

Disease Classification and Prediction using Ensemble Machine Learning Classification Algorithm



B.Meena Preethi, P.Radha

Abstract: In today's scenario, disease prediction plays an important role in medical field. Early detection of diseases is essential because of the fast food habits and life. In my previous study for predicting diseases using radiology test report, and to classify the disease as positive or negative three classifiers Naïve Bayes (NB), Support Vector Machine (SVM) and Modified Extreme Learning Machine (MELM) was used to increase the accuracy of results. To increase the efficiency of predicting the disease and to find which disease pricks the society, ensemble machine learning algorithm is used. The huge data from the healthcare industry were preprocessed, categorized and analyzed to find out and predict which patient to be treated and given priority and which hits the society the most. Ensemble machine learning's popularity in the medical industry is due to a variety of factors the Classifiers used are K Nearest Neighbors, Nearest Mean Classifier, Mean Feature Voting Classifier, KDtree KNN, Random Forest. To reduce the manual processes in medical field automating these processes has become important. Electronic medical records and significant advances in health care have given an opportunity to make find out which patients need to be given more importance. Several methodologies and techniques were used to preprocess the data in order to meet the study's requirements. To improve the performance of machine learning algorithms, feature selections were made using Tabu search. When ensemble prediction is combined with the Random Forest algorithm as the combiner, the results are more reliable. The aim of this study is to create a system to classify Medical records whether it is diseased or not and find out which disease rate has increased. This research will help the society to an individual to get treated easily and take preventive measures to avoid diseases.

Keywords: Machine Learning, K Nearest Neighbors, Nearest Mean Classifier, Mean Feature Voting Classifier, KDtree KNN, Random Forest

I. INTRODUCTION

Electronic Health Records (EHR) and Electronic Medical Records (EMR) are critical tools in the health-care industry for forecasting diseases and determining which diseases have the greatest impact on society. With rapid increase of access to a huge amount of patient data and files, healthcare providers are now feeling difficult to organize the data and prioritize which patient to be given more importance.

Manuscript received on March 08, 2020.
Revised Manuscript received on March 25, 2021.
Manuscript published on March 30, 2021.

B.Meena Preethi, Assistant Professor, Department of Software Systems, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India. Email: idmeenapreethibphd@gmail.com

P.Radha, Assistant Professor, PG & Research Department of Computer Science, Government Arts College, Coimbatore (Tamil Nadu), India. Email.id radhamuthu.cbe@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Retrieval Number: 100.1/ijrte.F5507039621
DOI:10.35940/ijrte.F5507.039621
Journal Website: www.ijrte.org

So to provide solution to the above problem, this research focuses on using ensemble classification machine learning algorithms to optimize the efficiency and quality of classification of data and to analyze the text data and bring out the results based on prediction of disease.

Data Mining

Data mining is the method of identifying trends in large data sets by combining machine learning, statistics, and database systems. It has been used in many health care organizations intensively and is becoming more popular and essential. It allows health systems to evaluate data regularly, maximize efficiencies, and implement best practices that enhance patient safety, lower costs, and save lives. [5]

Text Mining

Text Mining is equal to text analytics and it is used for deriving high-quality information from text. It's used to get useful information out of organized, semi structured, and unstructured data [7].

- i) Information retrieval
- ii) Lexical Analysis
- iii) Pattern recognition
- iv) Tagging

Machine Learning

Machine learning (ML) is a branch of artificial intelligence that is the scientific study of algorithms and mathematical models that are used to perform a task. Text mining creates a mathematical model from sample data, often referred to as "Training data." It employs unsupervised learning to make choices without the need for an explicit curriculum to complete the mission. Machine learning algorithms are commonly used in email filtering and computer vision, where a standard algorithm for efficiently performing the task is difficult to create. [8]

Ensemble Classification

Ensemble Learning / Classification is a method of solving a problem by integrating several machine learning (classifier) models that are strategically constructed. By integrating multiple classifier models, it aids in displaying and optimizing machine learning performance. As compared to a single model, this group of approaches allows for the creation of better expected results. [9].

Disease Classification and Prediction using Ensemble Machine Learning Classification Algorithm

II. LITERATURE SURVEY

S.No	Title of the Paper	Author	Journal	Methodologies	Results
1.	Disease detection using machine learning and biomedical mining	Bajj Nath Kaushik and Niharika, 2018.	Journal of Artificial Intelligence	To achieve the best performance, support vector machines (SVM), extreme learning (EL), and different combinations of Swarm Techniques are used.	To obtain reliable results, classifiers such as Entity Recognition and Information Extraction are used in conjunction with their DISEASE resource.
2.	Text mining techniques for clinical medical records	Isaac Chankai, Ann Prestrud, and Ari Brooks, Xiaohua Zhou and Hyoil Han	Research Gate	A graph-based approach is combined with an ID3-based decision tree and a feature extraction system based on natural language processing.	This initial approach to categorical fields has proven to be very successful so far.
3.	A Survey of Disease Diagnosis Machine Learning Algorithms	Maruf Pasha, Meherwar Fatima	Intelligent Learning Systems and Applications is a journal devoted to the study of intelligent learning systems and applications.	Functional Trees FT, Bayes Net, and Support Vector Machine	This has brought Bayes Net to 84.5% precision, gives SVM 85.1% accuracy and FT 84.5% accuracy.
4.	For Heart Disease Detection We use a hybrid system framework to use the machine learning algorithms	1 Muhammad Hammed Memon,1 Shah Nazir,2 Ruinan Sun,1 Amin Ul Haq,1 Jian Ping Li	Mobile systems of information	The support vector machine (SVM), the closest-neighbor (K-NN), neural network artificial (ANN), decision tree (DT), regression logistics (LR), AdaBoost (AB).	When selected by FS algorithm relief, logistic regression of 10 times cross validation classifiers showed 89% best accuracy.
5.	Classification of Text Survey Algorithms	ChengXiang Zhai, Charu C. Aggarwal	Business Media, Springer Science	Decision trees, rules, methods for bayes, nearest neighbour classification, SVM and neural network categorization	Decision-tree - Maximize the accuracy. Rule-based classifiers - predictive accuracy. Bayesian classifiers, t - weighed by the cost of the class the forecast is made for. The optimal hyperflugg separating the classes is determined in a cost-weighted way in linear classifiers. Weighed against costs of the various classes is the LLSF method.
6.	We can use Comparative Analysis of Clasifying Learning Classifiers and Deep Network Classifiers to predict Parkinson's disease	Muhammad Hammad Memon; Jalaluddin; Amin Ul Haq; Jianping Li	Wavelet Active media technology and data processing 15th International Computer Conference (ICCWAMTIP)	Vector machine support, logistic regression and deep neural networking support	23 characteristics and 195 cases. 70% for training and 30% for testing purposes. This classification was precise in neural performance compared to the classifying method of traditional machines
7.	Machine Learning Techniques Predicting Heart Disease	Kumar Babu D.Raghunath K.Veera Vidhya K.Usha Sree	Computer Science International Research Journal (IRJCS)	Gaussian Naïve Bayes Support Vector Machines K-Nearest Neighbors Decision Logistic Regression Trees	F-Score & Accuracy 0.8344 0.8 KNN 0.8 Tree decision 0.8344 0.9 Decision Regression of logistics 0.827 0.79 0.8211 0.78 Naïve Bayes 0.8476 0.82 SVM 0.9139 0.92 Random Forests Random forest performance is good from the above reports

III. PROPOSED METHODOLOGY

The Proposed methodology includes the following phases

- Preprocessing
- Feature Extraction
- Feature Selection
- Feature Classification

3.1 Preprocessing

In text mining techniques and applications, the preprocessing method is essential. The preprocessing method is the first step in text mining. Preprocessing requires a variety of steps.

- Categorization of text documents
- Tokenization of text (n –grams)
- Removal of stop words (Eg : is, was)
- Stemming (Eg: ing,ed)
- N-Grams Based Separation

3.11. Categorization of text documents

Text categorization is the process of arranging the documents according to the classes or categories from a predefined set of documents. The dataset used in this research is text documents which is the Radiology reports collected from Hospitals’ and Scan Centre. The Radiology reports in the form (.doc/.txt) file is

categorized based on classes to create a training set and saved in the training database.

3.12 Word (tokenization) extraction

Tokenization is the method of breaking down a string into smaller bits, such as sentences, phrases, keywords, and symbols. It’s used to extract the word (tokenize) from a file's text, which is made up of a series of strings. Tokens are grouped as a semantic unit and used as input for further processing including parsing. To tokenize the word in document d that is presented in the document package, natural language processing (NLP) is used.

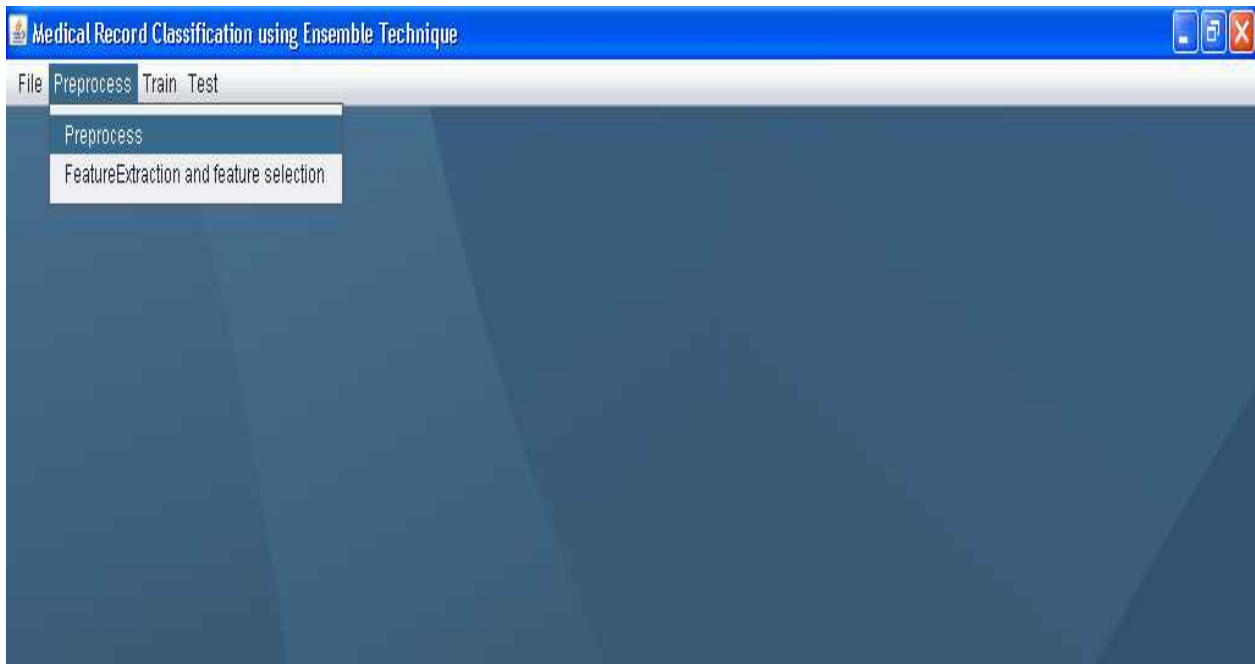


Figure 1. Preprocess in Medical Record Classification

3.13 Stop words

Prepositions, articles, and pronouns are common words in text documents. Stop words are used for these words. The most popular method for deciding a "stop list" is to sort the words by collection frequency and then create a stop list out of the most commonly used terms, which is then removed during indexing. Since stop words are not considered keywords in text mining applications, they are omitted from the documents.

A simple examples for stop words are:

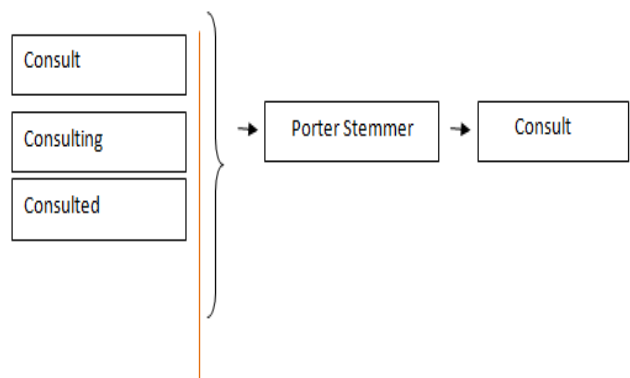
In	The	An	A	with	and	as	at	be	For
from	has	He	She	in	is	her	him	that	Those

3.14 Stemming

Stemming is a method for evaluating a word's root stem. The Porter stemming algorithm (or "porter" stemmer) is a tool for extracting morphological and inflexional endings from English words used by the average person. The term normalisation method, which is normally performed while setting up Information Retrieval systems, is the most common application. This approach is used to minimise the number of terms with dissimilar suffixes and the number of words with equivalent stems, saving time and memory space.

Stemming is the process of reducing a word to its simplest form, which includes all changes or roots of words, and is known as a lemma. In the field of information retrieval, process of removing suffixes automatically is useful. Porter stemmer discovers decent estimations to the stems of words, without essentially having a database of the actual words and stems.

Example:



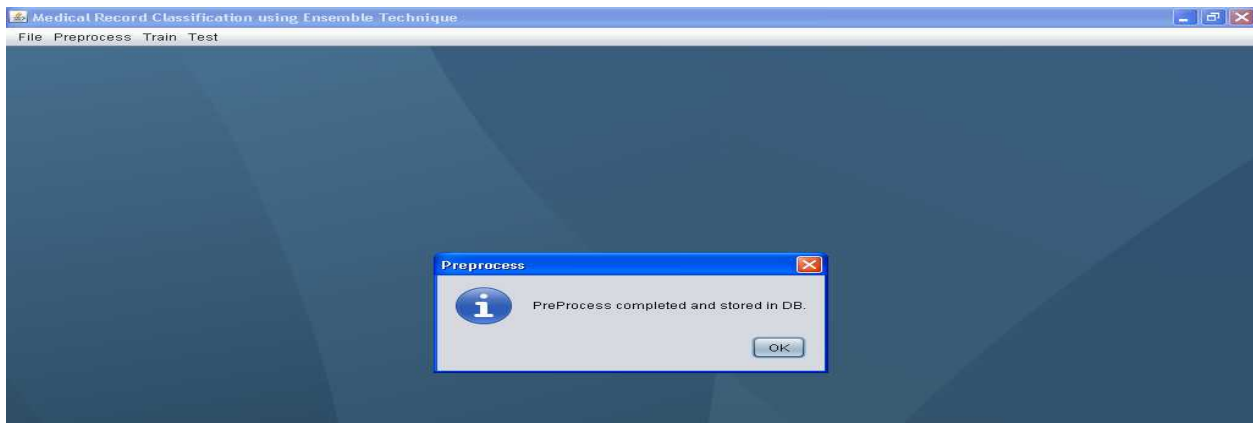


Figure 2. Stemming, Stop word removal performed on train files

3.5 N- grams

N-grams are used to distinguish features. The number n indicates how many words should be extracted from the dataset. A character n-gram is a set of n consecutive characters. The set of n-grams is denoted by the letter n which may be from (0,1,2,3,4) grams that can be generated for a given dataset (radiology report as document) the effect of moving an n-character window along the text (12). Table 1 displays the sample function extraction. After that, each n-occurrences gram's must be counted. This study proposes a new method for extracting n-grams. It allows for a significant reduction in the number of possible byte combinations while still allowing for the analysis of non-adjacent byte combinations. Only those bytes of the sliding window, which are used in n-gram extraction, are used as n grammes in this process, and the rest are ignored.

Table 1. Sample Feature Extraction

2 gram FE	3-gram FE	4-gram
'3d ct'	'3d ct pelvi'	'3d ct pelvi includ'
'ct pelvi'	'ct pelvi includ'	'ct pelvi includ both'
'pelvi includ'	'pelvi includ both'	'pelvi includ both hip'
'includ both'	'includ both hip'	'includ both hip joint'
'both hip'	'both hip joint'	'both hip joint indic'
'hip joint'	'hip joint indic'	'hip joint indic
'joint indic'	'joint indic	'joint indic techniqu'
'indic techniqu'	'indic techniqu'	'joint indic techniqu
'techniqu serial'	'indic techniqu	'serial'
'serial axial'	'serial'	'indic techniqu serial
'axial section'	'techniqu serial	'axial'
'section pelvi'	'axial'	'techniqu serial axial
'pelvi includ'	'serial axial	'section'
'includ both'	'section'	'serial axial section
'both hip'	'axial section pelvi'	'pelvi'
'hip joint'	'section pelvi	'axial section pelvi
'joint were'	'includ'	'includ'
'were studi'	'pelvi includ both'	'section pelvi includ
'studi without'	'includ both hip'	'both'
'without administr'	'both hip joint'	'pelvi includ both hip'
'administr iv'	'hip joint were'	'includ both hip joint'
'iv contrast'	'joint were studi'	'both hip joint were'
	'were studi without'	'hip joint were studi'
	'studi without	'joint were studi
	'administr'	'without'
	'were studi without	'were studi without
	'without administr	'administr'
	'iv'	'studi without
	'administr iv	'administr iv'
	'contrast'	'without administr iv
	'iv contrast media'	'contrast'

3.3 Feature Extraction

TFIDF feature extraction is used to remove features. The approach is TF-IDF is an information retrieval system that considers both the frequency (TF) and the Inverse Document Frequency (IDF) of a word (IDF). Each word in the document is assigned its own TF and IDF score. The TF*IDF weight of a word is the product of the TF and IDF scores for that term. Feature vectors are used to describe training sets. [13]. As shown in Fig.3, features are created for each text. It is divided into two parts.

1. TF Score (Term Frequency)

Considers text documents to be a jumble of words with no guarantee of their order. A text document containing ten occurrences of a word is more important than one containing term frequency. Relevance is not proportional to frequency if it is not 10 times more significant.

2. IDF Score (Inverse Document Frequency)

The frequency of the word in the meeting is used to weight and rate it. Infrequent terms are more descriptive than terms that are used often. And you want low positive weights for regularly occurring terms and high positive weights for rarely occurring terms.

Term Frequency(w) = (Number of times term w appears in a Document)
 Number of words in the text as a whole)

Inverse Document Frequency (IDF)(w) = $\log_e(\text{Total number of documents})$

$\frac{\text{Number of documents with term w in it}}$

$$W_{i,j} = TF_{i,j} * \log(N/DF_i)$$

Where N is No of Documents

TF is Term Frequency

IDF is Inverse Document Frequency

Consider a text with 100 words and five instances of the word "brain."

- TF = (5 / 100) = 0.05 is the word frequency (tf) for 'Brain.'
- Assume we have ten million records, and 1000 of them contain the word "brain." After that, IDF = $\log(10,000,000 / 1,000) = 4$ will be used to measure inverse document frequency (idf).
- As a result, the Tf-idf weight is the sum of these values. TF-IDF = 0.05 * 4 = 0.12.



3.5 Feature Classification

In this research, Ensemble Classification method is used to classify the records in the database. Ensemble Classification is an effective method used to combine set of classifiers whose individual results are combined by using (weighted or unweighted method) to provide new results. This paper brings out better results from the Text dataset which is an optimized method is used to classify Medical records whether it is diseased or not.

Recent hospitals are well-equipped with observing and additional data collection systems, resulting in massive amounts of data that are obstinately gathered by health checks and medical management. All of this has led to the fact that the medical sector is rapidly generating vast volumes of electronic data, which is becoming more difficult to handle. Prior to the advent of data mining, a number of statistical techniques for disease diagnosis modelling were used. It is currently difficult because data mining has been shown to be more successful and involved in finding useful patterns from large datasets.

By combining several models, ensemble methodology is used to create an empirical model. Ensemble methods are well-known for their ability to improve results. Because of their ability to precisely calculate class labels of modest and lightweight classes, ensemble methods for organised machine learning have become common. Statistics, pattern recognition, and machine learning scientists are exploring the use of ensemble methodology. The aim of the ensemble method is to improve the accuracy of a single classification or regression model [14]. The execution when the yields of several models are combined, ensemble mapping methods demonstrated greater precision than any single model. Ensemble models combine several hypotheses to prevent overfitting errors. When applied to a pancreatic cancer proteomic dataset, ensemble classifiers consistently outperformed single decision tree classifiers in terms of consuming superior accuracies and small predicting errors [15]. Other ensemble classifier features are used in data value assessment sensors, shellfish farm closure prediction and trigger detection, handwriting recognition, benthic habitat mapping, missing sensor data handling, and algae growth prediction.

The following set of classifiers used to classify the data are

- i) KNearest Neighbors,
- ii) Nearest Mean Classifier,
- iii) Mean Feature Voting classifier ,
- iv) KDtree KNN,
- v) Random Forest.

(Java ML tool is used to implement algorithms)

i)KNearest Neighbors

The K Nearest Neighbor algorithm is a supervised text categorization technique. This algorithm is used to classify and predict known data, with the target attribute/variable normally known ahead of time. It necessitates the use of numbered points. The k parameter in the k Nearest Neighbor algorithm is crucial for text categorization. The choice of the parameter k has a major effect on the machine act. Having a fixed k value can result in a bias on broad categories. An improved kNN algorithm that uses changed nearest neighbour records for different categories instead of a constant number for all categories.

Several samples from the closest neighbours are used to assess if a test document can be placed into a group of extra samples in the training package.

- Load the training and test data for the KNN algorithm to be implemented.
- Choose the data points that are closest to each other (i.e., the value of K [K=integer]).
- Using a range of techniques Calculate the distance between each row of training data and the test data.

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$



$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

Fig 5. KNearest Neighbor Euclidean Distance Formula

- Select them in ascending order based on the distance between them.
- The first K rows of the sorted array will be chosen.
- The most frequent class of these rows is used to assign a class to each test stage[16].

ii)Nearest Mean Classifier

The nearest centroid classifier is another name for the nearest mean classifier. One of the data processing algorithms is the nearest mean classifier. It is based on the pattern formation and recognition procedures. The preparation and identification sets are used in the identification process.

-  Relevant Document
-  Non-Relevant Document

Ideas:

Finding the most similar sentences in the document
Modified vectors

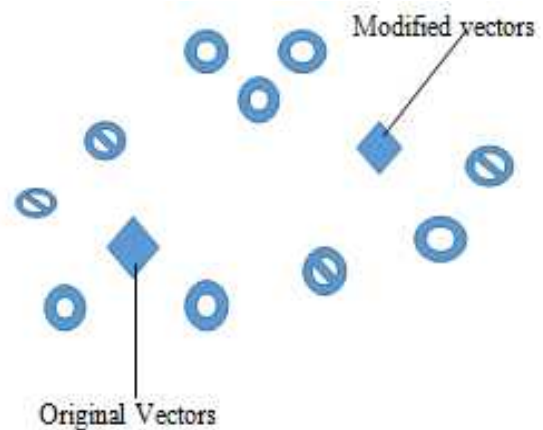


Fig 6. Near Mean Classifier

iii)MEAN FEATURE VOTING ENSEMBLE CLASSIFIER

The simplest and effective classifier algorithm is voting. Classification or regression problems uses voting algorithm. Voting mechanism works by generating two or more sub-models. Individually sub-model makes predictions which are combined, such as by taking the mean or the mode of the predictions, permitting each sub-model to vote according to the desired outcome.

Using a system known as single transferable vote, each voter ranks all of the alternatives in order of choice. All votes are counted in STV, and quota q is a set of rules.

$$q = \text{floor}\left(\frac{\text{total number of votes}}{\text{desired winner} + 1}\right) + 1$$

The most common quota is known as the Droop quota. A nominee is declared the winner whether he or she meets or exceeds the quota [17].

KD Tree KNN

KD Tree KNN algorithm is a data structure used to improve performance of finding nearest neighbor. Kd-tree search algorithm has less probability of resulting an approximate nearest neighbor[18]. KD Tree is efficient nearest neighbor searches and useful for layout analysis problem in document image analysis.

The two customization of k nearest neighbor search in kd tree are

- It returns only within the line
- Between line neighbor

VI) RANDOM FOREST

The supervised learning algorithm is used in classification and regression. Random forest is a supervised learning algorithm that is used to classify and predict data. It's also used to address classification issues. The random forest algorithm produces decision trees on data samples, then obtains predictions from each of them, and eventually, using voting, selects the best resolution. It's an ensemble approach that's better than a single decision tree because it combines the results to minimise over-fitting.

The random forest algorithm's steps are as follows:

- From given dataset, random sample selection is made.
- Decision tree is constructed for every sample and prediction result is acquired from every decision tree.
- Voting process take place for every predicted result
- Finally, most voted prediction result is considered as output

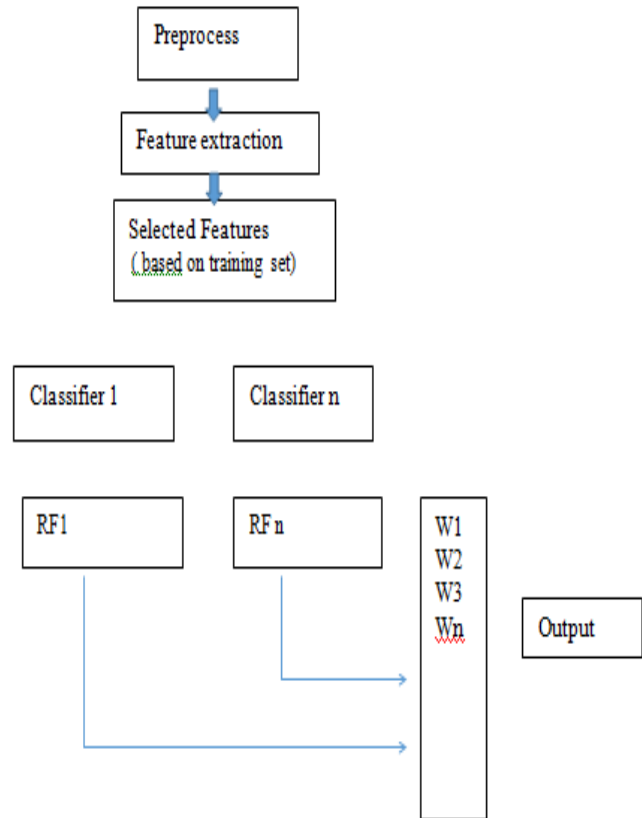
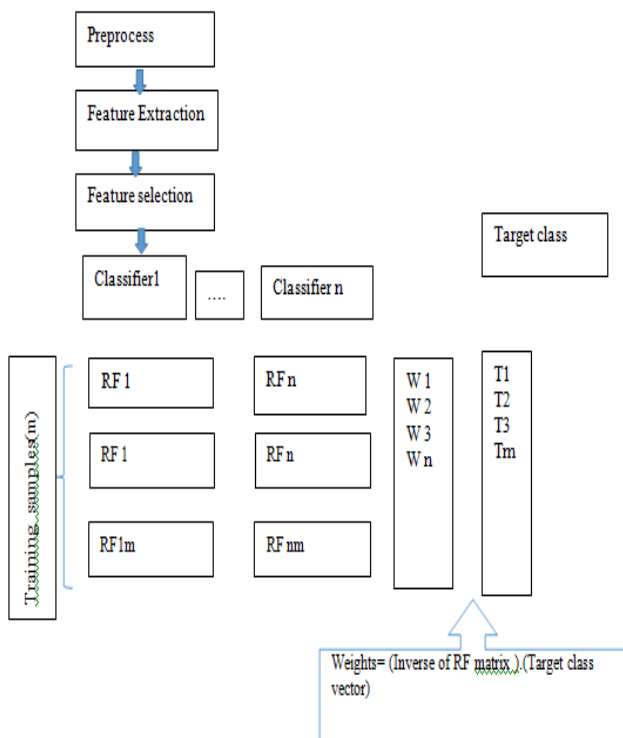


Fig 7. Random Forest

IV. IMPLEMENTATION

The Ensemble Classification method has used Classifiers used are KNearest Neighbors, Nearest Mean Classifier, Mean Feature Voting classifier ,KDtreeKNN, Random Forest techniques and implemented by the following procedure :

- Each classifier predict class for an instance.
- Relevance Factor for a classifier is calculated using tested instance and other vectors in training set.
- Relevance Factor matrix is calculated. Where value represents RF of each instance against classifiers.
- Using Pseudo Inverse calculate weight matrix W.
- For classifying a report, Create test vector(after preprocess steps).
- Pass test vector into ensemble classifier and calculate RF row matrix(alias Vector) for test instance.
- Perform Dot Product between W weight matrix(learned in training process) and test RF Vector (T) to calculate output value(Op=W.T).
- Based on threshold value Op is classified to respected Classes.
- Threshold value is set as 0.4 in this research.

Training Procedure:

Testing Procedure:

Calculation of Relevance Factor:

$$RF = \frac{\sum_{\varphi \in Pr} CS(\varphi, \tau)}{|Pr|} - \alpha \frac{\sum_{\varphi \in NPr} CS(\varphi, \tau)}{|NPr|}$$

Pr= Predicted class vectors
 τ = test class vector
 Npr=Non Predicted class vectors



Disease Classification and Prediction using Ensemble Machine Learning Classification Algorithm

α = parameter that adjusts the relative impact of positive and negative neighboring instances

Cosine Similarity:

$$CS(X, Y) = \frac{\sum_{i=1}^f X_i Y_i}{\sqrt{\sum_{i=1}^f X_i^2} \sqrt{\sum_{i=1}^f Y_i^2}} \text{ where } f = \text{no of features}$$

Pseudo inverse:

RF matrix:

$$R = \begin{pmatrix} RF_{1j} & \dots & RF_{1n} \\ \vdots & \dots & \vdots \\ RF_{mj} & \dots & RF_{mn} \end{pmatrix}$$

Target matrix:

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

Weight matrix:

$W = R^+ Y$; R^+ = Moore–Penrose inverse of R matrix

Output:

$O = W * R$; where R contains single test instance RF value

Algorithm:

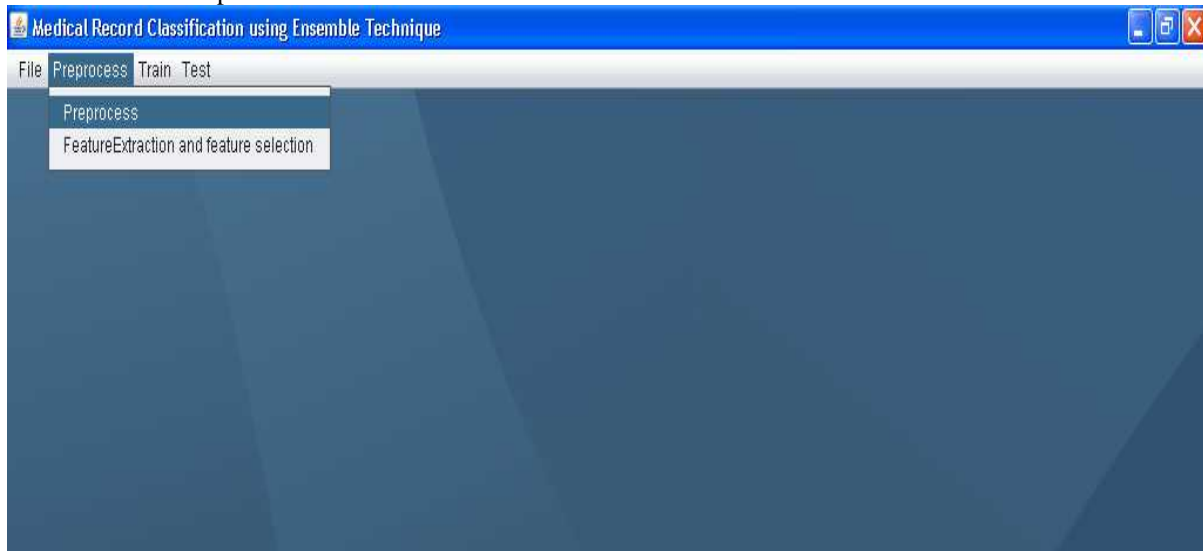
Step1: Perform preprocess steps

tokenization, Stemming (Porter Stemmer), Stop word removal.

Step3: Get instance from training Dataset

Step2: Classifier C_i from set of classifier{ }

1. PreProcess-> Preprocess



2. Stemming, Stopwords removal Performed on trainFiles.

Step 4: $prC = \text{classify instance using } C_i$ //prC-predicted class

Step 5: Calculate RF using prc and acC. //acC-actual Class of instance

Step 6: Repeat step 2 for next classifier

Step 7: Repeat Step 3 for next instance

Step 8: Collect RF for each class as attributes Make meta training set for each instance.

Step 9: Use Meta set to calculate weight for each classifier using **Pseudo inverse**.

Classifying:

Step 1: Perform Preprocess, Stemming Stop word removal for test set.

Step 2: Feature Extraction and Feature selection as per learning steps.

Step 3: Create RF instance for test file. Contains each classifier's Rf value as attribute.

Step 4: Use learned weight to classify.

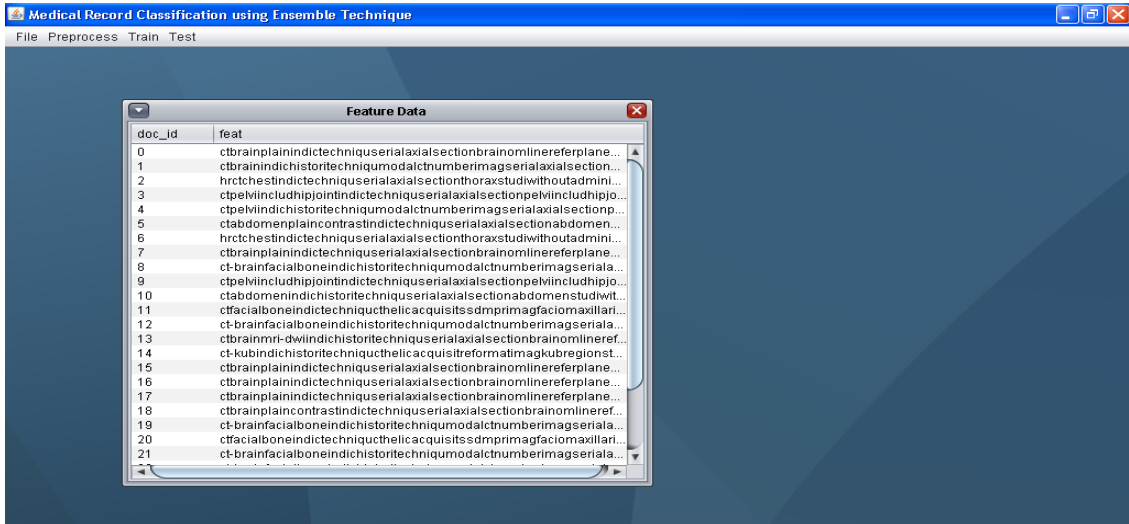
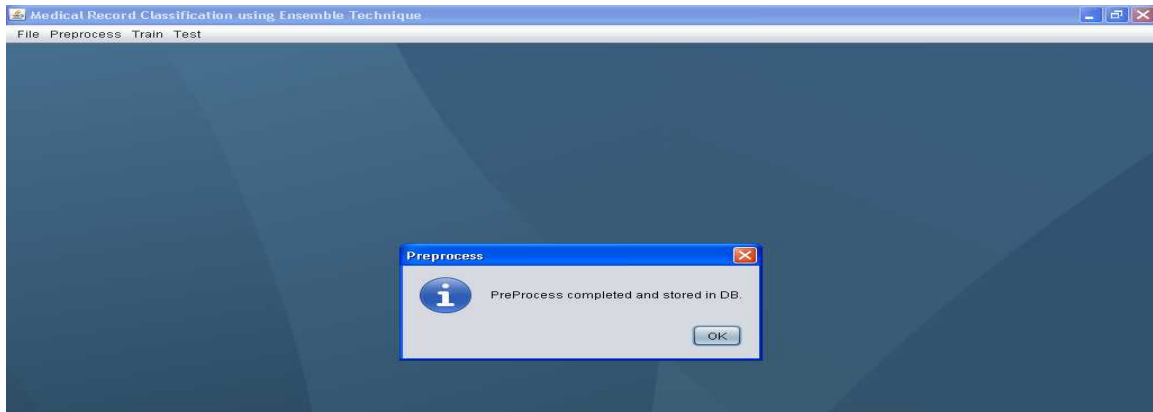
Step 5: $\delta = W * Rf$ instance.

Step 6: Based on threshold value(Th) classify test file.(if $\delta < \text{th}$ predict as class1)

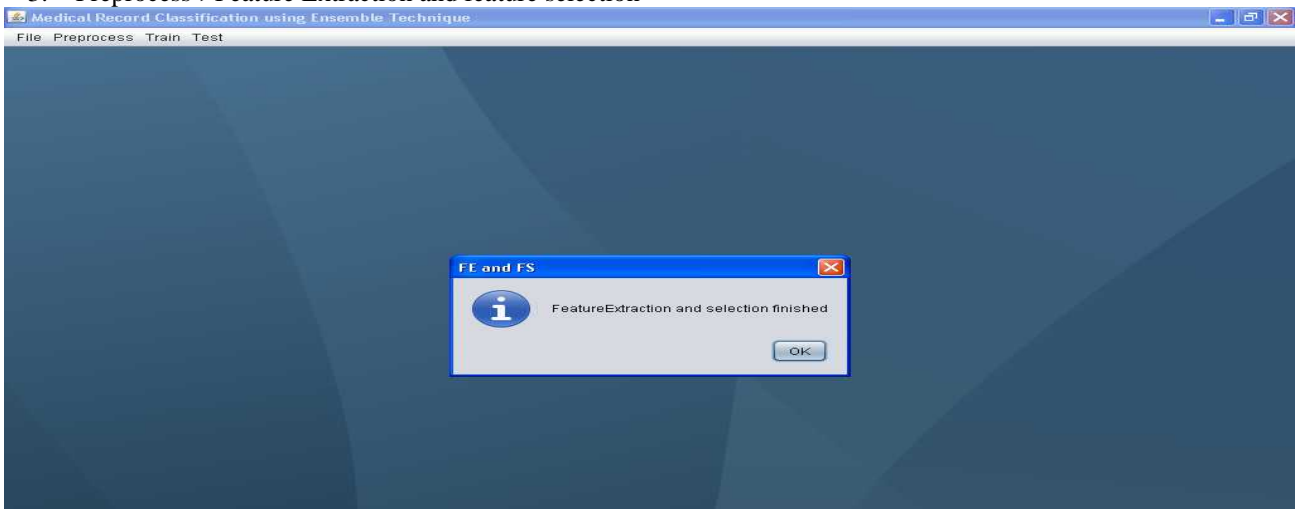
V. RESULTS

Disease classification and prediction datasets consists of samples of disease cases. The classification work is done and complete research methodology is divided into preprocessing, feature selection, feature extraction and classification.

The Following images shows the results of how the research is carried out:

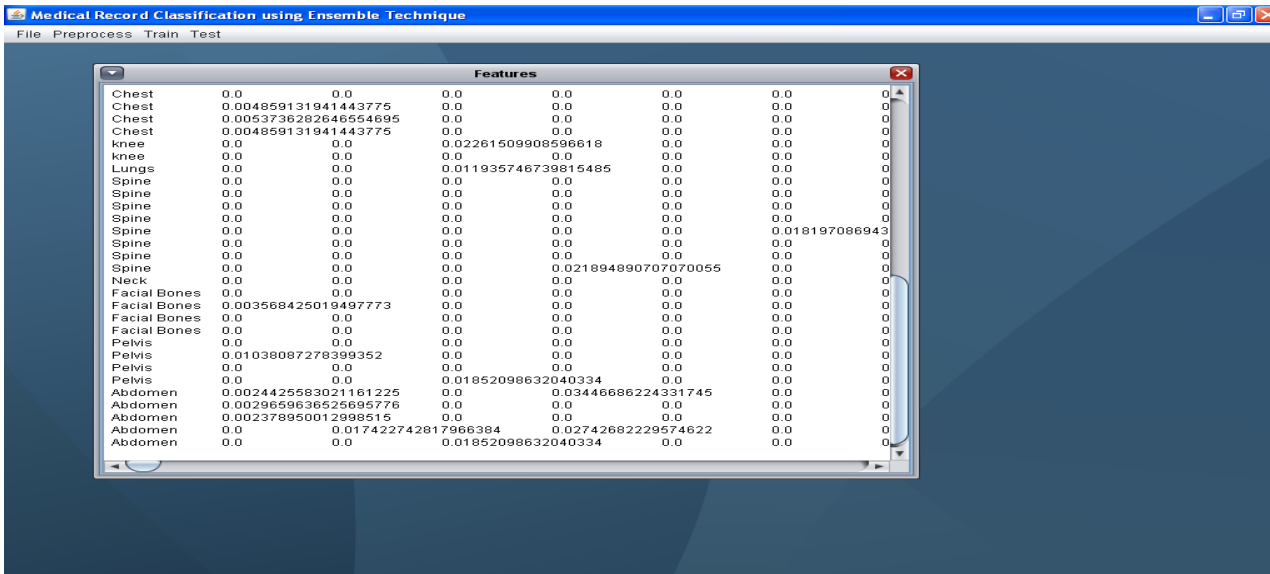


3. Preprocess->Feature Extraction and feature selection

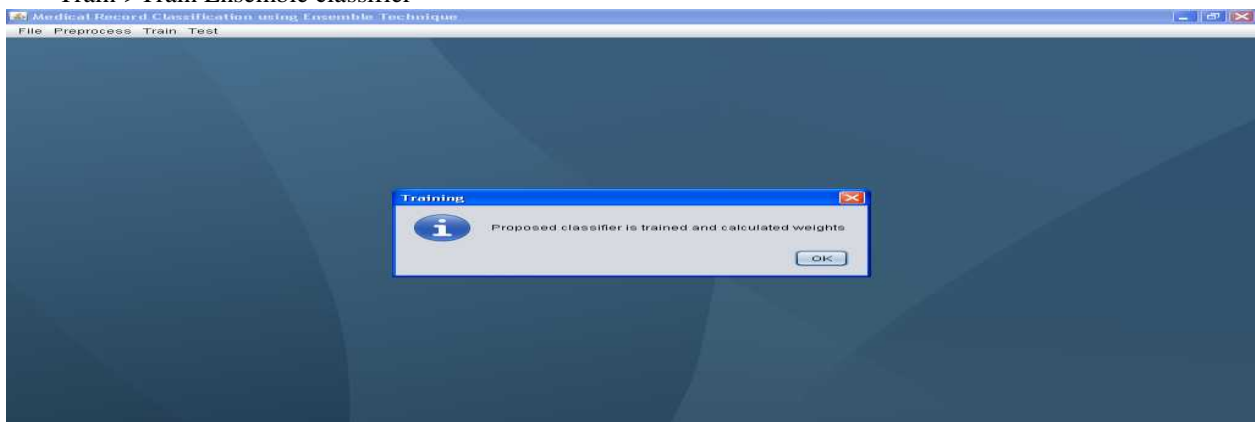


4. Feature sets:

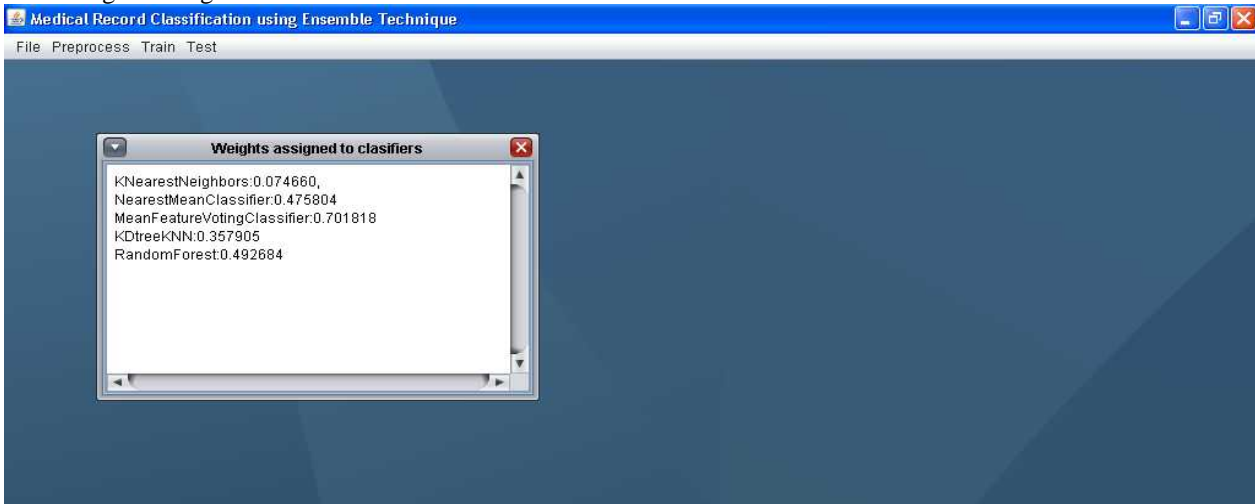
Disease Classification and Prediction using Ensemble Machine Learning Classification Algorithm



5. Train->Train Ensemble classifier



6. Weights Assigned to Classifiers



7. Test->Performance Evaluation and shows categorization

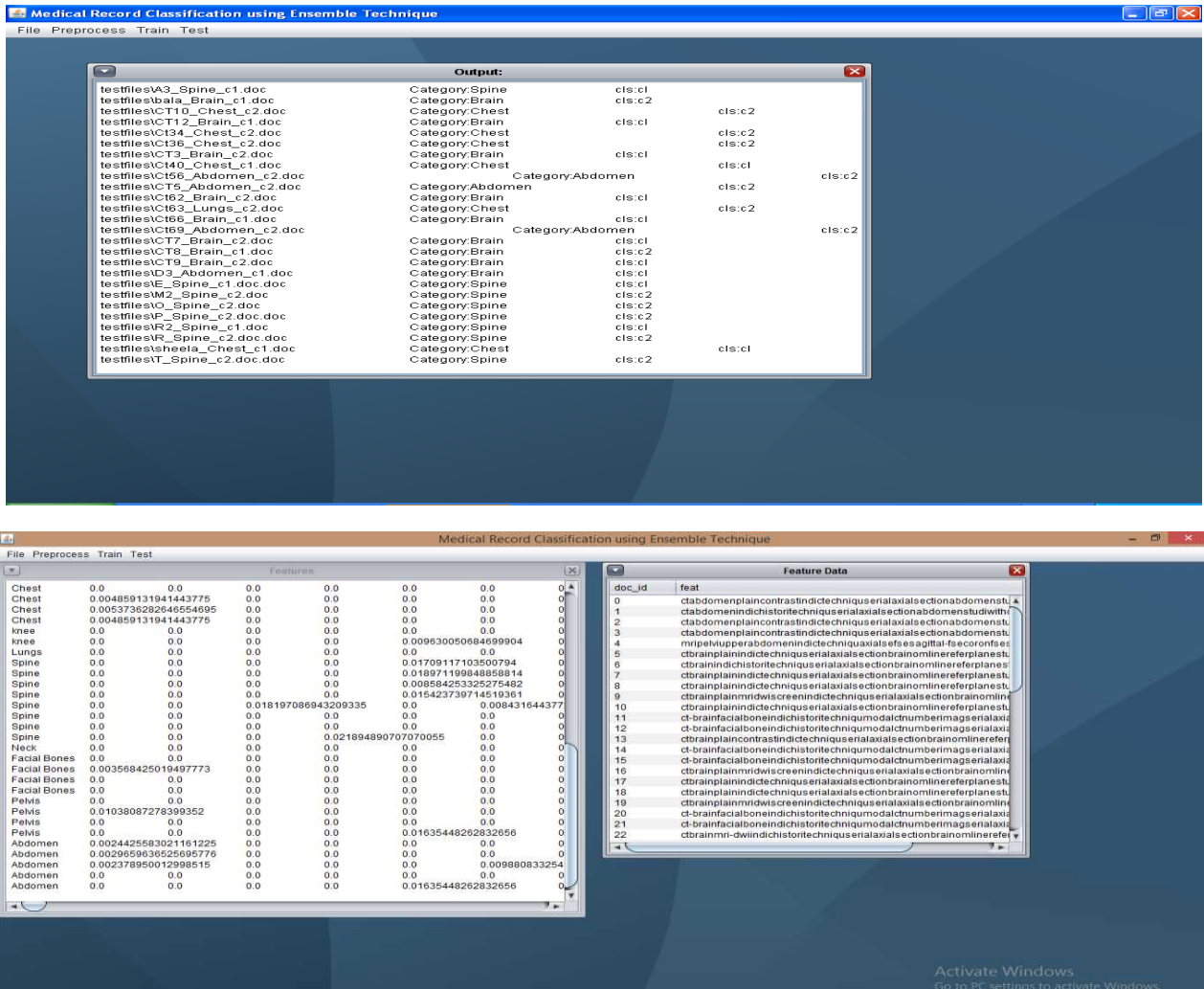


Fig 8. Performance Evaluation Using Ensemble Classifier

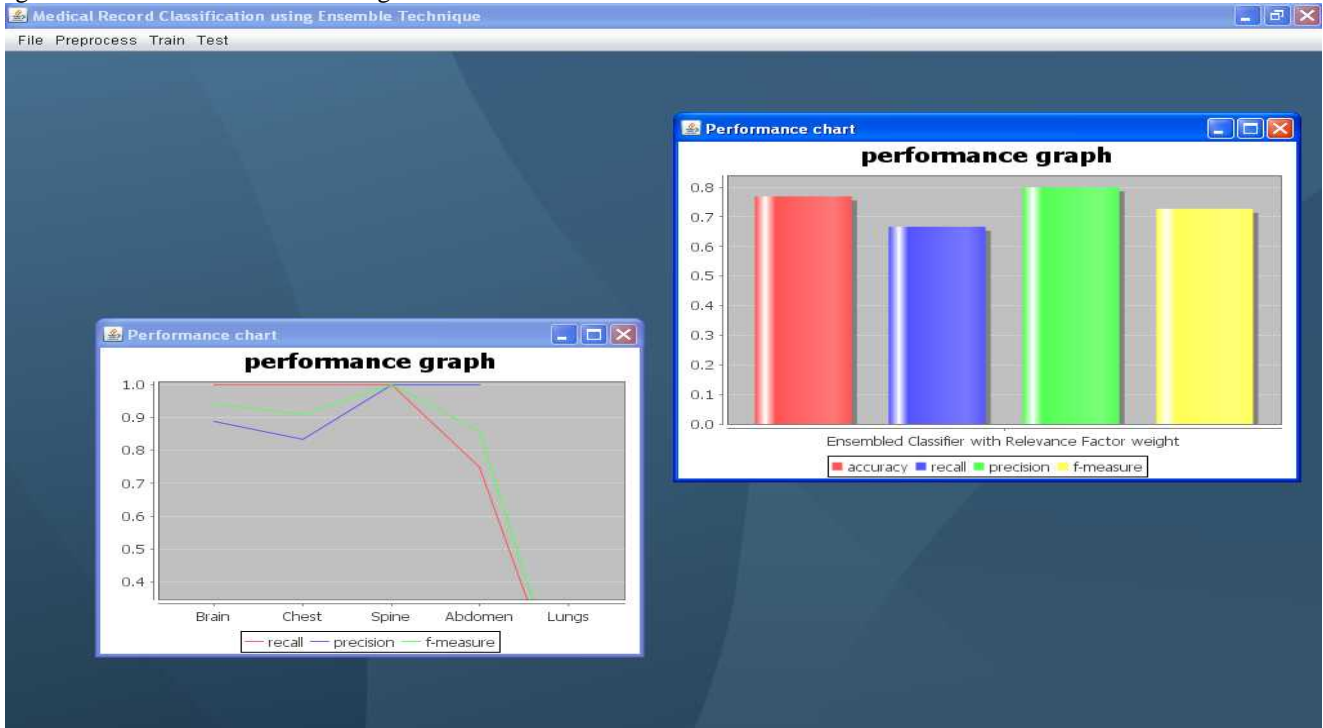


Fig 9: Ensemble classifier with Relevance Factor Weight and Performance Graph

Table 3 : Results on Analysis of Medical Records

No Per features	Precision	Recall	F score
50	0.6	0.68	0.67
70	0.7	0.77	0.78
80	0.75	0.85	0.81
90	0.8	0.75	0.75
95	0.65	0.88	0.72

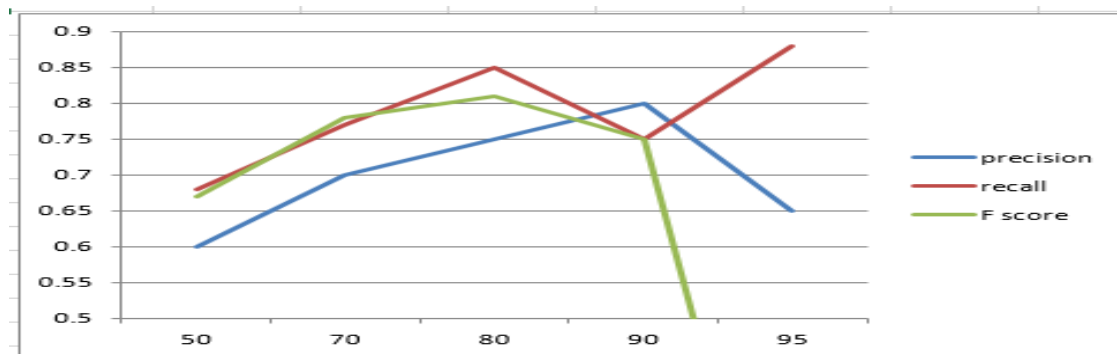


Fig. 10: Analysis Of Medical Record

The following are the assessment criteria that will help reflect on the results of the classifier that has on the minority class:

Recall or the true positive (TP) rate is:

When FN is False Negative, the true positive (TP) rate of the classifier is $TP/(TP+FN)$.

Precision:

Precision and recall are often in competition with one another. The proportion of positives that are categorised correctly named as Precision: $TP/(TP+FN)$. Precision is one-fourth of the ratio of correctly graded +ve instances to the total number of +ve instances.

F1 Score:

The F1 metric or score takes both recall and precision into account. It's also known as a fourth of the harmonic mean of recall and precision. It is used to determine the acknowledgment's efficiency.

In this research , this system will be helpful for health care organization to predict the diseases efficiently and prevent it. In figure 11, its shows the graph frequency of precision ,recall and fscore value according to the individual datasets. Ensemble machine learning classification algorithm is used for the efficient mechanism.

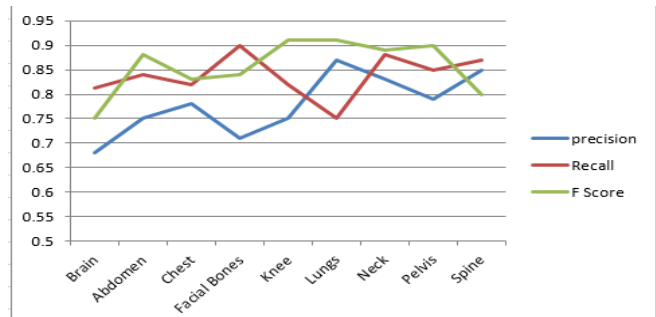


Fig. 11: Medical Records Analysis according to Datasets

VI. CONCLUSION

This research surveyed some ensemble machine learning classification algorithm to predict classify and predict the diseases. The research analyzed the classifiers such as K Nearest Neighbors, Nearest Mean Classifier, Mean Feature Voting Classifier, KDtree KNN, Random Forest. From the results, Datasets for disease is categorized with the help of precision, recall and Fscore The machine learning algorithm was effective in classifying whether or not the medical records were diseased. The value of this research is that it will assist in the identification of disease and the development of preventative measures to stop it.

REFERENCES

1. Simon Kocbek , Lawrence Cavedon David Martinez et al.,Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources Journal of Biomedical Informatics 64 (2016) 158–167.
2. <https://downloads.healthcatalyst.com/wp-content/uploads/2014/05/Healthcare-Data-Mining.pdf>
3. <https://blog.statsbot.co/ensemble-learning-d1dcd548e936>
4. <https://towardsdatascience.com/advanced-ensemble-classifiers-8d7372e74e40>
5. <https://www.healthcatalyst.com/data-mining-in-healthcare>

Table 3 : Results on Analysis of Medical Records

Dataset	Precision	Recall	F Score
Brain	0.68	0.812	0.75
Abdomen	0.75	0.84	0.88
Chest	0.78	0.82	0.83
Facial Bones	0.71	0.9	0.84
Knee	0.75	0.82	0.91
Lungs	0.87	0.75	0.91
Neck	0.83	0.88	0.89
Pelvis	0.79	0.85	0.9
Spine	0.85	0.87	0.8

6. <https://www.usfhealthonline.com/resources/key-concepts/data-mining-in-healthcare/>
7. https://en.wikipedia.org/wiki/Text_mining
8. https://en.wikipedia.org/wiki/Machine_learning
9. <https://www.geeksforgeeks.org/ensemble-classifier-data-mining/>
10. A hybrid ensemble for classification in multiclass datasets: An application to oilseed disease dataset, Computers and Electronics in Agriculture , Volume 124, June 2016, Pages 65-72
11. <https://www.hindawi.com/journals/misy/2018/3860146/>
12. Zhihua Wei, Duoqian Miao, Jean-Hugues Chauchat, and Caiming Zhong, "Feature Selection based on Chinese Text Classification Using Character N -Grams , Lecture Notes in Computer Science, Publication Date: 2008
13. Cha Yang Jun Wen , "Text Categorization Based on a Similarity Approach", Sruthi
14. Partalas, I., Tsoumakas, G., Hatzikos, E. V., & Vlahavas, I. (2008). Greedy regression ensemble selection : Theory and an application to water quality prediction. Information Sciences Journal, 178, 3867–3879. <https://doi.org/10.1016/j.ins.2008.05.025>
15. Ge, G., & Wong, G. W. (2008). Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. BMC Bioinformatics, 9, 275.
16. An Improved k-Nearest Neighbor Algorithm for Text Categorization1, Li Baoli1, Yu Shiwen1, and Lu Qin2
<https://arxiv.org/ftp/cs/papers/0306/0306099.pdf> **KNN**
<https://arxiv.org/abs/cs/0306099>
17. <https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d>
18. http://rasbt.github.io/mlxtend/user_guide/classifier/EnsembleVoteClassifier/
19. https://www.researchgate.net/publication/220980101_An_Improved_Algorithm_Finding_Nearest_Neighbor_Using_Kd-trees

AUTHOR PROFILES



Dr. P. Radha MCA., M.Phil., Ph.D., Assistant Professor in PG and Research Department of Computer Science, Government Arts College, Coimbatore. She has obtained her PG Degree in Alagappa University, Karaikudi and M.Phil Degree in Manonmaniam Sundaranar University, Thirunelveli. She completed her Doctorate degree from Alagappa University, Karaikudi, India in the year 2013. She has more

than 25 years of teaching experience. Her Specialization is Data Mining, Network Security, and Artificial Neural Networks. She published more than fifty Research papers in National, International journals and Conferences. She has organized various national seminars and acted as a resource person, she has acted as an editorial board member in computer science for various colleges as a representative of Bharathiar University. She has written articles for wide spectrum and organized a free certificate course and also published books. She has acted as a Domain Expert for State Board Higher Secondary (framing syllabus & Book writing) of Tamil Nadu for Computer Science, Computer Technology and Computer Applications.



Prof. B. Meena Preethi, M.Sc., M.Phil., is an Assistant Professor at Sri Krishna Arts and Science College in Coimbatore's Department of Software Systems. She has 11 years of teaching and administrative experience, and she graduated from Bharathiar University with a University gold medal. She is also the recipient of the 2008 "Best Outgoing Student Award." She has delivered over 30 research papers and has 28 research articles published in national and international journals to her name. Data Mining,

Network Security, and Artificial Neural Networks are her areas of expertise. Her research focus is on data mining in the medical sector.