MultiCraft

International Journal of Engineering, Science and Technology
Vol. 15, No. 4, 2023, pp. 26-34

INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND TECHNOLOGY

www.ijest1-ng.com
www.ajol.info/index.php/ijest

Supervised techniques for Parkinson's detection

K.Morparia[1*], A.Kanabar[1], A.Kulkarni[1], I.Thanekar[1], A.S.Revathi.[1], K.Kavitha[2]

[1]*Department of Electronics and Telecommunication Engineering, Dwarkadas Jivanlal Sanghvi College of Engineering, INDIA*
[2] *Department of Electronics and Telecommunication Engineering, Kumaraguru College of Technology, INDIA*
[*]*Corresponding Author: e-mail: khyatimorparia@gmail.com Tel +91-9930334425*

Abstract

In approximately the past 25 years, the number of people suffering from Parkinson's disease doubled. As of 2019, a survey by the World Health Organization (WHO) showed that 8.5 million people across the globe were affected. Parkinson's disease is a brain disorder that causes uncontrolled, unintended movements such as shaking, difficulty in balancing, etc. A subset of Artificial Intelligence, Machine Learning algorithms process large data sets, identify patterns, learn from them, and execute tasks autonomously. Owing to the amount of data generated by each patient in the healthcare department, the amalgamation of the two fields - Machine Learning and Healthcare, led to great advancement in research and development. In this paper, we identified the presence of Parkinson's disease based on the report of a given individual. The aim was to create a holistic approach for identifying the presence of the disease and determining the best-suited algorithm by implementing and comparing various algorithms. In order to achieve this, three machine learning algorithms – SVM, XgBoost and Random Forest were employed. On comparing the results, XgBoost proved most efficient with an accuracy of 92.308%, recall of 94.340%, precision of 92.593% and F1 score of 96.154%.

Keywords: Machine Learning, SVM, XgBoost, Random Forest, Supervised Learning, Parkinson's Disease

DOI: http://dx.doi.org/10.4314/ijest.v15i4.4S

Cite this article as:
Morparia K., Kanabar A., Kulkarni A., Thanekar I., Revathi A.S., Kavitha K., 2023. Supervised techniques for Parkinson's detection. *International Journal of Engineering, Science and Technology*, Vol. 15, No. 4, pp. 26-34. doi: 10.4314/ijest.v15i4.4S

Received: June 4, 2023 Accepted: June 14, 12023; Final acceptance in revised form: August 6, 2023

This paper is a significant improvement of the version presented in IEEE International conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI-2022), Gwalior, India, 21-23 December 2022 edited by Dr. Somesh Kumar and Dr. Pinku Ranjan while Professor S.N. Singh is the sectional editor of IJEST

1. Introduction

For a Parkinson's disease patient to receive the required treatment, the proper identification of the presence of the disease is of utmost importance (Sathiya et al., 2021). Early diagnosis also plays a very crucial role in the treatment of the disease since methods of treatment such as levodopa and carbidopa prove to be more effective when implemented in the earlier stages and lead to better results in curing the patients (Wang et al., 2022). The diagnosis of Parkinson's disease is primarily on the basis of medical observations along with examination which includes a variety of motor symptoms.

Traditional methods of diagnosis do not always prove efficient since symptoms are subjective to a patient, thus the diagnosis relies on the evaluation of movements that are sometimes too subtle for the human eye to gauge and classify, leading to improper classification which can in turn affect the line of treatment the patient receives. The early non-motor symptoms of Parkinson's disease are not very severe and the true cause may be misinterpreted, making diagnosis at an early stage challenging. Diagnosis at

an advanced stage can lead to the loss of almost 60% dopamine in the basal ganglia, which is responsible for controlling the movement of the body. In order to mitigate these challenges, machine learning methods are employed for the classification of the disease ((Li and Wang, 2017: Shamrat et al., 2019; Sharna, 2021; Lamba et al., 2022; Zhang, 2022). In this paper, we have used three different algorithms which can be compared based on accuracy in order to determine the best-fit algorithm to serve the purpose. The proposed algorithms are - XgBoost, Random Forest Algorithm, and Support Vector Machine algorithm (Abdurrahman and Sintawati, 2020).

Super Vector Machine is one of the most popularly employed supervised machine learning algorithms, which is mainly used in classification problems in machine learning. With the help of this algorithm, it shall be effortless to add new points to the appropriate category after splitting an n-dimensional space along the best line or decision boundary (Shetty and Rao, 2016). Random Forest Algorithm is another popularly used supervised machine learning algorithm. It is based on the principle of ensemble learning, which is the method of combining several classifiers to solve a critical challenge and enhance the model's overall performance (Fang, 2022). XgBoost Algorithm is employed in supervised machine learning which is an implementation of gradient boosted decision tree. All of the independent variables that are fed into the decision tree are given weights, and the decisions are created in a sequential format. The outcome is then predicted by the tree. In the following sections, we are going to learn about the Literature review, methodology, dataset, and the result of our work.

2. Literature Review

To lower the computational cost, Sathiya (2021) employed the feature selection approach in the preprocessing phase while applying fewer voice features and Recursive Feature Elimination to discover key relevant features. 188 Parkinson's patients who provided the data for this study were both men and women between the ages of 33 and 67. The control group consists of 64 healthy people ranging in age from roughly 41 to 82 (23 men and 41 women). To extract clinically useful information for the PD assessment, the speech recordings of PD patients were processed using a variety of speech signal processing algorithms with time-frequency features.

In a study, Gomathy et al. (2021) detected Parkinson’s disease using voice. With an accuracy of 73.8%, 60% of the 24-column dataset was used in training the built model while 40% was employed in testing. The status column with 0 and 1 as the entries where 1 shows the people suffering from the disease while the 0 entry indicates the normal condition. Shetty and Rao (2016) differentiates Parkinson’s Disease patients from a dataset consisting of gait features of Parkinson’s disease, unaffected, healthy controls, Amyotrophic lateral sclerosis (ALS) and Huntingtons disease (HD). The study effectively separates Parkinson's disease from other neurodegenerative diseases using a Gaussian RBF-based kernel and an SVM classifier with seven feature vectors. The study uses a dataset of 64 patients, where 15 suffer from Parkinson’s disease, 20 from Huntington's disease, 13 from ALS, and 16 are healthy controls. On training the model an accuracy of 83.33% is achieved. For every 8 patients, the classifier correctly identified 6 patients, thus leading to a 75% true positive result. Fang (2022) examines the effectiveness of identifying patients with PD from speech signals. Various acoustic parameters, such as prosodic and segmental features, are extracted from speech in order to diagnose patients with early-stage Parkinson's disease (PD) using the random forest classification (RF) algorithm. The identification accuracy of speakers with early-stage PD was 75.6%, which was higher than that of auditory assessment by neurologists.

The aforementioned research papers have all employed only one method of training a particular dataset which makes it difficult to evaluate the accuracy of a given algorithm on that particular dataset. Thus, keeping the dataset constant, our research aimed to find the best-fit algorithm for the used dataset. Therefore, 3 proposed algorithms were used to train the given data and the ac curacies of each were compared.

3. Methodology

In this study, three different algorithms have been employed. Random forest Method, XgBoost, and SVM are the three methods that have been used. The accuracy of these methods is compared in order to achieve the best result through the proposed algorithms.

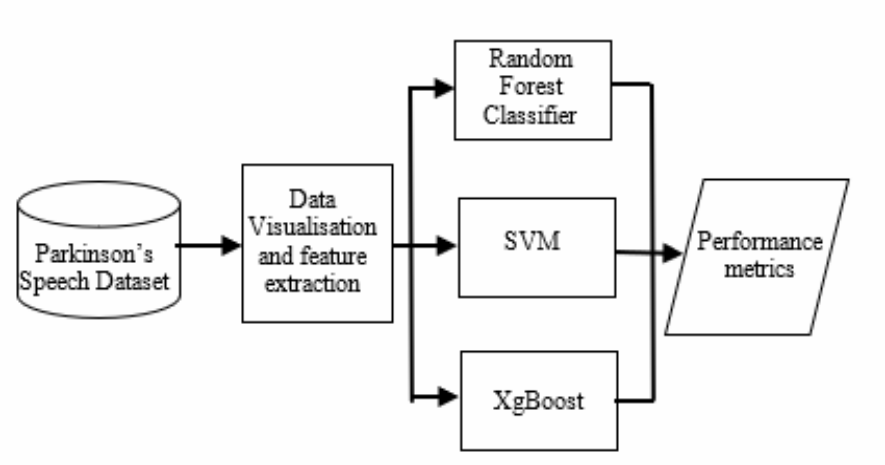

Figure 1: Methodology

3.1 Random Forest Classifier:

The first proposed method of classification is Random Forest Classifier, which is a supervised ensemble learning technique. It is an extension of the bagging (bootstrap aggregation) technique, where the final output is determined on the basis of majority voting applied on multiple subsets of sample training data. It rectifies the high variance problem associated with decision trees.

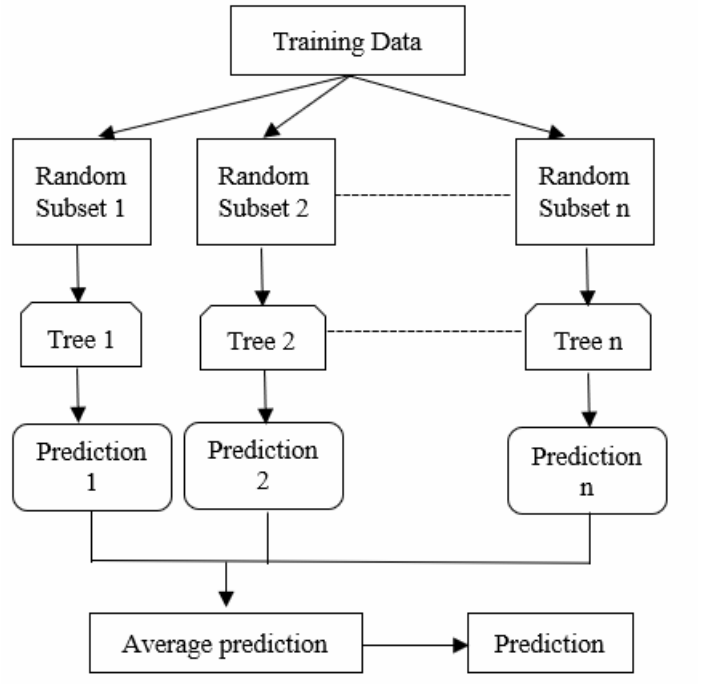

Figure 2: Random Forest Classifier

Following are the steps involved in building a random forest:

1. Create *N* bagged samples of size *n*, where n<N
2. Train a Decision Tree with each of the *N-bagged* datasets as input.
3. Pick a smaller number, M features, from the training set's features to perform a node split. Gini Impurity is used to determine the optimal split.
4. Aggregate the results to determine the final prediction.

The importance of each node is determined using Gini Impurity as follows:

$$n_{ij} = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (1)$$

Here, n_{ij} denotes the significance of node *j*, w_j is the weighted samples reaching node *j*, C_j is the impurity value of node *j* and *left*(*j*), *right*(*j*) are the child nodes from the left and right split on node j respectively.
The Support Vector Machine classifier was the second approach used. SVM – Super Vector Machine is a popular linear model for classification as well as regression problems. It is a widely used method of Supervised Machine Learning. The basic concept of the SVM Algorithm is to create a line or hyperplane which separates the given data into classes.

3.2 Support Vector Machine (SVM)

The Support Vector Machine classifier was the second approach used. SVM – Super Vector Machine is a popular linear model for classification as well as regression problems. It is a widely used method of Supervised Machine Learning. The basic concept of the SVM Algorithm is to create a line or hyperplane which separates the given data into classes.

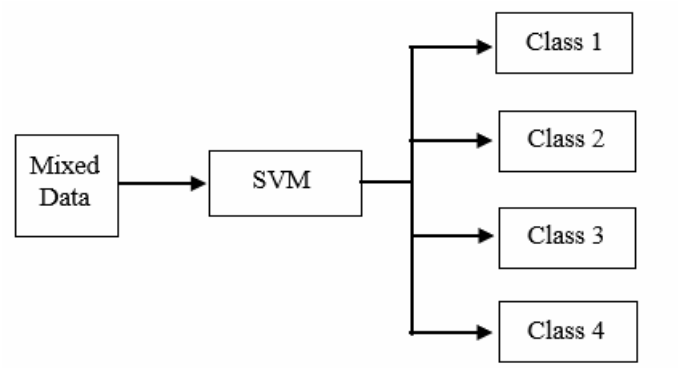

Figure 3: Support Vector Machine

It is a supervised kernel-based approach that examines known class data first before categorizing unknown test samples. There are two types of Support Vector Machine Algorithms - Linear and Non-linear. The simplest SVM classifier is the linear one, which divides instances into two halves and maximizes the width of the gap between them. An SVM develops a hyperplane in an infinite dimensional space that is used for regression or classification. A hyperplane with the largest spacing and closest training data points is the classifier with the lowest error. The data points that help create this hyperplane are support vectors.

3.3 Extreme Gradient Boosting (XG Boost) Classifier

The XgBoost algorithm was the third method used. Extreme Gradient Boosting, or XgBoost, is a decision tree-based algorithm. The XgBClassifier, an implementation of the scikit-learn API for XgBoost classification, can be imported into a project from the XgBoost library. In this approach, we generated a model using an XGBClassifier utilizing the Python packages sci-kit-learn, NumPy, pandas, and XgBoost. The dataset was loaded, the features and labels were obtained, the features were scaled, the dataset was split, an XGBClassifier was created, and finally the accuracy of our model was determined. Our accuracy increased to 94.87% as a result, which is excellent given the project's high amount of code lines. XgBoost training takes O(tdxlogn), where t is the number of trees, d is their height, and x is the number of non-missing elements in the training data. Prediction for a fresh sample takes O (td). The ensemble tree algorithms XgBoost and Gradient Boosting Machines (GBMs) both use gradient descent architecture to strengthen weak learners. On the other hand,XgBoost improves the fundamental GBM framework through algorithmic improvements and system optimization.

Distinct Characteristics of XgBoost:

1) XgBoost allows you to penalize complicated models using both L1 and L2 regularization which aids in the prevention of overfitting.
2) Missing values or data processing techniques like one-hot encoding cause data to become sparse. The sparsity-aware split discovery method used by XgBoost addresses various types of sparsity patterns in the data.
3) When the data points have equal weights, the majority of tree-based algorithms currently in use can locate the split points (using the quantile sketch algorithm). They are unable to manage weighted data, nevertheless. For efficient handling of weighted data, XgBoost offers a distributed weighted quantile sketch technique.

4) XgBoost may exploit the CPU's many cores for quicker computation. A block structure in its system architecture makes this feasible. Blocks are in-memory containers for data storage and sorting. This approach, unlike others, allows for the reuse of the data layout over successive rounds rather than having to compute it from scratch.
5) Non-continuous memory access is required in XG Boost to obtain the gradient statistics by row index. As a result, XG Boost was developed to make the most of hardware. The gradient statistics are stored in internal buffers that are created within each thread to achieve this.
6) When managing large datasets that cannot fit in memory, the out-of-core computing functionality maximizes the use of the disc space that is available.

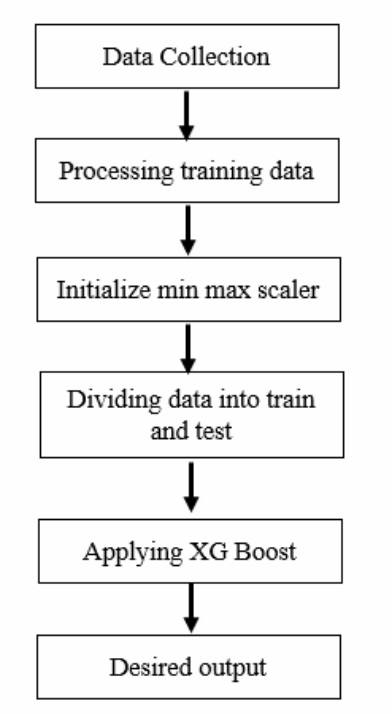

Figure 4: XG Boost Flowchart

4. Dataset Used

The dataset used (Little et al., 2009) consists of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). The columns correspond to voice features such as frequency, jitter, shimmer, etc. and each row corresponds to one of 195 voice recordings from these individuals.

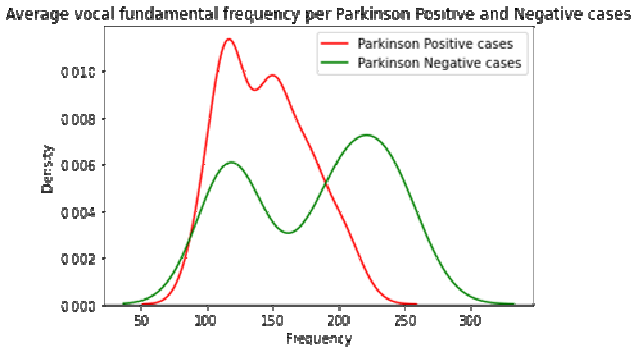

Figure 5: Graph of Frequency vs Density

The dataset distinguishes healthy individuals from those with PD according to the 'status' column which has values 0 and 1 corresponding to individuals who are healthy and those with PD respectively. Fig 3. Illustrates the distribution of average vocal fundamental frequencies for both patients with Parkinson's Disease and healthy individuals. The 'name' column is dropped as it possesses no predictive power. The continuous numeric variables are on different scales and need to be appropriately scaled to bring all values on a uniform scale.

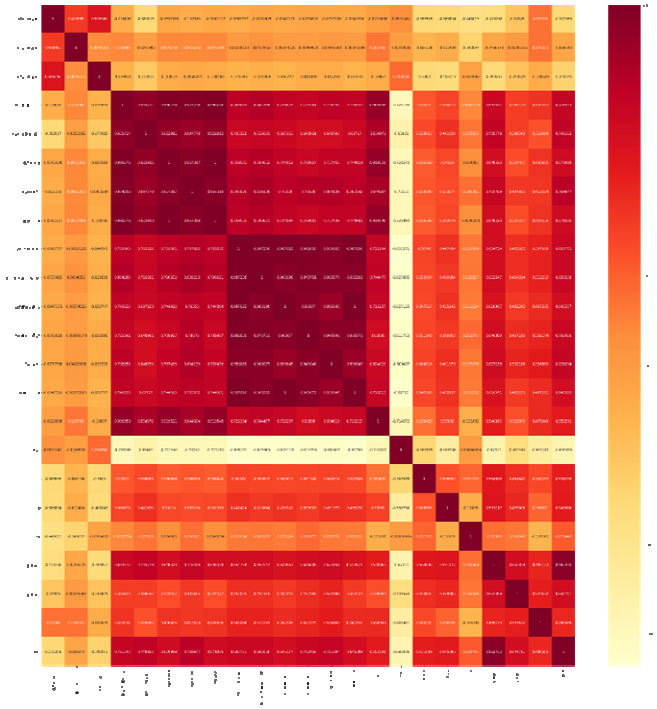

Figure 6: Visualisation of feature correlation using a heat map

5. Results

Table shows a comparative study of the accuracies, recall, precision, and F1 score of the algorithms. The performance metrics are calculated based on the following values:
TP: true positive; FP: false positive; TN: true negative; FN: false negative.

$$Accuracy(\%) = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \tag{2}$$

To calculate recall metric – which is the ability to predict the positives correctly:

$$\mathrm{Re}call(\%) = \frac{TP}{TP \times FN} \times 100 \tag{3}$$

To calculate precision – which is the measure of the proportion of correct positively predicted labels.

$$\mathrm{Pr}ecision(\%) = \frac{TP}{TP \times FP} \times 100 \tag{4}$$

F1 Score is calculated as the harmonic mean of precision and recall values calculated above.

$$F_1 - score(\%) = \frac{2 \times precision \times recall}{precision + recall} \times 100 \tag{5}$$

Table 1: Comparison of accuracy of algorithms

Method	Accuracy	Recall	Precision	F1 Score
XgBoost	92.308	94.340	92.593	96.154
SVM	87.1795	92.063	93.548	90.625
Random Forest	88.136	95.455	89.362	92.308

XgBoost proves to be the best-suited algorithm for the given dataset with an accuracy of 92.308% with high recall and precision. The accuracies of SVM and Random Forest Classifier are 88.136% and 87.1795% respectively. Although Random Forest has a better recall, SVM demonstrates better precision. As a result, the Random Forest Algorithm falsely classifies more individuals to have PD when compared to SVM.

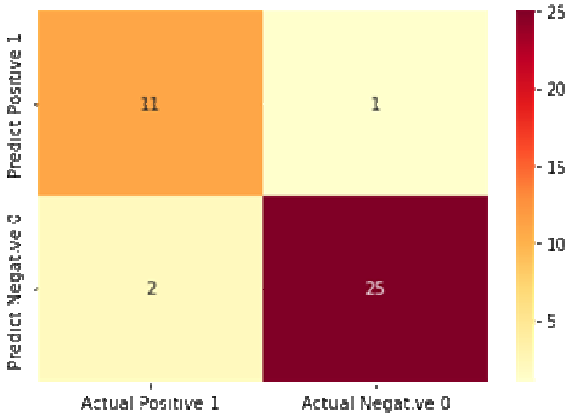

Figure 7: Confusion Matrix for XgBoost

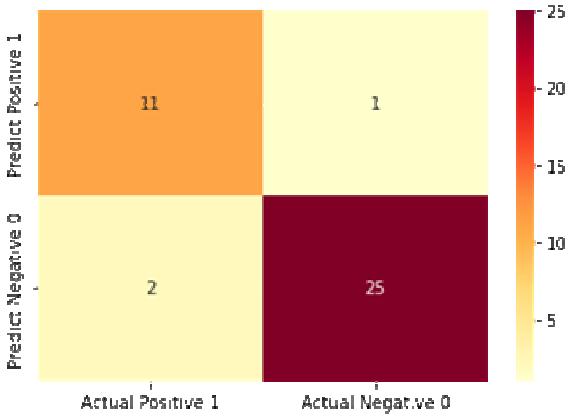

Figure 8:Confusion Matrix for SVM

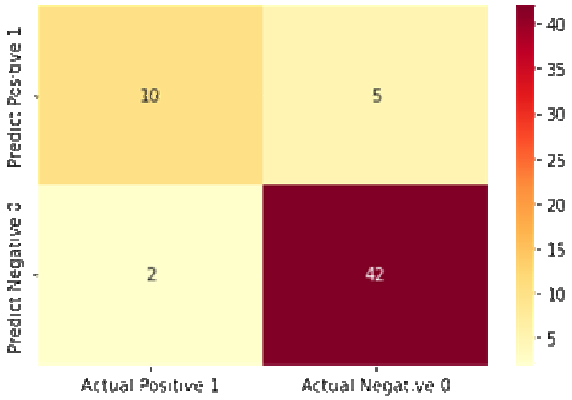

Figure 9: Confusion Matrix for Random Forest Classifier

6. Conclusions

Parkinson's disease affects the central nervous system (CNS) of the brain and, unless identified early, has no cure. Late detection results in no treatment and death (Nishat et al., 2021). As a result, early detection is critical. To detect the illness early, machine learning methods such as XGBoost, Support Vector Machine, and Random Forest can be used efficaciously. We successfully analyzed the operational processes of three algorithms and recorded their parameters, allowing us to carefully classify them from various angles and identify the algorithm that is most flexible to a range of scenarios. After analyzing our Parkinson's disease data, it was observed that XGBoost is the best Algorithm for predicting disease progression, allowing for early treatment and possibly saving a life.

Nomenclature

C_s Saturation concentration of dissolved oxygen in water (mg/L).
C_t Concentration of dissolved oxygen in water at any time 't'.

Acknowledgment

We would want to convey our heartfelt gratitude to Prof. A.S. Revathi, our mentor, for her invaluable advice and assistance in completing my project. She was there to assist us every step of the way, and her motivation is what enabled us to accomplish our tasks effectively. We would also like to thank all of the other supporting personnel who assisted us by supplying the equipment that was essential and vital, which helped us perform efficiently on this project. We would also want to thank Dwarkadas.J.Sanghvi College of Engineering for providing us with the resources and knowledge required for this project.

References

Abdurrahman G., Sintawati M., 2020. Implementation of XGBoost for classification of Parkinson's disease, *Journal of Physics Conference Series,* Vol. 1538, Article 012024, https://doi.org/10.1088/1742-6596/1538/1/012024

Fang P., 2022. Random forest algorithm based on speech for early identification of parkinson's disease, *Computational Intelligence and Neuroscience,* https://doi.org/ 10.1155/2022/3287068

Gomathy C.K., B.D. Kumarreddy, Varsha B., Varshini B., 2021. The parkinson's disease detection using machine learning techniques, *International Research Journal of Engineering and Technology*, Vol. 8, No. 10, pp. 440-444

Lamba R., Gulati T., Fahad H., Jain A., 2022. A hybrid system for Parkinson's disease diagnosis using machine learning techniques, *International Journal of Speech Technology,* Vol.25, pp. 583–593, https://doi.org/10.1007/s10772-021-09837-9

Li Y. and Wang P., 2017. Classification of parkinson's disease by decision tree based instance selection and ensemble learning algorithms, *Journal of Medical Imaging and Health Informatics*, Vol. 7, No. 2, pp. 444-452. https://doi.org/10.1166/jmihi.2017.2033

Little M.A., McSharry P.E., Hunter E.J., Ramig L.O., 2009. "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease, *IEEE Transactions on Biomedical Engineering*, Vol. 56, No. 4, pp. 1015 – 1022. https://doi.org/10.1109/TBME.2008.2005954.

Nishat M., Hasan T., Nasrulla S., Faisal F., 2021. Detection of parkinson's disease by employing boosting algorithms, *Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*

Sathiya T., Reenadevi R., Sathiyabhama B., 2021. Random forest classifier based detection of parkinson's disease, *Annals of the Romanian Society for Cell Biology*, Vol. 25, No. 5, pp. 2980-2987

Sharna R., 2021. Detecting parkinson's disease using machine learning, *International Journal of Innovations in Engineering Research and Technology*, Vol. 6. No. 7, pp. 267-269

Shamrat J., Asaduzzaman M. Rahman A.K.M.S., Hasan R.T.H., Tasnim Z., 2019. A comparative analysis of parkinson disease prediction using machine learning approaches, *International Journal of Scientific and Technology Research*, Vol. 8, No. 11, pp. 2576-2580

Shetty S. and Rao Y.S., 2016. SVM based machine learning approach to identify parkinson's disease using gait analysis, 2016 *International Conference on Inventive Computation Technologies (ICICT)*, 26-27 August 2016, Coimbatore, India https://doo.org/10.1109/INVENTIVE.2016.7824836

Wang X., Chen X., Wang Q., 2022. Early diagnosis of Parkinson's disease with speech pronunciation features based on XGBoost model, *IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI)*

Zhang J., 2022. Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of Parkinson's disease, *npj Parkinson's Disease*, Vol. 8, Article number: 13

Biographical notes

Prof. **A S..Revathi,** professor of, the Department of Electronics and Telecommunication, Dwarkadas J. Sanghvi College of Engineering, has 12 years of academic experience. She received her Master of Engineering degree in Communication Systems from Kumaraguru College of Technology. Her area of specialization includes Communication Networks/Systems, Image Processing, and Data Science. She is a member of the Indian Society of Technical Education (ISTE). Over the course of time, she has 6 research papers published internationally. Her research work has also been recognized by conferences, both nationally and internationally.

Dr. **K. Kavitha**, professor of, the Department of ECE, Kumaraguru College of Technology, has 23 years of academic and research experience. She is a member of professional bodies IEEE, ISTE, and HAM license holder. Her research interests include signal processing, communication systems, wireless communication, MIMO, OFDM, and wireless networks. She has published 3 book chapters, 50 papers in National/International conferences, and 20 papers in International/ National Journals.

K.Morparia is pursuing her Bachelor of Technology in Electronics and Telecommunications from Dwarkadas.J.Sanghvi College of Engineering, Mumbai. She is in her final year and has previous experience attending conferences, workshops, and publications. Her publications include a Face Recognition System that helps those in need. Her involvement in a college-affiliated team's coding department has improved her technical and problem-solving abilities, allowing her to delve into the world of AIML.

A.Kanabar is a student at the Dwarkadas.J.Sanghvi College of Engineering in Mumbai, where she is studying a Bachelor of Technology in Electronics and Telecommunications. One of her publications is a smart shopping cart that uses IoT. Her involvement in the marketing division of a team linked with a college has improved her problem-solving and interpersonal abilities, enabling her to succeed in related professions.

Aditi is a final year student, pursuing a Bachelor of Technology(B.Tech) in Electronics and Telecommunications from Dwarkadas J. Sanghvi College of Engineering, Mumbai. Her core competencies lie in Data Science focusing on Artificial Intelligence and Machine Learning. Being a former member of the college's robotics team, she has developed a knack for problem-solving and is always exploring new technologies and frameworks.

I.Thanekar is a student from Dwarkadas J Sanghvi College of Engineering from the Electronics and Telecommunication department. He has been a part of the Electronics system team in DJS Racing where he integrated electronic circuits compatible with our Electric vehicle. His recent publications include 'Overview of a transformer model', Schizophrenia classification using convolutional neural networks, and a research project on a GPS tracking safety device.