Detection of Parkinson Disease Using Clinical Voice Data Mining

Saloni, R.K. Sharma., and A. K. Gupta

***Abstract*—**Parkinson disease is the second most common neurological disorder. Approximately 90% of people with Parkinson have speech disorders. In this paper we have classified the healthy people and Parkinson suffering people using data mining of voice features. Support vector machine is used as a classifier. The accuracy of the classifier depends on the voice features and their count. Various subsets can be prepared with the available voice features. An algorithm is proposed to select the best subset and as a result 100% accuracy is achieved. DFA (Detrended fluctuation analysis) and PPE (pitch period entropy) are the very significant features in this classification.

Keywords— Classification; Data mining; Parkinson disease; Support Vector Machine; Voice Features

I. Introduction

DATA mining have great potential in disease detection for the advancement of medical field. Data mining is basically a tool for converting the raw data into some very useful information. Data mining provides ways to extract information transform and present the data in a useful format. It is used widely in many applications [1]. Parkinson disease is a progressive neurodegenerative disease. It is caused by the death of dopamine neurons which conveys the message from the brain to the rest body. PD patients show the symptoms like poverty of movement, slowness of movements and rigidity. Diagnosis of Parkinson disease is a difficult process. At the early stages its symptoms resemble with other medical conditions. No laboratory tests are available for the detection. To exclude the other medical conditions blood tests, MRI (magnetic resonance image), PETscan (positron emission tomography), SPECT (single photon emission computed tomography) are done. With the advancement of signal processing experts use some discriminative measures from the voice of people for PD detection. Detection can be done at an early stage of disease because vocal cord disorder starts early. In the Parkinson disease patient's voice have some abnormal deviations and extra oscillations are present. Sometimes even patients cannot speak the correct vocal sound. Parkinson incidence increases with age and is slightly higher in men than women. The Parkinson detection using clinical voice data mining is very reliable, easy and economic [2], [3].

Max Little developed software for differentiating the healthy and Parkinson disease patients voices. A large amount of data is collected for this purpose. PPE (Pitch period entropy), a measure of dysophonia, and also robust to the unwanted effects has been introduced. He achieved accuracy of 91.4% with SVM classifier [4]. Multi-Layer Perceptron neural network and Support Vector Machine with linear and puk kernel function were used for classifying Parkinson data set with an accuracy of 90% [5]. In Artificial neural network method, 70% of data was used for training, and 30% for testing .Using this approach, 93.2% accuracy was achieved. The data set consist of twenty three features [6]. By using maximum-relevance-minimum-redundancy criteria, features are selected on the basis of mutual information measures between the features [7]. In some cases, twenty three attributes are reduced to sixteen and 83.3% accuracy is achieved [8]. Various features subsets can be prepared and the subset which gives maximum accuracy is selected [2]. Genetic algorithm is used for feature selection. In genetic algorithm, solutions are represented by chromosomes until acceptable results are obtained. Crossover and mutation process is done to get new chromosomes. With genetic algorithm for feature selection and support vector machine for classification, 94.5% accuracy is achieved [9]. When genetic algorithm with KNN (k- nearest neighbour) classification method is applied, 98.2% performance is obtained [10].

For feature selection a correlation filter is used. Fuzzy C means clustering and pattern recognition is applied on selected features for classifying normal speakers and PD speakers [11]. Classification results of healthy and PD speakers are equally significant for both male and female [3]. Among four classifier models in WEKA software, naive bayes simple, naive bayes, decision table, NNge the decision table attains the best accuracy as 96% [1]. Relief feature selection and random tree classification combination provides 100% accuracy [12]. Fisher score attribute selection method is used for detection of effective attributes and 91.28% is obtained with feed forward

Saloni is with the Electronics and communication department, National Institute of technology, Kurukshetra (e-mail: er.saloni83@gmail.com).

R. K. Sharma is with the Electronics and communication department, National Institute of technology, Kurukshetra (e-mail: mail2drrks@gmail.com).

A. K. Gupta is with the Electronics and communication department, National Institute of technology, Kurukshetra (e-mail: anilg699@gmail.com).

neural network [13]. Genetic programming and expectation-maximization algorithm is used to create a learning feature function which classifies the two different groups [14]. Sixteen features are extracted from the data set using student's t-test. Multilayer Perceptron network and radial basis function network are used for classification. RBF (Radial Basis Function) gives better results [15]. The Fuzzy K-nearest neighbour (FKNN) with principal component analysis to construct the most discriminative new feature set is used for Parkinson diagnosis [16].

In this paper, we have used the feature dataset of Parkinson disease. Feature selection and classification is used to classify healthy and pathological datasets. An optimum feature subset is selected and an accuracy of 100% is observed.

II. Data set

The dataset used in this paper was created by Max little of the University of Oxford with the NCVS (National centre for voice and speech) collaboration. The dataset consists of phonation from 23 Parkinson and 8 control subjects. The sustained phonation of vowel 'a' was recorded for duration of 36 seconds. Phonations were recorded for six times and total 195 samples were recorded with a head mounted microphone (AKG CS420) positioned 8cm from lips. Age of the subjects range from 46-85 years. The recordings are made at a sampling frequency of 44.1 KHz with 16 bit resolution. All samples were digitally normalized in amplitude before feature calculations. The dataset is divided into two classes according to its "status" column which is set to 0 for healthy subjects and 1 for those PD [4]. Fig 1 represents the speech signal of healthy and Parkinson persons.

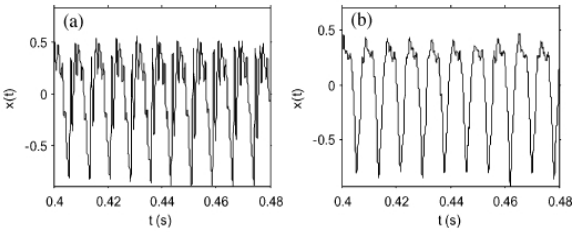

Fig.1. Two selected examples of speech signals. (a) Healthy. (b) Parkinson. The horizontal axis is time in seconds and the vertical axis is signal amplitude [4].

III. Feature Extraction

For all 195 phonations, features are extracted. For feature extraction only first half of the recordings are considered, because the second half of the recording is influenced by reduced lung pressure. Various traditional measures and nonstandard measures are extracted. A set of 22 features is prepared. Feature set consists of Fo(Hz), Fhi(Hz), Flo(Hz), Jitter(%), Jitter(Abs), RAP, PPQ, Jitter:DDP, Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, APQ , Shimmer:DDA, NHR, HNR, RPDE, DFA, spread1, spread2, D2, PPE.

The vocal fold vibration frequency is known as fundamental frequency. The perturbation in the frequency and amplitude in successive vocal fold cycles is termed as jitter and shimmer respectively. The noise-to-harmonic and harmonic-to-noise measures are measured using estimates of signal to noise by calculation of autocorrelation of each cycle. D2 is the correlation dimension between the signal and its first time delay embedded signal whereas the RPDE (recurrence period density) is the measure of periodicity of the reconstructed signal after embedding time delay. DFA (Detrended fluctuation analysis) is the log-log plot of the time scales L and amplitude variation F (L). Non linear measure of fundamental frequency variation is defined in terms of spread 1 and spread 2. The logarithmic scale of pitch sequence is explained as semitone pitch **p(t)** where **t** is the time. The entropy of relative semitone variation is known as pitch period entropy (PPE). All these parameters show variation for the healthy and parkinson's case. Next, features are selected among these to get best classification among the two groups.

IV. Feature selection

Feature are selected which have more separable values than others and a new feature data subset is prepared which contains 15 features as shown in Table 1 . Support vector machine classifier is used and their performance is evaluated. Classifier used is supervised classifiers and therefore dataset is divided into training and testing datasets.75% of the data is used for training purpose and rest 25% is for testing. Out of 195 observations, 146 are used for training (110 parkinson +36 healthy) and 49 are used for testing. Target data is also prepared. The classifiers that are used in this work are described below

Table 1. Details of 15 features of the PD data set.

Label	Attribute	Description
F1	MDVP:F0(Hz)	Average vocal fundamental frequency
F2	MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
F3	MDVP:Shimmer	Several measures of variation in amplitude
F4	MDVP:Shimmer(Hz)	
F5	Shimmer:APQ3	
F6	Shimmer:APQ5	
F7	MDVP:APQ	
F8	Shimmer:DDA	
F9	HNR	Measure of ratio of noise to tonal components in the voice
F10	NHR	
F11	RPDE	Nonlinear dynamical complexity measures
F12	D2	
F13	DFA	Signal fractal scaling exponent
F14	Spread2	Nonlinear measures of fundamental frequency
F15	PPE	

V. Support Vector Machine Classifier

Various classifiers are present which provide good results. ANN (artificial neural network), SVM (support vector machines), GMM (Gaussian Mixture Model) and HMM (Hidden Markov Model are mostly used in speech processing applications [17, 18, 19]. SVMs are set of related supervised learning methods used for classification and regression.

Support vector machines builds a model using set of training examples, each marked to its category and then used for classification. SVM simultaneously minimize the empirical classification error and maximize the geometric margin. SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. In all cases, a maximum margin hyperplane is selected which have the largest separation between the two classes. Maximum margin hyper plane is a plane from which distance to the nearest data point on both sides is maximized.

Classifiers performance can be compared using confusion matrix and some other parameters. The confusion matrix includes the following parameters. True positive (TP) term represent the Parkinson disease samples correctly classified and true negative (TN) represent healthy samples that are correctly classified. False negative (FN) means Parkinson samples classified as healthy and false positive (FP) means healthy sample as parkinson's sample.

Parameters are mathematically described as follows:

$$\text{Sensitivity}=\frac{TP}{TP+FN}*100 \quad (1)$$

$$\text{Specificity}=\frac{TN}{TN+FP}*100 \quad (2)$$

$$\text{Overall accuracy}=\frac{(TP+TN)}{(TP+TN+FP+FN)}*100 \quad (3)$$

Accuracy of the classifier depends on the feature data set that is fed to the classifier for classification during training and testing phases. We have selected 15 features out of 22 features by eliminating features which have little and no predictive information. Feature selection can appreciably improve the comprehensibility and lucidity of the resulting classifier [12].

To get highest accuracy in classifying healthy and pathological voice data base, an algorithm shown in Fig. 2 has been used. This algorithm tells what should be the feature subset size and which features should be used. All sizes of subsets from 1 to 15 are considered. Then all possible combinations of features for that particular subset size are taken. Support Vector Machine classifier is used and all the classifier performance parameters are calculated. Classifier accuracy varies with the size of the subset and also the features combination. This algorithm gives the highest accuracy, and therefore helps in selecting and choosing the feature combinations that most important to consider. The classification accuracy varies for different features combinations.

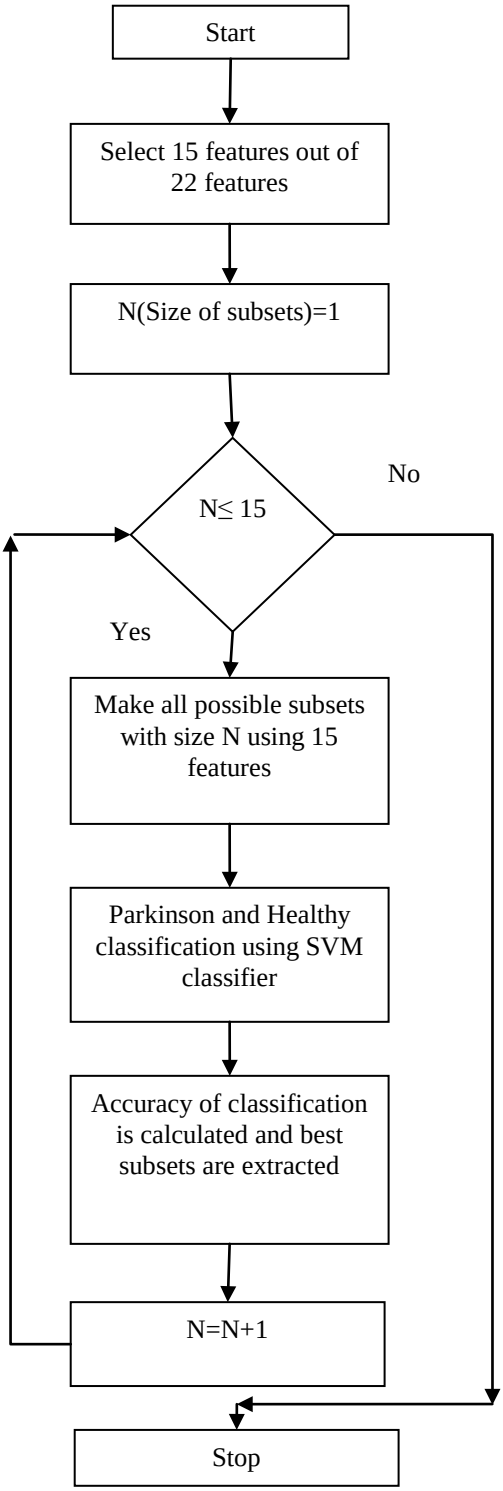

Fig. 2 Algorithm to get highest classification accuracy.

VI. Results

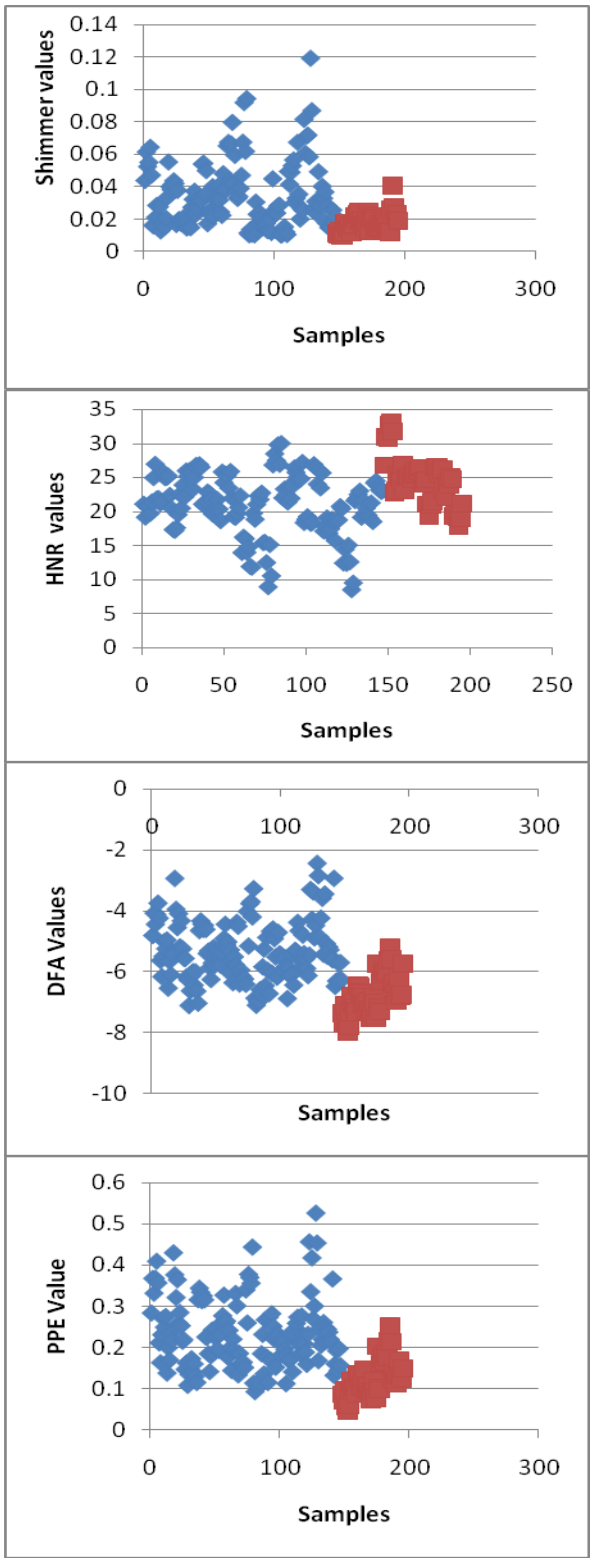

Fig. 3 Parameters value distribution
Parkinson people→ , Healthy people→

All selected 15 features are significant in classifying the healthy and Parkinson subjects. These features show different values for both cases. The vocal tremor and rigidity in voice is visible in parkison disease. Due to this, various traditional and non traditional features of voice get effected.The fundamental frequency gets reduced when voice of a person is suffering from Parkinson.vocal tremor is visible in jitter and shimmer values of the voice.Shimmer values are higher in parkinson subjects than the healthy one. The harmonic to noise ratio values are high for the healthy one. The non-traditional measures show appreciable differentiation between the two classes. All the three features DFA, Spread2 and PPE have higher values for parkinson subjects than the healthy one. All the 15 features are used in SVM; the classification accuracy comes out 94%. This accuracy got improved when we used subsets of this 15 features dataset. Table 2 shows the no of subsets of size ranging from 1 to 15 and their corresponding accuracies. 100% accuracy is achieved in a large no of cases. Using SVM classifier with subset size 7 or 8, we obtained maximum higher accuracy feature combination subsets. The complete information is presented in table. Comparision of different size subsets which gives 100% accuracy is shown in Fig 4. Different series represent the subset sizes. In three cases 100% accuracy is not achieved, when the sizes of the subsets are 1, 14 and 15. Here least accuracy of SVM classifier is 74%. Also, when large size subsets are used classifier accuracy ranges 90% to 100%. It shown in the Table 2 for subset size 12, 13 and 14 no entry is visible upper part of the table. 83.3% and 80.8% accuracy is achieved using artificial neural network using 16 and 22 features respectively [8]. Using genetic algorithm maximum accuracy is achieved when the number of features is 4 as compare to when number of features are 7 and 9 [10]. So the choice of proper size of subset is very important. The comparison of proposed work with the literature is shown in Table 3.

VII. Conclusion

Clinical data mining makes the diagnosis process computerized. So less human expertise is required. It becomes difficult for the Parkinson patients to visit the clinic again and again. In this process of diagnosis through voice analysis patients need not to visit the clinic. Decrease in count of dopamine neurons makes improper muscles movement and vocal parameters get changed from normal one. Parameters related to non linear measure of fundamental frequency are very significant.94% accuracy is achieved when all selected 15 features are taken. But when we reduce the no of features the 100% classification accuracy is achieved with various subsets. Voice features those expose the malfunctioning of nervous system are very efficient and improve the accuracy when included in the classifier subset. A proper subset size and features that included both things are very important to select properly to get good classification.

REFERENCES

[1] D. Narmadha, D. Marudhadevi and B. Santhi, "Parkinson Disease detection using soft computing" *World Applied Sciences Journal*, Vol. 29,pp 89-92, 2014.

[2] H. K. Rouzbahani and M. R. Daliri, "Diagnosis of Parkinson's Disease in Human using Voice Signals," *Neuroscience*, Vol 2, 2011, pp 12-19.

[3] N. Afza, M. Challa and J. Mungara, "Speech Processing algorithm for detection of Parkinson's disease"*International Journal of Engineering Research and Technology*, Vol 2, 2013, pp 1798-1803.

[4] Max A. Little, P. E. Macsharry, E. J. Hunter, J.Sielman, L. O. Raming, "Suitability Of Dysophonia Measurements for Telemonitoring of Parkinson's Disease, *IEEE Transaction on biomedical engg*,Vol. 56,2009, pp 1015-1022.

[5] A. David Gill and B. Magnus Johnson, "Diagnosing Parkinson by using Artficial Neural Network and Support Vector Machines," *Global Journal of Computer Science and technology*, vol. 9,2009, pp. 63-71 .

[6] F. S. Gharehchopogh and P. Mohammadi, " A Case Study of Parkinson Disease using Artificial Neural Network, *IJCA*, vol. 73, 2013, pp.1-6.

[7] C.O. sakar and O. Kursun, "Telediagnosis of Parkinson's Disease Using Measurements of Dysophonia, *J Med Syst*, Vol 34, 2010, pp. 591-599.

[8] A. Khemphila and V. Boonjing, "Parkinson Disease Classification Using Neural Network and Feature Selection" *WASET*, Vol 6, 2012, pp.15-18.

[9] M. Shahbakhi, D. T. Far and E. Tahami, " Speech Analysis for Diagnosis of Parkinson's Disease Using Genetic Algorithm and Support Vector Machine," *J. Biomedical Science and Engineering*, vol. 7, 2014, pp. 147-156.

[10] R. A. Shirvan and E. Tahami, "Voice Analysis for Detecting Parkiinson's Disease using Genetic Algorithm and KNN Classification Method"*Biomedical engg ICBME*, 2011,pp 278-283.

[11] I. Rustempasic, M. Can ,"Diagnosis of Parkinson's Disease using Fuzzy C-Means Clustering and Pattern Recognition" *Southeast Europe Journal of Soft Computing*, pp 42-49,Vol 2, 2013.

[12] R. Geetha Ramani, G. Sivagami, S. Gracia Jacob, "Feature Relevance Analyis and Classification of Parkinson Disease Tele-Monitoring Data Through Data Mining Techniques" *International Journal of Advanced Research in Computer Science and Software Engginering*, Vol 2, pp 298-304,2012.

[13] *I. Atacak, B. Gökpinar*, "A Computer –aided Diagnosis system for detection of parkinson disease using fisher score feature selection and neural network" *Global Journal on Technology*, Vol 4 ,2013,pp 639-643.

[14] P. Guo, P. Bhattacharya, N. Kharma, "Advances in Detecting Parkinson's Disease" *Medical Biometrics*, Vol. 6165, 2010, pp 306-314.

[15] U. Rani and M. Holi, "Analysis of Speech Characteristics of Neurological Disease and their Classification" Computing Communication & Networking Technologies (ICCCNT), 2012, pp 1-6.

[16] H. chen, "An efficient diagnosis system for detection of parkinson's disease using fuzzy k-nearest neighbor approach" *Expert System With Aplication*, Vol 40, 2013, pp 263-271.

[17] A. Benba, A. Jilbab, A. Hammouch, "Voiceprint analysis using Perceptual Linear Prediction and Support Vector Machines for detecting persons with Parkinson's disease" *Recent Advances in Biology, Biomedicine and Bioengineering*, pp 85-90.

[18] K. Daqrouq, A. Morfeq, M. Ajour and A.Alkhateeb, "Wavelet LPC with neural network for speaker identification system" WSEAS transaction on signal processing, Vol 9, 2013, pp 216-226.

[19] N. Sharma and H. om, "cascade correlation neural network model for classification of oral cancer" Vol 11, 2014,pp 45-51.

Saloni, received his M.Tech in Microelectronics and VLSI design from Kurukshetra University Kurukshetra. Currently, she is pursuing PhD at National Institute of technology in the department of Electronics and Communication. Her research interests include biomedical signal processing and VLSI design.

R.K.Sharma, received his M.Tech inelectronics and communication engineering and PhD degree in electronics and communication from Kurukshetra University Kurukshetra (through National Institute of Technology Kurukshetra), India in 1993 and 2007, respectively. Currently he is Professor with the Department of Electronics and Communication Engineering, NIT Kurukshetra, India. His main research interests are in the field of low power VLSI design, Voice profiling, Microprocessor and FPGA based systems.

A.K.Gupta received his M.Tech degree in electrical engineering and PhD degree in microelectronics from Indian Institute of Technology Kanpur, India in 1975 and 1987, respectively. He is Professor with the Department of Electronic and Communication Engineering NIT Kurukshetra, India. His main research interests are in the field of semiconductor device modeling, analog IC design, electron ic measurements and SOI.

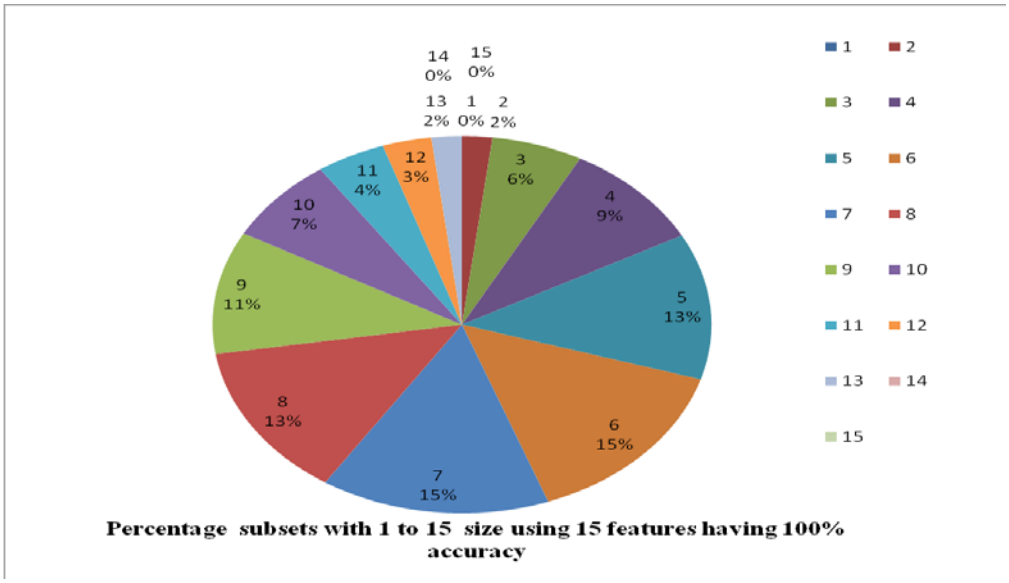

Fig.4. Percentage of count number of subsets with 1 to 15 in size having 100% accuracy

TABLE 2. Classification Accuracy using all subsets with size ranging from 1to 15

Size of subsets-→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Accuracy(%)	No of subsets having corresponding Accuracies														
74	10	36	90	126	116	61	23	7	1	0	0	0	0	0	0
76	0	5	9	34	92	148	112	42	6	1	0	0	0	0	0
78	2	5	13	33	69	94	136	136	89	35	9	1	0	0	0
80	0	6	23	71	132	204	220	164	80	20	2	0	0	0	0
82	0	10	49	119	188	205	144	59	13	2	0	0	0	0	0
84	0	2	16	50	92	105	95	71	43	17	5	1	0	0	0
86	1	4	13	48	104	149	143	135	85	45	15	3	0	0	0
88	0	2	17	41	91	135	153	122	84	40	15	5	1	0	0
90	0	0	4	22	64	139	227	267	250	182	88	24	3	0	0
92	0	0	4	29	85	208	376	495	477	345	199	87	24	3	0
94	0	0	5	60	213	464	717	826	733	523	297	130	41	9	1
96	0	1	29	104	299	607	859	903	740	437	177	47	10	1	0
98	2	32	157	505	1095	1773	2303	2382	1894	1142	499	143	24	2	0
100	0	2	26	123	363	712	927	826	509	214	59	14	2	0	0
Total no of subsets	15	105	455	1365	3003	5004	6435	6435	5004	3003	1365	455	105	15	1

Table. 3 Comparison table

Reference	[1]	[2]	[4]	[6]	[7]	[12]	[13]	[14]	[16]	Proposed work
Accuracy	96%	93.8%	91.4%	93.2%	92.7%	100%	91.2%	93.2%	96%	100%