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

Mining Voice Biomarkers for Analyzing the Severity Index of Neurodegenerative Parkinson’s Disease using SMOTE -RNN

Ramya N
Ph. D (Research Scholar),
Department of Computer Science,
Vellalar College for Women (Autonomous), Erode
Tamil Nadu, India

Dr. S.Devi Suganya
Assistant Professor
Department of Computer Science,
Vellalar College for Women (Autonomous), Erode,
Tamil Nadu, India

Abstract: -

Parkinson Disease (PD) occurs due to the loss of dopamine in the brains thalamic section that results in unconscious or oscillatory movement in the body. Normally Doctors diagnosis the PD disease clinically with their expertise and experience For this, patients need to take number of tests for diagnosis, but the time, these all tests still not sufficient to diagnosis Parkinson Disease effectively. This work emphasis on classify the severity of PD or idiopathic Parkinsonism.

The aim of this work was to mine the predict whether the patient has Parkinson’s Disease (PD) or not. Motor Diseases substantially characterize PD, and accordingly, a variety of data sets are recorded from the motor system. These data sets correspond of either physical actions of cases or neuroimaging data captured from their smarts. still, the complaint substantially begins times before the motor symptoms.

This research Artificial Neural Network system is used. It simulates through biological Neuro system. Neuron plays a main role, the back propagation algorithm is used to classify the pattern given in statistical classification using multi layer perceptrons.

Keywords: Parkinson’s Disease (PD), Dopamine, ANN Data Mining, Disease Prediction.

.

I. INTRODUCTION

Parkinson’s disease (PD) is a complex neurodegenerative disorder evident through classic motor impairment and complications, a mixture of non-motor symptoms, and progressive disability ,people affected by PD in figure 1.1

Recognizable signs (e.g., tremor or freezing) and performance actions are assessed by an expert rater, but the evaluation of symptoms (e.g., pain, nervousness) and supplementary individual aspects (e.g., quality of life, realization with care) needs the input of patients. However, it is recognized that perceptions on the patients’ condition frequently differ between patients themselves and their doctors making difficult sometimes to decide which of these evaluations is more reliable.

Figure 1.1:Parkinson’s Disease

1.2 Parkinson's Disease

Parkinson’s Disease (PD) is a neurodegenerative disorder which affects mostly older people (1 out of 100 over 75 year [3]). It is characterized by the progressive loss of dopaminergic neurons [4].Its main motor symptoms are:

- tremor (vibration to hands, arms, legs or jaws);
- muscle rigidity (stiffness);
- bradykinesia (slowness of voluntary movements);
- Postural and balance impairment.

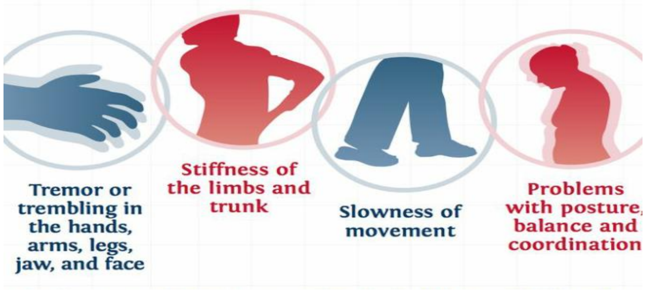
Tremor or trembling in the hands, arms, legs, jaw, and face
Stiffness of the limbs and trunk
Slowness of movement
Problems with posture, balance and coordination

Figure1.2: Symptoms of PD

These symptoms have a negative impact in quality of life, can severely limit motor abilities, and lead to adverse events such as falls.

Pharmacological remedy based on dopaminergic medications can minimize or reduce most of the symptoms but after prolonged periods of treatment it is normal to develop motor complications like:

- dyskinesia: involuntary movements;
- dystonia: involuntary muscle contractions;
- Fluctuations of symptoms severity: abrupt transitions from periods when the medication is effective (ON-periods) and periods when the symptoms are high although the subject is under medication (OFF-periods).

As it was shown here motor impairment have many different characteristics in PD.
Various stages in Parkinson's disease are,

- Primary - Due to unknown reasons
- Secondary - Dopamine deficiency
- Hereditary- Genetic origin
- Multiple system atrophy - Degeneration of parts other than mid brain

The structure of this paper is as follows: the reviews of previously done studies on PD detection have been substantiated in section 2. Section 3 presents the proposed methodology employed for PD detection. Section 4 entails the analysis of classification results and discussion. The paper was concluded finally in section 5.

II. LITERATURE SURVEY

Classification algorithms are most common applicable DM techniques applied for disease prediction. Some of them are listed below:

Parisi, et al(2018)(12) developed a brand-new artificial intelligence- grounded mongrel classifier to help in PD early opinion. The UCI ML database was used to acquire information on 68 cases' clinical scores and dyslexia assessment results. In order to distinguish between physiological and pathological data patterns, MLP(Multilayer Perceptron) generated weights were employed for point selection, and their modules were used to rank input characteristics according to their relative applicability. The original 27 characteristics were thus condensed to 20 chosen individual variables. LSVM(SVM Lagrange) is also used to classify the reduced point set. The MLP- LSVM, a point-grounded mongrel system, was estimated against classifiers from affiliated studies as well as commercially available software to determine its overall performance. The issues demonstrate that the proposed point- grounded system, the MLP- LSVM for 100 overall bracket delicacy and 100 area under the receiver operating characteristic wind, with relative confluence. briskly, showing that it has the implicit to help with the early clinical identification of PD.

From the review above, it may be observed that various Deep learning techniques have been applied in recent research works over voice based PD detection. In this proposed work originally preprocessing is performed on min maximum normalization to homogenize the input data. After preprocessing Synthetic Minority Oversampling TEchnique (SMOTE) is used to balance the input data. Important features are named using information gain model. Once the features subsets are named it sends to the classifier for Parkinson conditions discovery. Eventually Weighted intermittent Neural Network (WRNN) is applied in this work for Parkinson conditions discovery. Proposed system is estimated interms of accuracy, precision, recall and f – measure.

III. METHODLOGY

In this paper, three steps were suggested Parkinson's disease detection paradigm. Preprocessing using SMOTE-based data equalization[14] and min-max normalization-based data normalization come first. Feature selection based on information collection comes second. Last step is classification Weighted Recurrent Neural Network (WRNN) for detection of Parkinson disease. Figure 1: Overall architecture of the proposed model.

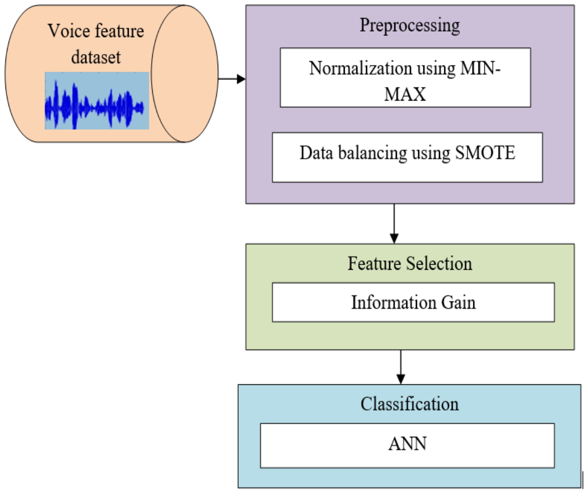

Figure 1: Overall architecture of the proposed model

Steps in Prediction Process:

- Step 0: Start
- Step 1: Load data set
- Step 2: Model Creation using Training set
- Step 3: Testing Model using validation set
- Step 4: Performance analysis
- Step 5: Selection of best Model
- Step 6: Stop

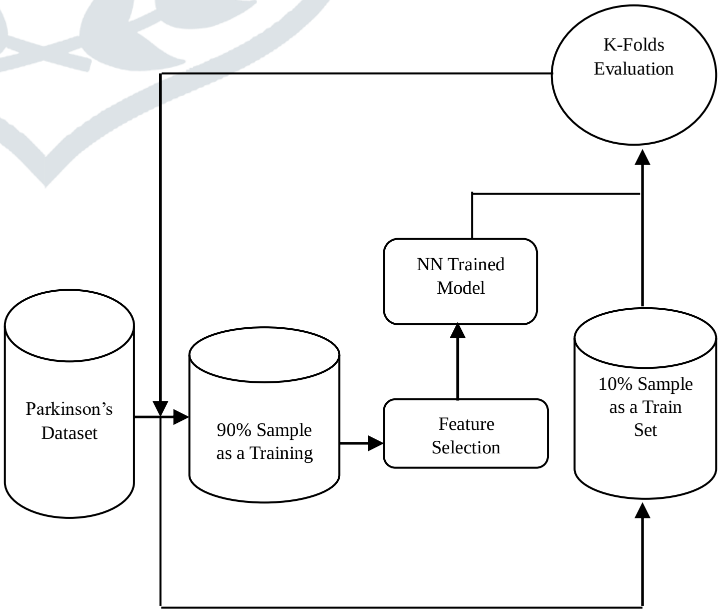

Figure 2 : Proposed Prediction Model

1) In first initialize the model of our proposed system
2) After the initialization the data is loaded to the database from the source(dataset) available.

3) Model is Created using Training set, in this 90% of the sample data is used.
4) Testing Model using validation set, using the NN trained data set.
5) Performance analysis is the finding the result for the model.
6) Selection of best Model

1. Data set

The first step in the process is data collection. Voice analysis, data are collected from UCI, a machine readable container containing voice data in both PD and healthy subjects. The database used has 195 instances and 22 symbols.[13] The objective of the dataset is to differentiate fit persons compared to the unhealthy using the "status" column, which is set to negative for fit persons and positive for those having the disease. The data were in the ASCII CSV format. Each column represents a specific voice measure, while each row represents one recording from patients.[15] For each patient, there are roughly six recordings with different specific voice measures. The first column in the dataset refers to the patient's name. Table 1 details the dataset features information. The statistical analysis for the dataset is illustrated in Table 2.

Column Name	**Description**
Name/ASCII	Patient's name/Record Num
MDVP-Fo (Hz)	Vocal fundamental (mean frequency)
MDVP Fhi (Hz)	Vocal fundamental (Max frequency)
MDVP Flo (Hz)	Vocal fundamental (Min frequency)
MDVP jitter (%)	Several measurements differ in fundamental frequency (i.e., RAP, MDVP, APQ, etc.)
MDVP Fhi (Hz)	Several measurements differ in amplitude (i.e., APQ5, MDVP: APQ, etc.)
NHR, HNR	The ratio of noise with regard to total components in voice
RPDE, D2	Nonlinear complexity measurements
DFA	Fractal scaling exponent
PPE, spread1, spread2	Three nonlinear methods for calculating fundamental frequency variation

Table 1: Data Set

Label	***Value***
Dataset Characteristics	Multivariate
Attribute Characteristics	Real
Number of Instances	197
Number of Attribute	23
Missing Values	None
Area	life

Table 2: Statistics of the dataset

2. Preprocessing

The next step is data preprocessing. The process of transforming the raw data into an understandable format is said to be data preprocessing. With this, we need to separate the raw data into two parts. The first part for training the model, we've used 70 percent of knowledge for training and second part for testing the model; we've used 30 percentages for testing.

a) Min-Max Normalization
b) SMOTE

(a) Min-Max Normalization

Min-max normalization (usually so-called feature scaling) does a linear transformation on the original data. This procedure gets all the scaled data in the range (0, 1). The formula to attain this is the following:

$$v' = \frac{v-\min_A}{\max_A-\min_A}(new\_max_A - new\_min_A) + new\_min_A \quad (1)$$

Where A implies Attribute data, Min (A), Max (A) stand for min. and max. Absolute values of A respectively, stands for new values of data entries, v implies old values of data entries, New_ max (A), new_ min (A) imply max and min values of ranges (i.e boundary values of required range).

(b) SMOTE

This paper shows how a technique called Synthetic Minority Oversampling Technique (SMOTE) deals with the class imbalance problem in PD stage-wise classification by improving minority class recognition. [17] The method is validated by quantifying the dissimilarity among samples generated viewing the non-existence of overlapping or duplication.

The minority class can be oversampled to handle unbalanced datasets. Oversampling is used in the SMOTE algorithm to rebalance the original training set [16]. Rather than simply replicating minority class instances ,SMOTE introduces synthetic examples. By interpolating between several minority class instances within a defined neighborhood, this new data is generated. The instructions to execute SMOTEmethods presented in algorithm[18] 1 are described as:

```
Algorithm 1:SMOTE algorithm
1. Input: minority class illustrations i.e., T; N; k.
   a. Output: synthetic minority class samples i.e., (N/100)*T
   b. Variables: Sample array of minority class
2. If N<100
      a. then ensure that T minorities are randomly selected
      b. T=(N/100)*T5
      c. N=100
3. Endif multiples of 100 are assumed to represent SMOTE
      a. N=(int)N/100
```

```
4.Fori=1 to T do Create an nn-array with the indices of k nearest neighbors for i
5. POPULATE(N,i,nnarray)
6. End for
7. End Function
```

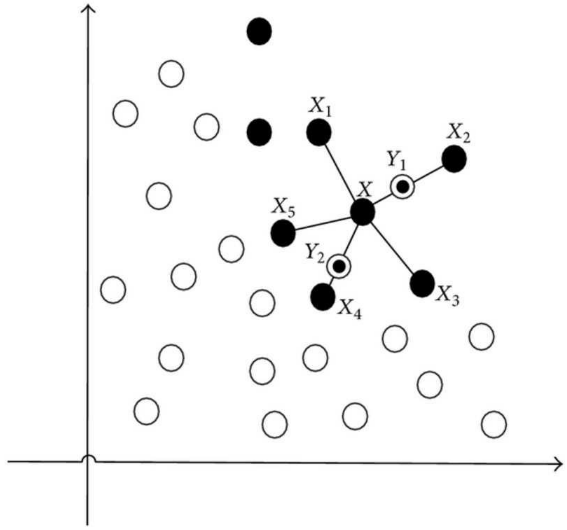

Figure 4. SMOTE algorithm representation diagram

(c) Feature Selection

The next step in our workflow is Feature selection. The process of reducing the number of input variables when developing a predictive model. There are various models that have been used till date by researchers and scientist. In our case we have defines the PD patient's samples from various patients so we have chosen such models which will classify or differentiates the unhealthy patient with the healthy one. There are two feature selection techniques available. They are supervised and unsupervised learning.

Information gain calculates the reduction in entropy or surprise from transubstantiating a dataset in some way. It's generally used in the construction of decision trees from a training dataset, by assessing the information gain for each variable, and opting the variable that maximizes the information gain, which in turn minimizes the entropy and stylish splits the dataset into groups for effective bracket. Information gain is calculated by comparing the entropy of the dataset ahead and after a metamorphosis. collective information calculates the statistical dependence between two variables and is the name given to information gain when applied to variable selection.

(d) Classification

In this study, WRNN is used to identify PD where the classifier receives chosen characteristics to identify the illness. A form of neural network called an RNN has several retired layers in between the input and affair layers. The major responsibility of the RNN is to precisely replicate the input data at the affair subcaste using the training data as a companion. The input and affair layers each have the same number of units as there are features in the training set. The number of units in the three retired orders was courteously determined during the trial to reduce the reconstruction error. still, not all types of features are sensitive to the initialization weights for the input layers of an RNN, which might affect in a slower optimisation process(19,20).

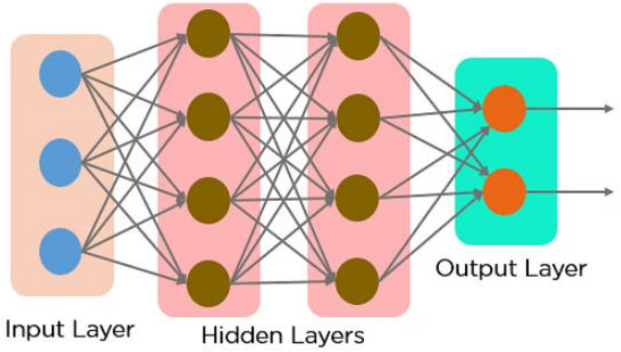

Figure 5. RNN Architecture

IV. RESULT

In the proposed methodology , the Parkinson's Disease Dataset containing voice parameters is used. The data is first skimmed and feature selection is performed on various classification algorithms like SVM, Naive Bayes and RNN. Also, after validating if a person has Parkinson's Disease or not, K-Means is applied to perform clustering in 3 clusters of Low, Medium and High Probability of the Disease.

	Prediction outcomes	
True class	Predicted with no PD	Predicted with PD
0: (no PD)	True Negatives (TN)	False positives (FP)
1: Patient with PD	False Negatives (FN)	True Positives (TP)

Table 3: Confusion matrix representation

A. Performance Metrics

Performance Criteria can be used to estimate a model's capability. It's possible to compare model prognostications with known values of dependent features in a dataset by using parameters such as accuracy, precision, recall, and f-measure

Metrics	Methods		
	SVM	Naive Bayes Classification	ANN
Accuracy (%)	90	96.7	98
Precision(%)	74	82	90
Recall(%)	88	92.42	93
F-measure (%)	76	87.01	89

Table 4. Performance comparsion of result

1) Precision

Precisions are percentages of outcomes which are appropriate and defined as

$$\text{Precision}=\frac{\text{Truepositive}}{\text{truepositive+falsepositive}} \quad (9)$$

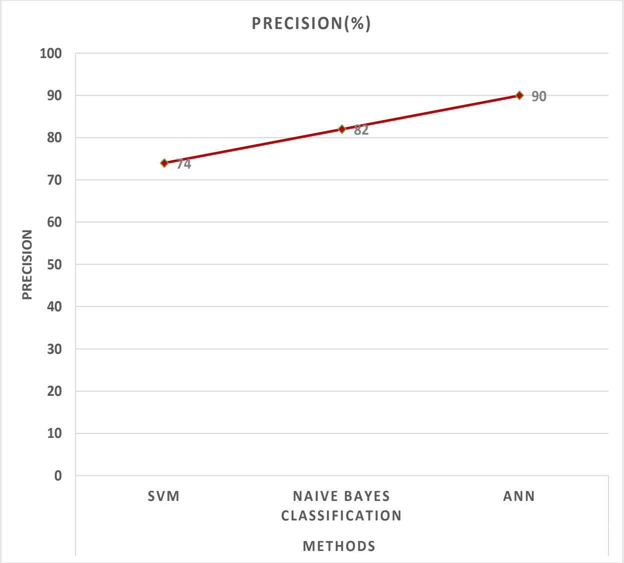

Figure 7. Precision Result

2) Recall

Recalls are percentages of total correctly classified appropriate results by the future algorithm and defined as

$$\text{Recall}=\frac{\text{Truepositive}}{\text{truepositive+FalseNegative}} \quad (10)$$

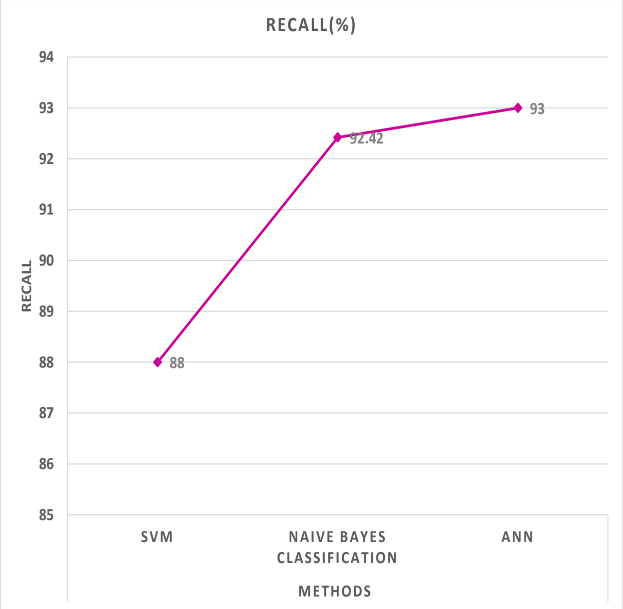

Figure 8. Recall Result

3) Accuracy

Accuracies are fractions of predictions effortlessly identified. Formally, accuracies can be defined as:

$$\text{Accuracy}=\frac{\text{Truepositive+TrueNegative}}{\text{Total}} \quad (11)$$

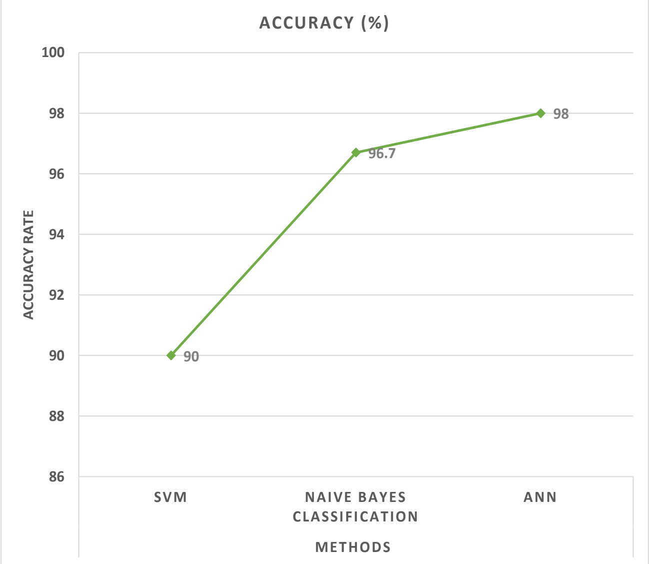

Figure 9. Accuracy Result

4) F measure

An F-score is calculated by taking the harmonic mean of a system's accuracy and recall scores.

$$2 \times [(\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})] \quad (12)$$

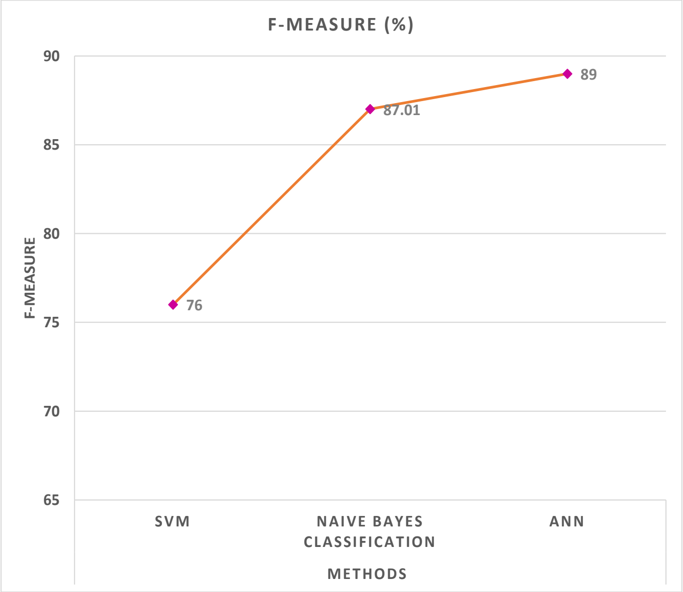

Figure 10. Precision Result

V. Conclusion

For proper medical care in neuropathy, PD opinion is pivotal. The thing of this exploration is to identify Parkinson's complaint at an early stage. Data balance is carried out in this study utilising SMOTE. In this study, the input data are normalised using the min- maximum system. Grounded on the information accession methodology, features are chosen. On the base of speech features, RNN is used to identify Parkinson's complaint. delicacy, perfection, recovery, and f- dimension of the suggested approach are assessed. The findings demonstrate that the suggested model provides 98 further accuracy than other current models. The suggested approach, still, selects features using a single model. The classifier's performance will be enhanced by employing set point selection as opposed to a single model, and this may be targeted in the future.

In this paper, an overview of Data Mining Techniques for Diagnostic Biomarkers for Parkinson’s disease is discussed.

The paper considers the literature survey, methodology and various algorithms that can be applied for diagnostic Parkinson's disease detection using DM techniques. Each and every technique has some merit and demerits.

REFERENCES

[1] Jellinger, K. A. (2015). Neuropathobiology of non-motor symptoms in Parkinson disease. J. Neural Transm. 122, 1429–1440. doi: 10.1007/s00702-015-1405-5. PubMed Abstract | CrossRef Full Text | Google Scholar

[2] D. Vilas et al. Assessment of α-synuclein in submandibular glands of patients with idiopathic rapid-eye-movement sleep behaviour disorder: a case-control study Lancet Neurol. (2016)

[3] R. Bharathi, T. Abirami, S. Dhanasekaran, Deepak Gupta, Ashish Khanna, Mohamed Elhoseny K. Shankar, "Energy Efficient Clustering with Disease Diagnosis Model for IoT based Sustainable Healthcare Systems", Sustainable Computing: Informatics and Systems, Volume 28, December 2020, Pages 1-28

[4] P. Durga et al, Diagnosis and Classification of Parkinsons Disease Using Data Mining Techniques, International Journal of Advanced Research Trends in Engineering and Technology, Vol. 3, Special Issue 14, March 2016.

[5] Rahul R. Zaveri, Prof. Pramila M. Chawan, Prediction of Parkinson's Disease using Data Mining: A Survey, International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 10 | Oct 2020.

[6] Carlo Ricciardi, et al, "Using gait analysis' parameters to classify Parkinsonism: A data mining approach" Computer Methods and Programs in Biomedicine vol. 180, Oct. 2019, 105033, https://doi.org/10.1016/j.cmpb.2019.105033.

[7] Mostafa, S.A., Mustapha, A., Mohammed, M.A., Hamed, R.I., Arunkumar, N., Abd Ghani, M.K., Jaber, M.M. and Khaleefah, S.H., 2019. Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease. Cognitive Systems Research, 54, pp.90-99.

[8] Lamba, R., Gulati, T., Alharbi, H.F. and Jain, A., 2021. A hybrid system for Parkinson's disease diagnosis using machine learning techniques. International Journal of Speech Technology, pp.1-11.

[9] Sharanyaa, S., Renjith, P.N. and Ramesh, K., 2020, December. Classification of Parkinson's disease using speech attributes with parametric and nonparametric machine learning techniques. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 437-442). IEEE.

[10] Dinesh, A. and He, J., 2017, November. Using machine learning to diagnose Parkinson's disease from voice recordings. In 2017 IEEE MIT Undergraduate Research Technology Conference (URTC) (pp. 1-4). IEEE.

[11] Agarwal, A., Chandrayan, S. and Sahu, S.S., 2016, March. Prediction of Parkinson's disease using speech signal with Extreme Learning Machine. In 2016 international conference on electrical, electronics, and optimization techniques (ICEEOT) (pp. 3776-3779). IEEE.

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

[13] Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam,H., Sakar, B. E., ... & Apaydin, H. (2019). A comparativeanalysis of speech signal processing algorithms forParkinson's disease classification and the use of thetunable Q-factor wavelet transform. Applied SoftComputing, 74, 255-263.

[14] Polat Kemal. A hybrid approach to Parkinson disease classification using speech signal: the combination of SMOTE and random Forests. In: 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT); 2019. https://doi.org/10.1109/ebbt.2019.8741725

[15] Iqra Nissar,, Danish Raza Rizvi, Sarfaraz Masood and Aqib Nazir Mir (2019). Voice-Based Detection of Parkinson's Disease through Ensemble Machine Learning Approach: A Performance Study. EAI Endorsed Transactionson Pervasive Health and Technology.

[16] Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data.BMC Bioinformatics,14,1–16

[17] Chintalapudi, N., Dhulipalla, V. R., Battineni, G., Rucco, C., &Amenta, F. (2023). Voice Biomarkers for Parkinson's Disease Prediction UsingMachine Learning Models with Improved Feature Reduction Techniques.Journal of Data Science and Intelligent Systems1(2), 92–98,https://doi.org/10.47852/bonviewJDSIS3202831

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer,"SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

[19] M. Cheng,W. J. Sori, F. Jiang, A. Khan, and S. Liu, "Recurrent neural network based classification of ECG signal features for obstruction of sleep apnea detection", In 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), vol. 2, pp. 199-202, 2017.

[20] P. Xu, and R. Sarikaya, "Contextual domain classification in spoken language understanding systems using recurrent neural network", In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 136-140, 2014.