PREDICTION OF PARKINSON DISEASE USING KNN ALGORITHM.

Mr. Indrajit Bandisode, Mr. Rushabh Bhanushali, Mr. Vineet Singh, Mr.Vishal Singh.
Mrs. Ashwini Deshmukh
Mahavir Education Trust Chowk, W.T Patil Marg, Next To Duke's Company, Chembur, Mumbai, Maharashtra 400088.
I.T Department

Abstract:- *Parkinson disease(PD) is a progressive neurodegenerative disorder that has no known cure and no known prevention. Parkinson's disease is primarily caused by low and falling dopamine levels. A person with Parkinson's has abnormally low dopamine levels. Dopamine-generating cells, known as dopaminergic neurons (types of nerve cells) in the midbrain reticular part of the brain have died. Millions of people mainly above the age of 30 are getting affected by this Terminal illness. Early detection of the disease can help to slow down its progression rate. Symptoms of the Parkinson's are tremors, rigidity of muscles, slow slurred speech, and vocal cord disorder which starts early and is easy to mark. The voice feature dataset is used for research purpose. we have used Jaudio for voice recognition and feature extraction and KNN algorithm for analytics and prediction purpose.*

Keywords:- *KNN: K-nearest Neighbour, PD: Parkinson's Disease, SVM: Support Vector Machine, CART:Classification And Regression Tree, MIR: Music information retrieval, DFA: Detrended fluctuation analysis, HNR: Harmonics-to-Noise Ratio NHR: Noise-to-Harmonics Ratio RPDE: recurrence pitch entropy density .*

I. INTRODUCTION:

In past few years a lot of research has been going on the Parkinson's disease because the healthcare related cost due to this disease is keeping on increasing as the longevity of the population is increasing in the developed countries[1]. Parkinson's disease (PD) is a degenerative disorder of the central nervous system which mainly affects the motor system. The Disease is main seen in the people of older age. Parkinson's signs and symptoms may include:

Tremor : A tremor, or shaking, usually begins in a limb, often your hand or fingers. You may a rub your thumb and forefinger back and forth, known as a pill-rolling tremor. Your hand may tremor when it's at rest.

Slowed movement : Over time, Parkinson's disease may slow your movement, making simple tasks difficult and time consuming. Your steps may become shorter when you walk. It may be difficult to get out of a chair. You may drag your feet as you try to walk.

Speech changes(dysphonia) : You may speak softly, quickly, slur or hesitate before talking. Your speech may be more of a monotone rather than with the usual inflections.

Rigid muscles : Muscle stiffness may occur in any part of your body. The stiff muscles can be painful and limit your range of motion.

Impaired posture and balance : Your posture may become stooped, or you may have balance problems as a result of Parkinson's disease.

Loss of automatic movements : You may have a decreased ability to perform unconscious movements, including blinking, smiling or swinging your arms when you walk.

Writing changes : It may become hard to write and your writing may appear small.

The Parkinson's disease is progressive in nature. As the disease progresses more than 90% of the patients has the speech disorder [2].

II. LITERATURE REVIEW:

Aich, Satyabrata proposed a new approach by comparing the performance metrics with different feature sets such as original feature sets as well as Principal component Analysis based feature reduction technique for selecting the feature sets. They have used non-linear based classification approach to compare the performance metrics. They have found an accuracy of 96.83% using random forest classifiers using PCA based feature sets. This analysis will help the clinicians to differentiate the PD group from healthy group based on the voice data[1].

Chandrayan, Spriha proposed a factor analysing System to select meaningful and dominant features from the speech signals, which are relevant for prediction of Parkinson's disease. They infer that along with the jitter variants, shimmer variants and noise to harmonic ratio, pitch period entropy (PPE), the recurrence period density entropy (RPDE), and spread parameters are important in identifying PD. For classification, Support Vector Machine (SVM) is used. The proposed model discriminates Parkinson afflicted individuals from healthy ones with an average accuracy, sensitivity and specificity of about 90%[2].

In this paper, they have predict if a person has Parkinson's disease using voice recording dataset of patients by using data mining algorithm decision tree (CART). The voice of the patients are recorded and is converted into voice attributes like jitter, shimmer, frequency by using PRAAT script. The voice recordings are tested to predict if a person has

Parkinson's disease and also to tell the condition of the disease[3].

Bourouhou have tried to apply three types of classifiers, k-near neighbour "k-NN", the Naive Bayes "NB" and support vector machines "SVM", on the same database to compare and to know which of the three classifiers will be the most efficient. The K-NN have a good detection performance with 70%, and high Support Vector Machines detection performance with an accuracy of 80% correct detection rate. But we remark that the Naïve Bayes has a low quality of detection performance 65%[4].

Benba, Achraf have used a PD dataset of 34 sustained vowel / a /, from 34 people including 17 PD patients. They then extracted from 1 to 20 coefficients of the Mel Frequency Cepstral Coefficients from each person. To extract the voiceprint from each voice sample, they compressed the frames by calculating their average value. For classification, they used Leave-One-Subject-Out validation-scheme along with the Support Vector Machines with its different types of kernels. The best classification accuracy achieved was 91.17% using the first 12 coefficients of the MFCC by Linear kernels SVM[5].

III. FEATURE ANALYSIS:

jAudio is a feature extraction system designed to meet the needs of MIR researchers by providing a collection of feature extraction algorithms bundled with both an easy-to use GUI and a commandline interface. The system accepts audio files as input and produce ExcelSheet.

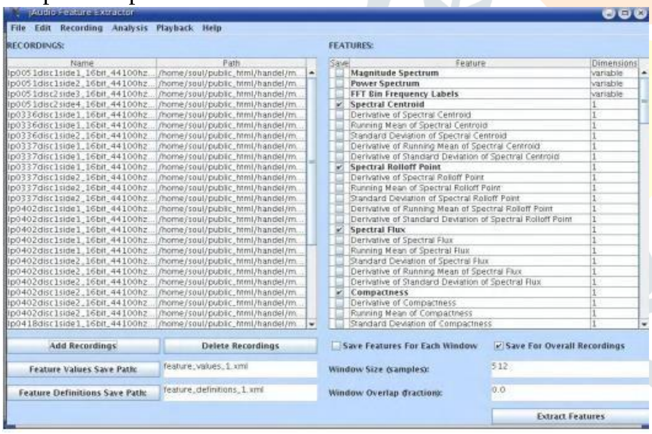


fig 1-jAudio Extractor

Extracting high-quality features is of critical importance for many MIR projects. Since these features are the only information that a classifier or other interpretive construct has about the original data, a failure in voice pattern recognition is not good. jAudio makes the tedious work for extracting the data of the voice features for the analytical purpose easy.

jAudio Characteristic

1) jAudio base

jAudio was written in Java in order to capitalize on Java's cross platform portability and design advantages. A custom low-level audio layer was implemented in order to supplement Java's limited core audio support and allow those writing jAudio features to deal directly with arrays of sample values rather than needing to concern themselves directly with low-level issues such as buffering and format conversions.

2) Intuitive interface

jAudio permits control of down sampling of the input signal, signal normalization, window size, window overlap, and control of which features are extracted and saved with an easy to use GUI (See Figure 1). The GUI also permits users to configure settings and save them for batch processing.

Support for multidimensional features jAudio has the capacity to accept features that provide an arbitrary number of dimensions. This is an extremely useful way to group related features calculated at once such as MFCC.

3) Handling dependencies

In order to reduce the complexity of calculations, it is often advantageous to reuse the results of an earlier calculation in other modules. jAudio provides a simple way for a feature class to declare which features it requires in order to be calculated. An example is the magnitude spectrum of a signal. It is used by a number of features, but only needs to be calculated once. Just before execution begins, jAudio reorders the execution of feature calculations such that every feature's calculation is executed only after all of its dependencies have been executed. Furthermore, unlike any other system, the user need not know the dependencies of the features selected. Any feature selected for output that has dependencies will automatically and silently calculate dependent features as needed without replication.

4) Support for multidimensional features

jAudio has the capacity to accept features that provide an arbitrary number of dimensions. This is an extremely useful way to group related features calculated at once such as MFCC. This is in contrast to the ARFF format from Weka where all features are unidimensional. Furthermore, the dimensionality of each feature is exported. This permits derivatives and other meta features to have the same number of dimensions as the feature they are calculated from, even though the same code is used for all features.

5) Meta features

Meta features are templates that can be applied against any feature to create new features. Examples of meta features include Derivative, Mean, and Standard Deviation. Each of these meta features are automatically applied to all features without the user needing to explicitly create these derivative features.

6) Extensibility

Effort was taken to make it as easy as possible to add new features and associated documentation to the system. An abstract class is provided that includes all the features needed to implement a feature.

IV. METHADOLOGY:

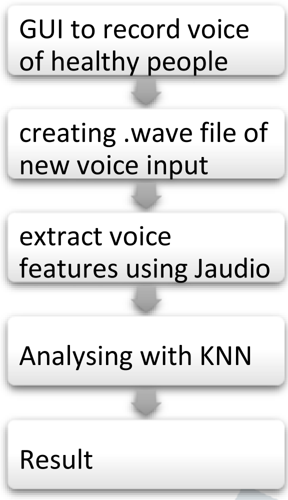
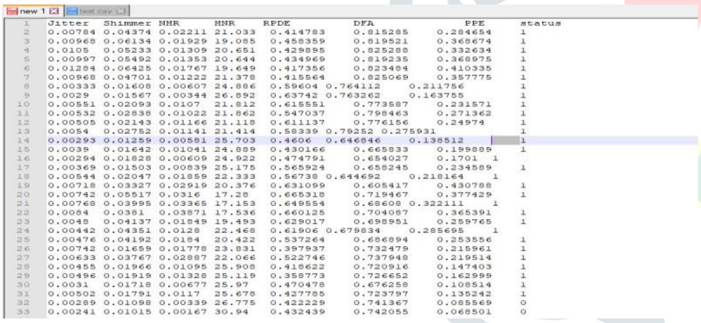


Fig 2-Workflow of project

We have build GUI using Java Eclipse IDE to record the voice of the user for analytical purpose. After the sound gets recorded a

.Wav file gets generated automatically and a test file also gets generated which consist of voice features (shitter, jitter, NHR etc) coefficient of the user voice whose prediction has to be made.



Jitter	Shimmer	NHR	HNR	RPDE	DFA	PPE	status
0.00784	0.04374	0.02211	21.033	0.414783	0.815285	0.284654	1
0.00968	0.06134	0.01929	19.085	0.458359	0.819521	0.368674	1
0.0105	0.05233	0.01309	20.651	0.429895	0.825288	0.332634	1
0.00997	0.05492	0.01353	20.644	0.434969	0.819235	0.368975	1
0.01284	0.06425	0.01767	19.649	0.417356	0.823484	0.410335	1
0.00968	0.04701	0.01222	21.378	0.415564	0.825069	0.357775	1
0.00333	0.01608	0.00607	24.886	0.59604	0.764112	0.211756	1
0.0029	0.01567	0.00344	26.892	0.63742	0.763262	0.163755	1
0.00551	0.02093	0.0107	21.812	0.615551	0.773587	0.231571	1
0.00532	0.02838	0.01022	21.862	0.547037	0.798463	0.271362	1
0.00505	0.02143	0.01166	21.118	0.611137	0.776156	0.24974	1
0.0054	0.02752	0.01141	21.414	0.58339	0.79252	0.275931	1
0.00293	0.01259	0.00581	25.703	0.4606	0.646846	0.138512	1
0.0039	0.01642	0.01041	24.889	0.430166	0.665833	0.199889	1
0.00294	0.01828	0.00609	24.922	0.474791	0.654027	0.1701	1
0.00369	0.01503	0.00839	25.175	0.565924	0.658245	0.234589	1
0.00544	0.02047	0.01859	22.333	0.56738	0.644692	0.218164	1
0.00718	0.03327	0.02919	20.376	0.631099	0.605417	0.430788	1
0.00742	0.05517	0.0316	17.28	0.665318	0.719467	0.377429	1
0.00768	0.03995	0.03365	17.153	0.649554	0.68608	0.322111	1
0.0084	0.0381	0.03871	17.536	0.660125	0.704087	0.365391	1
0.0048	0.04137	0.01849	19.493	0.629017	0.698951	0.259765	1
0.00442	0.04351	0.0128	22.468	0.61906	0.679834	0.285695	1
0.00476	0.04192	0.0184	20.422	0.537264	0.686894	0.253556	1
0.00742	0.01659	0.01778	23.831	0.397937	0.732479	0.215961	1
0.00633	0.03767	0.02887	22.066	0.522746	0.737948	0.219514	1
0.00455	0.01966	0.01095	25.908	0.418622	0.720916	0.147403	1
0.00496	0.01919	0.01328	25.119	0.358773	0.726652	0.162999	1
0.0031	0.01718	0.00677	25.97	0.470478	0.676258	0.108514	1
0.00502	0.01791	0.0117	25.678	0.427785	0.723797	0.135242	1
0.00289	0.01098	0.00339	26.775	0.422229	0.741367	0.085569	0
0.00241	0.01015	0.00167	30.94	0.432439	0.742055	0.068501	0

Fig 3-Dataset to train KNN Algorithm.

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD. The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column.

We have trained our KNN algorithm using the above mentioned Dataset.

Working of KNN algorithm:

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

Let's take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS) :

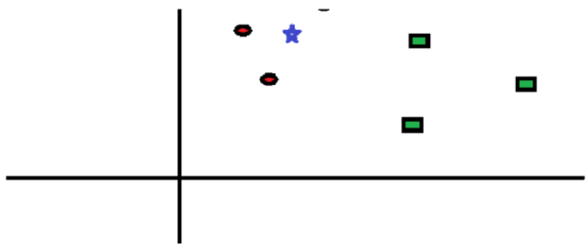


Fig 4-plotting of Data before KNN

You intend to find out the class of the blue star (BS) . BS can either be RC or GS and nothing else. The "K" is KNN algorithm is the nearest neighbours we wish to take vote from. Let's say K = 3. Hence, we will now make a circle with BS as centre just as big as to enclose only three datapoints on the plane. Refer to following diagram for more details:

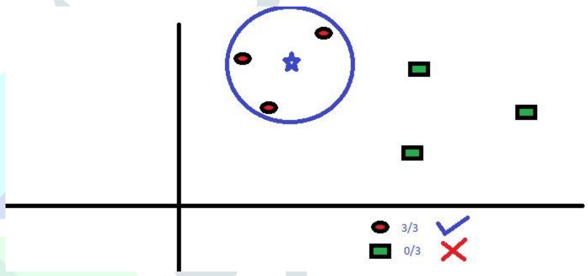


Fig 5- feature Similarity plotting of Data after KNN

The three closest points to BS is all RC. Hence, with good confidence level we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbour went to RC. The choice of the parameter K is very crucial in this algorithm. Next we will understand what are the factors to be considered to conclude the best K.

Pseudo code for KNN:

1) Load the data
2) Initialise the value of k
3)For getting the predicted class, iterate from 1 to total number of training data points
    3.1) Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.
    3.2) Sort the calculated distances in ascending order based on distance values
    3.3) Get top k rows from the sorted array
    3.4) Get the most frequent class of these rows
    3.5) Return the predicted class

V. CONCLUSION AND RESULT

Disorders characteristics of Parkinson's disease do not appear abruptly. It is a slow process whose first stages may go unnoticed. The extracted voice features(jitter, shimmer, HNR etc) from different participants contain many frames which take maximum processing time in the classification process, and prevent making correct diagnosis.

After analytical process of KNN algorithm is carried our system predicts and display where the user has Parkinson or not.

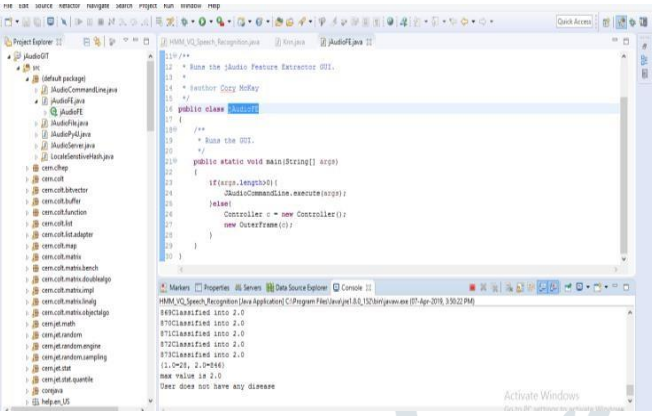


Fig 6-output for person with no Parkinson disease

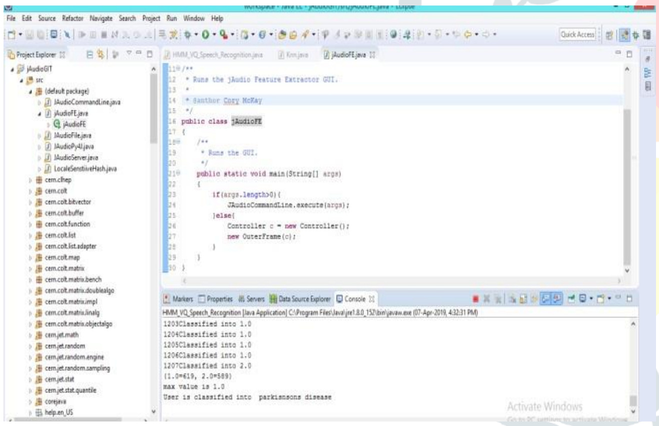


Fig 7-output of person may have Parkinson Disease

VI. REFERNCES

1) Aich, Satyabrata, et al. "A Nonlinear Decision Tree Based Classification Approach to Predict the Parkinson's Disease Using Different Feature Sets of Voice Data." 2018 20th International Conference on Advanced Communication Technology (ICACT), 201,.

2) Chandrayan, Spriha, et al. "Selection of Dominant Voice Features for Accurate Detection of Parkinson's Disease." 2017 Third International Conference on Bio signals, Images and Instrumentation (ICBSII), 2017.

3) Prediction of parkinson's diseae using data mining International Conference on Energy, Communication, Data Analytics and Soft ComputingYear: 2017.

4) Bourouhou, A., et al. "Comparison of Classification Methods to Detect the Parkinson Disease." 2016 International Conference on Electrical and Information Technologies (ICEIT), 2016.

5) Benba, Achraf, et al. "Voiceprints Analysis Using MFCC and SVM for Detecting Patients with Parkinson's Disease." 2015 International Conference on Electrical and Information Technologies (ICEIT), 2015.