Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Tuned Homogenous Ensemble Regressor Model for Early Diagnosis of Parkinson Disorder Based on Voice Features Modality

C. D. Anisha[1], N. Arulanand[2]

[1]Research Scholar, Dept of CSE, PSG College of Technology, Coimbatore, India
[2]Professor, Dept of CSE, PSG college of Technology, Coimbatore, India

E-mail: ani.c.dass@gmail.com[1], naa.cse@psgtech.ac.in[2]

Abstract

Parkinson Disorder (PD) is a neurological disorder which is progressive and degenerative in nature. There are no specific tests pertaining to the diagnosis of PD. The symptoms at an early stage are mild. The early diagnosis of PD is really essential to delay the progression of the disorder. Speech disorder namely dysphonia is experienced by approximately 90% of PD patients. The incorporation of Artificial Intelligence (AI) techniques integrated with non-invasive capture of speech data from patients in diagnosis system aids to provide a robust, reliable and accurate estimation of Unified Parkinson Disease Rating Scale (UPDRS) score which ease the decision-making process effective for healthcare professionals. The proposed system incorporates a novel tuned Homogenous Ensemble Regressor wherein the hyperparameters are chosen and tuned using various experiments. Tuned Extreme Gradient (XgBoost) Regressor and Tuned Random Forest (RF) Regressor are the two homogenous regressor model. The proposed system is compared with the Tuned Linear Regression (LR) model which is the single Regressor model. The proposed system is evaluated using the large database of voice features samples of 42 PD patients. The Mean Absolute Error (MAE) and Mean Squared Error (MAE) values are minimal for the proposed system and it shows that the errors of the proposed system are lower than the single classifier errors and existing similar system.

Keywords: Parkinson Disorder (PD), Artificial Intelligence (AI), homogenous regressor models

1. Introduction

Parkinson Disorder (PD) is a neurological disorder. It is progressive in nature. The main problems identified in the diagnosis system for PD are:

DOI: https://doi.org/10.36548/jaicn.2022.3.005
Received: 04.08.2022, received in revised form: 23.08.2022, accepted: 07.09.2022, published: 17.09.2022

- There are no specialized tests for PD.
- The tests mostly suggested namely Magnetic Resonance Imaging (MRI) and Dopamine Transporter scan (DaTscan) for PD diagnosis are not cost effective.
- It takes time to diagnose at an early stage of PD

The solutions to address the problems identified in the PD diagnosis is to develop a non-invasive, cost effective and time effective diagnosis system mainly to capture the early symptoms of PD. The method which helps to implement the proposed solution is to incorporate Artificial Intelligence (AI) into diagnosis system due to its intuitiveness nature. Approximately 90% of PD patients experience voice symptoms namely dysphonia at an early stage. The main idea is to develop an early diagnosis system by mapping the dysphonia symptom to the speech modality and the integration of speech modality to the Artificial Intelligence modal.

Artificial Intelligence (AI) is the emerging technology and it has a widespread application in many fields ranging from large scale system to the small-scale system. Machine Learning (ML) is the sub system of AI and Deep Learning (DL) is the sub system of Machine Learning (ML). There are three main types in Machine Learning (ML) models namely supervised learning, unsupervised learning and reinforcement learning. Classification and Regression are the two types of Machine Learning (ML) model. The implementation types of ML model are single model implementation and ensemble model implementation.

The common existing diagnosis system of Parkinson Disorder (PD) are:

- Usage of image scanning technique which is expensive and invasive in nature.
- Integration of traditional Machine Learning (ML) algorithms to voice modality for PD prediction.
- Qualitative examination and subjective examination by clinicians.

The remaining paper is organized as follows section 2 discusses the related works along with the advantages and disadvantages of the existing system, section 3 focuses on the methodologies of the proposed system wherein the workflow, architecture of the proposed Machine Learning (ML model) are presented, section 4 presents the results analysis , the insights and inferences procured from the results and section 5 presents the conclusion with the advantage of the proposed system , future focus and challenges of the system.

2. Related Works

Asgari et al [1], presents a regression analysis of voice data obtained through an experimental paradigm wherein the subjects are asked to make vowel sound, the recorded speech is processed through Hanning Window method, the speech related features are extracted from the processed speech samples, finally the regression analysis is done on the extracted data using Epsilon SVR and nuSVR and the regression model is evaluated using Mean Absolute Error (MAE).

R. Viswanathan et al [2], presents an investigation of voice-based analysis for the estimation of the most prominent score in Parkinson Disorder (PD) diagnosis known as Unified Parkinson's Disease Rating Scale (UPDRS). AdaBoost regressor model is used for automatic prediction for UPDRS score based on the extracted features of voice samples using Least Absolute Shrinkage and Selection Operator (LASSO).

Elmehdi BENMALEK et al [3], presents a Linear Regression and Neural Network (NN) based UPDRS prediction on the voice features which has been initially explored using Spearman rank correlation technique.

Erdogdu Sakar B et al [4], presents a two-step approach-based Machine Learning (ML) technique for the evaluation of voice features which aids in early diagnosis of Parkinson Disorder (PD) with respect to the estimation of Unified Parkinson Disorder (UPDRS) score. K Nearest Neighbour (K-NN), Support Vector Machine (SVM) and ELM are the Machine Learning (ML) algorithms used.

Athanasios Tsanas et al [5], presents a classical least squares and nonparametric classification and regression trees (CART) framework for analysis of voice features samples of PD patients.

Basil K Varghese et al [6], presents a comparative analysis of various Machine Learning (ML) models for UPDRS score prediction. The models used for comparison are Linear Regression (LR), Support Vector Machine (SVM), Decision Tree Regression and Resilient Back Propagation.

Kaan Yılancıoğlu [7], presents a regression analysis using Neural Network (NN) based on voice features for estimation of UPDRS score. The exclusion of Jitter voice features from the complete set of voice features used for prediction, will have prominent impact in the performance of the prediction model.

The main gaps found from the literature survey are:

- The incorporation of models without any tuning process which makes the model to produce high errors.
- There is absence of incorporation of ensemble model with tuning, only single model incorporation has been performed.

3. Research Methodologies

Figure 1 presents the architecture of the proposed system which depicts the process followed by the proposed system, input and output parameters in each process and the algorithms used in the proposed system.

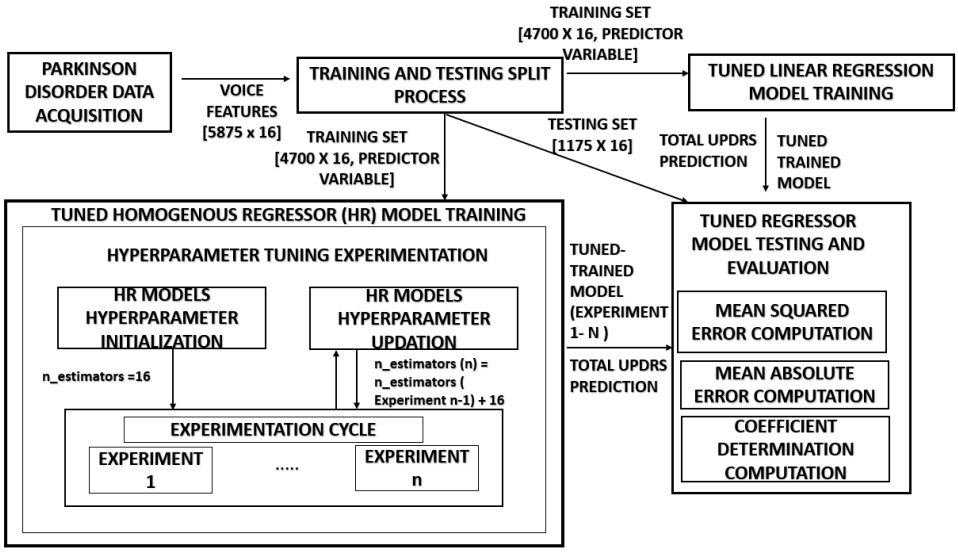

Figure 1. Architecture of the Proposed System

3.1 Data set Description

The dataset has been retrieved from UCI repository [5]. The dataset consists of voice samples features which has been extracted from voice samples procured from 42 Parkinson Disorder (PD) patients. The dataset consists of 5,875 recording samples, 200 recording samples from each patient along with 16 voice features, 3 features corresponding to the subject information, 1 feature about the time duration and 2 columns provides the scores namely motor Unified Parkinson's Disease Rating Scale (UPDRS) and total UPDRS.

The following are the 16 voice features namely:

- 5 features pertaining to the variations of "Jitter" (change in fundamental frequency)
- 6 features pertaining to the variations of "Shimmer" (change in amplitude)

- 2 features represent Noise Harmonic Ratio (NHR)/Harmonic Noise Ratio (HNR)
- 1 feature represent Recurrent Period Density Entropy (RPDE)
- 1 feature represent Detrended Fluctuation Analysis (DFA)
- 1 feature represent Pitch Period Entropy (PPE)

3.2 Training and Testing Set Split Process

The training and testing set is split in the ratio of 80: 20. It has been evident from the literature survey and empirical studies that 80: 20 ratio split provides best results as 80% of data is provided for training the model and remaining 20% of data is used to test the trained model. Table 1 presents the training and testing split of dataset used in this system.

Table 1. Training and Testing Split

Training set	Testing set
4700 Rows x 16 Columns x 1 Column (Predictor Value – Total Motor UPDRS Score)	1175 rows x 16 Columns (Predictor Value – Total Motor UPDRS Score)

3.3 Model Training

3.3.1 Tuned Linear Regression Model

Linear Regression (LR) model analysis aids in establishing relationships between independent and dependent variables [8]. n_jobs parameter value is set to 16 by mapping to the number of voices features present in the dataset.

3.3.2 Tuned Homogenous Regression (HR) Model Training

Homogenous Regression Models are ensemble models which consists of same type of models in the ensemble [9]. The two HR models considered in the proposed system are Extreme Gradient Boosting (XgBoost) Regressor Model and Random Forest Regressor Model.

a) Tuned Random Forest Regression Model

Random Forest Regression Model is a bagging technique approach wherein it creates multiple decision trees for training and prediction [10]. Finally, the predictions of various decision trees are combined by averaging.

b) Tuned XgBoost Regression Model

XgBoost Regression Model is the implementation of the Gradient Boosting Framework [11]. The training of the XgBoost Regression Model occurs in an iterative manner. It creates new trees which aids in computation of the residual errors of the previous trees which then makes the final prediction by integrating with previous trees. All trees which are built by the model are connected and it is a boosting approach which improves the Regression Tree iteration in each iteration by rectifying the errors of the previous iteration.

c) HR Model Hyperparameter Value Initialization

The two hyperparameter which is initialized for HR models are:

- n_estimators: (Common for both Models) which indicates the number of estimators (Regression Trees) to be considered [15]. This parameter value is initialized to 16 since it is mapped to the number of voices features present in the dataset. n_estimator =16 is the hyperparameter value for the first experiment.
- n_jobs: (only for XgBoost Regressor Model) which indicates the number of jobs that can be used for computation. This value is constant throughout all the experiments but instead of using the default parameter, the parameter value is set to 16 which is mapped based on the number of voices features present in the dataset.
- Random_state: (for Random Forest Regressor Model) which denotes the state of the randomness in the sample. It is set to 42 based on trial-and-error method and found that when this value is set to 42 it controls the randomness in splitting the samples and provides consistency in results for every execution of the model.

d) HR Model Hyperparameter Value Updation

After each experiment, the hyperparameter value of the regressor model is updated in the multiples of 16 which is fed to the next experiment. The formulation for updating the n_estimator value is:

n_estimator (current experiment) = n_estimator (previous experiment) + 16

Here current experiment denotes the next experiment which is to executed and the previous experiment denotes the experiment which is executed prior the current experiment. The constant number 16 is added to n_estimator for each experiment which is mapped to the number of features present in the dataset.

e) Tuned Regressor Model Testing and Evaluation

The evaluation metrics considered for the testing tuned- trained models are Mean Squared Error (MSE), Mean Absolute Error (MAE) and Co-efficient of Determination.

f) Mean Squared Error (MSE)

Mean Squared Error (MSE) provides the average of the squared difference between the actual value (A_n) and the Predicted value (P_n). [12]

The mathematical formula to compute the Mean Squared Error (MSE) is as follows:

$$\text{MSE} = {}^{1}/_{d} \sum_{n=1}^{d} (A_n - P_n)^2$$

Where,

d: number of data samples present in the test dataset

A_n: Actual Value

P_n: Predicted Value

g) Mean Absolute Error (MAE)

Mean Absolute Error provides the absolute value of the difference between actual the Predicted value (P_n) and the actual value (A_n).[13]

The mathematical formula to compute the Mean Absolute Error (MAE) is as follows:

$$\text{MAE} = \frac{1}{d} \sum_{n=1}^{d} |P_n - A_n|$$

Where,

d: number of data samples present in the test dataset

Pn: Predicted Value

An: Actual Value

h) Co-efficient of Determination

This evaluation metric provides the variation of the dependent variable with respect to the independent variable. This metric is also known as R squared.[14] If the Co-efficient of Determination is 1, then it indicates that the dependent variable prediction based on independent variable is possible without any error.

4. Results Analysis and Discussion

Table 2,3 presents the experimental results of Random Forest Regressor model and XgBoost Regressor model. The hyperparameter considered for the experiment is n_estimators which indicates the number of estimators which in this case is the number of Regression trees to be included in the model for training. The above specified parameter plays a key role in accurate prediction to be made by the model. The n_estimators is initialized to 16 by mapping to the total number of voices features present in the dataset and this parameter value is increased in the multiples of 16 in the consecutive experiments. The other hyperparameter which is set based on the voice features present in the dataset is n_jobs, this parameter value which is 16 is constant throughout the experiments. It is clearly evident from the experiments that the number of estimators when increased reduces the error of the prediction. The least error of Random Forest Regressor model is obtained in the experiment 20 when n_estimators =320, MAE =6.493 and MSE=70.809. The least error of XgBoost Regressor model is obtained in the experiment when n_estimators =336, MAE =7.094 and MSE=78.694. The final outcome of the experiments is the Tuned Homogenous Ensemble Regressor Models namely Tuned Random Forest Regressor model and Tuned XgBoost Regressor model because the hyperparameter n_estimator is tuned through the experiments 1-21 for each model and the best value of the n_estimator is found from the conducted experiments.

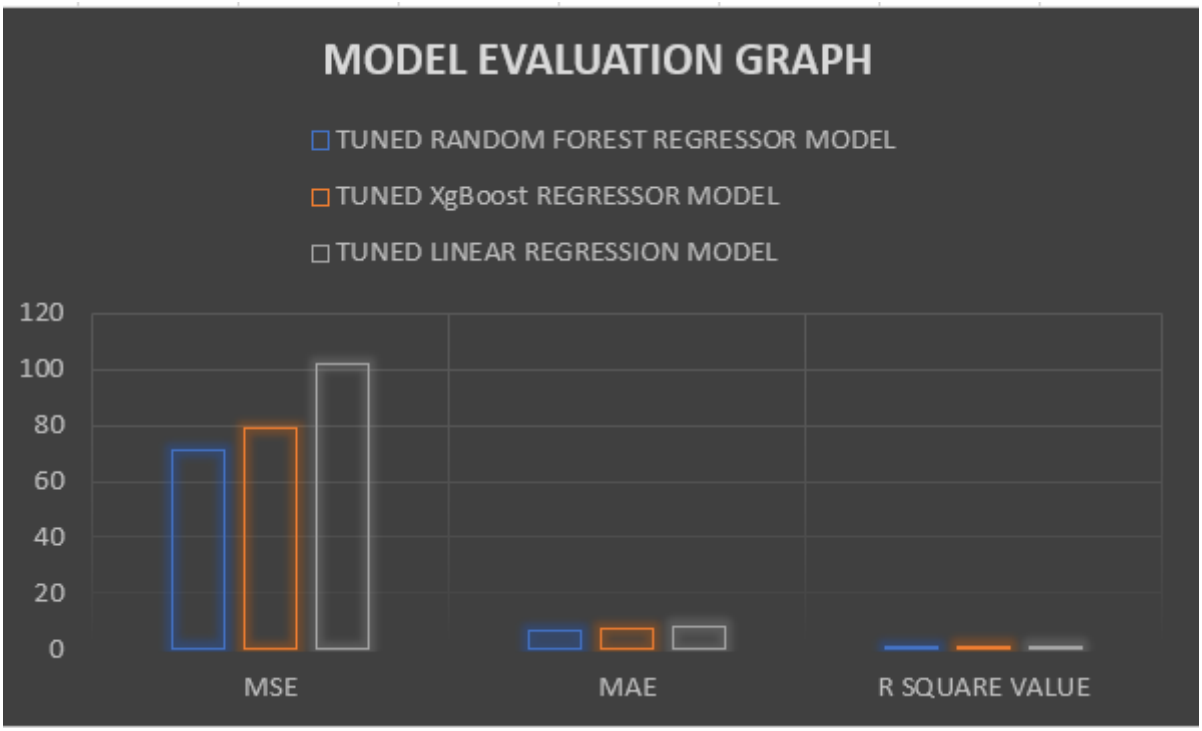

Figure 2. Model Evaluation Graph

From the Figure 2 which presents the model evaluation graph it is clearly evident that the Mean Squared Error (MSE) and Mean Absolute Error (MAE) is less for the proposed Tuned Homogenous Regressor (HR) Models compared to the Tuned Linear Regression Model.

Table 2. Random Forest Regressor Experiment Results

Experiment	n_estimators	MAE	MSE
1	16	6.727	76.908
2	32	6.605	73.174
3	48	6.573	72.915
4	64	6.557	72.490
5	80	6.557	72.324
6	96	6.542	71.750
7	112	6.533	71.581
8	128	6.528	71.471
9	144	6.536	71.597
10	160	6.522	71.482
11	176	6.521	71.245
12	192	6.518	71.262
13	208	6.514	71.134
14	224	6.511	70.993
15	240	6.497	70.862
16	256	6.499	70.877
17	272	6.498	70.827
18	288	6.500	70.887
19	304	6.495	70.850
20	320	6.493	70.809
21	336	6.499	70.901

Table 3. XgBoost Regressor Experiment Results

Experiment	n_estimators	MAE	MSE
1	16	8.363	119.825
2	32	7.658	90.032
3	48	7.603	87.456
4	64	7.491	85.183
5	80	7.421	83.854
6	96	7.362	82.968
7	112	7.324	81.294
8	128	7.296	81.887
9	144	7.264	81.451
10	160	7.232	80.732

11	176	7.207	80.469
12	192	7.197	80.325
13	208	7.179	80.182
14	224	7.173	80.072
15	240	7.162	79.968
16	256	7.167	80.0167
17	272	7.137	79.544
18	288	7.137	79.607
19	304	7.120	79.138
20	320	7.107	78.913
21	336	7.094	78.694

5. Conclusion

Parkinson Disorder (PD) is a neurological disorder which has the characteristic of progression and degeneration. The most common symptom shown in the PD patients at an early is the voice-based disorder which is known as dysphonia. The proposed system aims to develop a Machine Learning (ML) model which is reliable in providing the early diagnosis of PD based on the voice features. The proposed system incorporates the use of Homogenous Regressor Models namely Random Forest Regressor model and Extreme Gradient Boosting (XgBoost) Regressor models for the PD diagnosis system based on UPDRS score, because these models are high in performance due to ability of converting a week learner into strong learner, the weak learner considered in this system is the Regression trees whose performance is boosted through these models. The advantage of the proposed system is the introduction of experimentation method for tuning the hyperparameter of the homogenous regressor model wherein the number of estimators is tuned and the best parameter is obtained. The final outcome of this hyperparameter experimentation is the Tuned Homogenous Regressor Models which are robust and accurate in nature with less error compared to Tuned Linear Regressor Model and existing systems. The best Mean Absolute Error (MAE) and Mean Squared Error (MSE) are 6.493 and 70.809 respectively obtained by the Random Forest Regressor Model and it produced less error compared to XgBoost Regressor model and Tuned Linear Regression Model. The best Mean Absolute Error (MAE) and Mean Squared Error (MSE) obtained by the XgBoost Regressor Model are 7.107 and 78.913 respectively and it produced less error compared to Tuned Linear Regression Model. The future work focuses on forming a Hybrid Regressor Model for PD Prediction and also to leverage the possibility of increasing the modalities for diagnosis which provides more robust diagnosis system.

References

[1] Asgari M, Shafran I. Predicting severity of Parkinson's disease from speech. Annu Int Conf IEEE Eng Med Biol Soc. 2010; 2010:5201-4. doi: 10.1109/IEMBS.2010.5626104. PMID: 21095825; PMCID: PMC7889280.

[2] R. Viswanathan, S. P. Arjunan, P. Kempster, S. Raghav and D. Kumar, "Estimation of Parkinson's disease severity from voice features of vowels and consonant," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020, pp. 3666-3669, doi: 10.1109/EMBC44109.2020.9175395.

[3] Elmehdi BENMALEK et al, UPDRS tracking using linear regression and neural network for Parkinson's disease prediction, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 4, Issue 6, November - December 2015.

[4] Erdogdu Sakar B, Serbes G, Sakar CO (2017) Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. PLoS ONE 12(8): e0182428. https://doi.org/10.1371/ journal. pone.0182428.

[5] Athanasios Tsanas et al, Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests, IEEE, 30 July 2009.

[6] Basil K Varghese, Geraldine Bessie Amali D, Uma Devi K S. Prediction of Parkinson's Disease using Machine Learning Techniques on Speech dataset. Research J. Pharm. and Tech 2019; 12(2):644-648. doi: 10.5958/0974-360X.2019.00114.8

[7] Kaan Yılancıoğlu "Vocal Cord Measures Based Artificial Neural Network Approach for Prediction of Parkinson' s Disease Status", Sdü Sağlık Bilimleri Enstitüsü Dergisi / Cilt 8 Sayı 2 / 2017

[8] Montgomery, Douglas G, et all., (2012). Introduction to Linear Regression Analysis (fifth edition). Jhon Willey & Sons Inc: Canada.

[9] Sabzevari, M., Martínez-Muñoz, G. & Suárez, A. Building heterogeneous ensembles by pooling homogeneous ensembles. Int. J. Mach. Learn. & Cyber. 13, 551–558 (2022). https://doi.org/10.1007/s13042-021-01442-1

[10] Gerard Biau, "Analysis of a Random Forests Model", Journal of Machine Learning Research 13 (2012) 1063-1095.

[11] Tianqi Chen, "XGBoost: A Scalable Tree Boosting System", KDD '16, August 13-17, 2016, San Francisco, CA, USA c 2016 ACM. ISBN 978-1-4503-4232-2/16/08. . . $15.00 DOI: http://dx.doi.org/10.1145/2939672.2939785

[12] Lehmann, E. L.; Casella, George (1998). Theory of Point Estimation (2nd ed.). New York: Springer. ISBN 978-0-387-98502-2. MR 1639875

[13] Pontius Jr., Robert Gilmore; Thontteh, Olufunmilayo; Chen, Hao (2008). "Components of information for multiple resolution comparison between maps that share a real variable". Environmental and Ecological Statistics. 15 (2): 111–142. doi:10.1007/s10651-007-0043-y

[14] Draper, N. R.; Smith, H. (1998). Applied Regression Analysis. Wiley-Interscience. ISBN 978-0-471-17082-2.

[15] Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.