International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Parkinson’s Disease Detection: Comparative Analysis Between Basic Machine Learning Classifiers and Ensemble Models

Urva Bhatnaggar [a], *Mokshda Dubey* [a], *Udit Saran Bhatnagar* [a], *Aparna Pandey* [b]*

[a] *Student, Department of Computer Science & Engineering, Bhilai Institute of Technology Raipur, Raipur, India*
[b] *Asst. Prof., Department of Computer Science & Engineering, Bhilai Institute of Technology Raipur, Raipur, India*

ABSTRACT

To ensure logical and reliable treatment and examination of Parkinson's disease (PD), it is crucial to have quantitative, reliable, and repeatable estimates of disease stage and severity. In the past 50 years, subjective emotional evaluations have dominated PD research, hindering its progress. However, advancements in machine learning and artificial intelligence (AI) have provided new opportunities to address these challenges. This article presents a comprehensive study utilizing the Parkinson's disease dataset and employing the basic machine learning classifiers and various ensemble algorithms for accurate classification of PD patients. The Parkinson's disease dataset comprises 23 features and 197 instances. Feature extraction strategies such as correlation matrix were employed to identify relevant features. Additionally, supervised ML algorithms including Decision Trees, Support Vector Machine (SVM), and ensemble models like Bagging, and XGBoosting were utilized alongside the KNN classifier for accurate classification and evaluation of the disease. The successful implementation of machine learning techniques and the KNN classifier in this study has immense significance for the diagnosis and treatment of Parkinson's disease. By providing quantitative and objective assessments, computer-based findings can augment and complement clinical evaluations, ensuring improved accuracy and reliability. This approach has the potential to transform PD research and enhance patient care by enabling early detection, better treatment planning, and more personalized therapies. The KNN classifier has a higher accuracy of 94.87% in a dataset with 23 attributions.

Keywords: *Parkinson's disease, Machine learning Classifiers, KNN classifier, Ensemble Models.*

1. Introduction

Parkinson's disease (PD), a progressive neurodegenerative disorder, significantly impacts motor control [1]. First described by Dr. James Parkinson in 1817, this condition extends beyond movement-related issues, often involving various non-motor symptoms. Central to this disease is the degeneration of dopamine-producing neurons situated in the substantia nigra region of the brain. Characterized by a gradual loss of motor function, Parkinson's disease affects the central nervous system and both motor and non-motor systems. The onset of symptoms is typically slow, with non-motor indicators becoming increasingly prevalent as the disease advances [2]. Early signs commonly include tremors, rigidity, bradykinesia (slowness of movement), and difficulties with walking. Patients may also experience cognitive impairment, behavioral changes, sleep disturbances, and sensory issues. Additionally, Parkinson's disease dementia frequently occurs in advanced stages. Motor symptoms of Parkinson's disease arise from the demise of nerve cells within the substantia nigra. This midbrain region supplies dopamine to the basal ganglia, and the decrease in dopamine levels leads to disrupted movement regulation [3]. The exact cause of nerve cell death remains poorly understood, but the accumulation of a protein called alpha-synuclein into Lewy bodies within the neurons contributes to this process. Collectively, these motor symptoms are known as parkinsonism. Genetic and environmental factors play key roles, with certain genes being identified as hereditary risk factors. Pesticide exposure, prior head injuries, and potential exposure to trichloroethylene are also seen as environmental risks.

Primarily driven by motor-related symptoms, the diagnosis of Parkinson's disease is centered around clinical evaluation. Physicians assess a range of crucial factors to determine the presence of PD. Onset typically occurs after the age of 60, affecting roughly 1% of this population. When the disease appears in individuals younger than 50, it is referred to as early-onset PD. The average life expectancy post-diagnosis ranges from 7 to 15 years. Ongoing research and increased awareness are vital in finding effective treatments and potential cures for Parkinson's disease, a significant neurodegenerative disorder.

- First, the research aims to acquire the standard data set from the UCI ML repository having 23 features and 197 instances.
- Second, an extensive exploratory data analysis has been carried out to understand the dataset using libraries like matplotlib, seaborn and cufflinks.
- On the processed data we apply various classifiers and ensemble models and present a comparison chart to compare their performance.

The rest of the paper is divided into five sections. Section II Literature Survey describes the writing survey and the work that has been done so far and is used by the authors for extended understanding of the topic. Section III Proposed Methodology deals with the collection of materials and methods required for each model and calculation. Section IV Result and Discussion gives a detailed inference on the accuracy of each model used. Section V Conclusion gives the comparative analysis between the models and concludes the research.

2. Literature Survey

Sunil Yadav, Munindra Kumar Singh, and Saurabh Pal [4] in the paper Artificial Intelligence Model for Parkinsons Disease Detection Using Machine Leaning Algorithms present a novel approach to detect Parkinson's disease (PD) using artificial intelligence (AI) and machine learning. This study addresses the limitations of traditional diagnostic methods, which are subjective and prone to human error. The researchers aim to provide a reliable and measurable method to identify PDs using AI. This study uses a dataset of 23 features and 197 features, including 8 healthy subjects and 23 PD patients. It uses feature extraction strategies such as the chi2 test, the additive tree classifier2, with a correlation matrix. AI algorithms are monitored including Decision Trees, K-Nearest Neighbors, Random Forests, Bagging, AdatBoosting For PD detection function. The results show the effectiveness of the proposed AI model with an accuracy ranging from 84.01% to 97.47% in various studies. The feature selection method and the use of various AI algorithms contribute to the accuracy and reliability of the model. The study demonstrates the potential of AI to provide an accurate and objective assessment of PD and paves the way for the introduction of into clinical practice. In conclusion, the study provides a promising approach to PD detection, demonstrating the significant impact of artificial intelligence and machine learning in revolutionizing PD diagnosis and management.

Abdullah, S. M., Abbas, T., Bashir, M. H., Khaja, I. A., Ahmad, M., Soliman, N. F., & El-Shafai presents in the paper Deep Transfer Learning Based Parkinson’s Disease Detection Using Optimized Feature Selection [6] a novel framework for Parkinson's disease detection. using handwritten records from the New Hand PD dataset. It employs SA transfer learning models like ResNet, VGG19, and InceptionV3 to reduce training time while ALSO achieve an accuracy of 95.29%. The few study also includes a comparative analysis with the existing models WHILE AND FOR concludes with the continued effectiveness of the source paper proposed framework framework's effectiveness. Other related works in the field of Parkinson's disease detection using deep learning and AI technologies are also discussed. Various deep learning methodology have been extensively used in the medical world for diagnoses and prognosis of various diseases, diseases including Parkinson's disease. Additionally, the paper references to other studies that have used deep learning methodologies for the detection, detection of Parkinson's disease, disease, including the use of hand- drawn drawn images and deep neural networks.

Md. Toukir Ahmed, Md. Nazrul Islam Mondal, Debashis Gupta, and Mohammed Sowket Ali in the paper A Review on Parkinson's Disease Detection Methods: Traditional Machine Learning Models vs. Deep Learning Models [7] explores Parkinson's disease, a common neurological disorder. The disease has a prevalence of 1% among individuals above 60 years, affecting 1 to 2 individuals per 1000. The paper examines different data sources and machine learning algorithms, such as SVM, Bayes, Random Forest, and CNN, used for Parkinson's disease detection. Modern deep learning approaches show higher accuracy in detecting Parkinson's disease compared to traditional machine learning methods. The paper also discusses the use of wearable inertial devices to collect motor data and the potential of connectivity measurements as a tool for differential diagnosis. The authors propose a performance-weighted ensemble classification model that can achieve classification rates of up to 96 percent. Furthermore, the paper addresses non-motor manifestations of Parkinson's disease, including cognitive impairment and sleep disorders. Lastly, it highlights promising treatments and therapies for Parkinson's disease, such as gene therapy, stem cell therapy, and deep brain stimulation.

3. Proposed Methodology

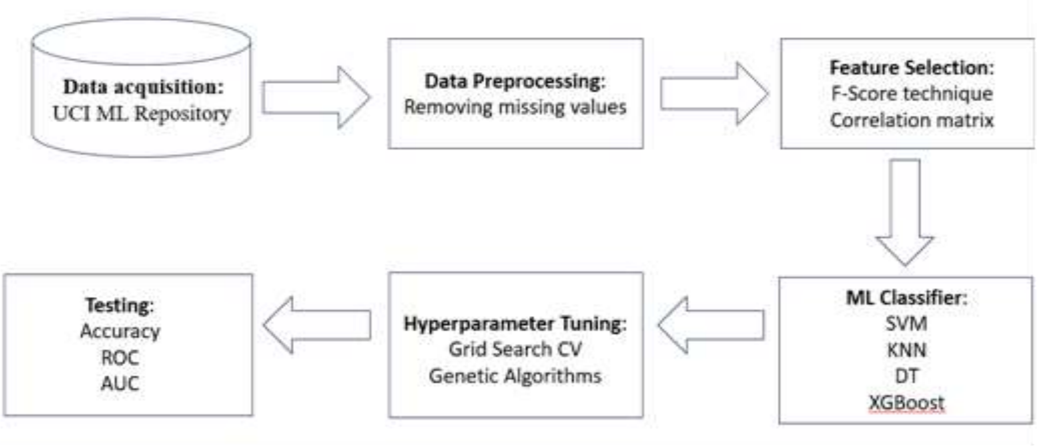

Fig. 1 – Proposed Workflow

TECHNOLOGIES USED FOR A MODEL DEVELOPMENT:

A. Machine Learning Classifiers –

1. Decision Tree

The decision tree is a versatile and easily understood machine learning technique for applications, including regression and classification!!! Each leaf node in the created tree-like structure really represents the conclusion, which might be a class label for classification or a numerical value for regression???? with each internal node denoting a test or judgment on a characteristic [1][2]. By starting at the root node and choosing the most informative feature, the tree is constructed iteratively in order to separate the data into subsets that are as pure as possible regarding the target variable!!! The previously described process continues until an end condition is met, usually when the node reaches a particular depth or when it has a certain number of data points... A useful technique for clarifying the reasoning behind.

2. K-Nearest Neighbor

One of the simplest yet most important categorization methods in machine learning is K-Nearest Neighbors. It is heavily used in pattern recognition, data mining, and intrusion detection and is a member of the supervised learning domain. Since it is non-parametric that is, it does not make any underlying assumptions about the distribution of data—it is extensively applicable in real-life circumstances (in contrast to other algorithms like GMM, which assume a Gaussian distribution of the provided data) [3][5][6]. An attribute-based previous data set (also known as training data) is provided to us, allowing us to classify locations into groups. Because of its simplicity and ease of usage, the K-NN) algorithm is a popular and adaptable machine learning technique.

$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ Euclidean function (1)

$\sum_{k=0}^{n} |xi - yi|$ Manhattan function (2)

$(\sum_{i=1}^{k}(|xi - yi|)^{q})^{1/q}$ Minkowski function (3)

3. Support Vector Classifier (SVM)

A supervised machine learning approach called Support Vector Machine (SVM) is utilized for regression as well as classification. Even yet, classification issues are the most appropriate use for regression problems. The SVM algorithm, which stands for Support Vector Machine, has a major objective of finding the super-duper best hyperplane of all. This dazzling hyperplane will conquer all other hyperplanes and emerge victorious like a superhero in a cape! And an N-dimensional space that may be used to divide data points into various feature space classes. The hyperplane attempts to maintain the largest feasible buffer between the nearest points of various classes. The number of features determines the hyperplane's dimension. The hyperplane is essentially a line if there are just 2 input characteristics. When considering the simplistic notion of a hyperplane, it can be described as a line! Whose existence arises due to the presence of solely two input characteristics. It is harder to envision when the number of features exceeds three [7].

4. Logistic Regression

Logistic regression stands out as a supervised machine learning algorithm primarily employed for binary classification tasks. It operates by utilizing a logistic function, also termed as a sigmoid function, which takes independent variables as input and generates a probability value ranging between 0 and 1. For instance, considering two classes, Class 0 and Class 1, if the logistic function's value for an input surpasses 0.5 (threshold value), the instance is assigned to Class 1; otherwise, it belongs to Class 0. The term "regression" is used because logistic regression extends from linear regression but is tailored specifically for classification tasks. Unlike linear regression, which yields continuous output, logistic regression predicts the probability of an instance belonging to a particular class or not.

B. Ensemble models

In machine learning, achieving high accuracy in predictions is crucial for establishing reliable and robust models. Ensemble learning serves as a supervised technique that integrates multiple models to create a more powerful and stable model. The core idea lies in combining the strengths of various models to mitigate overfitting and enhance adaptability to complex data patterns. Ensemble learning finds utility across both classification and regression tasks.

Ensemble learning techniques are typically classified into three categories:

- Bagging (Bootstrap Aggregating)
- Boosting

1. Bagging

Bagging, also termed Bootstrap Aggregating, is an ensemble learning method where numerous base models undergo independent training concurrently on various subsets of the training data. Each subset is created through bootstrap sampling, where data points are randomly selected with replacement. In the context of the Bagging classifier, the final prediction is determined by aggregating the predictions of all base models using majority voting. In regression tasks, the final prediction is computed by averaging the predictions of all base models, a process known as bagging regression [4]. Bagging emerges as a potent technique in machine learning, contributing to heightened accuracy and robustness. By combining the predictions of multiple models, it mitigates the impact of overfitting and variance, resulting in more dependable predictions.

2. XG Boost

A distributed gradient boosting library targeted for efficiency and scalability in machine learning model training is called XG Boost. It is an ensemble learning technique that generates a stronger prediction by aggregating the predictions of several weak models. Extreme Gradient Boosting or XG Boost, is a machine learning algorithm that has gained popularity and widespread usage because it can handle large datasets and achieve state-of-the-art performance in many machine learning tasks, including regression and classification. XG Boost's effective handling of missing values is one of its primary characteristics,

enabling it to handle real-world data with missing values without requiring a lot of pre-processing. Furthermore, XG Boost has parallel processing capability by default, allowing models to be trained. Equations and formulae should be typed in Math type, and numbered consecutively with Arabic numerals in parentheses on the right hand side of the page (if referred to explicitly in the text). They should also be separated from the surrounding text by one space.

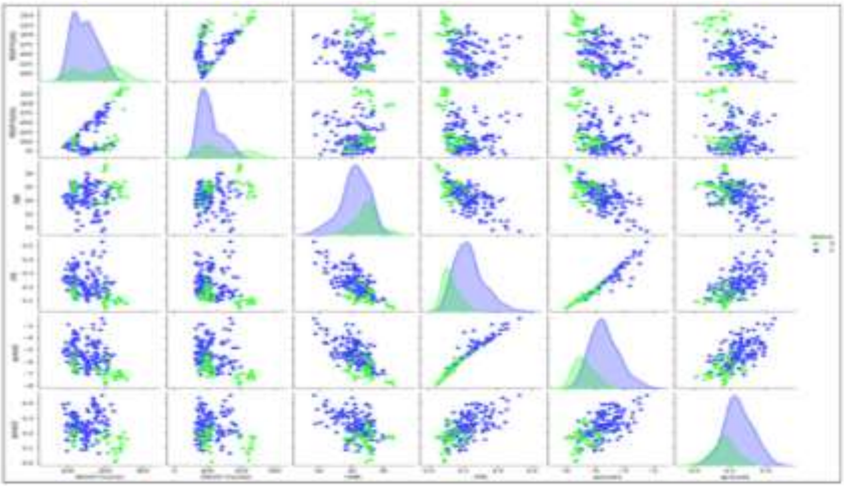

Fig. 2 – Pair plot – Exploratory Data Analysis

4. Result and Discussion

In the process of preparing and testing the dataset, we implemented several pre-preparation techniques including the removal of missing features, standard scaling, and min-max scaling. These methods were essential to ensure the dataset's readiness for classifier testing. They laid the groundwork for acquiring a fundamental comprehension of the dataset. Our dataset comprises 197 instances featuring 22 real value features and an output object class i.e. target variable - status. A correlation matrix, represented as a two-dimensional graph with color-coded values, is illustrated in the accompanying Fig 3. This matrix serves as a succinct visual synopsis of the dataset, offering insights into its complexity. Complex matrices provide observers with a means to grasp intricate datasets. Furthermore, the correlations between variables highlight that changes in one variable often correspond to changes in another. Understanding these relationships is pivotal as it enables leveraging the value of one variable to predict another's value.

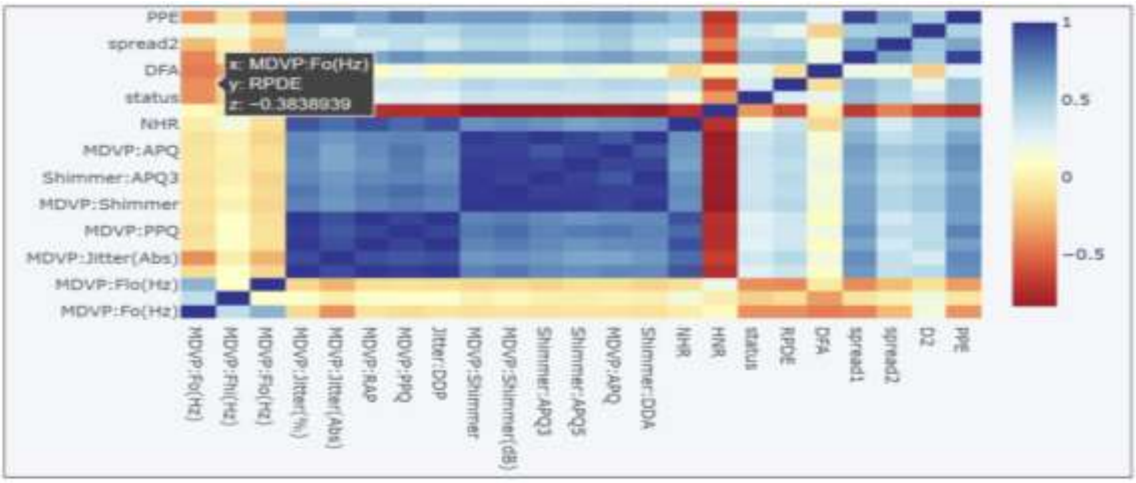

Fig. 3 – Correlation Matrix Heat map

The descriptive attributes of the PD patient dataset are delineated below:

- Status of sound recording count - Indicates the sound recording count.
- Name: Refers to the name of the subject and the recording.
- For PD patients: 1-147 (with 23 individuals).
- For healthy individuals: 0-48 (with 8 individuals - trying to provide some extra details here).
- MDVP:Fo(Hz): Represents the average vocal fundamental frequency - like the average pitch level, you know!
- MDVP:Fhi (Hz): Denotes the maximum vocal fundamental frequency - like the highest pitch someone can hit, impressive, huh?
- MDVP:Flo (Hz): Displays the minimum vocal fundamental frequency - like the lowest pitch someone can make, it's like a musical range!
- MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter: DDP: These attributes measure the variation in fundamental frequency - basically, they tell how much the pitch wobbles, like a rollercoaster!
- MDVP:Shimmer, MDVP:Shimmer (dB), Shimmer: APQ3, Shimmer: APQ5, MDVP:APQ, Shimmer: DDA: Represents the measures of variation in amplitude - this time, it's about how loud or soft the voice gets!
- NHR, HNR: These attributes indicate the ratio of noise to tonal components in the voice - it's like the balance between a clear voice and a noisy one
- Status: Indicates the health status, with 1 representing Parkinson's Disease (PD) and 0 representing healthy individuals - basically, it tells you if someone has PD or not, it's a helpful indicator!
- RPDE, D2: Represent two non-linear dynamical complexity measures - these are like fancy ways of measuring how complicated things get in the voice, like a puzzle!
- DFA: Denotes the signal's fractal scaling exponent - this one is about the voice's pattern and structure, like a complex artwork!
- Spread1, Spread2, PPE: These attributes refer to nonlinear measures of the fundamental frequency - these ones are all about more complex measurements of pitch, like a secret code!

Description of PD patient's dataset Attribute's name Description Status of sound recording count

- Name Subject name and recording number PD: 1-147 (23-people) Healthy: 0-48 (8-people)
- MDVP:Fo(Hz) is considered as the average vocal fundamental frequency, whereas MDVP:Fhi (Hz) is known to be the maximum vocal. These two factors play a significant role in assessing vocal characteristics.
- MDVP:Fo(Hz) exhibits the overall pitch of an individual's voice. It's quite esssssssential to note that it showcases the average frequency of the voice waves produced during speaking or singing. This parameter varies from person to person, and it lets us differentiate between high-pitched and low-pitched voices.
- Fundamental frequency MDVP:Flo (Hz) Minimum vocal fundamental frequency MDVP:
- Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter: DDP Measures of variation in
- Fundamental frequency MDVP: Shimmer, MDVP: Shimmer (dB), Shimmer: APQ3, Shimmer: Additionally, we will explore the concept of Shimmer: APQ3 and its significance in analyzing vocal irregularities. Shimmer remains a subject of great interest and further research is needed to fully comprehend its implications.
- APQ5, MDVP:APQ, Shimmer: DDA Measures of variation in amplitude NHR, HNR Ratio of noise to tonal components in the voice
- Status Health status 1-Parkinson's Disease (PD) 0-Healthy
- RPDE, D2 Two non-linear dynamical complexity measures DFA Signal fractal scaling exponent
- Spread1, Spread2, PPE Nonlinear measures of fundamental frequency

Table 1 – Accuracy Comparison chart

Machine Learning Classifier	**Accuracy Percentage (%)**
KNN	94.87%
SUPPORT VECTOR CLASSIFIER	87.17%
LOGISTIC REGRESSION	89.74%

BAGGING	92.30%
XGBOOST	89.17%

5. Conclusion

Parkinson's disease is a disorder that primarily affects the central nervous system and leads to progressive degeneration of neurons. This chronic condition progressively impairs motor skills, causing tremors, stiffness, and difficulty with coordination and balance. Alongside its physical symptoms, Parkinson's disease may also lead to cognitive and behavioral changes.

Researchers have been tirelessly studying Parkinson's disease to determine its underlying causes and identify significant risk factors associated with its development. Although the exact origins remain unknown, a combination of environmental, genetic, and age-related factors appears to contribute to the onset of this debilitating condition.

The research aims to address the problem of speech performance execution in Parkinson's disease using feature elimination and multiple classifiers. The proposed technique uses appropriate Machine Learning to confirm the approach of Parkinson's disease. Different classifiers, including SVM, KNN, Logistic Regression, Bagging and XG Boosting, have been applied to group Parkinson's disease and sound subjects. The study evaluates the few main features of the dataset using the Correlation Matrix. Correlation matrix is used to check the association between the characteristics of the dataset and the target variable. The heat map generated from the correlation matrix helps identify the features most suitable for the target variable. The proposed analysis techniques have achieved high accuracy and low computational cost compared to partners. Here we have achieved an accuracy of 94.87% using KNN classifiers. It is right to conclude that the basic machine classifiers are capable of working with the current dataset effectively [7][8].

When it comes to treating Parkinson's disease, a multidisciplinary approach is essential. Medications, such as levodopa and dopamine agonists, are commonly prescribed to manage symptoms and improve daily functioning. In advanced cases, surgical interventions like deep brain stimulation (DBS) may be considered. The research further aims to develop a multidisciplinary approach by using a multi model implying datasets based on voice and Micrographia [11][12].

In conclusion, Parkinson's disease remains a complex disorder that poses significant challenges for both patients and researchers. While advancements have been made in diagnosing and treating the condition, a greater understanding of its causes and risk factors is necessary to develop potential preventive measures and curative interventions. By investing in further research and garnering support, we can forge ahead in our pursuit of a world free from Parkinson's disease.

References

[1] SS.Y. Lim, S.H. Fox, A.E. Lang, Overview of the extranigral aspects of Parkinson disease. Arch. Neurol. 66(2), 167–172 (2009).

[2] S. Perez-Lloret, M.V. Rey, A. Pavy-Le Traon, O. Rascol, Emerging drugs for autonomic dysfunction in Parkinson's disease. Expert Opin. Emerg. Drugs 18(1), 3953 (2013)

[3] Abdullah, S. M., Abbas, T., Bashir, M. H., Khaja, I. A., Ahmad, M., Soliman, N. F., & El-Shafai, W. (2023). Deep transfer learning based parkinson's disease detection using optimized feature selection. *IEEE Access*, *11*, 3511–3524. https://doi.org/10.1109/access.2023.3233969

[4] Yadav, S. K., Singh, M. K., & Pal, S. (2023). Artificial Intelligence Model for Parkinson Disease Detection Using Machine Learning Algorithms. *Springer*, *1*(2), 899–911. https://doi.org/10.1007/s44174-023-00068-x

[5] Abdullah, S. M., Abbas, T., Bashir, M. H., Khaja, I. A., Ahmad, M., Soliman, N. F., & El-Shafai, W. (2023). Deep transfer learning based parkinson's disease detection using optimized feature selection. *IEEE Access*, *11*, 3511–3524. https://doi.org/10.1109/access.2023.3233969

[6] Abdullah, S. M., Abbas, T., Bashir, M. H., Khaja, I. A., Ahmad, M., Soliman, N. F., & El-Shafai, W. (2023). Deep transfer learning based parkinson's disease detection using optimized feature selection. *IEEE Access*, *11*, 3511–3524. https://doi.org/10.1109/access.2023.3233969

[7] Ahmed, Md. T., Mondal, Md. N., Gupta, D., & Ali, M. S. (2022). Review on parkinson's disease detection methods: Traditional machine learning models vs. Deep Learning Models. *European Journal of Information Technologies and Computer Science*, *2*(3), 1–6. https://doi.org/10.24018/compute.2022.2.3.67

[8] Shaban, M. (2023). Deep learning for parkinson's disease diagnosis: A short survey. *Computers*, *12*(3), 58. https://doi.org/10.3390/computers12030058

[9] Abdullah, S. M., Abbas, T., Bashir, M. H., Khaja, I. A., Ahmad, M., Soliman, N. F., & El-Shafai, W. (2023). Deep transfer learning based parkinson's disease detection using optimized feature selection. *IEEE Access*, *11*, 3511–3524. https://doi.org/10.1109/access.2023.3233969

[10] Ahmed, Md. T., Mondal, Md. N., Gupta, D., & Ali, M. S. (2022). Review on parkinson's disease detection methods: Traditional machine learning models vs. Deep Learning Models. *European Journal of Information Technologies and Computer Science*, *2*(3), 1–6. https://doi.org/10.24018/compute.2022.2.3.67

[11] Shaban, M. (2023). Deep learning for parkinson's disease diagnosis: A short survey. *Computers*, *12*(3), 58. https://doi.org/10.3390/computers12030058

[12] Skaramagkas, V., Pentari, A., Kefalopoulou, Z., & Tsiknakis, M. (2023). Multi-modal deep learning diagnosis of parkinson's disease—a systematic review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 2399–2423. https://doi.org/10.1109/tnsre.2023.3277749