International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Statistical and Unsupervised MLs Analysis on Parkinson's Disease Data set Acquired from A.P. India

T. PanduRanga Vital, P. Shiny, S. E. Ashish, T. Sai Kumar

Abstract*: *In recent years, the voice analysis is the important work for identifying the neurological diseases like Parkinson's disease (PD). PD is the subsequent general neurodegenerative disorder after Alzheimer's lacking of dopamine in mid brain. In most people, symptoms appear at the age of 50 years or over. In this research, one thousand two hundred vowel-sounded (/.a/.e/.i/.o/.u) voice records are collected from A.P., India for analyzing people with PD from that of healthy people. The records constitute the data of 40 PD patients and 36 non-PD people who are having their age between 50 and 85. Those voice recordings are processed and relevant features or characters are extracted. Here, the data set contains features of both people with PD and healthy to distinguish performance. In this, we analyzed the PD dataset with statistical and unsupervised machine learning analysis. The efficient clustering k-means algorithm represents the Centroids of each attribute of the PD voice data set in two clusters (cluster 0, cluster 1). Another used unsupervised ML algorithm, hierarchal clustering clusters the data set in row wise (attribute wise) as well column wise(data wise) and analyze the projections of attributes and their rankings with using PCA(Principle component analysis). Parkinson's disease (PD), Unsupervised Machine Learning, Voice, PCA

Index Terms*: *Parkinson's disease (PD), Unsupervised Machine Learning, Voice, PCA

I. INTRODUCTION

PD was distinguished by James Parkinson in 1817. PD is viewed as a neural issues those outcomes from the joined impacts of numerous hazard and defensive components, including hereditary and ecological ones. The symptoms of PD grow bit by bit. In the majority of people, symptoms show up at 50 old years or over. Researchers have connected low or falling dimensions of dopamine, a synapse, with PD. An individual with PD may have bunches of protein in their cerebrum known as Loewy bodies. PD can likewise be caused because of hereditary elements. As dopamine levels fall in an individual with PD, their side effects bit by bit become progressively serious.

Unsupervised machine learning algorithms derives patterns from a dataset without any reference to labelled, or known, outcomes. Dissimilar to managed ML, unsupervised techniques can't be connected legitimately to a regression or a grouping issue as we don't have the foggiest idea how the output seems, by all accounts, to be, by making it unfeasible for you to prepare the calculation. This kind of learning can likewise be utilized for finding the basic model of the data. Unsupervised ML aims to show the unknown patterns in the data. The best time to use this is when you don't have any desired output. Some of the applications of this type of ML include clustering, anomaly detection, Association mining, latent variable models. K-means clustering, an unsupervised learning algorithm, is used when you have uncategorized data. The aim of this algorithm is to detect clusters in the data set, where 'k' is the number of clusters. The algorithm performs repetitively to assign every data point to one of K clusters based on the attributes that are given and similarity. To calculate the similarity of each data point and cluster, we use Euclidean distance as a measure. Fuzzy C-Means is a clustering algorithm which makes the data point available to one or more clusters such that data points in the same cluster are as similar as possible. This is used in pattern recognition. Clusters are identified using some similarity measures.

Hierarchical clustering mainly focuses on forming a hierarchy of clusters. This can be of two types, agglomerative and divisive. The output is a set of clusters where everyone is different from others. In troublesome hierarchical clustering, every point of data are doled out a solitary cluster and after that partitioned to two least comparable clusters. This procedure proceeds until every datum point has its own cluster. In aggleromative hierarchical clustering, every datum point is treated as a singleton-cluster and clusters are consolidated dependent on likeness. This procedure proceeds till every one of the clusters structure into one.

II. RELATED WORK

We reviewed so many papers from reputed journals. Some of the papers are projected in this part as review. Aich et al., [1] Proposed a methodology by contrasting the performance measurements and distinctive capabilities,

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
T. PanduRanga Vital*, Dept. of Computer Science and Engineering, Aditya Institute of Technology and Management (Autonomous), Tekkali-532 201, Andhra Pradesh, India
P. Shiny, Dept. of Computer Science and Engineering, Aditya Institute of Technology and Management, Tekkali-532 201, Andhra Pradesh, India
S. E. Ashish, Dept. of Computer Science and Engineering, Aditya Institute of Technology and Management, Tekkali-532 201, Andhra Pradesh, India
T. Sai Kumar, Dept. of Computer Science and Engineering, Aditya Institute of Technology and Management, Tekkali-532 201, Andhra Pradesh, India

for different data sets, unique capabilities just as important segment investigation based element decrease system for choosing the capabilities and used non-straight based order way to deal with look at the execution measurements and have picked up an performance of 96.83%. This examination enables the clinicians for separating the PD to aggregate from sound gathering dependent on the voice information. Lizbeth Naranjo Carlos et al., developed an expert system for diagnosing PD. This expert system takes out the important features from the voice samples and advanced statistical approach is used for pattern recognition. The importance of the work lies in use and development of Bayesian approach for the dependent nature of data in feature-based design using Gibbs sampling. To calculate the performance of the proposed system, an experiment with voice records is performed to distinguish healthy people from those of PD which includes 80 subjects where half are affected by PD. In spite of the reduced subjects, the system can discriminate healthy people from that of PD affected people [2]. Gürüler, H. et al., researched on Parkinson's disease (PD) that it is a degenerative, central nervous system issue. A Parkinson dataset involving the features taken from speech examples are used to find PD. PD characteristics are weighted using the KMCFW technique. Test results have revealed that the proposed system, entitled KMCFW– CVANN, remarkably outperforms all the other models listed and gets the highest classification results mentioned so far, with an accuracy of 99.52%. The method ensures the conclusion that the complex-valued model's ability to classify diseased people from healthy people is high [3]. Hui-LingChen et al., used fuzzy k-nearest neighbor (FKNN) for diagnosing Parkinson's disease (PD) by comparing FKNN-based framework with the support vector machines (SVM) based methodologies. The strength has been thoroughly assessed on a PD date set regarding order precision, sensitivity, specificity and the area under the receiver operating characteristic (ROC) curve (AUC). Murat Gök et al., [5] examined that PD affects the central nervous system which further leads to difficulties in motor functionality. Different computational tools are to be developed for diagnosing PD in the early stages based on some symptoms. In this paper, several classification algorithms are applied and a model is built based on a feature set for the diagnosis of PD. All the performances of the classification algorithms are examined on a PD data set. The new models outperform the old ones in terms of accuracy of (98.46%) and ROC (0.99) by applying k-Nearest Neighbour algorithm [4]. The table1 shows the contribution of different authors research work on Parkinson's disease and related diseases.

Table 1 : Contribution of authors work on PD Voice analysis

Reference	Area of application	Year
Aich et.al.	A nonlinear decision tree based classification approach to predict the Parkinson's disease using different feature sets of voice data	2018
Lizbeth Naranjo Carlos et.al.	Addressing voice recording replications for Parkinson's disease detection	2016
Gürüler, H. et.al.	A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method	2016
Hui-LingChen et.al.	An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbour approach	2013
Murat Gök et.al.	An ensemble of k-nearest neighbours algorithm for detection of Parkinson's disease	2013
Indira Rustempasic et.al.	Diagnosis of Parkinson's Disease using Fuzzy C-Means Clustering and Pattern Recognition	2013
Kemal Polat et.al.	Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering	2011
Hamid-Karimi Rouzbahani et.al.	Diagnosis of Parkinson's Disease in Human Using Voice Signals	2011
Stephanie M. van Rooden et.al.	The identification of Parkinson's disease subtypes using cluster analysis: A systematic review	2010
Rhonda J. Holmes et.al.	Voice characteristics in the progression of Parkinson's disease	2010
E.C. de Angelis et.al.	Effect of voice rehabilitation on oral communication of Parkinson's disease patients	2009

Gürüler, H. et al., researched on Parkinson's disease (PD) that it is a degenerative, central nervous system issue. A Parkinson dataset involving the features taken from speech examples are used to find PD. PD characteristics are weighted using the KMCFW technique. Test results have revealed that the proposed system, entitled KMCFW–CVANN, remarkably outperforms all the other models listed and gets the highest classification results mentioned so far, with an accuracy of 99.52%. The method ensures the conclusion that the complex-valued model's ability to classify diseased people from healthy people is high [3]. Hui-LingChen et al., used fuzzy k-nearest neighbor (FKNN) for diagnosing Parkinson's disease (PD) by comparing FKNN-based framework with the support vector machines (SVM) based methodologies. The strength has been thoroughly assessed on a PD date set regarding order precision, sensitivity, specificity and the area under the receiver operating characteristic (ROC) curve (AUC).

Murat Gök et al., [5] examined that PD affects the central nervous system which further leads to difficulties in motor functionality. Different computational tools are to be developed for diagnosing PD in the early stages based on some symptoms. In this paper, several classification algorithms are applied and a model is built based on a feature set for the diagnosis of PD. All the performances of the classification algorithms are examined on a PD data set. The new models outperform the old ones in terms of accuracy of (98.46%) and ROC (0.99) by applying k-Nearest Neighbour algorithm [4].

Indira Rustempasic et al., analyzed that Parkinson's disease is a neurodegenerative disease which is a public health issue globally. Various recent studies have shown that one of the early indicator of PD is voice, and so, Parkinson's data set is taken that contains different features of voice which are responsible for the detection of PD. The PD data set is given to the Fuzzy C-Means clustering (FCM) and pattern recognition and the performance is examined. Here, the data set contains features of both people with PD and healthy people to distinguish performance. More accurate results can be obtained by classifying the data first and then applying these two patterns i.e., FCM and pattern recognition [6]. Kemal Polat et al., [7] used Fuzzy C-Means clustering for the diagnosing the Parkinson's disease(PD). In the PD dataset, practical values of some featured attributes for distinguishing people form PD with that of the healthy people are considered as the input for FCM. The main aim of FCM is assign each data point to clusters so that all the data points in a cluster are similar based on a similarity measure. The PD data set is given to KNN system where several k values are compared with each other. The result of k-values in the KNN for distinguishing the PD is based on the best value of k. The results have found that the combination of both FCM and KNN has obtained the best results for PD classification. Hamid Karimi Rouzbahani et al., [8] investigated the voice features of people with and without Parkinson's disease (PD)and extracted required features. A total of 31 people are investigated to collect the data set. The voice signals were recorded, processed and then feature extraction is done using different feature selection methods for obtaining the best results in diagnosis of PD. These features are then given to various classifiers to check for the symptoms of PD. The performances obtained from the classification algorithms and were compared to each other and the best performance was obtained using the KNN classifier with a rate of 0.9382. Stephanie et al., [9] studied that there may be existing several sub types of the disease based on the differences between patients with Parkinson's disease(PD). To classify the patients into sub-types, Cluster Analysis(CA) is used. Seven studies were found which matched the required criteria and the differences in these studies gave the comparison of the results The clusters depicts that people of older age are highly prone to PD where people of young age have a slow disease progression. Rhonda et al., [10] Performed a study considering the voice characteristics of patients with PD based on the severity of the disease. For the diagnosis, voice characteristics of 30 people with PD of early stage and 30 people with PD of later stage were compared with the data set of 30 normal people. Compared to normal data, voices of people having PD have lower mean values and reduced maximum frequency ranges. The present PD data have excess jitter and a reduced frequency in male and variability in female. While many of these voice features did not appear to decline with disease progression, mono pitch, breathiness and mono loudness, reduced maximum frequency range and low loudness were all bad in the later stages of PD. Angelis et al., [11] Studied that in Parkinson's disease, voice disorders are general which leads to social isolation twenty patients are examined before and after the voice rehabilitation program that consisted of 13 programs during 1 month with a stress on an increase in sphincter activity. It produced a decrease in strained voice and increase in the vocal intensity and monotonous speech with the elimination of swallowing alterations. These data show a greater efficiency after the voice rehabilitation reflecting a more structural oral communication.

III. METHODOLOGY

The voice data is collected from various places and resources centers (hospital and clinical centers of neurological-diseases) from individual with Parkinson's disease and non-PD peoples from state of Andhra Pradesh, India. The dataset has been analyzed by using Data Mining Techniques. The attributes: 29 (letter, age, gender, Median_pitch, Mean_pitch, pitch_Standard _deviation, Minimum_pitch, Maximum_pitch, Number_of_pulses, Number_of_periods,pul_Mean_period,Standard_deviation_ o f_period, Fraction_of_ locally_unvoiced_frames, Number_of_ voice_breaks, Degree_of_ voice_breaks, jit_local_per, jit_loc_absol, jit_rap_per, jit_ppq5_per, jit_ddp_per, shi_local_per, shi_loc_dB, shi_apq3_per, shi_apq5_per,shi_apq11_per,shi_dda_per, har_Mean_autocorrelation,har-Mean-noise to harmonics ratio, har Mean harmonics to noise ratio) and Class: pd_or_npd). The figure 1 shows the data mining model for clustering and visualization. The collected records constitute the data of 40 PD people and 36 non-PD people who are having their age between 50-85 with 29 attributes and one class (Vowel , Age, Gender, pitch, pulses, frames, voice-breaks, jitter, harmonic noise, PD or NPD class status of pd(1 (yes) for non-pd (0 (no))). The data is pre-processed with using DM techniques then create the data set for processing the visualizations and cluster analysis.

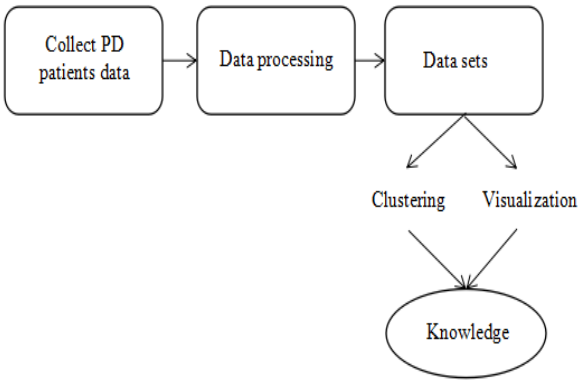

Figure 1: Data mining model for clustering

The *.wave files are decomposed in 26 hidden characteristics of voice with using frequencyc (F), impedance (I) and modulations. Below formulas show the hidden characteristics of voice.

The equations 1 to 4 represent the jitter hidden voice values and the equations 4 to 8 represent the shimmer calculations. The equation 9 represents the harmonic value.

$$jitter = \frac{1}{N_p - 1}\sum_{l=1}^{N_p} |T_l - T_{l-1}| \quad (1)$$

$$Jitter_{Relative} = \frac{\frac{1}{N_p - 1}\sum_{l=1}^{N_p} |T_l - T_{l-1}|}{\frac{1}{N_p}\sum_{l=1}^{N_p} T_l} \times 100 \quad (2)$$

$$Jitter_{rap} = \frac{\frac{1}{N_p - 1}\sum_{l=1}^{N_p - 1} \left| T_l - \left(\frac{1}{3}\sum_{m=l-1}^{l+1} T_m \right) \right|}{\frac{1}{N_p}\sum_{l=1}^{N_p} T_l} \times 100 \quad (3)$$

$$Jitter_{ppq5} = \frac{\frac{1}{N_p - 1}\sum_{l=2}^{N_p - 2} \left| T_l - \left(\frac{1}{5}\sum_{m=l-2}^{l+2} T_m \right) \right|}{\frac{1}{N_p}\sum_{l=1}^{N_p} T_l} \times 100 \quad (4)$$

$$Shimmer_{dB} = \frac{1}{N_p - 1}\sum_{l=1}^{N_p - 1} \left| 20 \times \log \left(\frac{A_{l-1}}{A_l} \right) \right| \quad (5)$$

$$Shimmer_{Relative} = \frac{\frac{1}{N_p - 1}\sum_{l=1}^{N_p - 1} |A_l - A_{l+1}|}{\frac{1}{N_p}\sum_{l=1}^{N_p} A_l} \times 100 \quad (6)$$

$$\text{Shimmer}_{apq3} = \frac{\frac{1}{N_p - 1}\sum_{l=1}^{N_p - 1} \left| A_l - \left(\frac{1}{3}\sum_{m=l-1}^{l+1} A_m \right) \right|}{\frac{1}{N_p}\sum_{l=1}^{N_p} A_l} \times 100 \quad (7)$$

$$Shimmer_{ppq5} = \frac{\frac{1}{N_p - 1}\sum_{l=2}^{N_p - 2} \left| A_l - \left(\frac{1}{5}\sum_{m=l-2}^{l+2} A_m \right) \right|}{\frac{1}{N_p}\sum_{l=1}^{N_p} A_l} \times 100 \quad (8)$$

$$Harmonic_{NoiseRatio} = 10 \times \log_{10} \frac{V_{AC}(T)}{V_{AC}(0) - V_{AC}(T)} \quad (9)$$

K-Means Clustering Algorithm:

K-Means is the simplest unsupervised learning algorithm used for solving a clustering problem. It is a method for vector quantization and also used as a clustering algorithm in data mining. It performs the task of grouping together similar data points to a cluster. The data points are clustered based on the similarities between them. The method pursues a straightforward and a simple method to group the information focuses into bunches dependent on a comparability measure [12][14]. Here, in this calculation, Euclidean separation is considered as the similitude measure.

Algorithm:

1. To start, we select 'k' number of distinct clusters depending upon the data and randomly initialize their centres.

2. The Euclidean distance between the point and each cluster centre is computed, and then the point is classified to the corresponding cluster whose centre is nearest to it.

3. Supported with these classified data points, we estimated the cluster centre by captivating the mean of all the data points in the cluster set.

4. Iterate the steps 2 and 3 until the cluster centres don't change much between consecutive iterations.

Agglomerative Hierarchical Clustering Algorithm:

Hierarchical clustering basically centers on shaping a hierarchy of clusters. This can be of two kinds, divisive and agglomerative. The o/p is a lot of clusters where everybody is not the same as others. In troublesome various hierarchical clustering, every one of the information indicates are relegated a solitary cluster and after that parceled to two least comparable clusters [15]. This procedure proceeds until every datum point has its own cluster. In aggleromative clustering, every datum point is treated as a singleton-cluster and clusters are consolidated dependent on similarity [16]. This procedure proceeds till every one of the clusters form into one.

Algorithm:

Let X = {x1, x2, x3, ..., xn} be the set of data points. D = d[(i,j)] be the N*N proximity matrix. 0,1,....., (n-1) are the sequence numbers assigned to the clusters.

L(k) is the height of the kth cluster. Proximity = d [(r),(s)]between clusters (r) and (s).

1. To start with, consider the clustering having sequence number n = 0 and the level L(0) = 0.
2. Find the cluster pairs having least distance. Suppose they are (r), (s), then d [(r),(s)]= min d[(i),(j)].
3. Increment the Sequence number by 1, n=n+1. Merge the clusters (r), (s) to a single clustering to build the next cluster n. Set it's clustering level to L(n)=d[(r),(s)].
4. Now, modify the proximity matrix, D by replacing the rows and columns of the clusters (r), (s) with the newly built clusters. Compute the distances between each of the new clusters (r,s) and each of the old cluster (k) is denoted as d[(k), (r,s)] = min(d[(k),(r)], d[(k),(s)]).

Check whether all the data points are in the same cluster. If yes, stop. Else, repeat from step 2.

IV. RESULTS AND DISCUSSION

Statistical Analysis PD and non-PD voice records: The Figures 2 and 3 show the PD and Non-PD pulses. The Figure 2 shows the Parkinson's Diseased (P.D) voice pulse. It shows the breaking the voice in PD between pulse to pulse of the voice. The figure 3 shows about the non-PD voice in plain with out harmonic noice errors. As per voice pulses visulizations, there are lot of difference between PD and non-PD that there are many differences between PD and Non-PD voice parameters. Mainly there are immpact on pitch values, harmonic noice values, jitter and shimmer values.

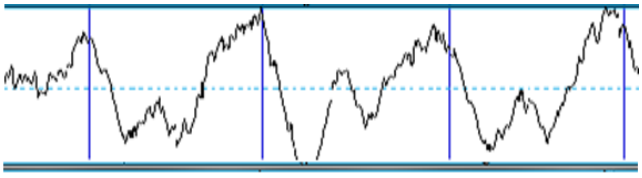

Fig 2: PD Voice Pulses

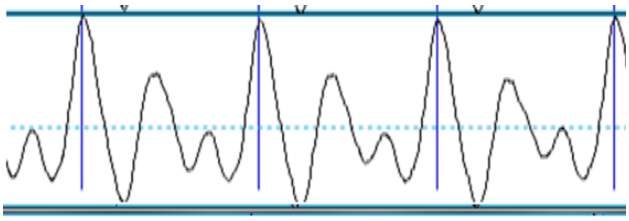

Fig 3. Non-PD Voice Pulses

The figure 4 shows the age attribute stattistical analysis. The acquired data is very simmilar values on attribute of age that the age difference is in ± 3 yeares. The acquired data of PD age is between 56(min) to 83(max) and non-PD age is is between 59(min) to 84(max). The mean values of PD and non-PD are 66.38 ± 8.07 and 69.48±7.42 respectively. The median values of PD and non-PD are 64.00 and 66.00 respectively.

The figure 5 shows the Pitch Median attribute stattistical analysis. The acquired data of PD Pitch Median is between 6(min) to 318(max) and non-PD Pitch Median is between 18.00
(min) to 1704.00 (max). The mean values of PD and non-PD are 54.37 ±29.84 and 81.30±120.65 respectively. The median values of PD and non-PD are 53 and 66 respectively.

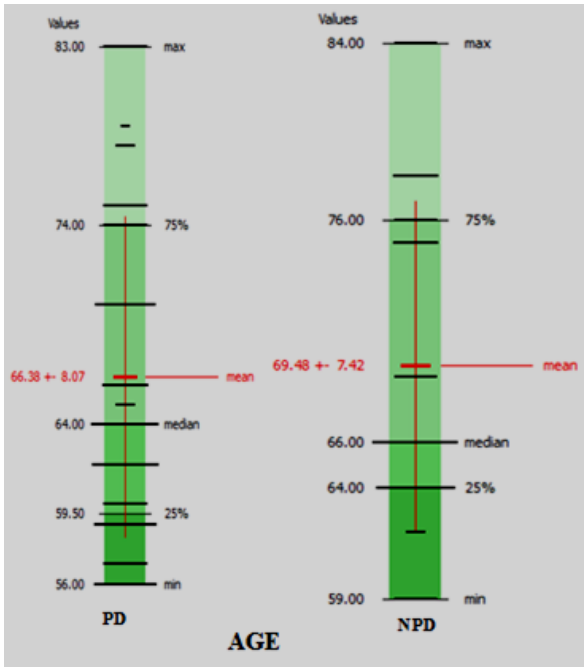

Figure 4: Comparing Age between PD and NPD Statistics

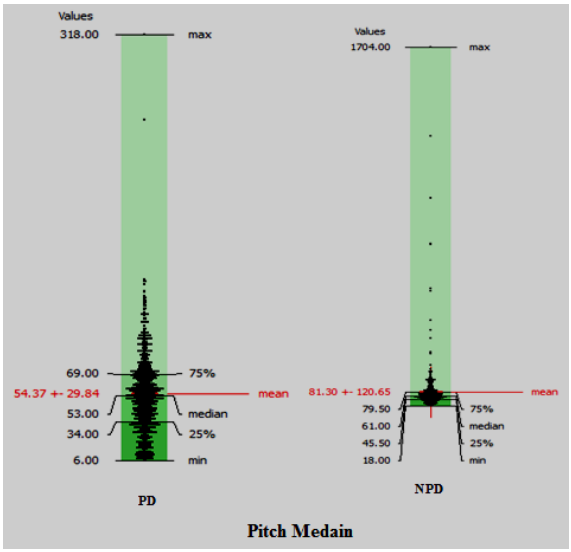

Figure 5: Comparing Pitch Median between PD and NPD Statistics

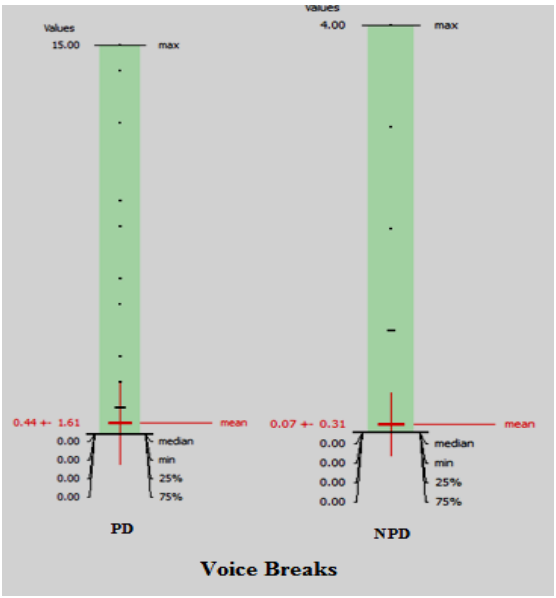

Figure 6: Comparing Voice Breaks between PD and NPD Statistics

The figure 6 shows Comparing Voice Breaks between PD and NPD Statistics. The acquired data of PD Voice Breaks values are in between 0(min) to 15(max) and non-PD Voice Breaks values are between 0 (min) to 4 (max). The mean values of PD and non-PD are 0.44±1.61and 0.07±0.31 respectively. The median values of PD and non-PD are 0.0 and 0.0 respectively. The table 2 shows the every attribute statistical analysis of PD versus non-PD. It shows the detailed analysis of every attribute analysis. Some of the attributes are statistically shown the difference or marginal values are very high that it indicates the impact factor in PD diagnosis.

Table 2 : Attribute Statistical Analysis of The PD And NPD

Attributes	PD Voice Statistical Analysis				Non-PD Voice Statistical Analysis			
	Min	Max	Mean	Median	Min	Max	Mean	Median
Age	56.0	84.0	68.6±8.33	66.0	56.00	83.00	66.90±8.04	65.00
Median pitch	78.87	283.50	171.01±41.96	.71.45	99.9	448.60	227.36±61.16	217.22
Mean pitch	84.38	283.92	173.22±39.42	.71.63	103.1	419.20	231.47±57.13	218.27
Standard deviation	0.15	130.71	23.39±25.98	15.16	1.22	170.61	35.94±31.75	23.53
Minimum pitch	66.25	281.62	138.79±42.0	.39.49	49.77	320.98	177.15±53.52	176.55
Maximum pitch	96.99	627.03	230.54±99.63	209..3	108.8	526.42	294.06±99.74	252.68
Number of pulses	6.00	1704.0	75.57±138.95	51.0	7.00	138.00	62.90±24.70	59.00
Number of periods	5.0	1694.0	73.58±136.92	49.0	6.00	134.00	61.39±24.30	58.00
Mean period	0.00	0.01	0.01±0.00	0.01	0.00	0.01	0.00±0.00	0.00
UnvoicedFramesFraction	0.00	72.32	3.96±10.02	0.81	0.00	67.14	10.97±16.07	2.44
Number of voice breaks	0.00	21.0	0.49±2.02	0.00	0.00	3.00	0.09±0.31	0.00
Degree of voice breaks	0.00	73.76	2.79±9.99	0.00	0.00	19.80	0.67±2.49	0.00
Jitter (local)	0.00	13.28	1.87±1.49	1.62	0.16	7.43	1.72±1.16	1.50
Jitter (local, absolute)	0.00	0.04	0.00±0.00	0.00	0.00	0.11	0.00±0.01	0.00
Jitter (rap)	0.03	6.73	0.83±0.78	0.69	0.04	4.59	0.76±0.63	0.61
Jitter (ppq5)	0.04	8.60	0.93±0.91	0.73	0.07	3.88	0.81±0.68	0.62
Jitter (ddp)	0.09	899.0	3.88±33.94	2.07	0.13	13.75	2.28±1.92	13.75
Shimmer (local)	1.01	29.38	8.29±3.96	7.79	0.87	13.17	2.28±1.92	1.82
Shimmer (local, dB)	0.09	420	1.42±15.84	0.80	0.13	2.28	0.90±0.38	0.85
Shimmer (apq3)	0.24	14.59	3.62±2.22	3.29	0.39	18.83	3.38±2.29	2.83
Shimmer (apq5)	0.34	17.50	4.85±2.74	4.30	0.91	21.70	4.81±2.89	4.05
Shimmer (apq11)	0.52	29.35	7.74±4.44	6.94	1.30	42.44	8.92±5.78	7.70
Shimmer (dda)	0.71	43.77	10.99±6.67	9.86	1.17	71.49	10.19±7.09	8.51
Mean autocorrelation	0.04	1.00	0.90±0.08	0.92	0.62	1.0	0.90±0.66	0.91
Mean noise-to-harmonics	0.00	16.76	0.17±0.64	0.10	0.00	0.97	0.14±0.12	0.12
Mean harmonics-to-noise	2.83	30.19	14.26	5.33	0.23	29.19	13.99±4.96	13.33

K-Means Cluster analysis: Table 1 shows the main Cluster Centroids using K-Means Clustering algorithm. The model is constructed with full training data and takes the 0.59 seconds time. There are two clustered instances that they are Cluster 0 and Cluster 1. The full data is 1200 records the cluster 0 contains 736 datum points as well the cluster 1 contains 464 data points. The full data Centroids are age(67.8917), gender 1(Male), pih_medn (320.3621), pih_mean(197.489), pih_sd (28.9672) , pih_min(154.7737),harmean_nse2har(0.1555),harmean_dBhar2nse(14.1375) and the class attribute pd_npd is 1 for PD. The cluster 0 Centroids are age(65.9688), gender 2 (Female), pih_medn (400.0593), pih_mean (193.9459), pih_sd (32.9616), pih_min (152.2296), harmean_nse2har (0.1626), harmean dBhar2nse (12.6195) and the class attribute pd_npd is 1 for PD. The cluster 1 Centroids are age(65.9688), gender 1(Male), pih_medn (400.0593), pih_mean (198.9286), pih_sd (32.9616), pih_min (152.2296), harmean_nse2har (0.1626), harmean_dBhar2nse (12.6195) and the class attribute pd_npd is 2 for NPD.

Table 3: K-Means Cluster analysis (Centroids of each cluster)

Attribute	Full-Data (1200)	Cluster 0 (736)	Cluster 1 (464)
letter	4	1	4
age	67.8917	65.9688	70.9418
gender	1	1	2
pih_medn	320.3621	400.0593	193.9459
pih_mean	197.489	198.9286	195.2055
pih_sd	28.9672	32.9616	22.6313
pih_min	154.7737	152.2296	158.8093
pih_max	257.0035	267.902	239.7162
pul_num	70.29	70.1563	70.5022
pul_period	68.4975	68.163	69.028
pul_meanp	0.0055	0.0055	0.0054
pul_sdp	0.0008	0.0009	0.0006
voi_unvo	6.8767	7.807	5.401
voi_nov	0.3242	0.4063	0.194
voi_degno	61.2596	99.2072	1.067
jit_local_per	1.9468	2.2116	1.5268
jit_loc_absol	0.0007	0.0009	0.0003
jit_rap_per	0.8251	0.9081	0.6935
jit_ppq5_per	9.314	14.7062	0.761
jit_ddp_per	11.6617	17.7078	2.0715
shi_local_per	8.6256	9.3098	7.5405
shi_loc_dB	1.8352	1.4833	2.3934
shi_apq3_per	3.5542	3.9235	2.9685
shi_apq5_per	4.8305	5.2544	4.1582
shi_apq11_per	9.8768	8.7938	11.5945
shi_dda_per	10.6585	11.7651	8.9031
harmean_autocorr	1450.4897	1090.0905	2022.1573
harmean_nse2har	0.1555	0.1626	0.1443
harmean_dBhar2nse	14.1375	12.6195	16.5453
pd_npd	1	1	2

Figure 7 shows K-Means clustering related to Cluster and pd_npd values. Cluster values (cluster 0 and cluster 1) are located on x-axis and class attribute PD_NPD values are located on y-axis. The red color data points specify the Parkinson's disease class data points and blue colored Non Parkinson's disease class data points.

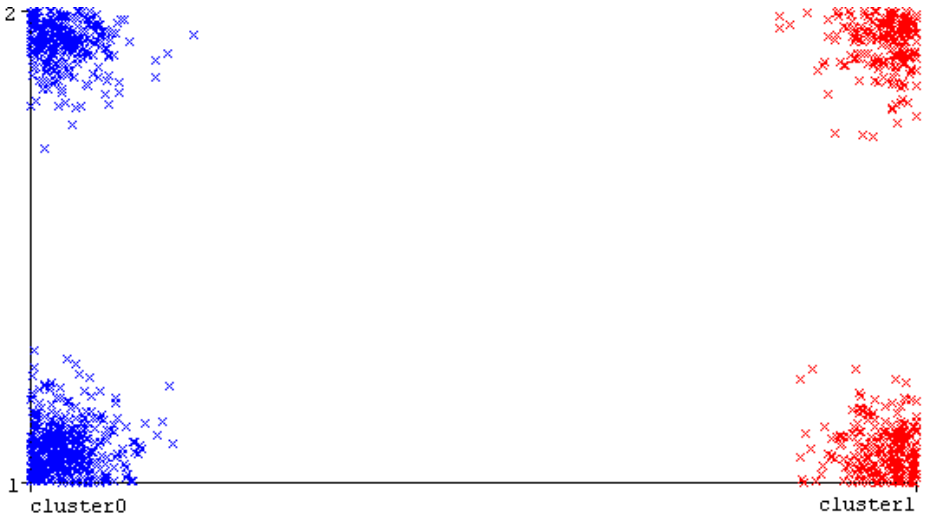

Figure 7: K-Means Cluster analyses with clusters versus class attribute PD_NPD

Hierarchical Clustering:

Vertical (Attribute wise) Complete Linkage Hierarchical Clustering: The figure 8 shows the complete linkage with the attribute (column) wise hierarchal clustering. The distance measures on continuous attributes and used the Pearson's correlation equation. The minimum distance of attributes is shi_apq3_per and shi_dda_per that the distance is 0.005190 and next minimum distance is 0.023474 between jit_local_per and jit_ddp_per. These values are in the blue colour cluster. The figure 8 contains the three clusters (blue, pink and green) at cut line distance 0.56. The blue colored cluster grouped in pulses, voiced-breaks, un-voiced frames, jitter and shimmer attributes. The height of this cluster is 0.546373. The pink colored cluster height is 0.51144 and the green colored cluster height value is 0.559787.

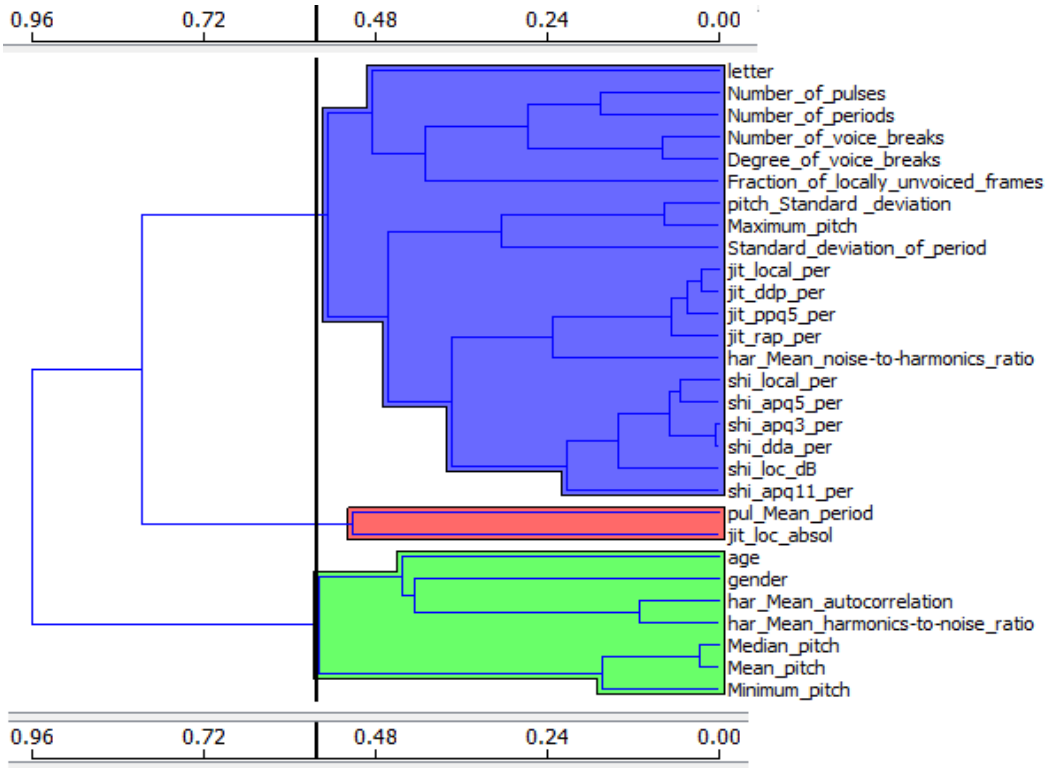

Figure 8: Hierarchical Cluster with respect all Attributes of Parkinson's disease dataset

Horizontal (Row wise) Complete Linkage Hierarchical Clustering: The figure 9 shows the complete linkage with the Horizontal (Row wise) wise hierarchal clustering with respect to mean pitch attributes values. The height measures with using Euclidean distance equation. The minimum distance in blue colored cluster is 0.406328 and next least height is 0.402567, as well the least height of the pink colored cluster is 0.628820 and next least height is 0.706418. The highest height of the blue colored cluster is 2.719270 and pink colored cluster eight is 2.755209. The cut line is appeared at the height of 2.76.

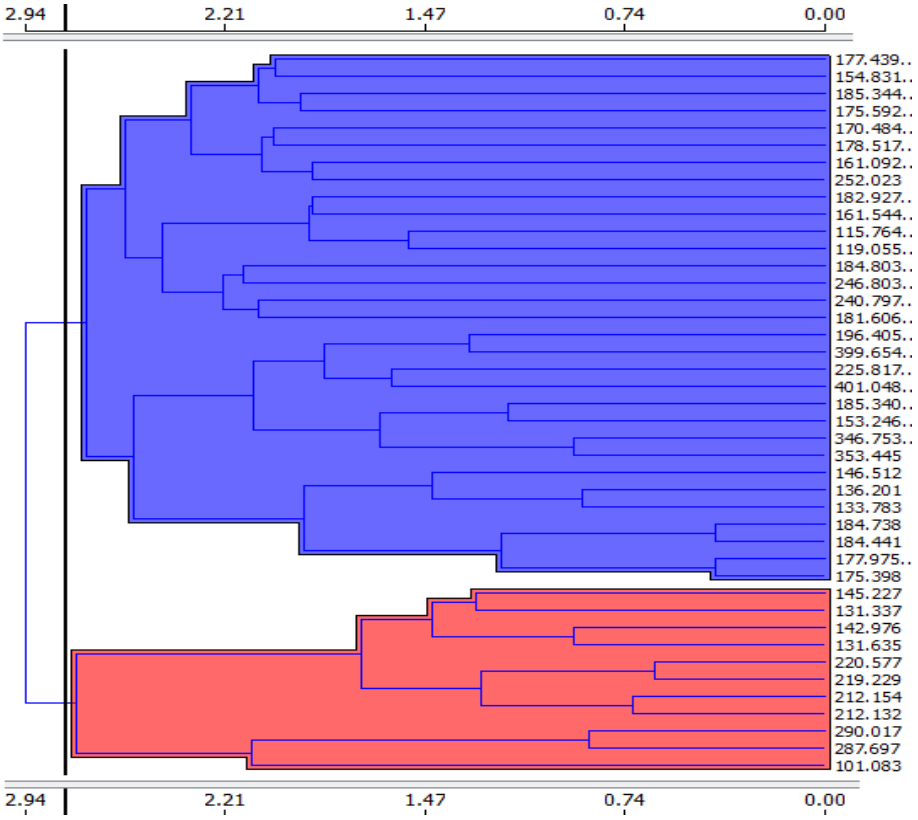

Figure 9: Horizontal Complete Linkage Hierarchical Clustering with respect to Mean –Pitch

Figure 10 shows the cluster projection attributes with respect to Mean pitch, Minimum pitch, Maximum pitch, pulse mean period and Fraction of locally unvoiced frames. The blue elements show non_PD and the red elements show the PD data points. Most of diseased patients have high Fraction of locally_unvoiced_frames. These elements are under the hierarchical clusters, 0 and 1. The projections are shown in the figure that before applying PCA. These five projection attributes are listed in first five ranks for identification of PD with 98.23% accuracy. The circle ("O") symbols represent the 'Male' data points of Gender attribute as well the cross symbol ("X") attributes represents the Female Data points. In the Projections it clearly show this gender attribute with class PD or Non-PD class with red and blue colors.

Figure 10: Cluster projections before Appling PCA

Figure 11 shows the cluster projections attribute data points with respect to Mean pitch, Minimum pitch, Maximum pitch, Pulse mean period and Fraction of locally unvoiced frames after applying the PCA. The blue elements show non-PD and the red elements show the PD data points. Most of diseased patients have high Fraction of locally_unvoiced_frames. These elements are under the hierarchical clusters, 0 and 1.

Figure 11: Cluster Projections after Applying PCA

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. CONCLUSION

PD is aging and second neurological disease in the world. Identification of PD with voice is very important and less cost process. In this work, we focused on statistical analysis on hidden values of the PD and no-PD voice. We have compared the PD and healthy people voice and measures acoustic errors of the voice. The good unsupervised ML algorithms like K-Means and hierarchy clusters show the good results with projection values for identification of related voice attributes of PD. The PCA projection values show very accurate than normal projection values. It shows high factored attribute values related to PD and non-PD voice data points. Further, we will implement classification algorithms with high quality attributes for predicting the PD

VI. EDITORIAL POLICY

The submitting author is responsible for obtaining agreement of all coauthors and any consent required from sponsors before submitting a paper. It is the obligation of the authors to cite relevant prior work.

Authors of rejected papers may revise and resubmit them to the journal again.

REFERENCES

1. Aich, S., Younga, K., Hui, K. L., Al-Absi, A. A., & Sain, M. (2018, February). A nonlinear decision tree based classification approach to predict the Parkinson's disease using different feature sets of voice data. In 2018 20th International Conference on Advanced Communication Technology (ICACT)(pp. 638-642). IEEE.
2. Naranjo, L., Pérez, C. J., Campos-Roca, Y., & Martín, J. (2016). Addressing voice recording replications for Parkinson's disease detection. Expert Systems with Applications, 46, 286-292.
3. Gürüler, H. (2017). A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method. Neural Computing and Applications,
4. 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,
5. Gök, M. (2015). An ensemble of k-nearest neighbours algorithm for detection of Parkinson's disease. International Journal of Systems Science, 46(6), 1108-1112.
6. Rustempasic, I., & Can, M. (2013). Diagnosis of parkinson's disease using fuzzy c-means clustering and pattern recognition. Southeast Europe Journal of Soft Computing, 2(1).http://dx.doi.org/10.21533/scjournal.v2i1.44
7. Polat, K. (2012). Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering. International Journal of Systems Science, 43(4), 597-609.
8. Karimi Rouzbahani, H., & Daliri, M. R. (2011). Diagnosis of Parkinson's disease in human using voice signals. Basic and Clinical Neuroscience, 2(3), 12-20.http://bcn.iums.ac.ir/article-1-96-en.html
9. van Rooden, S. M., Heiser, W. J., Kok, J. N., Verbaan, D., van Hilten, J. J., & Marinus, J. (2010). The identification of Parkinson's disease subtypes using cluster analysis: a systematic review. Movement disorders, 25(8), 969-978.
10. J. Holmes, R., M. Oates, J., J. Phyland, D., & J. Hughes, A. (2000). Voice characteristics in the progression of Parkinson's disease. International Journal of Language & Communication Disorders, 35(3), 407-418.
11. De Angelis, E. C., Mourao, L. F., Ferraz, H. B., Behlau, M. S., Pontes, P. A. L., & Andrade, L. A. F. (1997). Effect of voice rehabilitation on oral communication of Parkinson's disease patients. Acta Neurologica Scandinavica, 96(4), 199-205.
12. Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304.
13. Guo, P. F., Bhattacharya, P., & Kharma, N. (2010, June). Advances in detecting Parkinson's disease. In International Conference on Medical Biometrics (pp. 306-314). Springer, Berlin, Heidelberg.
14. Jia, C., Zuo, Y., & Zou, Q. (2018). O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. Bioinformatics, 34(12), 2029-2036.
15. Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., & Mathieu, C. (2018). Hierarchical clustering: Objective functions and algorithms. In Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 378-397). Society for Industrial and Applied Mathematics.
16. Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., & Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. Trends in Food Science & Technology, 72, 83-90.
17. Cao, J., Liang, M., Li, Y., Chen, J., Li, H., Liu, R. W., & Liu, J. (2018, March). PCA-based hierarchical clustering of AIS trajectories with automatic extraction of clusters. In 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA) (pp. 448-452). IEEE.

AUTHORS PROFILE

Dr. PanduRanga Vital Terlapu pursed Bachelor of Science in Computer Science from Andhra University of A.P, India in 1995 and Master of computer Application from Andhra University in year 1998. He completed his M. Tech in Computer Science and Engineering from Acharya Nagarjuna University of A.P, India and he completed his Ph.D in Computer Science and Engineering from GITAM University of A.P, India. He has 19 years of teaching and 13 years of research experience. He is currently working as Associate Professor in Department of Computer Science and Engineering, Aditya Institute of Technology and Management (AITAM), India. He is a member of ACM, Life Time Membership from International Computer Science and Engineering Society (ICSES), USA and Life Time Membership from Indian Society for Technical Education (ISTE), New Delhi, India. He has published more than 30 research papers in reputed international journals including SCOPUS indexed and a conference including Sringer, Elsevier and it's also available online. He is reviewer of reputed journals like Springer, Elsevier and IEEE. His main research work focuses on Machine Learning, Deep Learning and Data Mining, Data and Big Data Analytics, IoT and Computational Intelligence, Voice Analysis and Voice Processing, Bioinformatics.

Ms Pidugu shiny is pursuing IV Year B.Tech Computer Science and Engineering, Aditya Institute of Technology and Management from Javaharlal Nehru University, Kakinada, India. She is member of CSI. She completed more than 6 real-time major and minor projects. She selected as Software Engineer in CTS (MNC Software Company). She attended more than 8 workshops and conferences on ML and DL. Her interested topics are Machine Learning, Deep Learning and Data Mining and Voice Analysis and Voice recognization.

Mr. S. E. Ashish is pursuing IV Year B.Tech Computer Science and Engineering, Aditya Institute of Technology and Management from Javaharlal Nehru University, Kakinada, India. He is member of CSI. He completed more than 3 real-time major and minor projects on ML and DL. He selected as Software Engineer in DQ Solutions. His interested topics are Data Analytics, Machine Learning and Voice Analysis.

T. Sai Kumar is pursuing IV Year B.Tech Computer Science and Engineering, Aditya Institute of Technology and Management from Javaharlal Nehru University, Kakinada, India. She is member of CSI. He completed more than 3 real-time major and minor projects on ML. His interested topics are Big Data Analytics, Machine Learning and Voice Analysis.