Journal of Engineering, Science and Mathematics

JESM

Volume 02, Number 01, Pages 1-10, Year 2023
https://jesm.in/archives/

Research Paper

Applying Machine Learning Algorithms to Analyse Parkinson's Disease in the Age Group of 50+

Rijwan Khan, Mansi Gupta, Khushi Patel, Kashish Gupta

Department of Computer Science and Engineering ABES institute of technology, Ghaziabad, India

Correspondence should be addressed to Rijwan Khan; rijwankhan786@gmail.com

Handling Editor: Surbhi Gupta

The use of machine learning techniques in telemedicine to identify Parkinson's disease (PD) in its early stages is explored in this paper. PD is a neurodegenerative condition that primarily affects older people. Early detection is essential for effective management and treatment, but for patients, physical visits can be difficult due to mobility and communication issues. The study used the MDVP (Multidimensional Voice Program) audio data from thirty PD patients and healthy individuals to train four classification results from Support Vector Machine (SVM), Random Forest, K-Nearest Neighbour (KNN), Logistic Regression, AdaBoost classifier, Decision tree classifiers, and stacking classifier of ensemble learning technique were compared. On balanced data, the stacking classifier ensemble learning technique has 97% detection accuracy. Furthermore, these outcomes outperform recent literature-based research. The most reliable Machine Learning (ML) method for the detection of Parkinson's disease was discovered.

Keywords: *Parkinson's disease, K-Nearest Neighbours, Logistic Regression, AdaBoost, Ensemble techniques, Support vector machine (SVM), Stacking classifier, balanced dataset.*

1. Introduction

Parkinson's disorder is an illness of the neurological system that impacts people over the age group 50 [1], and PD is an illness of the nerve system that affects body mobility. It's a long-term and progressive condition. UPDRS is an acronym that means "Unified Parkinson's Disease Rating Scale" and it's the clinic's main tool for detecting PD. [2]. It is an example of neurodegenerative disease that occurs when nerve cells in the brain gradually lose function and eventually die. In our aging culture, neurodegenerative illness is one of the most severe health issues [3]. It primarily affects the body's motor system that supports motor function in the nervous system; with the increase in illness or time, it generates non-motor symptoms. PD diagnosis based on speech signals such as voice disorder (Dysphonia) is a significant symptom of PD [4] [5] [27]. The problem of PD patients has increased in this COVID-19 pandemic situation because people cannot take more care [6]. A lot of deep learning, machine learning, and AI-based techniques are defined to detect symptoms [7]. Recently, rather than these techniques, various kinds of sensors have been used as a diagnostic tool to diagnose PD signs [8]. As the condition worsens over time, Parkinson's early detection is needed for detecting the signs. Recent Parkinson's disease telemedicine investigations focused on vocal disorder-based systems, with machine learning classifiers for Dysphonia, a voice condition applied to PD [9-12]. Patients with PD handwriting also degrade with time [13]. Machine learning models based on the MR images of PD patients also give good accuracy by creating a brain heat map [14]. AI systems are gaining popularity in medicine because they can manage large amounts of data and make accurate predictions [15]. In 1817, an English physician named James Parkinson was the first to say that 10 million people globally have Parkinson's disorder annually [16].

Parkinson's disease symptoms are characterized by a person's movement [1-26]:

- Motor Symptoms: Motor symptoms are movement related. Four prominent deficits are postural instability, rigidity, tremor, and bradykinesia.

- Non-motor Symptoms: Symptoms that have nothing to do with physical movement are non-motor symptoms; fatigue, hypotension, restless legs, dementia, depression, pain, eye problems, foot care, mouth and dental issues, and speech and communication problems are some common symptoms of this.

These symptoms may vary from one individual to another; it is a progressive disease, so the condition becomes more worsened with time. The Hoehn and Yahr Scale (HY) groups Parkinson's disease symptoms into five phases: the mildest stage; after this stage, changes in movement and posture are also noticeable in the following steps, and symptoms begin to obstruct your regular activities in the following stage assistance is required to carry out daily tasks and in the last stage the patients are entirely confined to their beds [17]. Table 1 summarizes the diagnosis of Parkinson's disease based on physiological markers.

2. Proposed Methodology

The suggested technique gathers audio information regarding vocal modulation in Parkinson's patients from Positive Point wise Mutual Information (PPMI) and University of California, Irvine (UCI) [23]. Dataset offers details on vowel phonation's jitter, shimmer, and MDVP. Pre-processing, analysis, and visualisation of the data are used to fully comprehend the properties. 70% of the data is used to train six models: SVM, AdaBoost, Random Forest Regressor, Decision Tree classifier, Logistic Regression, and K Nearest Neighbors using the ensemble learning technique called the stacking classifier. Models are created to classify audio data as PD or healthy based on differences in frequency. Models are tested on 30% of the data, and their performance is assessed using the area under the ROC CURVE, the metrics for sensitivity, precision, and accuracy, and a misclassification rate.

Figure 1 shows the general procedure that has been used. It shows the steps involved in ingesting data from the PPMI database, dividing the data into sets for training and testing, training seven distinct models on the collected information, and evaluating the conclusions from the test data.

Table 1, Literature survey on Parkinson's disorder:

Author	Dataset	Subjects	Approach	Accuracy (%)
Aditi Govindu et.al[1]	Audio data	195	Principal Component Analysis (PCA)	LR 83.67, RF 91.83, SVM 85.71, KNN 85.71
Shahid A.H. et. al. [2]	Telemonitoring dataset	42 (28 men,14 women)	Principle component analysis based DNN model.	*
Kaur S. et. al. [3]	(i) Speech Dataset (ii) Diabetes Dataset (iii) Breast Cancer Dataset, (iv)Telemonitoring dataset	(i) 196 (by 31 individual, 23 impaired by PD) (ii) 768 females (iii) 699 females (iv)188(107 men,81women)	Optimized deep learning model for classification, Grid search optimization	91.69
Valdovinos et. al [4]	*	*	Uses decentralized approach	*
YANHAO XIONG et. al. [5]	Speech Dataset	188(107 men,81 women)	Adaptive WSO Algorithm, Adaptive GWO, Sparse Autoencoder Neural Network.	SVM 85, LR 87, NB 82, GBM 88, RF 81, LDA 95
Zahid Laiba et. al. [6]	Speech Dataset	50(25 men,25 women)	Transfer Learning, Random Forest, Multilayer Perceptron	Transfer Learning, Random Forest, Multilayer Perceptron.

M. Jyotiyana et. al. [7]	Speech Dataset	42	Deep Neural Network based classification model	94.87
C. Quan et. Al. [8]	Speech Signals	45(25 men,20 women)	End-to-end DL employing CNN model and bidirectional LSTM with dynamic speech characteristics	*
O.Asmae et. Al. [9]	Vocal Phonation's	31	ANN classification, KNN classification	96.7
A. H. Butt et. al. [12]	*	114(79 men,35 women)	Bi-direction Long ShortTerm Neural Network (BLSTM)	82.4
Filippo Cavalloa et. al. [13]	*	90(71 men,19 women)	SVM, RF, NB	RF 95, SVM 97
Saha Roshni et. al. [14]	MRI Dataset	54	Convolution al Neural Network Model	97.63(without batch normalization), 97. 91(with batch normalization)
S. Raval et. al. [15]	Tappy, handwriting dataset	Tappy, handwriting dataset	Hard Voting, RF (Random Forest Classifier), and Adaptive Boosting (AB) (HV)	99.79
Ball, Nicole et. al. [22]	*	*	By studying how environmental variables interact with and impact the brain, we can discover PD's underlying cause(s).	*

This study intends to determine which PD categorization criteria are most important, as well as the effects of imbalanced medical data. These criteria have guided the implementation of two strategies: training on the entire dataset, which acts as a baseline examination for classifying PD, and training them on the 235 records acquired after dataset balancing. The following is a description of each approach's underlying algorithms: An algorithm implementing method 1: With 19 attributes of data, the models are going through training.

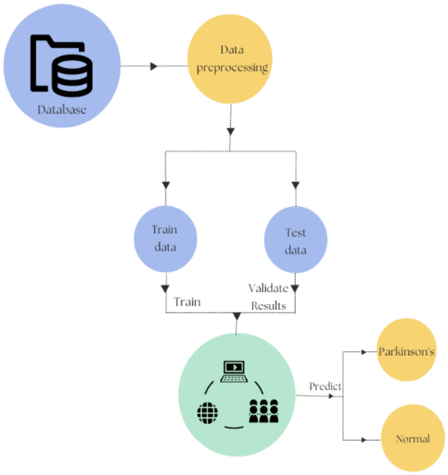

Figure 1. Proposed Architecture

- Assemble the PPMI and UCI databases' Multidimensional Voice Program (MDVP) audio data.

- Do data assessment to find data skew, imbalance, and variable distribution.

- Use Standard Scaler to adjust the data to a standard range.

- Train logistic regression, SVM, KNN, Adaptive boost, decision tree, classifier, random forest, and ensemble models by dividing the dataset into training and testing sets, with the training data making up 70% of the total.

- Using area under the ROC CURVE, the metrics for sensitivity, precision, and accuracy and misclassification rate, compare the classification findings.

An algorithm implementing method 2: Correcting imbalances in a dataset

- Assemble the PPMI and UCI databases' MDVP audio data.

- Do data assessment to find data skew, imbalance, and variable distribution.

- As seen in Figure 2(a), the unbalanced dataset, with 48 records of average persons and 148 records of Persons with parkinson (PWP). As seen in Figure 2(b), the imbalance is corrected by oversampling the minority class to a total of 148 records, and under sampling the majority class to a total of 48 records.

- Use Standard Scaler to scale the data to a common range.

- Train logistic regression, SVM, KNN, Adaptive boost, decision tree, classifier, random forest, and ensemble models by dividing the dataset into training and testing sets, with the training data making up 70% of the total.

- Using area under the ROC CURVE, the metrics for sensitivity, precision, and accuracy and misclassification rate, compare the classification findings.

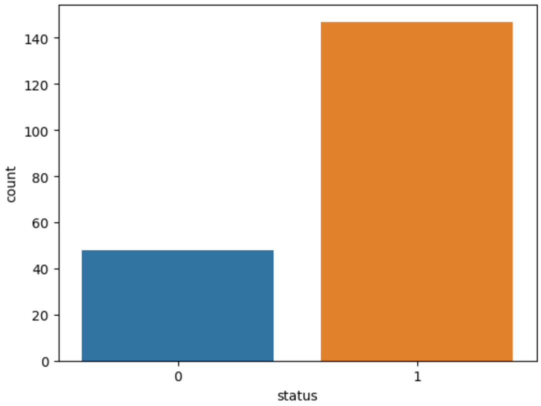

Figure 1. (a) Unbalanced data with 48 normal records

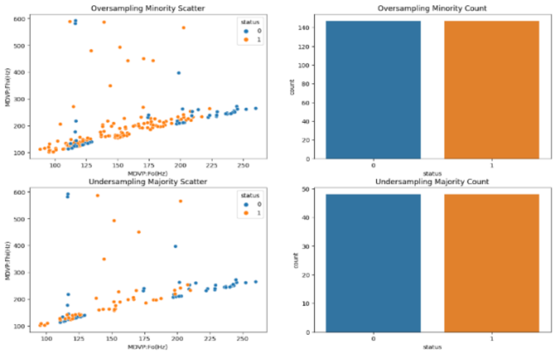

Figure 2. (b) After sampling balanced findings

3. Dataset

Authors used openly accessible data for this. 31 individuals' biomedical voice measurements [19] have been accumulated, where 23 patients have PD. Table 2 provides an explanation of the primary features' properties. Patients vary in age from 46 up to 85, whereas 23-year-olds provide normal readings. Each individual had an average of 6 phonation's, each lasting between 1 and 36 seconds, recorded 195 times.

Table 2. Component information for the PD patient's diagnosis [30, 34].

Attribute	Description
Name	Data is stored in ASCII CSV format along with the patient's name and recording number.
MDVP: Fo (Hz)	Periodic fundamental frequency.
MDVP: Fhi (Hz)	Fundamental frequency upper bound or highest voice modulation threshold.
MDVP: Flo (Hz)	Lower limit or minimal vocal fundamental frequency.
MDVP: Jitter, Abs, RAP, PPQ, DDP	These are several multidimensional voice programme (MDVP) measurements from Kay Pentax. Traditional measurements compare the frequency of vocal fold vibrations during pitch with the vibration time at the beginning of the following cycle, known as the pitch mark.
Jitter and Shimmer	Following averaging, measurements of the absolute difference between each cycle's frequencies.
NHR and HNR	Signal to noise and tonal ratio measures that indicate the robustness of the environment to noise.
Status	0 denotes a healthy person, and 1 PWP.
D2	Using fractal objects, correlation dimension is utilized to detect dysphonia in speech. It is a dynamic, nonlinear property.
RPDE	Period of Recurrence Density Entropy measures the degree of periodicity in a signal.
DFA	The degree of random self-similarity of noise in speech signals, detrended fluctuation analysis, or DFA, is used to measure it.
Spread 1, Spread 2	Analysis of the degree or variety of speech variations in relation to MDVP: Fo (Hz).

4. Results

The Stacking classifier achieves 95% accuracy and 0.94 precision for classifying Parkinson's disease using vowel phonation data. SVM is utilised by the stacking the classifier in question as its meta classifier. The MDVP dataset's 22 attributes were given equal weight; hence, the SVM model's results are flawless. The SVM model's findings are also highlighted in this study, which has a precision of 0.94 and an accuracy of 93%. Both the ensemble technique and SVM are powerful models with good outlier performance. The stacking model also works well for balanced datasets since it favours categorization into two categories without making data assumptions and achieves 97% accuracy,.96 precision, and.96 recall. Hence, in order to categorise the disease's progression, we advise using the ensemble technique. While a diagnosis that only uses audio data is insufficient for the categorization of Parkinson's disease, we recommend using audio and rapid eye movement sleep data in the future to enhance the results. Authors anticipate that these results will support the utilization of cell phone audio recording for telemedicine-based PD classification.

5. Conclusion

The final diagnosis of diseases based on numerous patient testing is a significant issue in the medical field. Early disease detection makes proper management of the condition possible. For various stages of the disease, there are various types of treatment.

Ultimately, the use of machine learning in the diagnosis of Parkinson's disease has demonstrated considerable promise in terms of reaching high accuracy rates. This study has the potential to greatly improve early detection and personalised treatment regimens for Parkinson's sufferers. The accuracy and F1 scoring system discussed above lead us to the conclusion that in both approaches the Stacking Model outperforms other models. In approach 1, the mean accuracy of the stacking model is 93%, while the standard deviation is 5%. Moreover, the stacking model has a mean F1-Score of 94% with a 5% standard deviation. With a 97% accuracy rate, here machine learning algorithms can clearly identify between healthy persons and those suffering from Parkinson's disease.

References

[1]. Aditi Govindu, Sushila Palwe, Early detection of Parkinson's disease using machine learning, Procedia Computer Science, Volume 218,2023, Pages 249-261, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2023.01.007.

[2]. Shahid, A.H., Singh, M.P. A deep learning approach for prediction of Parkinson's disease progression. Biomed. Eng. Lett. 10, 227–239 (2020). https://doi.org/10.1007/s13534-020-00156-7. PMID: 32477610 PMCID: PMC7235154

[3]. Kaur S., Aggarwal, H. & Rani, R. Hyperparameter optimization of deep learning model for prediction of Parkinson's disease. Machine Vision and Application 31,32 (2020) - Springer 10.1007/s00138-020-01078-1.

[4]. Valdovinos, B.Y., Modica, J.S. & Schneider, R.B. Moving Forward from the COVID-19 Pandemic: Needed Changes in Movement Disorder Care and Research. Curr Neurol Neurosci Rep 2022 – Springer. https://doi.org/10.1007/s11910-022-01178-7

[5]. Xiong, Y. and Lu, Y., 2020. Deep feature extraction from the vocal vectors using sparse autoencoders for Parkinson's classification. IEEE Access, 8, pp.27821-27830. DOI: 10.1109/ACCESS.2020.2968177. 2020 - ieeexplore.ieee.org

[6]. L. Zahid Et Al., "A Spectrogram-based Deep Feature Assisted Computer-Aided Diagnostic System for Parkinson's Disease," in IEEE Access, Vol. 8, pp. 35482-35495, 2020 - ieeexplore.ieee.org

[7]. Jyotiyana, M, Kesswani, N., & Kumar, M. (2021). A deep learning approach for classification and diagnosis of Parkinson's disease 2022 – Springer

[8]. C. Quan, K. Ren, and Z. Luo, "a Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech." IEEE Access, vol. 9, pp. 10239-10252, 2021 - ieeexplore.ieee.org. DOI: 10.1109/ACCESS.2021.3051432.

[9]. O. Asmae, R. Abdelhadi, C. Bounchaib, S. Sara and K. Taheddine, "Parkinson's Disease Identification Using KNN and ANN Algorithms based on Voice Disorder,"2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), DOI: 10.1109/IRASET48871.2020.9092228. 2020 - ieeexplore.ieee.org

[10]. Esmaeilzadeh, Soheil, Yao Yang, and Ehsan Adeli. "End-to-end Parkinson's disease diagnosis using brain MRI images by 3d-cnn." arXiv preprint arXiv:1806.05233 (2018) - arxiv.org

[11]. DeMaagd G, PharmD BCPS, Ashok Philip P (2015) Parkinson's disease and its management 2015 - ncbi.nlm.nih.gov https://doi.org/10.1136/bmj.308.6923.281

[12]. A. H. Butt, F. Cavallo, C. Maremmani and E. Rovini, "Biomechanical parameters assessment for the classification of

Parkinson Disease using Bidirectional Long Short-Term Memory*," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020 - ieeexplore.ieee.org. DOI: 10.1109/EMBC44109.2020.9176051 .

[13]. Cavallo, Filippo; Moschetti, Alessandra; Esposito, Dario; Maremmani, Carlo; Rovini, Erika (2019). Upper limb motor pre-clinical assessment in Parkinson's disease using machine learning. doi:10.1016/j.parkreldis.2019.02.028 . 2019 – Elsevier

[14]. Saha, Roshni. "Classification of Parkinson’s disease using MRI data and deep learning convolution neural networks." Creative Components卌241卌 R Saha - (2019). IOWA State University. https://core.ac.uk/download/pdf/227994087.pdf

[15]. S. Raval, R. Balar and V. Patel, "A Comparative Study of Early Detection of Parkinson's Disease using Machine Learning Techniques," 2020 4th International Conference on Trends in Electronics and 9 Informatics (ICOEI) (48184), 2020, pp. 509-516, doi: 10.1109/ICOEI48184.2020.9142956 2020 - ieeexplore.ieee.org

[16]. Drissi, Taoufiq Belhoussine, et al. "Diagnosis of Parkinson’s disease based on wavelet transform and mel frequency cepstral coefficients." Int. J. Adv. Comput. Sci. Appl 10 (2019). 125-132, 2019 - researchgate.net

[17]. R. Torres, M. Huerta, R. Gonzalez, R Clotet, J. Bermeo and G. Vayas, "Sensors for Parkinson's disease evaluation," 2017 International Caribbean Conference on Devices, Circuits and Systems (ICCDCS), 2017, pp. 121-124, 2017 - ieeexplore.ieee.org

[18]. Sakar, C. Okan; Serbes Gorkem; Gunduz, Aysegul; Tunc, Hunkar C.; Nizam Hatice; Sakar Betul Erdogdu; Tutuncu Melih Aydin, Tarkan; Isenkul, M.Erdem; Apaydin, Hulya (2018). A comparative analysis of speech signal processing algorithms for Parkinson’s disease classification and the use of the tunable Q-factor wavelet transform. Applied Soft Computing, (), S1568494618305799 2019 – Elsevier

[19]. M.S.Islam, I. Parvez, Hai Deng, And P. Goswami,” performance Comparison Of Heterogeneous Classifiers for Detection of Parkinson’s Disease Using voice disorder (dysphonia)," 2014 International Conference on Informatics, Electronics & Vision (ICIEV), 2014, pp, DOI: 10.1109/ICIEV.2014.6850849. 2014 - ieeexplore.ieee.org

[20]. L. Ali, C. Zhu, N. A. Golilarz, A. Javeed, M. Zhou and Y. Liu, "Reliable Parkinson’s Disease Detection by Analyzing Handwritten Drawings: Construction of an Unbiased Cascaded Learning System Based on Feature Selection and Adaptive Boosting Model," in IEEE Access, vol. 7, pp. 116480-116489, 2019,

[21]. Saravanan, S., Ramkumar, K., Adalarasu, K. et al. A Systematic Review of Artificial Intelligence (AI) Based Approaches for the Diagnosis of Parkinson’s Disease. Methods in Engineering, https://doi.org/10.1007/s11831-022-09710-1. 2022 - Springer

[22]. Ball, Nicole; Teo, Wei-Peng; Chandra, Shaneel; Chapman, James. Parkinson's Disease and the Environment. Frontiers in neurology, https://doi.org/10.3389/fneur.2019.00218. 2019 - frontiersin.org

[23]. Rung-Ching, Chen. “Random forest and support vector machine on

features selection for regression analysis.” Int. J. Innov. Comput. Inf. “Control 15 No,” 2019.

[24]. Goyal, J., Khandnor, P., & Aseri, T. C. Classification, Prediction, and Monitoring of Parkinson’s disease using Computer Assisted Technologies: A Comparative Analysis. Engineering Applications of Artificial Intelligence, 96, 103955. doi: 10.1016/j.engappai. 2020- Elsevier

[25]. Richens JG, Lee CM, Johri S Improving the accuracy of medical diagnosis with causal machine learning. https://doi.org/10.1038/s41467-020-17419-7.Nat Commun11: Nature communications, 2020

[26]. Ray Dorsey E, Elbaz A Global, regional, and national burden of Parkinson’s disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. https://doi.org/10.1016/S1474-4422 .2018 - Elsevier

[27]. Jakel, Rebekah J., and Mark Stacy. "Parkinson’s disease psychosis." Research and Reviews in Parkinsonism “.2014- researchgate.net

[28]. A., M. S., Liu H., E. Lyons K., Pahwa R., Liu W., F. Nobre F., Nadal. "Comparison among probabilistic neural network J., and vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait ort. “Journal of Biomechanics 43 No,” 2010.

[29]. Pramod, Kumar Mishra. “Performance Analysis of Machine Learning Based Optimized Feature Selection Approaches for Breast Cancer Diagnosis,” 2021.

[30]. Mehedi, Hasan, Krishno Sarkar Ajay, and Khan. "Classification of Parkinson’s Disease using Speech Signal with Machine Learning and Deep Learning Approaches Fayez. “European Journal of Electrical Engineering and Computer Science 7 No,” 2023.

[31]. Wang Q, Fu Y, Shao B, Chang L, Ren K, Chen Z, Ling Y. Early detection of Parkinson's disease from multiple signal speech: Based on Mandarin language dataset. Front Aging Neurosci. 2022 Nov 10;14:1036588. doi: 10.3389/fnagi.2022.1036588. PMID: 36438003; PMCID: PMC9691375.

[32]. Zhao, Huan & Wang, Ruixue & Lei, Yaguo & Liao, Wei-Hsin & Cao, Hongmei & Cao, Junyi. (2021). Severity Level Diagnosis of Parkinson’s Disease by Ensemble K-Nearest Neighbor Under Imbalanced Data. Expert Systems with Applications. 189. 116113. 10.1016/j.eswa.2021.116113.

[33]. Yifeng Yang, Long Wei, Ying Hu, Yan Wu, Liangyun Hu, Shengdong Nie,Classification of Parkinson's disease based on multi-modal features and stacking ensemble learning, Journal of Neuroscience Methods,Volume 350,2021, 109019, ISSN 0165-0270, https://doi.org/10.1016/j.jneumeth.2020.109019.

[34]. Little, Max & Mcsharry, Patrick & Roberts, Stephen & Costello, Declan & Moroz, Irene. (2007). Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. Biomedical engineering online. 6. 23. 10.1186/1475-925X-6-23.

[35]. *Advantages and disadvantages of logistic regression*. (n.d.). https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

[36]. *bagging algorithm in python*. (n.d.). https://www.section.io/engineering-

education/implementing-bagging-algorithms-in-python/

[37]. *How to Usemodel Stacking to Improve Machine Learning Predictions*, (n.d.). medium.com. https://medium.com/geekculture/how-to-use-model-stacking-to-improve-machine-learning-predictions-d113278612d4

[38]. Salih, Yusra Mohammed M., and Snwr Jamal Mohammed. "Parkinson’s Disease Detection by Processing Different ANN Architecture Using Vocal Dataset." Doi: 10.23918/eajse.v9i1p161