GS International Conference on Computer Science and Engineering 2020 (GSIS 2020), AUGUST 2, 2020

Voice signal data clustering of Parkinson disease using un-supervised learning

M.Shakeel khan [1st]

Department of Computer Science
Riphah international university Lahore
(HEC)
17mcs1562@gmail.com

Salma yousaf [2nd]

Department of Computer Science
Riphah international university Lahore
(HEC)
Salma.yousaf42@gmail.com

Hira jamil [3rd], Haiqa mansoor [4th]
Department of Computer Science
Riphah international university Lahore
(HEC)
jamilhira09@gmail.com

haiqamansoor1996@gmail.com

Abstract**—— A major public health issue is Parkinson's disease. We attempted to distinguish between healthy individuals and individuals with Parkinson's disease in this study. The most prevalent nerve disorder is Parkinson's disease (PD), which is related to symptoms of the disease and coral instability. Various reviews have revealed that sound is one of the first indicators of PD, and therefore, the Parkinson's database contains what is used to sound the human being's medically. The main purpose of this document is to automatically detect the voice signal frequencies/speech feature in relation and more effective voice feature and checking the highest ratio/value of vice signal those who are affected the human brain and cause of the disease. The techniques in which using the un-supervised learning using the K-means algorithms and get the clustered the related voice feature and analyzed him and checking the graph comparison using visualization arrays of the values so this analysis is helping the bio-medical field for better understanding the voice signal of the Parkinson disease. its very useful in future of the treatment and diagnosis of the medial field era.**

Keywords—un-supervised learning, k-means clustering visualization graphs.

I.INTRODUCTION

Parkinson's disease (DP) that affects neurons in the brain that cause dopamine, which includes muscle endurance, movement, and head and language changes. Parkinson's disease affects the sounds of men, which affects them in a nervous way in the conversation. DC is second in nerve diseases for Alzheimer's disease. An increase is expected. In the coming years, practical treatment needs to be tested and detected. A good ratio product for the most part since a field indicator, using the pan-assessment is an essential part of the analysis of older disease control. This machine is used to learn and sort algorithms to predict and study Parkinson's disease [4]. Ideal database features are moved as a partnership for the model, and predictions are obtained results [5].

Performance prediction is a range of glowing cells in the average brain that are responsible for the generation of neurotransmitter dopamine. [2] This piece of brain plays a vital role in controlling movement [5] and also appears to communicate in the way of dependency. When you see a brain in the motion section, the material is placed in the center of the brain, as well as at the top of the brain [8] we have also completed the analysis of the problem, which can be accomplished with final and quantitative results. This article is organized as follows. Parkinson's data parameters include conversations in machine-larning algorithms, interpretations of their results, and compatibility of future monitoring applications. [15]..

II.UN-SUPERVISED LEARNING

The A type of offensive learning machine is used to attract infra-charged data from the learning algorithm label Answer Input Data made of data. The most common lyse learning method is cluster analysis, which is used for analyzing motivational data to find hidden pattern or cluster in data. Unwanted Learning Machine is a type of learning that searches for already undetected pattern in statistics without pre-existing labels and with minimal human supervision. Unlike monitoring learning, which is commonly known as human label statistics, undesirable learning, self-organization, can be used to model the probability density on input. This machine is one of three main types of learning, as well as monitoring learning and learning. Semi-protected learning, a related type, uses surveillance and offensive techniques.

Two of the key methods used in learning to be offensive are analysis of key components and clusters. Cluster analysis group is used in learning to dislike or to divide the set of data with shared rows to the linear algorithmic relationships. Cluster analysis is a branch of machine learning that has not been labeled data, rating or rating. Instead of answering feedback, cluster analysis identifies the shares in the data and is based on the presence or absence of such shares in each new data sheet. This approach detects the points of extraordinary statistics that do not fit in any group. So in which I'm using K medium-grade algorithms for pleasure clustering.

A. K-mens Clustering

k-means clustering is a vector method of quantification that causes signal processing that aims to separate n-k cluster observations in which each observation belongs to the cluster with the closest cluster centers or cluster center) serving as a cluster prototype. The result is a share of the data space in Verona cells. It is popular for cluster analysis in data retrieval. k-means that clustering minimizes cluster deviations (equal square distances) but not regular distances of Euclid, which would be Weber's most difficult problem: average square error optimization, while only the geometric median reduces equal distances. For example, better easily solutions can be found using k-medians and k-methods

The problem is difficult from a computer point of view (NP-difficult; however, effective heuristic algorithms quickly converge to local optimism. They are usually similar to the algorithm to maximize expectations for Gaussian distribution mixes through an iterative approach to refinement used by both k-environments and Gausic mixing modeling. while maximizing expectations allows clusters to have different shapes.

The algorithm has a loose link to the nearest neighboring classifier, a popular machine learning classification technique that is often confused with K-environments because of the name. Applying the nearby 1 nearest classifier to cluster centers obtained by k-means classifies new data in existing clusters. This is known as the closest-centric classifier or Roclio algorithms

III.LITERATURE REVIEW

(Indira Rustempasic, 2015) Parkinson's disease is a huge global health problem. The main purpose of this document is to automatically determine whether a person's speech/voice is influenced by the police. We reviewed the performance of the Fiji C Middle Group (FCM) and model identification methods on Parkinson's disease database. The first method is the main purpose of the performance of the distinction between the two classes, when you try to differentiate between the people who talk and the speakers. This method can be greatly improved by first data rating and then testing new data using both models. That way, the second method here is to identify the used motives. Experimental results show that the combination of the Fiji Model methodology and the recognition of the base has been promised for DB classification.

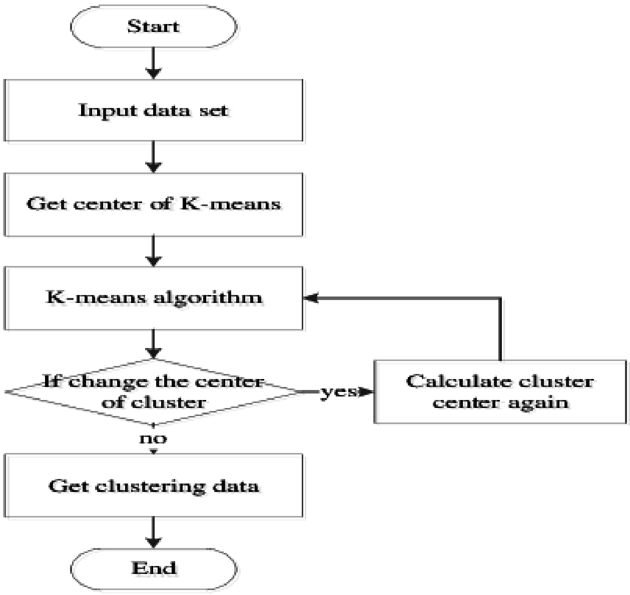
(Gu'ru' 1, 2016) In this study, we developed a new cyber-diagnostic system to solve the problem of PD diagnosis. The main innovation of this article is the proposed approach, which includes a set of happy-based k-(KMCFW WP) and an artificial nervous network of complex value (CVANN). A Parkinson's database which was used to diagnose THE PD sedition features derived from speech and sound samples. THE FEATURES of THE PAD are weighted using the KMCFW methodology. New features are converted into complex numbers. This functionality values are presented as an entry for CVANN. The performance and impact of the proposed system was assessed against the severity database using five different diagnostic methods. Experimental results show that the proposed cybersystem, called KMCFW-CVANN, significantly demonstrated other methods detailed in literature and reported for the highest date of achieving the highest rating results, with a rating accuracy of 99.52%. Therefore, a more accurate dnayagnasus of the proposed system OF PD is promised.

(In 2012) Wide data base collection and retrieval are a core area of bioinformatics, collecting new valuable knowledge in the medical world, where medical diagnosis is laborious and very important. The most prevalent nerve disorder is Parkinson's disease (PD), which is related to symptoms of the disease and coral instability. The series of time was achieved using the Raysha Signal as well as the first Xbox™ and Wii devices™ and was later analyzed using a linear and unusual analysis devices obtained from the device. The data obtained was classified using K-Model Pleasure and other data-mineing algorithms. This article highlights the importance of specific parameters of non-linear dynamics in Parkinson's and analysis. By analyzing the flaws and setbacks for patients with PDs, this paper provides insight into the most important information used for the specialist PD screening system.

(Timothy J. Wroge1, 2019) Biomarkeri made of human voice can provide insight into neurological disorders, such as Parkinson's (PD) diseases, because of their thorough cognitive and neurological functions. The PD is a progressive neurodegenerational disease that affect about 1 million people in the United States, with approximately 60000 new clinical diagnosis each year. We provide evidence to confirm this concept here using a set of voice data collected from people with and without. This article explores the effectiveness of using a monitored algorithm for classification, such as depth networks, to accurately diagnose people with disease. Our maximum accuracy of 85% of the learning machines exceeds the average accuracy of the clinical diagnosis of non-experts (73.8%) And average precision of the motion disorders experts (79.6% unsealed, 83.9% after the pathological examination of the autopsy was performed as the basic truth

IV.METHODOLOGY

In this article in which I'm using the Parkinson disease voice signal dataset in this datasets which the patients voice signal so I'm analysis of the voice and using k-means clustering algorithms and classifies the datasets features and purpose of the bio medical field so now we see the datasets feature and used for the purpose now showing below the datasets.



1) *Datasets.*

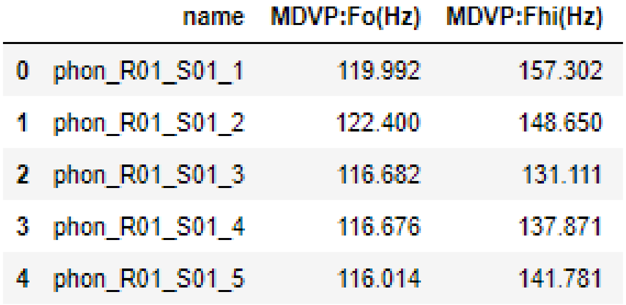
This In the database which, I use this database to function the voice signal and use case analysis and analysis. Now the database displays in the tables form. For better understanding and in the group of attributes and voice frequency. The database used in this study, including speech samples, was created a little bit more than the collaboration of the National Sound and Speech Center at the University of Colorado and Oxford University. It was obtained from UCI (Machine Larning Depot) The database consists of 195 clinical health-related sound measurements with 8 healthy subjects and 23 DDs.

TABLE I. EXTRACTED FEATURES FROM SPEECH RECORDINGS

Feature	Group
Shimmer (dda) Shimmer (local) Shimmer (apq3) Shimmer (apq11) Shimmer (apq5) Shimmer (local,dB)	**Amplitude Parameters**
Number of pulses Mean period Number of periods Standard deviation of period	**Pulse Parameters**
Jitter (ddp) Jitter (local) Jitter (rap) Jitter (local, absolute) Jitter (ppq5)	**Frequency Parameters**
Number of voice breaks Fraction of locally unvoiced frames Degree of voice breaks	**Voicing Parameters**
Mean pitch Median pitch Standard Deviation Maximum pitch Minimum pitch	**Pitch Parameters**
Harmonic-to-Noise Noise-to-Harmonic Autocorrelation	**Harmonicity Parameters**

2) *K-means clustering*

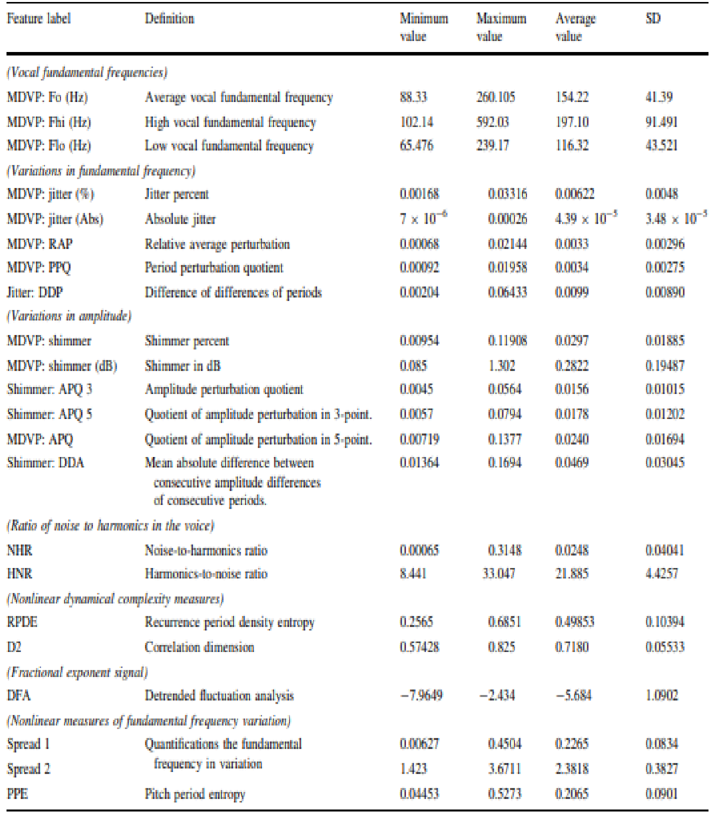
This algorithms is a clustering algorithms in which making the vector in cluster groups for classified the datasets feature for the medical diagnosis and easy to understand for the voice signal frequencies so now in showing the my clustering results and clearly for understanding the feature of the datasets .



	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)
0	phon_R01_S01_1	119.992	157.302
1	phon_R01_S01_2	122.400	148.650
2	phon_R01_S01_3	116.682	131.111
3	phon_R01_S01_4	116.676	137.871
4	phon_R01_S01_5	116.014	141.781

Figure 1: This voice signal is extract the k-means for clustering analysis and setting the arrays values.

Now Parkinson's Disease Min or Max's voice signal function and diagnosis are showing the average quality. Clustering algorithms are used not only for classification, but also for data compression, property weight and data reduction. Grouping is the most common choice method for use frequency

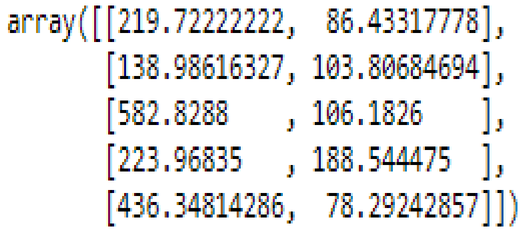


Feature label	Definition	Minimum value	Maximum value	Average value	SD
(Vocal fundamental frequencies)					
MDVP: Fo (Hz)	Average vocal fundamental frequency	88.33	260.105	154.22	41.39
MDVP: Fhi (Hz)	High vocal fundamental frequency	102.14	592.03	197.10	91.491
MDVP: Flo (Hz)	Low vocal fundamental frequency	65.476	239.17	116.32	43.521
(Variations in fundamental frequency)					
MDVP: jitter (%)	Jitter percent	0.00168	0.03316	0.00622	0.0048
MDVP: jitter (Abs)	Absolute jitter	7×10^{-6}	0.00026	4.39×10^{-5}	3.48×10^{-5}
MDVP: RAP	Relative average perturbation	0.00068	0.02144	0.0033	0.00296
MDVP: PPQ	Period perturbation quotient	0.00092	0.01958	0.0034	0.00275
Jitter: DDP	Difference of differences of periods	0.00204	0.06433	0.0099	0.00890
(Variations in amplitude)					
MDVP: shimmer	Shimmer percent	0.00954	0.11908	0.0297	0.01885
MDVP: shimmer (dB)	Shimmer in dB	0.085	1.302	0.2822	0.19487
Shimmer: APQ 3	Amplitude perturbation quotient	0.0045	0.0564	0.0156	0.01015
Shimmer: APQ 5	Quotient of amplitude perturbation in 3-point.	0.0057	0.0794	0.0178	0.01202
MDVP: APQ	Quotient of amplitude perturbation in 5-point.	0.00719	0.1377	0.0240	0.01694
Shimmer: DDA	Mean absolute difference between consecutive amplitude differences of consecutive periods.	0.01364	0.1694	0.0469	0.03045
(Ratio of noise to harmonics in the voice)					
NHR	Noise-to-harmonics ratio	0.00065	0.3148	0.0248	0.04041
HNR	Harmonics-to-noise ratio	8.441	33.047	21.885	4.4257
(Nonlinear dynamical complexity measures)					
RPDE	Recurrence period density entropy	0.2565	0.6851	0.49853	0.10394
D2	Correlation dimension	0.57428	0.825	0.7180	0.05533
(Fractional exponent signal)					
DFA	Detrended fluctuation analysis	−7.9649	−2.434	−5.684	1.0902
(Nonlinear measures of fundamental frequency variation)					
Spread 1	Quantifications the fundamental frequency in variation	0.00627	0.4504	0.2265	0.0834
Spread 2		1.423	3.6711	2.3818	0.3827
PPE	Pitch period entropy	0.04453	0.5273	0.2065	0.0901

Table 2:statistical values of the Parkinson disease datasets

a) *Cluster arrays values*

This is the clustering array values of the 2-voice signals (MDVP:FLO(HZ) and MDVP:Fhi(HZ)) this no of arrys in which showing the relation of the clause and in 2 columns forms.



```
array([[219.72222222,  86.43317778],
       [138.98616327, 103.80684694],
       [582.8288    , 106.1826    ],
       [223.96835   , 188.544475  ],
       [436.34814286,  78.29242857]])
```

Figure 3:

b) Cluster plotting

In this results I’m showing the tow different cluster in the plot and assign the different colors for the understanding the pd dataset voice signals. in which assign and making the 5 different cluster 0 cluster 1 cluster 2 cluster 3 cluster 4 are assigning the colors and now clearly showing in which highest cluster value of the orange color and after blue cluster is less value of the other cluster is low frequencies of effecting the disease.

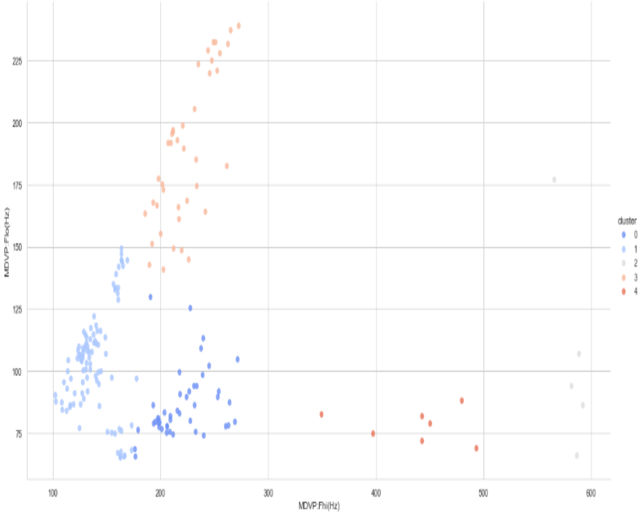


Figure 4

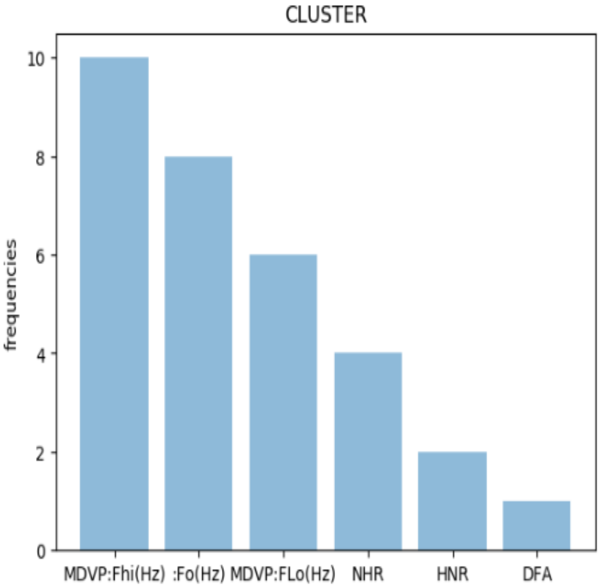


Figure 5: Is the bar graph in which in putting 6-cluster of voice signal feature so the now you clearly visible and view of the graph in which the highest signal of the Fhi(HZ) and 2nd is the Fo(HZ) and 3rd is Flo(HZ) is highest values of the other signals feature so this graph making the six cluster and showing the highest value and frequencies of my extracting the feature for analysis .

V. RESULTS

Parkinson's disease (PD) is the most prevalent neurological disease synonymous with signs of shaking and postural dysfunction. Various experiments have found that speech is one of the early markers of PD and Parkinson's dataset containing human biomedical voice is also used. In the k-means clustering in which I specifically display the cluster in plot those that make the heist ratio of the 5 different colours of frequencies in which the red colour of frequencies is heist ratio, and remove the data sets function and make the deferential classes and show the comprehension tables. In the table 2 in which clearly the showing the 1st highest vocal frequencies is MDVP: FLO(HZ) in which mini-values 88.33 nad max values is 260.105 and std value is 41.39 and average value is the 154.22 and the 2nd frequencies is the MDVP Fhi(HZ) in which the max-value is 592.03 and min-value 102 .14 and average value is 197.10 and std values is 91.49 this two frequencies I’m making for extract and getting the arrays values of comparison.

VI. CONCLUSION/DISCUSSION

In this paper study is based on the Parkinson disease voice frequencies to extract and getting the some feature of the ranges values and checking the different cluster and ratio point in the plotting so in which im using the cluster techniques and making the relative groups of the frequency and telling the purpose of the voice signal and how the values I effect on the parkinsonism and in which we are checking and classified the data using k-means algorithm with making the tables and showing the grou[ps of the datasets and table 2 in which in showing the max , min, average , and standard values of the voice frequencies' for the getting the highest value s those are effected and if the value and ratio is high thins Parkinson disease is effected on the brain and at the end the results in showing the clustering in the bar graph ratio point of the view so this research is very help full for the medical field and the doctor is easily to understand and describe the values and monitor for diagnosis the pd is the 2nd Alzheimer disease those are effected on the mid brains.

FUTURE WORK

In future of the research and this method and techniques is more effective for the bio-medical and cardiology center in diagnosis of the Parkinson disease for better understanding view of the treatments.

REFERENCES

[1] 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).

[2] 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, 28(7), 1657-1666.

[3] Pohoaţă, S., Geman, O., & Graur, A. (2012). *Dual tasking: gait and tremor in Parkinson's disease–acquisition,* processing and clustering. In Proc. of the National Symposium of Theoretical Electrical Engineering, SNET.

[4] 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.

[5] Frid, A., Safra, E. J., Hazan, H., Lokey, L. L., Hilu, D., Manevitz, L., ... & Sapir, S. (2014, June). *Computational diagnosis of Parkinson's Disease directly from natural speech using machine learning techniques.* In 2014 IEEE International Conference on Software Science, Technology and Engineering (pp. 50-53). IEEE.

[6] Hazan, H., Hilu, D., Manevitz, L., Ramig, L. O., & Sapir, S. (2012, November). Early diagnosis of Parkinson's disease via machine learning on speech data. *In 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel (pp. 1-4). IEEE.*

[7] Cantürk, İ., & Karabiber, F. (2016). *A machine learning system for the diagnosis of Parkinson's disease from* speech signals and its application to multiple speech signal types. Arabian Journal for Science and Engineering, 41(12), 5049-5059.

[8] Tsanas, A. (2012). *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear* speech signal processing and statistical machine learning (Doctoral dissertation, Oxford University, UK).

[9] Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). *Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease*. IEEE transactions on biomedical engineering, 59(5), 1264-1271.

[10] Shahbakhi, M., Far, D. T., & Tahami, E. (2014). Speech *analysis for diagnosis of parkinson's disease using genetic algorithm and support vector machine*. Journal of Biomedical Science and Engineering, 2014.