Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Parkinsons Disease Classification

Chandrasegar T, Vaishwik Vishwakarma, Rajwardhan Sikarwar, Rityuvraaj Patil

SCORE, VIT, Vellore, India
Email: chandrasegar.t@vit.ac.in

Parkinson's ailment is a degenerative ailment that affects movement, making early diagnosis critical for higher treatment and management. In this study, we assess the running performance of various machine studying fashions—Logistic Regression, Random Forest, XGBoost, aid Vector Classifier, K-Nearest buddies, Naive Bayes, and selection trees—on voice facts for Parkinson's ailment category. techniques consisting of (EDA), function scaling, and pass-validation have been conducted to enhance model overall performance. To address the magnificence imbalance, we used SMOTE, and version assessment became primarily based on accuracy, precision, F1-score metrics. XGBoost appeared because of the nice version, accomplishing 93.33% accuracy and an F1-score of 95.65%, successfully distinguishing PD sufferers and healthy people. This observation demonstrates the potential of ML in assisting early analysis of Parkinson's sickness, imparting clinicians with a precious tool to improve patient results.
Keywords: Machine Learning,Voice Data,Logistic Regression,Random Forest,XGBoost,Support Vector Classifier (SVC),K-Nearest Neighbors (KNN).

1. Introduction

Parkinson's disorder (PD) is a chronic and step by step worsening neurological disease that influences motor feature. People with PD regularly revel in symptoms such as tremors, muscle stiffness, slow movement, and problems with balance. because the ailment advances, it extensively reduces patients' satisfactory of lifestyles, necessitating complex hospital therapy. Early detection of Parkinson's disease is critical for enhancing remedy outcomes, as well-timed intervention can assist manage signs and gradual ailment development. However, diagnosing PD in its early degrees stays challenging due to the overlap of symptoms with other neurological situations, making accurate diagnosis difficult.

Recent improvements in device mastering offer promising solutions to enhance the accuracy and performance of PD prognosis. by using reading non-invasive biomarkers, consisting of voice recordings and other measurable physiological characteristics, gadget studying

fashions can become aware of subtle styles indicative of early PD. Those fashions provide a more goal, automated diagnostic approach, reducing the subjectivity and variability inherent in traditional clinical strategies. more objective, automated diagnostic approach, reducing the subjectivity and variability inherent in traditional clinical strategies.

This research investigates the overall performance of various system gaining knowledge of algorithms in classifying Parkinson's disorder the use of voice dimension facts. The examine assesses algorithms such as Logistic Regression, Random Forest, XGBoost, guide Vector Classifier (SVC), k-Nearest friends (KNN), Naive Bayes, and selection trees. Our technique contains techniques like exploratory facts evaluation (EDA), function scaling, move-validation, and sophistication balancing the use of SMOTE to optimize version overall performance. through comparing metrics together with accuracy, precision, consider, F1-score. we intention to discover the best set of rules for Parkinson's category. This observation gives precious facts about the development of AI-driven diagnostic equipment, which can assist healthcare specialists in the early detection of PD and enhance patient effects.

Parkison Dataset

This examine utilizes a dataset comprising 195 voice recording samples, each with twenty-two wonderful attributes associated with vocal characteristics. The target variable, categorized as popularity, suggests whether a person is recognized with Parkinson's disorder (repute = 1) or is healthful (popularity = zero). The dataset frequently captures fundamental voice functions, which are critical in differentiating individuals with Parkinson’s ailment from healthy controls, focusing on speech irregularities.

Key Features:

Frequency Measurements:

MDVP

(Hz): The average essential frequency (pitch) of the voice.

MDVP

(Hz): The maximum fundamental frequency.

MDVP

(Hz): The minimal essential frequency.

Jitter Measurements (Pitch variability):

MDVP

(%): the proportion of frequency perturbation.

MDVP

(Abs): the absolute jitter price.

MDVP

: Relative amplitude perturbation.

MDVP

: 5-factor period perturbation quotient.

Jitter

: The average absolute distinction among durations.

Shimmer Measurements (Amplitude variability):

MDVP

: Amplitude perturbation percent.

MDVP

(dB): Shimmer measured in decibels.

Shimmer

: three-point amplitude perturbation quotient.

Shimmer

: five-factor amplitude perturbation quotient.

MDVP

: Amplitude perturbation quotient.

Shimmer

: average absolute distinction of amplitude variations among periods.

sign-to-Noise Ratio:

NHR: Noise-to-harmonics ratio, a measure of signal clarity.

HNR: Harmonics-to-noise ratio, indicating the degree of vocal periodicity.

Nonlinear Dynamics:

RPDE: Recurrence duration density entropy, capturing irregularities in vocal alerts.

DFA: Detrended fluctuation analysis, used to evaluate signal self-similarity.

PPE: Pitch period entropy, reflecting the disorder in the voice signal.

Unfold Measurements:

spread1 and spread2: features related to the distribution and form of the voice's resonant frequencies (formants).

The dataset includes twenty-three columns, together with 22 features representing various factors of voice signal patterns, and a goal column (fame) for classifying Parkinson's sickness status. among the features relate to pitch and amplitude irregularities, which might be regularly located in individuals with Parkinson's sickness.

Statistics precis:

• The dataset consists of 195 samples, with 147 individuals recognized with Parkinson's ailment (status = 1) and forty-eight wholesome controls (fame = zero).

• No missing values are gift, ensuring completeness for analysis.

• The functions are usually continuous variables, making scaling necessary for correct gadget getting to know version performance.

This dataset affords a beneficial useful resource for studying various supervised system studying models to predict the presence of Parkinson's disorder, serving as a benchmark for growing automatic diagnostic gear.

A. Random Forest

It creates multiple decision tree fashions via a technique known as bagging, after which combines their consequences to achieve the most correct results. In each tree, variable selection takes place at each breakup. amongst these trees, the only with the bottom fashionable deviation is chosen as the first-rate. Random woodland (RF) has frequently appeared as fairly powerful for classification responsibilities.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \qquad (1)$$

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}} \qquad (2)$$

A. Logistic Regression

A category algorithm is applied for each binary and multiclass classification duties. It is miles usually applied whilst the output includes classes, along with in sentiment evaluation, photo, and video type, among other use instances.

$$P(x) = \frac{1}{1+e^{-(\beta_0+\beta_1)}} \qquad (3)$$

B. Naïve Bayes

A classification algorithm computes the probability of each class and makes selections by using evaluating these probabilities using Bayes' theorem. It operates beneath the assumption that all functions are independent of every other, given the class. in spite of its simplistic assumption of function independence, this set of rules is pretty effective for various type obligations, consisting of textual content classification and unsolicited mail filtering.

$$P\left(\frac{y}{X}\right) = \frac{P\left(\frac{X}{y}\right)P(y)}{P(X)} \qquad (4)$$

C. KNN

K-Nearest Neighbors is an easy gadget getting to know algorithms categorized under non-parametric methods. it really works through determining the value of a question point primarily based at the 'k' closest acquaintances, in which the prediction is made either by way of choosing the maximum frequent elegance or calculating the average cost of those neighbors.

D. SVC

Support Vector Classifier (SVC) is a system learning set of rules designed for classification obligations. It operates through locating the most fulfilling keeping apart hyperplane that maximizes the margin among impressive lessons inside the records, even in a high-dimensional space. This technique lets in for effective classification via creating a clean boundary between categories.

E. XG Boost

This system of getting to know set of rules, derived from selection timber, is especially recognized for both its speed and performance for both category and regression obligations. It enhances model accuracy through techniques like function choice, selection tree pruning, and leveraging parallel computing for faster processing.

F. Decision Tree

It represents all ability choice to get the preferred outcome. First, features are located the usage of entropy or facts advantage, and based totally at the results, the most relevant input capabilities are identified. The dataset is then broken up according to the chosen capabilities, with positive constraints implemented at some point of the technique.

$$E(s) = \sum_{i=1}^{c} -p_i \log_2 p_i \qquad (6)$$

Confusion Matrix:

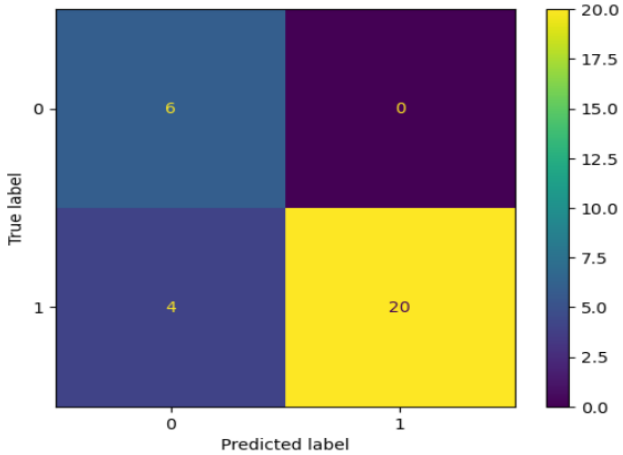

Figure 1: Random Forest Classifier

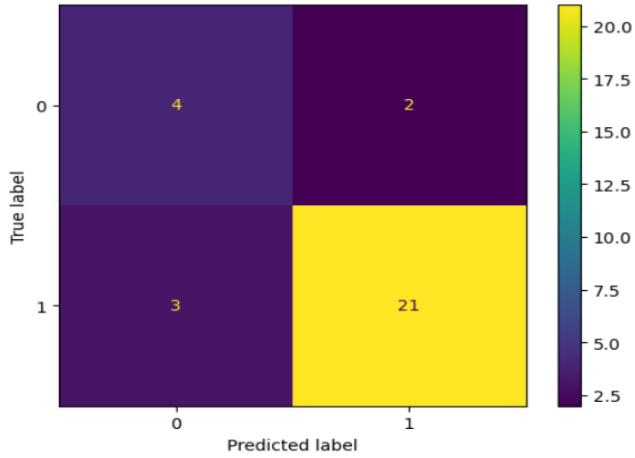

Figure 2: Logistic Regression

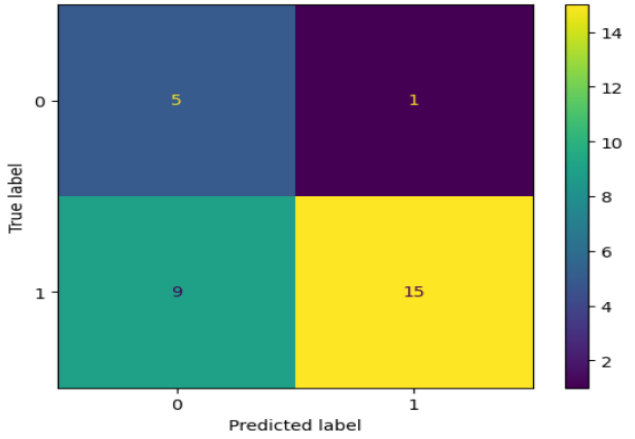

Figure 3: Naïve Bayes

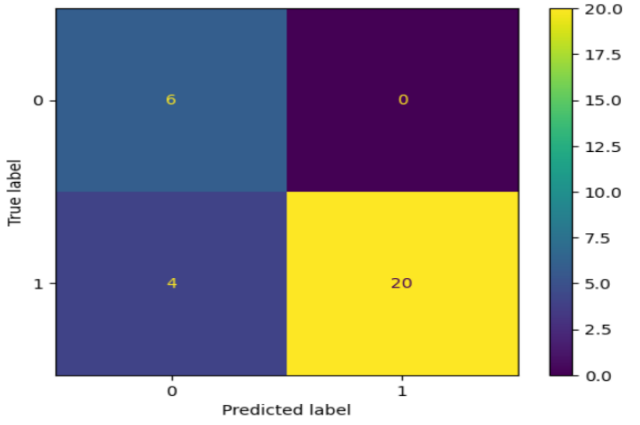

Figure 4: K-Nearest Neighbour

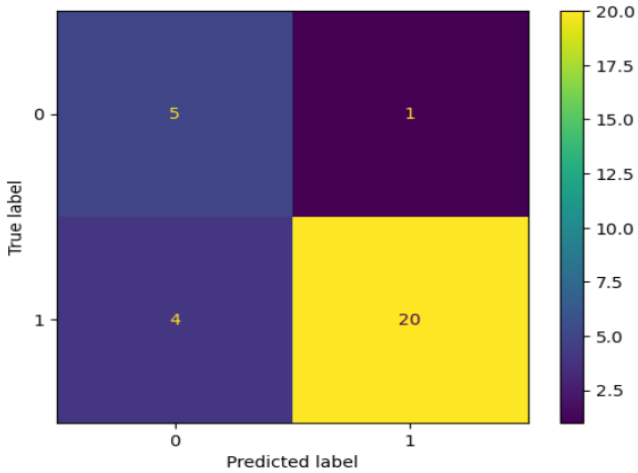

Figure 5: Decision Tree

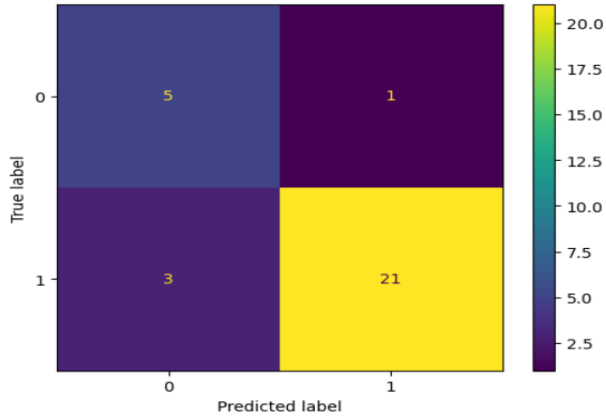

Figure 6: SVC

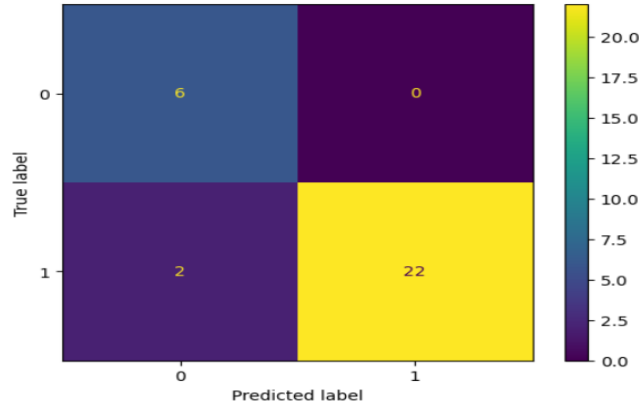

Figure 7: XG Boost

2. Results and discussions

In this task, we utilized a dataset containing voice measurements from individuals stumble on in the event that they have any Parkinson's ailment possibilities the usage of diverse gadget studying algorithms. The models examined encompass Logistic Regression, Random woodland, XGBoost, assist Vector Classifier (SVC), ok-Nearest neighbours (KNN), Naive Bayes, and choice bushes. the important thing metrics for version assessment have been accuracy, precision, recall, F1-rating.

version overall performance:

Model Performance:

1. Logistic Regression:

Logistic Regression done reasonable overall performance however showed barely decrease accuracy in comparison to other models like XGBoost. The precision and don't forget values have been robust, indicating the model's strength to correctly become aware of both tremendous and bad instances efficaciously.

2. Random Forest:

This model finished nicely in terms of precision, indicating a excessive wide variety of actual positives. but, take into account became slightly decrease, implying some false negatives. Random woodland is green at type tasks and its ensemble nature improves average overall performance.

3. XGBoost:

XGBoost outperformed changed into found to be appearing higher than different fashions with its high accuracy and F1-rating. The capacity to deal with magnificence imbalance and optimize studying rates makes it a robust model for Parkinson's ailment prediction. This version excelled in each precision and consider, ensuring a balanced performance.

4. Support Vector Classifier (SVC):

SVC confirmed strong overall performance with a high F1-score and accuracy. The version effectively located the most beneficial hyperplane to separate the instructions with a terrific margin. however, it did not surpass XGBoost in accuracy.

5. K-Nearest Neighbors (KNN):

KNN displayed performance much like Random forest, with perfect precision however barely decrease bear in mind. The set of rule's simplicity and non-parametric nature made it an affordable desire for this classification task, although it struggled with more complicated decision barriers.

6. Naive Bayes:

Naive Bayes had the bottom overall performance amongst all fashions. notwithstanding its simplicity, it was not able to take care of the intricacies of this dataset correctly, largely because of its assumption of characteristic independence, which won't maintain actual on this context.

7. Decision Trees:

o Decision Trees performed similarly to Logistic Regression, with solid precision and recall scores. However, it was slightly less robust compared to ensemble methods like Random Forest and XGBoost.

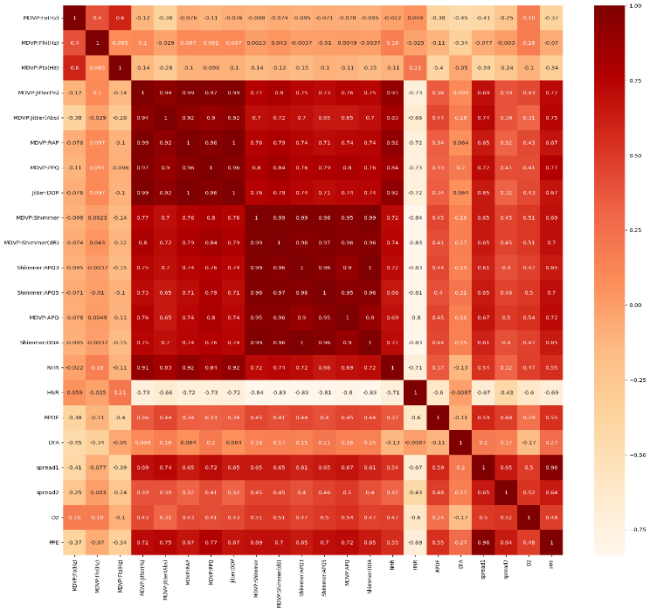

Figure 7: HeatMap

10-fold Cross-Validation Analysis

The go-validation outcomes display the robustness of the XGBoost version, which consistently outperformed different algorithms across all 10 folds. It achieved the very best common accuracy (ninety two.36%), precision (99.13%), and F1-rating (ninety four.fifty seven%). Random forest and SVC additionally performed properly, however XGBoost's capability to address feature interactions and class imbalance gave it an part.

go-validation facilitates make certain that the version isn't always overfitting and is acting properly on unseen statistics, supplying a complete and reliable assessment of model performance in this study.

This thorough validation procedure confirms that XGBoost is the great-acceptable algorithm for Parkinson's disorder category in this dataset, handing over the maximum constant consequences throughout a couple of metrics.

Figure 8: 10-fold Cross-Validation Analysis

G. Models vs. Performance

Models are the mathematical or computational structures worried in statistics evaluation to version, predict or classify events within the real world based on input facts whereas, performance is a measure of ways nicely the fashions sporting out the previously mentioned targets. one of the vital criteria in model building is the alternate between the models' sophistication and their accuracy.

MODELS	ACCURACY	PRECISION	RECALL	F1-SCORE
RANDOM FOREST	86.66%	100%	83.33%	90.90%
LOGISTIC REGRESSION	83.33%	91.3%	87.5%	89.36%
NAÏVE BAYES	66.66%	93.75%	62.5%	75%
KNN	86.66%	100%	83.33%	90.90%
SVC	86.66%	95.45%	87.5%	91.30%
XG BOOST	93%	100%	91.66%	95.65%
DECISION TREE	83.33%	95.23%	83.33%	88.88%

Table 3: Tabulation Of Analysis

3. Conclusion

This research explored several ML algorithms to classify Parkinson's disease the usage of voice measurements. The fashions evaluated covered Logistic Regression, Random wooded area, XGBoost, SVC, KNN, Naive Bayes, and selection bushes, with performance assessed through numerous performance metrics.

Amongst these, XGBoost stood out because it was the pinnacle performer, handing over the first-class accuracy (93.33%) and F1-score of 95. sixty-five%, demonstrating its capacity to deal with complex datasets successfully. Each Random wooded area and SVC additionally supplied sturdy outcomes, showcasing the electricity of tree-based and margin-based

classifiers in this context. In contrast, easier algorithms like Naive Bayes and KNN struggled to perform as nicely, likely because of complexity of the features.

The use of 10-fold pass-validation, XGBoost always outperformed the opposite fashions, confirming its reliability in distinguishing Parkinson's patients from healthy individuals. In end, this evaluation throws mild on the importance of the using ML models for the evaluation of critical diseases like Parkinsons.

References

[1] Kundu, M., Nashiry, M. A., Dipongkor, A. K., Sumi, S. S., & Hossain, M. A. (2021). An optimized machine learning approach for predicting Parkinson's disease. Int. J. Mod. Educ. Comput. Sci.(IJMECS), 13(4), 68-74.

[2] Karabayir, I., Goldman, S. M., Pappu, S., & Akbilgic, O. (2020). Gradient boosting for Parkinson's disease diagnosis from voice recordings. BMC Medical Informatics and Decision Making, 20, 1-7.

[3] Solana-Lavalle, G., Galán-Hernández, J. C., & Rosas-Romero, R. (2020). Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. Biocybernetics and Biomedical Engineering, 40(1), 505-516.

[4] Ali, L., Chakraborty, C., He, Z., Cao, W., Imrana, Y., & Rodrigues, J. J. (2023). A novel sample and feature dependent ensemble approach for Parkinson's disease detection. Neural Computing and Applications, 35(22), 15997-16010.

[5] Lamba, R., Gulati, T., Alharbi, H. F., & Jain, A. (2022). A hybrid system for Parkinson's disease diagnosis using machine learning techniques. International Journal of Speech Technology, 1-11.

[6] Lamba, R., Gulati, T., & Jain, A. (2022). A hybrid feature selection approach for parkinson's detection based on mutual information gain and recursive feature elimination. Arabian Journal for Science and Engineering, 47(8), 10263-10276.

[7] Ge, W., Lueck, C., Suominen, H., & Apthorp, D. (2023). Has machine learning over-promised in healthcare?: A critical analysis and a proposal for improved evaluation, with evidence from parkinson's disease. Artificial Intelligence in Medicine, 139, 102524.

[8] Chen, H. L., Huang, C. C., Yu, X. G., Xu, X., Sun, X., Wang, G., & Wang, S. J. (2013). An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert systems with applications, 40(1), 263-271.

[9] Sabeena, B., Sivakumari, S., & Teressa, D. M. (2022). Optimization-Based Ensemble Feature Selection Algorithm and Deep Learning Classifier for Parkinson's Disease. Journal of Healthcare Engineering, 2022(1), 1487212.

[10] Ashour, A. S., Nour, M. K. A., Polat, K., Guo, Y., Alsaggaf, W., & El-Attar, A. (2020). A novel framework of two successive feature selection levels using weight-based procedure for voice-loss detection in Parkinson's disease. Ieee Access, 8, 76193-76203.

[11] Lamba, R., Gulati, T., Al-Dhlan, K. A., & Jain, A. (2021). A systematic approach to diagnose Parkinson's disease through kinematic features extracted from handwritten drawings. Journal of Reliable Intelligent Environments, 1-10.

[12] Parisi, L., RaviChandran, N., & Manaog, M. L. (2018). Feature-driven machine learning to improve early diagnosis of Parkinson's disease. Expert Systems with Applications, 110, 182-190.

[13] Qasim, H. M., Ata, O., Ansari, M. A., Alomary, M. N., Alghamdi, S., & Almehmadi, M. (2021). Hybrid feature selection framework for the Parkinson imbalanced dataset prediction problem. Medicina, 57(11), 1217.

[14] Solana-Lavalle, G., & Rosas-Romero, R. (2021). Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation. Biomedical Signal Processing and Control, 66, 102415.
[15] Ul Haq, A., Li, J., Memon, M. H., Ali, Z., Abbas, S. Z., & Nazir, S. (2020). Recognition of the Parkinson's disease using a hybrid feature selection approach. Journal of Intelligent & Fuzzy Systems, 39(1), 1319-1339.
[16] Zhang, T., Zhang, Y., Sun, H., & Shan, H. (2021). Parkinson disease detection using energy direction features based on EMD from voice signal. Biocybernetics and Biomedical Engineering, 41(1), 127-141.
[17] Jatoth, C., Neelima, E., Mayuri, A. V. R., & Annaluri, S. R. (2022). Effective monitoring and prediction of Parkinson disease in Smart Cities using intelligent health care system. Microprocessors and Microsystems, 92, 104547.
[18] Ramani, R. G., & Sivagami, G. (2011). Parkinson disease classification using data mining algorithms. International journal of computer applications, 32(9), 17-22.
[19] Wodzinski, M., Skalski, A., Hemmerling, D., Orozco-Arroyave, J. R., & Nöth, E. (2019, July). Deep learning approach to Parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification. In 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 717-720). IEEE.
[20] Hariharan, M., Polat, K., & Sindhu, R. (2014). A new hybrid intelligent system for accurate detection of Parkinson's disease. Computer methods and programs in biomedicine, 113(3), 904-913.
[21] Sajal, M. S. R., Ehsan, M. T., Vaidyanathan, R., Wang, S., Aziz, T., & Mamun, K. A. A. (2020). Telemonitoring Parkinson's disease using machine learning by combining tremor and voice analysis. Brain informatics, 7(1), 12.
[22] Gourav Bathla, Sanoj Kumar, Harish Garg, Deepika Saini. "Artificial Intelligence in Healthcare - Emphasis on Diabetes, Hypertension, and Depression Management", CRC Press, 2024
[23] Talbi, Mohammed. "Safeguarding IoT Networks Using Machine Learning for Intrusion Detection & Prevention", The George Washington University, 2024
[24] Lahmiri, Salim. "Parkinson's disease detection based on dysphonia measurements", Physica A Statistical Mechanics and its Applications, 2017.
[25] Mobile Radio Communications and 5G Networks", Springer Science and Business Media LLC, 2024
[25] Md.Ariful Islam, Md.Ziaul Hasan Majumder, Md.Alomgeer Hussein, Khondoker Murad Hossain, Md.Sohel Miah. "A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets", Heliyon, 2024
[26] T. Chandrasegar, P. Viswanathan. "Dimensionality reduction of a phishing attack using decision tree classifier", 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), 2019
[27] Ferreiro-Vera, Carlos, Feliciano Priego-Capote, Mónica Calderón-Santiago, and María D. Luque de Castro. "Global metabolomic profiling of human serum from obese individuals by liquid chromatography–time-of f light/mass spectrometry to evaluate the intake of breakfasts prepared with heated edible oils", Food Chemistry, 2013.
[28] Mickalson, Sarah B.. "Detecting Data Poisoning Attacks on Smart Farm Devices Using Machine Learning", The George Washington University, 2024
[28] Stuart H Rubin, Lydia Bouzar-Benlabiod. "Reuse in Intelligent Systems", CRC Press, 2020
[29] Elmehdi Benmalek, Jamal Elmhamdi, Abdelilah Jilbab. "Voice Assessments for Detecting Patients with Parkinson's Diseases in Different Stages", International Journal of Electrical and Computer Engineering (IJECE), 2018

[30] Jaiteg Singh, S B Goyal, Rajesh Kumar Kaushal, Naveen Kumar, Sukhjit Singh Sehra. "Applied Data Science and Smart Systems Proceedings of 2nd International Conference on Applied Data Science and Smart Systems 2023 (ADSSS 2023) 15-16 Dec, 2023, Rajpura, India", CRC Press, 2024

[31] Sadanori Konishi. "Introduction to Multivariate Analysis - Linear and Nonlinear Modeling", CRC Press, 2014

[32] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 56(4), 1015-1022.

[33] Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1998). Speech impairment in a large sample of patients with Parkinson's disease. Behavioural Neurology, 11(3), 131-137.

[34] Md.Ariful Islam, Md.Ziaul Hasan Majumder, Md.Alomgeer Hussein, Khondoker Murad Hossain, Md.Sohel Miah. "A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets", Heliyon, 2024

[35] S. Sharanyaa, P N. Renjith, K. Ramesh. "Classification of Parkinson's Disease using Speech Attributes with Parametric and Nonparametric Machine Learning Techniques", 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020