanjit, Upadhyaya, Shuchita, 2012/01/01, Outlier Detection: Applications And Techniques, 9, International Journal of Computer Science Issues

In this paper we make an attempt to bring together various outlier detection techniques, in a structured and generic description. With this exercise, we hope to attain a better understanding of the different directions of research on outlier analysis for ourselves as well as for beginners in this research field who could then pick up the links to different areas of applications in details.

13. H. Wang, M. J. Bah and M. Hammad, "Progress in Outlier Detection Techniques: A Survey," in IEEE Access, vol. 7, pp. 107964-108000, 2019, doi: 10.1109/ACCESS.2019.2932769.

This paper gives current progress of outlier detection techniques and provides a better understanding of the different outlier detection methods. The open research issues and challenges at the end will provide researchers with a clear path for the future of outlier detection methods.

14. BOOK, Liu, Fei Tony, Ting, Kai, Zhou, Zhi-Hua, 2009/01/19, 413 , 422, Isolation Forest, 10.1109/ICDM.2008.17

This paper proposes a fundamentally different model-based method that explicitly isolates anomalies instead of profiles normal points. To our best knowledge, the concept of isolation has not been explored in current literature. The use of isolation enables the proposed method, iForest, to exploit sub-sampling to an extent that is not feasible in existing methods, creating an algorithm which has a linear time complexity with a low constant and a low memory requirement. Our empirical evaluation shows that iForest performs favourably to ORCA, a near-linear time complexity distance-based method, LOF and random forests in terms of AUC and processing time, and especially in large data sets. iForest also works well in high dimensional problems which have a large number of irrelevant attributes, and in situations where training set does not contain any anomalies.

15. Evgeniou, Theodoros, Pontil, Massimiliano, 2001/01/01, 249, 257, Support Vector Machines: Theory and Applications, 2049, 10.1007/3-540-44673-7_12

The goal of the chapter is twofold: to present an overview of the background theory and current understanding of SVM, and to discuss the papers presented as well as the issues that arose during the workshop. Support Vector Machines (SVM) have been recently developed in the framework of statistical learning theory, and have been successfully applied to a number of applications, ranging from time series prediction, to face recognition, to biological data processing for medical diagnosis. Their theoretical foundations and their experimental success encourage further research on their characteristics, as well as their further use.

16. Y. Ma and X. Zhao, "POD: A Parallel Outlier Detection Algorithm Using Weighted kNN," in *IEEE Access*, vol. 9, pp. 81765-81777, 2021, doi: 10.1109/ACCESS.2021.3085605.

The method first applies information entropy to calculate each attribute weight, and then uses the Z-order curve to encode high-dimensional data into Z-value. The weighted kNN of each object are searched according to its Z-value. Meanwhile, a novel outlier detection algorithm is presented based on the minimum distance and average distance between each object and its weighted kNN. On this basis, we propose a parallel outlier detection algorithm called POD to improve the efficiency of the outlier detection.

17. Anila M , Dr. G. Pradeepini, 2020, A Review on Parkinson’s Disease Diagnosis using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 06 (June 2020),

This paper is a survey of predicting Parkinson disease using machine learning algorithms, various new technologies applied, and their accuracies achieved.

18. T. M. Thang and J. Kim, "The Anomaly Detection by Using DBSCAN Clustering with Multiple Parameters," ***2011 International Conference on Information Science and Applications*****, 2011, pp. 1-5, doi: 10.1109/ICISA.2011.5772437**

In this paper, we propose a new way of finding DBSCAN's parameters and applying DBSCAN with those parameters. Each cluster may have different epsilon and minpts values in our algorithm. The algorithm is called DBSCAN-MP. We also propose a mechanism of updating normal behavior by updating size or creating new clusters when network environment is changing overtime. We evaluate proposed algorithm using the KDD Cup 1999 dataset. The result shows that the performance is improved compare to other clustering algorithms.

19. Fourure, Damien, Javaid, Muhammad, Posocco, Nicolas, Tihon, Simon, 2021/09/10, 3, 18, 978-3-030-86513-9, Anomaly Detection: How to Artificially Increase Your F1-Score with a Biased Evaluation Protocol, 10.1007/978-3-030-86514-6_1

In this paper, we show that F1-score and AVPR are highly sensitive to the contamination rate. One consequence is that it is possible to artificially increase their values by modifying the train-test split procedure. This leads to misleading comparisons between algorithms in the literature, especially when the evaluation protocol is not well detailed. Moreover, we show that the F1-score and the AVPR cannot be used to compare performances on different datasets as they do not reflect the intrinsic difficulty of modeling such data. Based on these observations, we claim that F1-score and AVPR should not be used as metrics for anomaly detection. We recommend a generic evaluation procedure for unsupervised anomaly detection, including the use of other metrics such as the AUC, which are more robust to arbitrary choices in the evaluation protocol.

20. Kaur, Parmeet, 2016/12/01, 693, 696, Outlier Detection Using Kmeans and Fuzzy Min Max Neural Network in Network Data, 10.1109/CICN.2016.142

In this paper, we propose a kmean clustering and neural network as novel to detect the outlier in network analysis. Especially in a social network, k means clustering and neural network is used to find the community overlapped user in the network as well as it finds more kclique which describe the strong coupling of data. In this paper, we propose that this method is efficient to find out outlier in social network analyses. Moreover, we show the effectiveness of this new method using the experiments data.

PROBLEM IDENTIFICATION :

The traditional methods for data mining include classification, clustering, and association rule mining. Most of the time, knowledge discovery is performed through these methods but, in the last few years, the use of outlier mining has emerged as a promising technique. There has been a limited amount of research related to the use of these methods in detecting anomalous data. This project focuses on detecting anomalous features of Parkinson's dataset. We cross check performance of the outlier detection model on real time dataset. Realtime data stream analysis is been performed to check the OSE model plot on the real time dataset.

OBJECTIVE :

1.Data Pre-Processing

Since data will likely be imperfect, containing inconsistencies and redundancies is not directly applicable for starting a data mining process. The bigger amounts of data collected require more sophisticated mechanisms to analyze it. Data preprocessing is able to adapt the data to the requirements posed by each data mining algorithm, enabling it to process data that would be unfeasible.

2.Data Modelling:

To detect an outlier is by graphing the features or the data points. Visualization is one of the best and easiest ways to have an inference about the overall data and the outliers. Scatter plots and box plots are the most preferred visualization tools to detect outliers. The outlier can also be detected by using methods like Isolation technique, SVM and NN.

3.Outlier Analysis:

By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Visualization of the insight gain is represented and monitored using visualization tools like tableau, power BI, Azure analysis.

4.Performance Analysis:

The models are evaluated based on performance parameters such as confusion matrix, accuracy score, area under curve of Receiver Operating Characteristic (AUC-ROC) of the model, Mean Square Error of the model and other performance parameters.

METHODOLOGY:

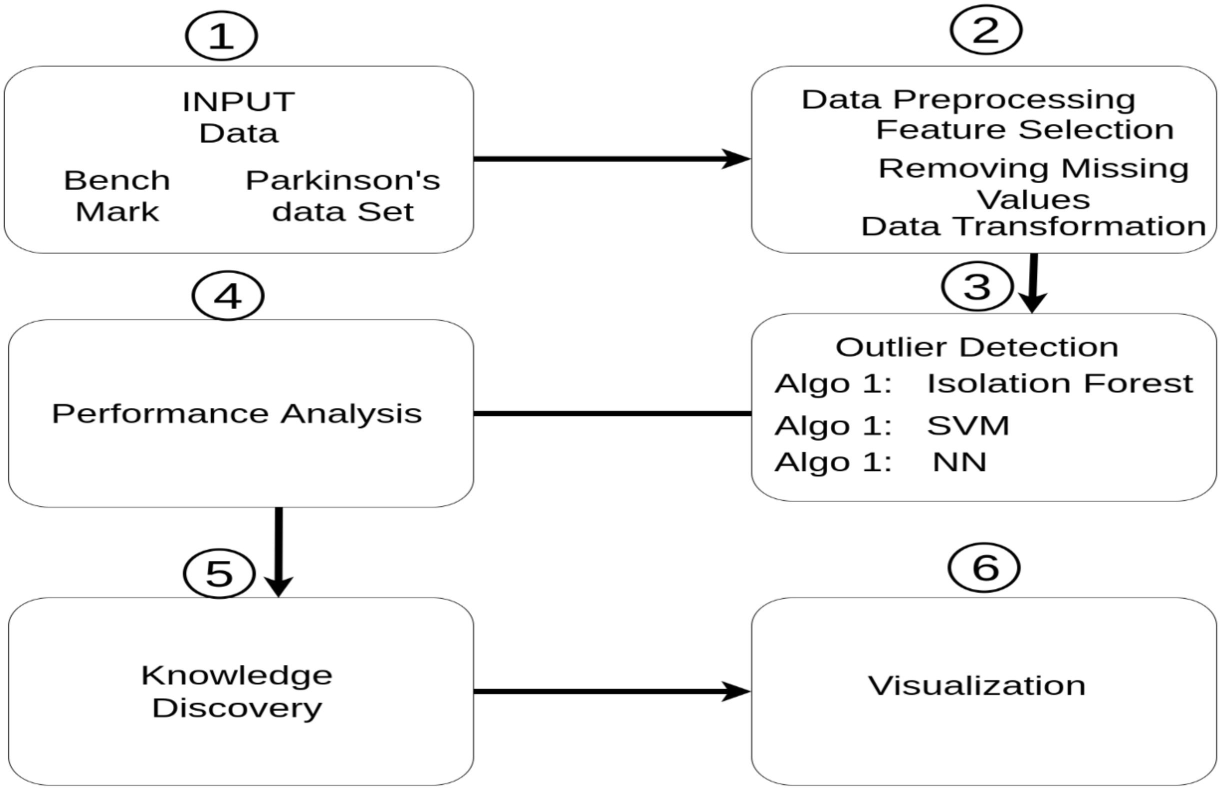

Figure 6. System Design

1. **Input Data**: The Data is inputted to the system by two different methods. Bench dataset: The Data is collected by various online websites like Kaggle, UCI, etc. The data is used as a bench data set where this static dataset is preprocessed and used to get insight of the data creating the standards Real time DataStream: The DataStream generated from real time api is sent to the model. The model is tested by keeping the bench dataset as standards and generating the output by undergoing the artificial intelligence methods

2. **Data Preprocessing**: Since data will likely be imperfect, containing inconsistencies and redundancies is not directly applicable for starting a data mining process. The bigger amounts of data collected require more sophisticated mechanisms to analyze it. Data preprocessing is able to adapt the data to the requirements posed by each data mining algorithm,enabling it to process data that would be unfeasible otherwise. He former includes data transformation, integration, cleaning and normalization; final data set obtained can be regarded as a reliable and suitable source for any data mining algorithm applied,afterwards.

3. **Outlier Detection**: To detect an outlier is by graphing the features or the data points. Visualization is one of the best and easiest ways to have an inference about the overall data

and the outliers. Scatter plots and box plots are the most preferred visualization tools to detect outliers. The outlier can also be detected by using methods like Isolation technique, SVM and NN

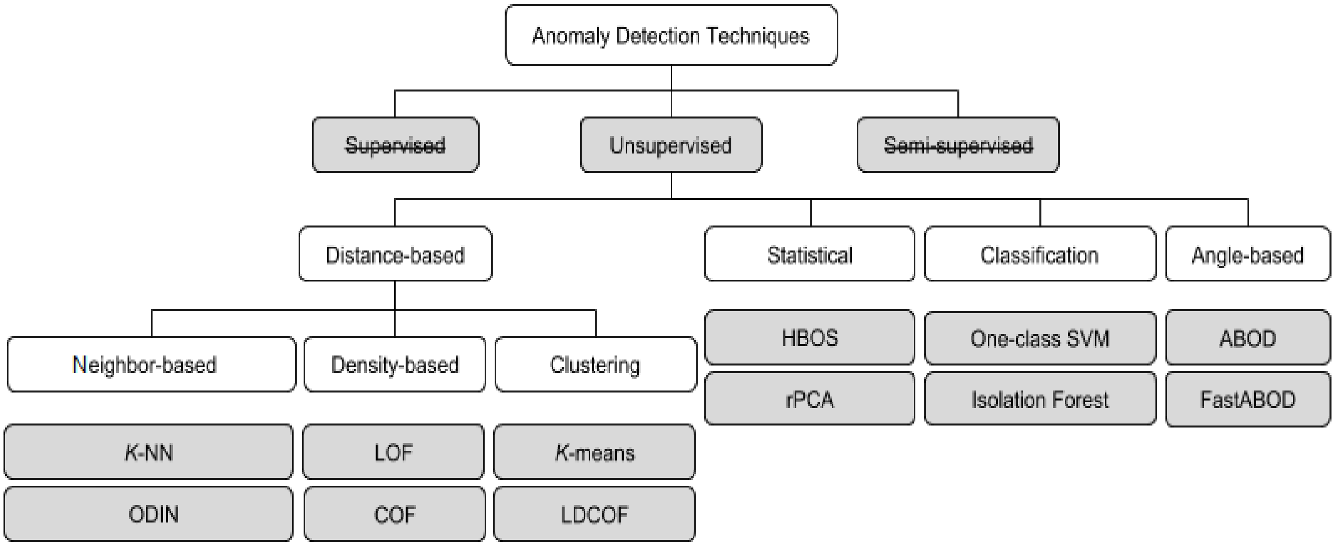

Figure 7.Outlier Detection Methods

4. **Performance Analysis:** Performance analysis at an early stage in software development is crucial. Until now, many researchers have developed statistical techniques. Various Machine Learning (ML) models used for classification of subjects as having average stress or high stress. Percentage are recorded for all subjects after they are introduced to stress inducing stimuli. From the PPG signals Pulse Rate Variability (PRV) parameters are calculated and on the basis of these PRV parameters the subject is either classified as having average stress or high stress. A dataset is framed from the PRV parameters of the subjects and ML models namely Logistic Regression, Support Vector Machine (SVM), Decision Tree and Random Forest are trained and tested, for classification of subjects as average stress or high stress. The models are evaluated based on performance parameters such as confusion matrix, accuracy score, area under curve of Receiver Operating Characteristic (AUC-ROC) of the model, Mean Square Error of the model and other performance parameters.

5. **Knowledge Discovery**: Knowledge discovery may be defined as the development of new tacit or explicit knowledge from data and information or from the synthesis of prior knowledge. This happens through communication, integration, and systemization of multiple streams of explicit knowledge.

6. **Visualization:** Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Visualization of the insight gain is represented and monitored using visualization tools like tableau, power BI, Azure analysis.

EXPECTED OUTCOME OF THE PROPOSED RESEARCH:

- The outcome of the model will contain a list of graphical diagrams and knowledge which is gained by using the outlier mining. We will perform accuracy testing on the outcome to validate if the input given result proper status of Parkinson's disease.
- We will be using various visualization tools to evaluate the algorithm used and ensure that the meaning of the visual graph is insightful. We will validate the bench dataset using the real time approach.
- Our Anomaly detection model will come across as a very useful application, to a wide range of doctors studying on Parkinson's disease to understand the outliers in the patience and help them to diagnose properly.

CONCLUSION:

- The main strength of this project is that identification of potential outliers as it is important for increasing the accuracy of the model. An outlier may indicate bad data .The outlier mining can give meaning insight about the dataset and helps in analysis the dataset more accurately. By understanding the errors we can analyse the reason behind the outlier obtained from the data source
- By analysing the outlier data points of Parkinson's disease, the doctors can analyse the Parkinson's patients more efficiently.
- By the interactive process the data inputted from the sensors can be analysed quickly and no input sensed is neglected as outlier uses every bad data sensed to give more accurate results.

REFERENCES:

1. https://www.ibm.com/search?lang=en&cc=in&q=ai%20ml
2. http://www.aiml.foundation/
3. https://trends.google.com/trends/?geo=IN
4. https://www.mindmeister.com/?utm_source=google&utm_medium=cpc&utm_campaign =ind_en_search&utm_content=mm&gclid=
5. Kaur, Parmeet, 2016/12/01, 693, 696, Outlier Detection Using Kmeans and Fuzzy Min Max Neural Network in Network Data, 10.1109/CICN.2016.142
6. Mathew, Mevin John, and Jomon Baiju. "MACHINE LEARNING TECHNIQUE BASED PARKINSON'S DISEASE DETECTION FROM SPIRAL AND VOICE INPUTS." European Journal of Molecular & Clinical Medicine 7.4: 2020.
7. T. M. Thang and J. Kim, "The Anomaly Detection by Using DBSCAN Clustering with Multiple Parameters," 2011 International Conference on Information Science and Applications, 2011, pp. 1-5, doi: 10.1109/ICISA.2011.5772437
8. Anila M , Dr. G. Pradeepini, 2020, A Review on Parkinson's Disease Diagnosis using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 06 (June 2020),
9. Y. Ma and X. Zhao, "POD: A Parallel Outlier Detection Algorithm Using Weighted kNN," in IEEE Access, vol. 9, pp. 81765-81777, 2021, doi: 10.1109/ACCESS.2021.3085605.
10. Evgeniou, Theodoros, Pontil, Massimiliano, 2001/01/01, 249, 257, Support Vector Machines: Theory and Applications, 2049, 10.1007/3-540-44673-7_12
11. BOOK, Liu, Fei Tony, Ting, Kai, Zhou, Zhi-Hua, 2009/01/19, 413 , 422, Isolation Forest, 10.1109/ICDM.2008.17
12. H. Wang, M. J. Bah and M. Hammad, "Progress in Outlier Detection Techniques: A Survey," in IEEE Access, vol. 7, pp. 107964-108000, 2019, doi: 10.1109/ACCESS.2019.2932769.
13. JOUR, Singh, Karanjit, Upadhyaya, Shuchita, 2012/01/01, Outlier Detection: Applications And Techniques, 9, International Journal of Computer Science Issues
14. Park, C. H. (2019). Outlier and anomaly pattern detection on data streams. The Journal of Supercomputing, 75(9), 6118-6128.
15. Djenouri, Y., Belhadi, A., Lin, J. C. W., Djenouri, D., & Cano, A. (2019). A survey on urban traffic anomalies detection algorithms. IEEE Access, 7, 12192-12205.

16. Singh, K., & Upadhyaya, S. (2012). Outlier detection: applications and techniques. International Journal of Computer Science Issues (IJCSI), 9(1), 307
17. Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. Ieee Access, 7, 107964-108000.

18. Wang, C., Liu, Z., Gao, H., & Fu, Y. (2019). Applying anomaly pattern score for outlier detection. IEEE Access, 7, 16008-16020.

19. JOUR, Maitin, Ana, García-Tejedor, Alvaro, Romero Muñoz, Juan Pablo, 2020/12/03, 8662,

20. T1 - Machine Learning Approaches for Detecting Parkinson’s Disease from EEG Analysis: A Systematic Review, 10, 0.3390/app10238662, Applied Sciences

21. W. Wang, J. Lee, F. Harrou and Y. Sun, "Early Detection of Parkinson’s Disease Using Deep Learning and Machine Learning," in IEEE Access, vol. 8, pp. 147635-147646, 2020, doi: 10.1109/ACCESS.2020.3016062.

22. McDonnell MN, Rischbieth B, Schammer TT, Seaforth C, Shaw AJ, Phillips AC. Lee Silverman Voice Treatment (LSVT)-BIG to improve motor function in people with Parkinson's disease: a systematic review and meta-analysis. Clin Rehabil. 2018 May;32(5):607-618. doi: 10.1177/0269215517734385. Epub 2017 Oct 5. PMID: 28980476.

23. Mei J, Desrosiers C, Frasnelli J. Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature. Front Aging Neurosci. 2021 May 6;13:633752. doi: 10.3389/fnagi.2021.633752. PMID: 34025389; PMCID: PMC8134676.

24. T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins and R. H. Ghomi, "Parkinson’s Disease Diagnosis Using Machine Learning and Voice," 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), 2018, pp. 1-7, doi: 10.1109/SPMB.2018.8615607.