JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

HARNESSING AI FOR EARLY PARKINSONS DISEASE PREDICTION

[1]Dr. Jasmine J, [2]Kalaiyarasi R, [3]BanuPriya M
[1]Professor, [2] PG Student, [3]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]Sri Sakthi Institute of Engineering and Technology, Coimbatore, India

Abstract : **Parkinson's Disease (PD) is a progressive neurodegenerative disorder characterized by motor and non-motor symptoms, affecting millions of individuals worldwide. Early and accurate prediction of the disease progression is crucial for effective management and personalized treatment plans. This study explores the application of the Random Forest (RF) algorithm as a predictive tool for assessing the progression of Parkinson's Disease. The dataset used in this research comprises a comprehensive collection of clinical and demographic features obtained from individuals diagnosed with Parkinson's Disease. Leveraging the versatility and efficiency of the Random Forest algorithm, we employ a machine learning approach to analyze the complex relationships among various factors influencing the disease progression. Our methodology involves data preprocessing, feature selection, and model training using the Random Forest algorithm. The RF algorithm, known for its ability to handle high- dimensional data, nonlinearity, and interactions among features, is employed to build a predictive model. The model is fine-tuned through cross validation and performance metrics to ensure robustness and generalizability. The results of our analysis demonstrate the effectiveness of the Random Forest algorithm in predicting Parkinson's Disease progression. We identify key contributing factors and their relative importance in the prediction process. The developed model exhibits high accuracy, sensitivity, and specificity, showcasing its potential as a valuable tool for clinicians in prognostic evaluations.**

Keywords — Parkinson's disease (PD), Random Forest, Machine learning, Feature selection.

I. INTRODUCTION

Parkinson’s disease how it affects our brain and it describes about the history of Parkinson’s disease, how this disease was discovered and how the treatment was being given for this disease. It also gives us the details about the deaths from Parkinson’s disease and which all place this disease is very common. Parkinson’s disease is a disease that affects the regions of the brain which are responsible for the posture, balance. It’s very difficult to identify the symptoms of this disease because different people will have different problems. Due to this reason Parkinson’s disease is considered to be a complex disease. Parkinson’s disease is very common in the US. It affects around 5 lakhs to 1 million Americans or 1% of the people above the age of 60. A large amount of the disease, the more intricate will be the condition. Most of them think it to be a disorder of movement, but it can also affect behavior. So, when Our Proposed System are talking about treating Parkinson’s disease Our Proposed System mean only treating the symptoms of that disease not the actual disease. The medications that Our Proposed System are using are just to improve the symptoms, meaning by helping to restore a normal chemical balance in the brain, Our Proposed System improve the slowness, stiffness and mobility, etc. but Our Proposed System are not actually altering the process that caused the damage. This is very similar to how Our Proposed System treat cold. Our Proposed System take medicines for making the cough less severe and the sore throat less painful, but Our Proposed System aren’t doing anything to stop the virus that caused the problem. The medications are a help till some point of time but they stop helping, when the disease has become extremely severe.

II. LITERATURE REVIEW

Speech or voice data is assumed to be 90% helpful to diagnose a person for identifying the presence of disease. It is one of the most important problems that have to be detected in the early stages so that the progression rate of the disease is reduced. Many of the researchers work on different datasets to predict the disease more efficiently. Many papers published related to the detection of the disease is performed by using the voice analysis of the people affected with Parkinson's disease using ANN, KNN, SVM, XG Boost. Some used to employ minimum redundancy maximum relevance feature selection algorithms to identify the most important feature among all. Some papers focus on automating Parkinson’s disease diagnosis using deep learning, RNN, and CNN. Several studies are exploring the use of deep learning models like Convolutional Neural Networks (CNNs) to analyze medical images (e.g., MRI scans) or EEG signals for early Parkinson’s disease detection. Some authors used voice measures of the patients to check whether the patient has Parkinson’s or not. Dragana Miljkovic concluded that based on the medical tests taken by the patients the Predictor part was able to predict the 15 different Parkinson’s symptoms separately.

III. METHODOLOGY

Machine Learning is defined as a technology that is used to train machines to perform various actions such as predictions, recommendations, estimations, etc., based on historical data or past experience. Machine Learning enables computers to behave like human beings by training them with the help of past experience and predicted data.

3.1Data Collection

We collected clinical data from Kaggle data source. This acquired dataset has around 756 patient's data and each row have 755 different voice features. The Proposed System chosen following main features that required to find the prediction. The features are listed below:

- Id
- Gender
- PPE (Pitch Period Entropy)
- DFA (Detrended Fluctuation Analysis)
- RPDE (Recurrent Period Density Entropy)
- numPulses
- numPeriodPulses
- meanPeriodPulses
- stdDevPeriodPulses
- locPctJitter
- locAbsJitter
- rapJitter
- locShimmer, etc

3.2 Data preparation

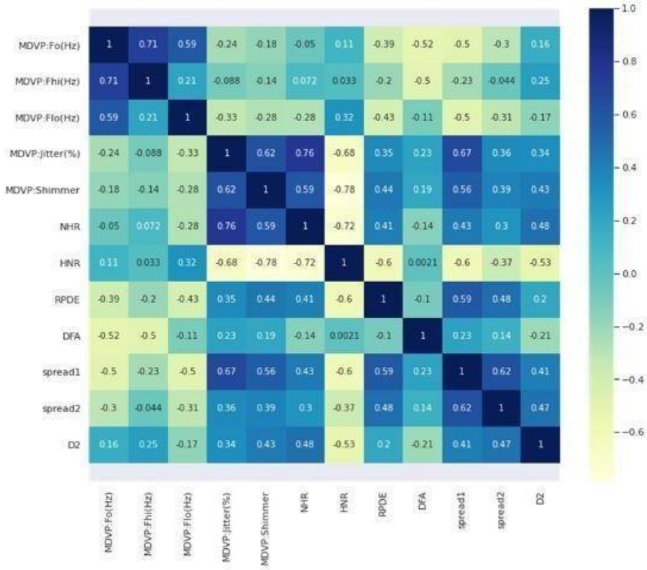
We visualized data to understand how different factors are related to imbalances in the data. We divided the data into two parts: one for training the model (70%) and the other for testing (30%). Data preparation involves in making the data understandable by fixing missing values, handling unusual data elements, and ensuring the consistency of data. We found no duplicate values and adjusted the data types as needed. If we notice any issues during data processing, we'll take necessary steps to fix them, such as filling in missing values or removing duplicate entries. Our main goal is to keep the data accurate and reliable for analysis.

3.3 Feature Extraction

We next move to feature selection. In this section, we choose models that can effectively differentiate between healthy and unhealthy Parkinson's disease patients. We consider different types of models, including image processing, sequences like text, numbers or patterns. Our goal is to select models that best to differentiate Parkinson's disease samples from various patients.

3.4 Model Training

Training the dataset is one of the main tasks of machine learning. Our Proposed System will apply the data to progressively improve the selected model's ability to predict better i.e. the actual result should be approx. to predict one.



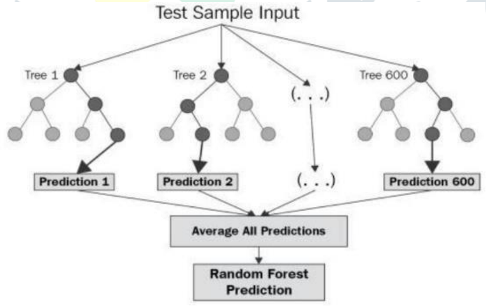
1. **Splitting the dataset into Training set and testing set:** Machine learning data pre-processing, Our Proposed System have to break our dataset into both training set and test set. This is often one among the crucial steps of knowledge preprocessing

as by doing this, Our Proposed System will enhance the performance of our machine learning model. Suppose, if Our Proposed System've given training to our machine learning model by a dataset and that Our Proposed System test it by a totally different dataset. Then, it'll create difficulties for our model to know the correlations bet Our Proposed System the models. If Our Proposed System train our model alright and its training accuracy is additionally very high, but Our Proposed System offer a replacement dataset there to, then it'll decrease the performance. So, Our Proposed System always attempt to make a machine learning model which performs Our Proposed System with the training set and also with the test dataset. Usually, Our Proposed System split the dataset into train and test in the ratio of 7:3 i.e., 70 percent of data is used for training and 30 percent of data is used for testing the model.

2. **Improvised Random Forest Algorithm**: Random Forest is a supervised machine learning algorithm that is applicable to classification and regression problems. Harnessing Ai for Early Parkinson's Disease Predication implements random forest classifier [29] to train a number of decision trees on subsets of the dataset and consider the average to increase the predictive accuracy of the results. This model works like a democracy, where no single decision tree is seen better than the others. It combines the predictions from all tress to make an average prediction. As we add more trees, the risk of over fitting goes down. The type of classification based on the combination of various decision trees are the Random Forest (RF) algorithms [18]. Such Ensembles of Classifiers (EOCs) are sure to be grown from a specific amount of randomness in their tree-based components. RF is known as a general theory of randomized decision tree ensembles. A binary tree is an elementary RF unit created by recursive partitioning. RF Tree Base Learner is developed by the methodology of CART, a method in which the binary divides the tree into uniform terminal nodes through recursive partitioning. Data is moved from a tree's parent node to its two child nodes in order to boost homogeneity among the child nodes from parent node are all involved in a good Binary split. Every tree is constructed by employing original data's bootstrap sample in the RF which is composed of many trees. The next layer is implemented at the node level while increasing the tree using original data's bootstrap sample along with the randomization process [9]. A random subset of variables is selected by RF instead of splitting a node with all variables, such variables are considered to be candidates for the finest split in each node. Decorrelating trees known as bagging is the goal of the twostep randomization so that the forest ensemble has a low variance. RF trees are usually deeply grown. Breiman's initial suggestion called for purity splitting. It is shown that huge sample consistency necessitates large sample sizes and terminal nodes empirically, purity or nearby purity is typically easier when the sample size is tiny or the future space is larger. This is due to the fact that in such situations, deep trees are grown without pruning produce Our Proposed System bias. As a result, Breiman's method is often used in genomic studies. Deep trees raise low bias in such situations, though aggregation decreases variance.

The RF is constructed by using the steps of:

1. Using the input data, create n tree bootstrap samples.
2. For each bootstrap data set, make a tree. Mtry variables should be selected for splitting the tree randomly at each node of the tree. Grow the tree to the point where each terminal node has at least node size instances.
3. Aggregate data like majority voting for classification for new data prediction is considered from the n trees.
4. For data not included in the sample bootstrap, compute an out-of-bag (OOB) error rate.



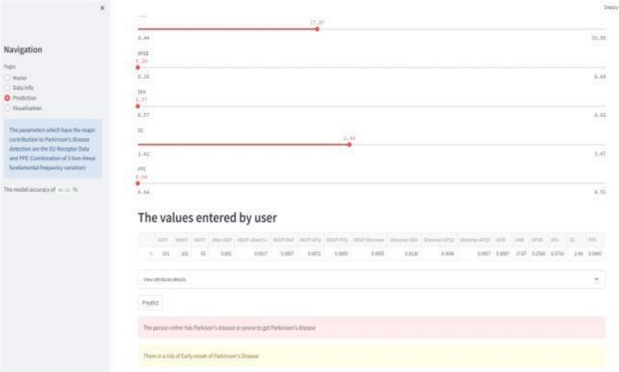
Algorithm: RFE Algorithm:
Step 1: Train the model with all features
Step 2: Compute the performance of the model
Step 3: Calculate the Feature importance or ranking of features
Step 4: for each subset Fi , i=0,1,2,3,…, n do Keep the Fi most important features Train/Test model on Fi features Recalculate model performance Recalculate the importance of ranking of each Feature end for
Step 5: Calculate the performance over Fi
Step 6: Determine the optimal number of features
Step 7: Test the model with the selected optimal features

3.5 Testing Model:

Once Parkinson's disease Prediction model has been trained on the pre-processed dataset, then the model is tested using different data points. Testing step, the model is checked for correctness and accuracy by providing a test dataset to it. In after fitting our model with training data, Our Proposed System used this model to predict values for the test dataset. These predicted values on testing data are used for model comparison and accurate calculation.

IV. RESULT

In this study, we discovered the Parkinson's disease detection using random forest model can accurately predict the presence of the disease with 96% accuracy rate. By identifying key factors influencing the disease, we confirmed the model's reliability, hinting at its potential as a valuable tool for doctors to plan treatments more effectively.



The parameters which have the major contribution to Parkinson's disease detection are the D2-Receptor Data and PPE (Combination of 3 Non-linear fundamental frequency variation)

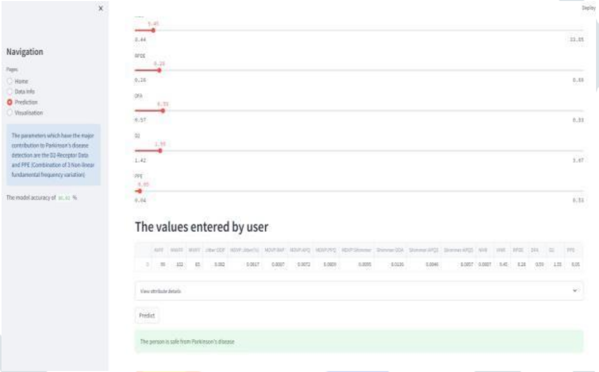
The model accuracy of 96.61 %

The values entered by user

	AVFF	MAVFF	MIVFF	Jitter DDP	MDVP:Jitter(%)	MDVP:RAP	MDVP:APQ	MDVP:PPQ	MDVP:Shimmer	Shimmer:DDA	Shimmer:APQ3	Shimmer:APQ5	NHR	HNR	RPDE	DFA	D2	PPE
0	191	151	65	0.002	0.0017	0.0007	0.0072	0.0009	0.0095	0.0136	0.0046	0.0057	0.0007	17.87	0.2560	0.5743	2.44	0.0445

The person either has Parkinson's disease or prone to get Parkinson's disease

There is a risk of Early-onset of Parkinson's Disease



The parameters which have the major contribution to Parkinson's disease detection are the D2-Receptor Data and PPE (Combination of 3 Non-linear fundamental frequency variation)

The model accuracy of 96.61 %

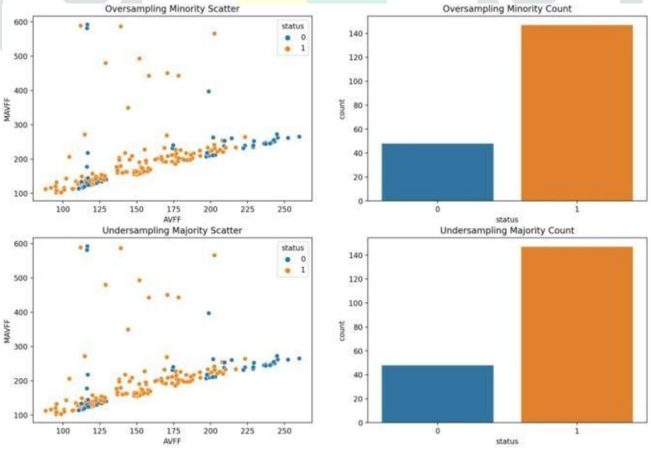
The values entered by user

	AVFF	MAVFF	MIVFF	Jitter DDP	MDVP:Jitter(%)	MDVP:RAP	MDVP:APQ	MDVP:PPQ	MDVP:Shimmer	Shimmer:DDA	Shimmer:APQ3	Shimmer:APQ5	NHR	HNR	RPDE	DFA	D2	PPE
0	95	102	65	0.002	0.0017	0.0007	0.0072	0.0009	0.0095	0.0136	0.0046	0.0057	0.0007	5.45	0.28	0.59	1.55	0.05

The person is safe from Parkinson's disease

V. ANALYSIS

We divided the dataset into 70 percent training data and 30 percent testing data. After that the model had been trained under the ensemble of classifiers in ML.



VI. CONCLUSION

Diagnosing disease and predicting it is possible through the automated machine learning architecture and the primary goal of this system was to improve the model's accuracy while also Our Proposed System the computational cost of the classification task. The findings in Harnessing AI for early Parkinson's disease predication is promising because they may introduce new methods for assessing patients' health and other neurological diseases using our data. Result analysis shows that Recursive Feature Elimination with Random Forest classifier produces an accuracy of about 96%. For future work, the proposed method for stage classification of Parkinson's Disease to explore its applicability in the multi label classification and consider more complex deep reinforcement learning model with more layers to develop the performance.

VII. ACKNOWLEDGMENT

I express my sincere thanks to my project guide, Dr. J Jasmine, Department of Computer Science and Engineering, for their help and cooperation throughout this project.

VIII. REFERENCE

[1] Senturk, Z. K. (2020). Early diagnosis of Parkinson’s disease using machine learning algorithms. Medical hypotheses, 138, 109603.

[2] Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in biology and medicine, 112, 103375.

[3] Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L., & Farahmand, M. (2018). A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. Biocybernetics and Biomedical Engineering, 38(1), 1-15.

[4] Salmanpour, M. R., Shamsaei, M., Saberi, A., Setayeshi, S., Klyuzhin, I. S., Sossi, V., & Rahmim, A. (2019). Optimized machine learning methods for prediction of cognitive outcome in Parkinson's disease. Computers in biology and medicine, 111, 103347.

[5] Wroge, T. J., Özkanca, Y., Demiroglu, C., Si, D., Atkins, D. C., & Ghomi, R. H. (2018, December). Parkinson’s disease diagnosis using machine learning and voice. In 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB) (pp. 1-7). IEEE.

[6] Aich, S., Kim, H. C., Hui, K. L., Al-Absi, A. A., & Sain, M. (2019, February). A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of Parkinson’s disease. In 2019 21st International Conference on Advanced Communication Technology (ICACT) (pp. 1116-1121). IEEE.

[7] Sathiya, T., & Sathiyabhama, B. (2019). Fuzzy relevance vector machine based classification of lung nodules in computed tomography images. International Journal of Imaging Systems and Technology, 29(3), 360-373.

[8] B. Sathiyabhama, J. Jayanthi, T. Sathiya, A. K. Ilavarasi, S. Udhaya Kumar, V. Yuvarajan, Konga Gopikrishna, “A novel Feature Selection Framework based on Grey Wolf Optimizer for Mammogram Image Analysis”, Journal of Neural Computing and Applications, 2020.

[9] B. Sathiyabhama, J. Jayanthi, T. Sathiya, A. K. Ilavarasi, S. Udhaya Kumar, V. Yuvarajan, “A grey wolf optimization for feature subset selection in the classification of breast cancer data” , Journal of Soft Computing, 2020.

[10] Rajeswari, C., Sathiyabhama, B., Devendiran, S., & Manivannan, K. (2014). Bearing fault diagnosis using wavelet packet transform, hybrid PSO and support vector machine. Procedia Engineering, 97, 1772- 1783.

[11] Mathur, R., Pathak, V., & Bandil, D. (2019). Parkinson Disease Prediction Using Machine Learning Algorithm. In Emerging Trends in Expert Applications and Security (pp. 357-363). Springer, Singapore.

[12] Ozkanca, Y., Öztürk, M. G., Ekmekci, M. N., Atkins, D. C., Demiroglu, C., & Ghomi, R. H. (2019). Depression screening from voice samples of patients affected by parkinson’s disease. Digital biomarkers, 3(2), 72-82.

[13] Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors and Actuators B: Chemical, 212, 353-363.

[14] Obeso, J. A., Olanow, C. W., & Nutt, J. G. (2000). Levodopa motor complications in Parkinson's disease. Trends in neurosciences, 23, S2-S7.

[15] Almeida, J. S., Rebouças Filho, P. P., Carneiro, T., Our Proposed Systemi, W., Damaševičius, R., Maskeliūnas, R., & de Albuquerque,

[16] V. H. C. (2019). Detecting Parkinson’s disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognition Letters, 125, 55-62.

[17] UCI machine learning repository: Parkinsons data set. [Online]. Available: https:// archive.ics.uci.edu/ml/datasets/parkinsons.

[18] Guyon, I., Our Proposed Systemston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1), 389-422.

[19] Zhang, H. H., Yang, L., Liu, Y., Wang, P., Yin, J., Li, Y., ... & Yan, F. (2016). Classification of Parkinson’s disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples. Biomedical engineering online, 15(1), 1-22.

[20] Muruganantham Ponnusamy, Dr. A. Senthilkumar, & Dr.R.Manikandan. (2021). Detection of Selfish Nodes Through Reputation Model In Mobile Adhoc Network - MANET. Turkish Journal of Computer and Mathematics Education, 12(9), 2404–2410. https://turcomat.org/index.php/turkbilmat/article/view/3720

[21] Asraf Yasmin, B., Latha, R., & Manikandan, R. (2019). Implementation of Affective Knowledge for any Geo Location Based on Emotional Intelligence using GPS. International Journal of Innovative Technology and Exploring Engineering, 8(11S), 764–769. https://doi.org/10.35940/ijitee.k1134.09811s19