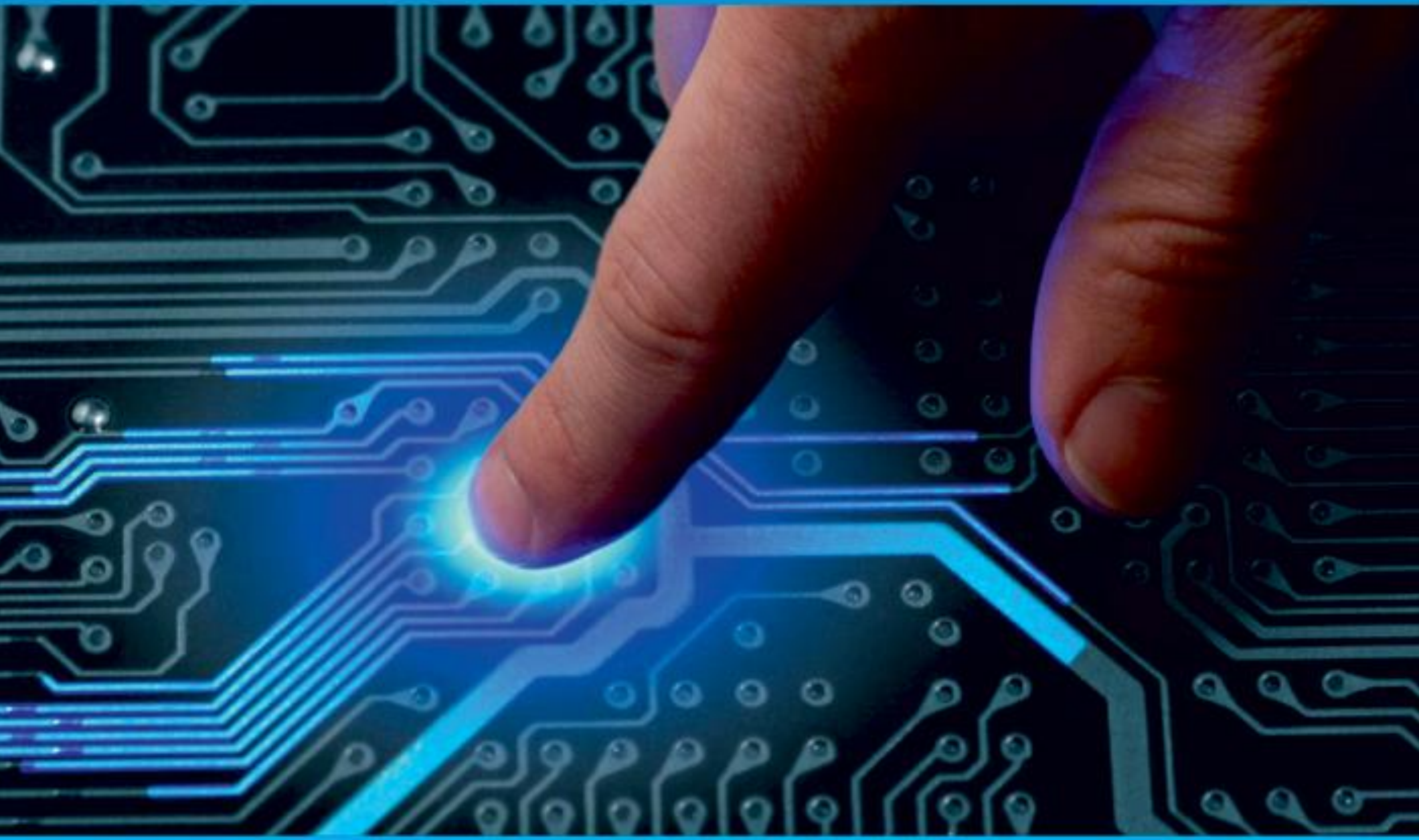
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204031 |

Diagnosis of Parkinson's Disease Using Ensemble Machine Learning Classifier

Navita

Assistant Professor, Department of Computer Science, Govt. P.G College for Women, Rohtak, India

ABSTRACT: In this growing era, technology plays a pivotal role in changing our lifestyle, enhancing healthcare and maintaining assets and resources in an efficient manner. In the healthcare sector, use of technology is growing day by day in order to address various emerging diseases and their symptoms. Among all growing diseases, Parkinson's disease (PD) is considering more attention towards it. It is the world's 2nd most common neurodegenerative disease that is affecting people worldwide. Its major symptoms appear in the form of termor in hands, legs and also stiffness in other body parts that highly affects the person daily life routine. Currently, there doesn't exist any medical cure and treatment till now. Its early treatment is possible only through early detection of this disease. Therefore this study introduces a model for early diagnosis of PD by utilizing ensemble classifier.

KEYWORDS: Ensemble classifier, Parkinson's disease, neurodegenerative

I. INTRODUCTION

Worldwide people are suffering from various neurological disorder like dementia, Alzheimer's disease, and Parkinson's Disease (PD), etc. PD is the world 2nd most common neurological disease that has affected around ten million people worldwide [1]. As per the WHO report in the past 25 years occurrence of Parkinson's disease (PD) has doubled. Disability and death among people are increasing gradually due to PD than any other neurological disorder (Launch of WHO's Parkinson Disease Technical Brief). Degradation of dopamine cells is the only reason behind the Parkisnon's disease. Dopamine cells acts as a messenger between nerve system and brain that acts as a messenger for handling the overall movement of the human body. A consequent reduction in dopamine cells generates unconscious shivering or agitation in the affected parts of the body and causes two types of symptoms: motor symptoms (tremors, bradykinesia slowness in movement) and non motor symptoms(like stiffness, rigidity, constipation, painful cramps, sleep disturbance, and extreme daytime sleepiness, etc.) [3]. Symptoms of PD progresses very slowly and differ from person to person. Up to the appearance of symptoms 50 to 80% of dopamine neurons have already died. Usual diagnosis of PD is done using few invasive procedures together with empirical testing and examinations. Procedures for invasive diagnosis of Parkinson's disease (PD) are very costly, ineffective, and need very sophisticated, inaccurate equipment. It takes new methods to diagnose Parkinson's disease. Therefore, techniques for ensuring therapies and diagnosing sickness should be modified to be simpler, less costly, and more dependable.

Therefore the integration of most non invasive technology offers a new way diagnosing PD among patients. Many researchers have worked with smart wearable sensors (like acoustic sensors, plantar sensors, force sensors, IMU(Inertial Measurement Unit), accelerometers, etc[4]) and employed ML techniques for the diagnosis of PD. It has been identified that vocal disorder is the major issue that can be evaluated for early diagnosis of PD[5]. Therefore this study aims to develop a model for early diagnosis of PD based on voice dataset, reducing the expenditures in diagnosing PD.

Key objective of this study are:
1. To develop a ML based model for early detection of PD.
2. This study also aims to review the work done by different researchers in this direction
3. Evaluating the model performance using different performance evaluators.
4. In last, comparison is done between the performances of different techniques used for development of the model.

II. RELATED STUDY

Several researchers have worked on different ML techniques for the classification of PD. This section makes a review about work made by different researchers in early detection of PD. Pahuja and Nagabhushan [6] perform a comparative analysis between different supervised ML techniques used for detection of PD. SVM, K-NN, and ANN are the selected techniques on which comparison is done and computed results shows that a feed-forward ANN combined with the Levenberg–Marquardt method achieved the best results when applied to the voice dataset. Ali et al. [7] proposed a multimodal framework using SVM and regression techniques for diagnosis of Parkinson's disease and implemented it on voice dataset. The framework proposed with SVM achieved the best performance. A comparative analysis between four different ML techniques named

SVM, RF, K-NN and LR models is performed by Govindu et al.[8] and find RF classifier model as the best classifier with 91.83% of accuracy and 0.95 of sensitivity. Another analysis is also done by Alshammri et al. [9] aong different ML techniques Machine SVM, RF, DT, KNN, and MLP and found SVM as the best classifier with 95% accuracy, 96% recall, precision 98%, and f1-score 97%. Another researcher Pramanik et cal. [10] develop a model by inducing various preprocessing techniques, such as data standardization,, dimensionality reduction etc. to enhance the efficiency of data and improving model performance. Different Machine Learning (ML) classifiers k-NN, SVM, RF, AdaBoost, LR) were used for PD classification. The developed model scored the best accuracy score of 94.10%. Vigneswari et al. [11] apply different ML techniques for the diagnosis of PD using voice signals and resulted that the Gradient Boosting algorithm achieved the best test accuracy score of 91.53%.

III. MATERIALS AND METHODOLOGY

Materials and methodology followed for the proposed model is described in this section. The model proposed for the detection of PD is depicted in Figure 1. In the proposed model first of all collected dataset is preprocessed and balanced using SMOTE method. After that train dataset is divided into 4 different train sets and each ML model is trained differently using different train dataset. Prediction made by each model is accessed and final prediction is done using averaging weighting score of each model.

3.1 Data Set Description

The implementation of the proposed model is done using voice dataset collected from 31 people out of which 23 people are suffering from PD (sixteen males and seven females) and 8 were Healthy people using microphone for recordings. The dataset is freely at the UCI Machine Learning Repository" [12]". A detailed description about the dataset features is described by Table 1.

3.1.1 Data Preprocessing

Preprocessing makes the model to learn the features in an efficient manner and eliminate redundant information [13]. The dataset was into Jupyter notebook as a "CSV file" using Python package. And screened the data in order to remove duplicates or null entries and then individually scaled the features using standard scalar represented by equation 1.

$$Standard\ scalar = \frac{X_i - mean(X)}{stdev(X)} \quad (1)$$

Table 1. Data Set Feature Description

Feature_name	Description
Name	Name ASCII name of subject and recording number (categorical variables)
MDVP:Fhi(Hz)	Max. vocal primary frequency (Numerical variables).
MDVP:Fo(Hz)	Avg. vocal primary frequency (Numerical variables)
MDVP:Flo(Hz)	Min. vocal primary frequency (Numerical variables).
MDVP:Jitter(Abs)	multiple measures of variation in primary frequency (Numerical variables).
MDVP:Jitter(%)	
MDVP: PPQ	
Jitter:DDP	
MDVP: RAP	
MDVP: Shimmer	Multiple measures of variation in amplitude (Numerical variables).
Shimmer: APQ3	
MDVP:Shimmer(dB)	
Shimmer: APQ5	
MDVP: APQ	
Shimmer: DDA	
NHR	ratio of noise to total components in the voice (Numerical variables)
HNR	
Status	0: HC and 1: PD
RPDE	complexity measures (Numerical variables).
D2	(Numerical variables: Signal fractal scaling exponent).

DFA	
Spread 1	Non linearity measures of fundamental frequency variation
Spread2	
PPE	

After that, the "status" column and found a great imbalancing among dataset i.e: PD: 147 and HC: 48. In order to avoid imbalancing, we apply Synthetic Minority Oversampling Technique (SMOTE) method to balance the dataset. This method improves minority data depending on earlier samples by building a linear co-relation using close features. Further split the data in to training and test dataset with a ratio of 70:30%. Model is trained using train data and tested on test data.

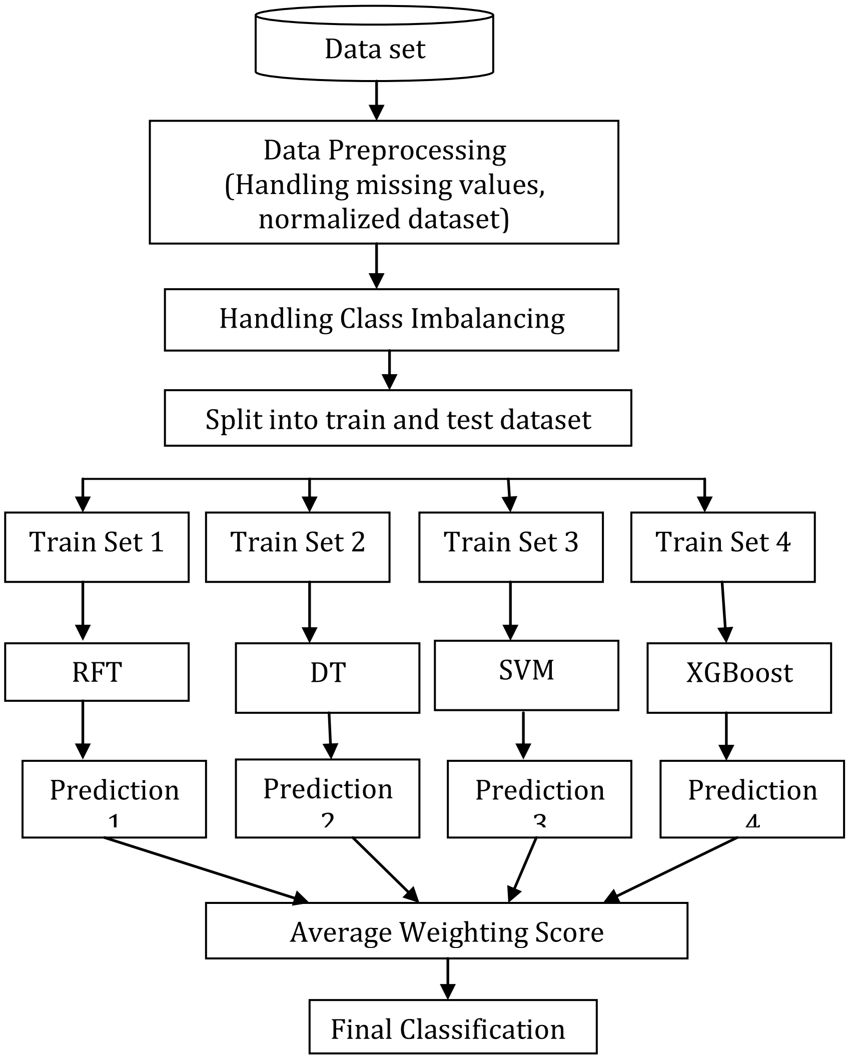


Figure 1. Proposed Parkinson's disease Diagnosis model

3.3 Machine Learning Techniques

3.3.1 Decision Tree (DT)

The Decision Tree algorithm is a key component of the proposed model in the paper. It constructs a tree-like structure where internal nodes make decisions based on features, and leaf nodes represent class labels. The algorithm splits the data recursively, maximizing information gain or minimizing Gini impurity[14].

$$Split\ crit = \arg\max(Info\ Gain, Gini\ Index) \quad (2)$$

The Information Gain measures the reduction in entropy or impurity after the split. The decision tree algorithm offers interpretability and handles both categorical and numerical features.

3.3.2 Support Vector Machine (SVM)

It is a powerful supervised ML techniques commonly used in various domains, including healthcare systems. In the context of this paper, SVM is utilized as one of the techniques. SVM is utilized as one of the techniques for predicting and classifying Parkinson’s disease using voice dataset. Mathematically, SVM aims to find an optimal hyperplane that can separate the data points into different classes while maximizing the margin, which represents the distance between the hyperplane and the closest data points from each class. The hyperplane can be represented by the equation 1:

$$w^T * x + b = 0, \qquad (3)$$

Where, w is the weight vector perpendicular to the hyperplane, x is the input feature vector, and b is the bias term. The use of SVM in the proposed model leverages its capability to handle both linear and non-linear data patterns, making it suitable for identifying complex voice signals.

3.3.3 Random Forest (RFT)

RFT is another important algorithm used in the proposed model that combines multiple decision trees to make predictions[15]. Each decision tree is trained on a random subset of the data and features, and the final prediction is determined by majority voting.

3.3.4 Extreme Gradient Boosting Machine (XGBoost)

It is a powerful ML technique employed in the proposed model. XGBoost is an optimized gradient boosting framework that utilizes an ensemble of weak prediction models to make accurate predictions[16]. The objective function of XGBoost combines a loss function and a regularization term to measure the model's performance and complexity, respectively. It aims to minimize the overall loss while penalizing complex models. XGBoost builds an ensemble of weak prediction models sequentially. Each model is trained to correct the errors done by the earlier models. XGBoost employs decision trees as weak models. The decision trees are grown in a greedy manner, optimizing the objective function to find the best splits at each node. XGBoost includes regularization techniques such as L1 regularization (Lasso) and L2 regularization (Ridge) to prevent overfitting and control model complexity. The regularization term is added to the objective function, influencing the learning process of the weak models. By utilizing the XGBoost algorithm, the proposed model leverages the power of boosting and gradient optimization to effectively classify the people into healthy and PD class.

3.4 Performance Evaluators

The proposed model utilizes several performance evaluators to assess its effectiveness. These evaluators provide quantitative measures to evaluate the model's performance and its ability to classify the people into PD and Healthy class. The accuracy, sensitivity, specificity, F1-Score are used as performance evaluators in the study:

IV.RESULTS AND DISCUSSION

The proposed model is implemented on ‘UCI Voice dataset in Jupyter Notebook. The collected dataset contains voice recordings of people. The data is highly utilized for diagnosing Parkinson’s disease. Results obtained through the implementation of the proposed model are computed in terms of precision, accuracy, and recall as described in Table 2.

Table 2. Proposed Model performance results

ML-Techniques	Accuracy	Recall	Precision	F1_Score
SVM	97.42	96.2	96.1	96
DT	96.12	95.5	97	95.7
RFT	97	98	96.2	96.8
XGBT	97.8	97.32	97.1	96.36
Proposed Model (Average Weighting Classifier)	98.7	97.8	98.3	97.7

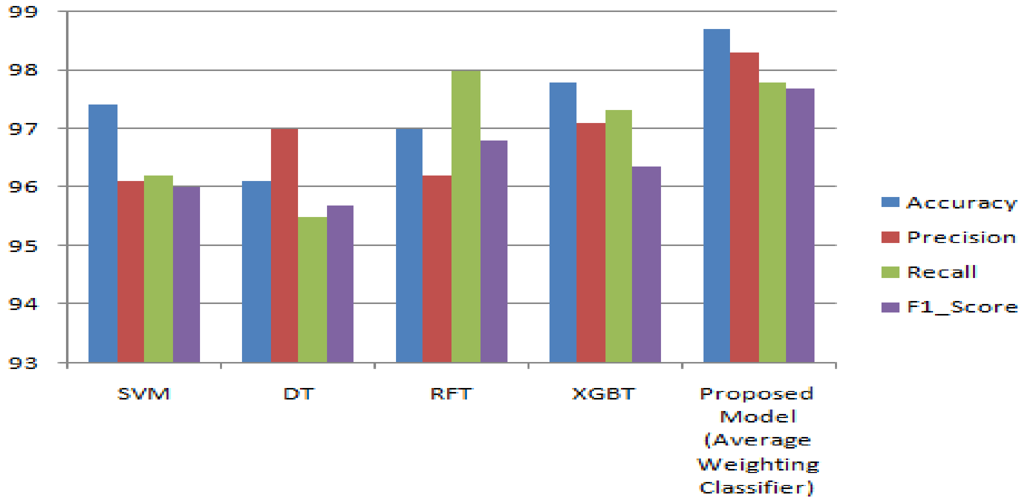


Figure 2. Performance results of different classifiers

Performances of different ML techniques and its corresponding results are graphically represented in Figure 2. The results reveal that the proposed model attained the best accuracy score with the majority voting classifier which is 98.7%.

V. CONCLUSION

The study concludes PD is the world 2nd most neurological diseases highly impacting the aged people. The current study emphasizes the utilization of smart sensor to collect and monitor significant biomarkers of PD that are beneficial in diagnosing and monitoring the progression of PD that will reduce regular visits to hospitals. The current work also emphasizes on utilization of ML techniques for effective analysis of data collected from these sensors. In the current work, a ML model is proposed for effective diagnosis of PD from voice signals. Voice signals were recorded using microphone and collected from UCI repository. The experimental results show that the model proposed with average weighting classifier scored the highest accuracy score of 98.7% in comparison to other classification techniques. The developed model helps in decreasing treatment expenditure by offering early diagnostics.

REFERENCES

[1] World Health Organization, *Parkinson disease: a public health approach: technical brief*. Geneva: World Health Organization, 2022. Accessed: Nov. 28, 2022. [Online]. Available: https://apps.who.int/iris/handle/10665/355973

[2] "Launch of WHO's Parkinson disease technical brief." Accessed: Mar. 16, 2024. [Online]. Available: https://www.who.int/news/item/14-06-2022-launch-of-who-s-parkinson-disease-technical-brief

[3] A. Dadu *et al.*, "Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts," *npj Parkinsons Dis.*, vol. 8, no. 1, p. 172, Dec. 2022, doi: 10.1038/s41531-022-00439-z.

[4] H. Zhao *et al.*, "Wearable sensors and features for diagnosis of neurodegenerative diseases: A systematic review," *DIGITAL HEALTH*, vol. 9, p. 205520762311735, Jan. 2023, doi: 10.1177/20552076231173569.

[5] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal

and incipient Parkinson’s disease: A longitudinal case study,” *Brain and Cognition*, vol. 56, no. 1, pp. 24–29, Oct. 2004, doi: 10.1016/j.bandc.2004.05.002.

[6] G. Pahuja and T. N. Nagabhushan, “A Comparative Study of Existing Machine Learning Approaches for Parkinson’s Disease Detection,” *IETE Journal of Research*, vol. 67, no. 1, pp. 4–14, Jan. 2021, doi: 10.1080/03772063.2018.1531730.

[7] L. Ali, S. U. Khan, M. Arshad, S. Ali, and M. Anwar, “A Multi-model Framework for Evaluating Type of Speech Samples having Complementary Information about Parkinson’s Disease,” in *2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, Swat, Pakistan: IEEE, Jul. 2019, pp. 1–5. doi: 10.1109/ICECCE47252.2019.8940696.

[8] A. Govindu and S. Palwe, “Early detection of Parkinson’s disease using machine learning,” *Procedia Computer Science*, vol. 218, pp. 249–261, 2023, doi: 10.1016/j.procs.2023.01.007.

[9] R. Alshammri, G. Alharbi, E. Alharbi, and I. Almubark, “Machine learning approaches to identify Parkinson’s disease using voice signal features,” *Front. Artif. Intell.*, vol. 6, p. 1084001, Mar. 2023, doi: 10.3389/frai.2023.1084001.

[10] A. Pramanik and A. Sarker, “Parkinson’s Disease Detection from Voice and Speech Data Using Machine Learning,” in *Proceedings of International Joint Conference on Advances in Computational Intelligence*, M. S. Uddin and J. C. Bansal, Eds., in Algorithms for Intelligent Systems. , Singapore: Springer Singapore, 2021, pp. 445–456. doi: 10.1007/978-981-16-0586-4_36.

[11] D. A. Vigneswari and J. Aravinth, “Parkinson’s disease Diagnosis using Voice Signals by Machine Learning Approach,” in *2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, Bangalore, India: IEEE, Aug. 2021, pp. 869–872. doi: 10.1109/RTEICT52294.2021.9573689.

[12] “UCI Machine Learning Repository: Parkinsons Data Set.” Accessed: Dec. 04, 2022. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/parkinsons

[13] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, “Data Preprocessing for Supervised Leaning,” vol. 1, no. 1, p. 8, 2006.

[14] F. M. Awan, Y. Saleem, R. Minerva, and N. Crespi, “A comparative analysis of machine/deep learning models for parking space availability prediction,” *Sensors*, vol. 20, no. 1, p. 322, 2020.

[15] E. Scornet, “Random Forests and Kernel Methods,” *IEEE Trans. Inform. Theory*, vol. 62, no. 3, pp. 1485–1500, Mar. 2016, doi: 10.1109/TIT.2016.2514489.

[16] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 0 ed. Chapman and Hall/CRC, 2012. doi: 10.1201/b12207.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

📱 9940 572 462 ✆ 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details