# Effective Feature Selection and Multi Disease Prediction Using Data Mining Techniques

[1]Mrs.K.Sindhu, [2]Mr.Yaswanth Kumar Alapati
[1]Assistant Professor,[2]Assistant professor
[1]Department of Information Technology, [2]Department of Information Technology
[1]VFSTR, vadlamudi, Guntur, Andhra Pradesh, India
[2]RVR&JC College Of Engineering, Guntur, Andhra Pradesh, India

*Abstract :*Data is growing exponentially all over the world day by day at a very rapid rate. Health care centre, hospital and diagnostic centre's generates large amounts of data that is too difficult to analyze and the data is heterogeneous and high dimensional. All the features in dataset may not be relevant to the prediction of disease. To improve the performance the dimensionality of data set has to be reduced using Feature Selection and dimensionality reduction plays important role in Medical Data Mining. This paper proposes a novel technique for feature selection and to detect different diseases efficiently and effectively.

***IndexTerms*–Feature Selection, Medical Data Mining, Dimensionality Reduction**

## I. INTRODUCTION

Health care industry stores a lot of patients information and that information is utilized for clinical data, prevent medication error and improve patient's outcome[2].Patient's data is recorded and this data is growing rapidly. Medical data is rich in information but weak in knowledge. Generally diseases can be identified by medical specialists based on the test results. Diagnosing a disease is critical task which needs high experience. Medical sector is producing huge amounts of data which is difficult to analyze.

The objective of data mining process is to find the useful data from the complex data. Data Mining is capable of handling and analyzing large amounts of Data. So data mining algorithms and techniques such as association rule mining, classification, clustering and regression are very useful in Health Care Industry. Data mining pristinely emanated from statistics and machine learning as an interdisciplinary field [7], but then it was grown a lot that in 2001 it was considered as one of the top 10 leading technologies which will transmute the world.

Sometimes the data produced by Health Care Industry is High-Dimensional. In High Dimensional data each sample is defined by a lot of variety of measurements. Processing of High-Dimensional data will take more time and accuracy is also reduced. The major problem with high dimensional data is curse of dimensionality and more samples are needed to perform any task like classification, clustering and association rule mining.

Data Mining is a powerful tool to identify hidden patterns and relationship between patterns. Data mining techniques have a wide scope of applicability in the field of disease diagnosis [9]. Data mining techniques can be applied in the prediction of diseases like heart disease, cancer and kidney related disease. Patients get better medical advices and Disease can be treated efficiently if it is identified in early stage.

In this paper we have considered the data sets of different diseases such as Chronic Kidney Disease, Breast Cancer, Lung Cancer and Heart Disease. With Data Mining Techniques and Feature Selection i.e. either by reducing the number of features or by selecting only important features from the dataset the accuracy of classifier may be improved and with reduced feature set construction of classifier model will take less time.

## II. LITERATURE SURVEY

DoronShalvi and Nicholas DeClarissuggested an approach to medical data mining using unsupervised Neural Network [3]. Prediction of Heart Disease using Naïve Bayes and Decision Tree algorithms has been defined by Priyanka N, Dr.PushpaRaviKumar[10]. In their research they compared the Efficiency of the Naïve Bayes and Decision Tree algorithms in order to decide which algorithm is suitable for predicting Heart Disease.

A decision support system for prediction of heart diseases using different data mining algorithms likebagging and Naïve Bayes Algorithm has been proposed by Tu [8]. Dr.S.Vijayarani compared the performance of Naïve Bayes and SVM classification algorithms for the prediction of Kidney Disease [4].

A prediction system for heart disease and kidney failure using A-priori and K-means was developed by AnuChaudhary et al[1].In 2015, Ruey Key [11], implemented three different neural network models for chronic kidney disease detection like back propagation neural network (BPN), generalized feed forward neural network (GRNN) and modular neural network (MNN). In his research work he combined all three neural network algorithms with Genetic Algorithm. Among these three algorithms Back Propagation Neural Network (BPN) yields to better accuracy.

Hai Wang, et. al. [5] carried out a work called "The acquisition of medical knowledge through data Mining." This paper discusses the important role of medical experts in the mining of medical data and Presents a model for the acquisition of medical knowledge through data mining.

Harsh Vazirani, et. al. [6] carried out a work, "Use of Modular Neural Network for Heart Disease". In his research he concentrated on diagnosis of the heart disease and he used two types of methods for diagnosis manual and automatic diagnosis method using intelligent expert system and modular neural network. Two classification methods Back Propagation Neural Network and Radial Basis Function Neural Network are used for diagnosis of Heart Disease.

A lot of work has been done for diagnosing and prediction of different diseases by applying different data mining algorithms. Vikaschaurasi and saurabh pal done research on prediction of heart disease using data mining techniques and the authors used three data mining algorithms such as CART,ID3 and Decision Tree and they have used 10-fold cross validation to estimate the accuracy..

## III. PROPOSED METHOD

All the features in the data set may not be important to determine whether patient is having a particular disease or not. Feature Selection (or) Feature Subset Selection is pre-processing technique to identify the significant attributes from original data set. There are different Feature Selection methods such as CFS, Relief-F. Relief-F method assigns weight to each Feature and Relief-F finds only relevant Features but not Redundant features. CFS finds only redundant features but fails to find relevant features.

The steps in the proposed method are Feature Subset Selection and Detection of Disease.

### 3.1 Feature Subset Selection

Feature Selection methods aims at selecting most useful features for constructing a prediction model. The objective of Feature Subset Selection is removing both irrelevant and redundant features. With feature subset selection curse of dimensionality can be avoided and data collection process is simplified. The steps involved in the feature Subset Selection are i) Removing irrelevant Features ii) Removing Redundant Features

### 3.1.1 Removing Irrelevant Features

Before Removing Irrelevant Features apply MDL discretization technique for the continuous valued attributes. For a data set D with n features F={$F_1$,$F_2$,….,$F_n$} compute SU($F_i$,C) value for each feature. Remove all the features whose SU($F_i$,C) is less than the threshold.

Symmetric Uncertainty calculated using the formula

$$SU(X,Y) = \frac{2 * Gain(\frac{X}{Y})}{H(X) + H(Y)}$$

H(X) is the Entropy of variable X and H(Y) is the Entropy of Y.

### 3.1.2 Removing Redundant Features

To identify redundant features Clustering Based method is used. After finding the relevant features construct a completely connected weighted graph G is constructed by taking relevant features as vertices and correlation between the features as a weight of the edge. Then construct the minimum spanning tree and partition the minimum spanning tree into clusters by removing inconsistent edges from the spanning tree. The features in each cluster are redundant so select representative feature from the cluster. All cluster representative features together forms feature subset.

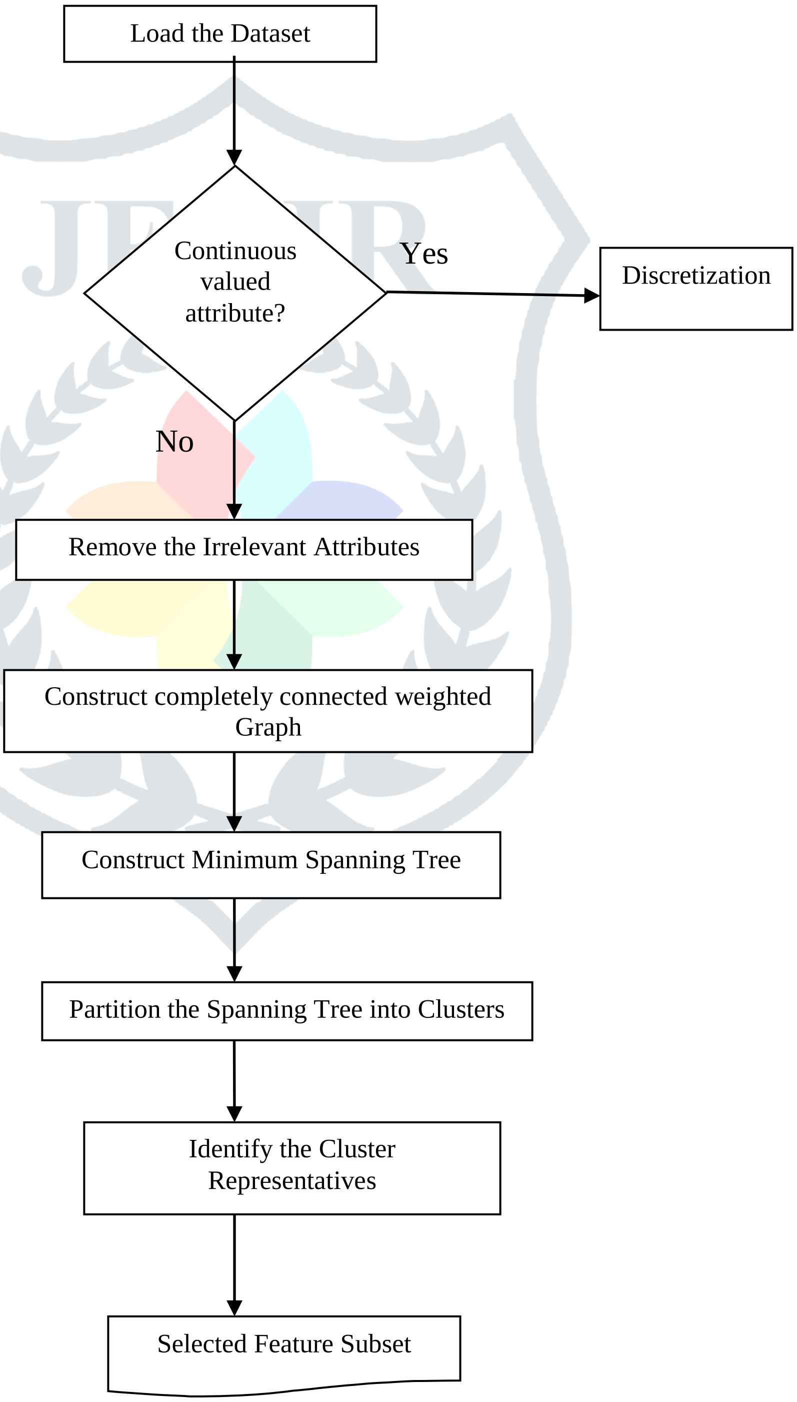
Load the Dataset
Continuous valued attribute?
Yes
Discretization
No
Remove the Irrelevant Attributes
Construct completely connected weighted Graph
Construct Minimum Spanning Tree
Partition the Spanning Tree into Clusters
Identify the Cluster Representatives
Selected Feature Subset

Figure 3.1: Frame work of the Feature Subset Selection Process

### 3.2 Detection of Disease

Different classification algorithms are used to predict different diseases. In this paper multiple classification algorithms like ID3 with Genetic Algorithm, Multi Layer Perceptron Neural Networks, Naive Bayes and Logistic Regression are used to build a model to detect any disease.

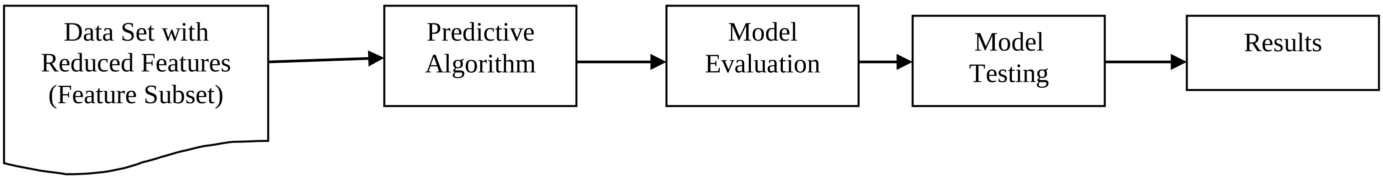

Figure 3.2: Framework of the proposed method

### 3.3 Data and Sources of Data

Data related to different kinds of diseases collected from UCI Repository and TunedIT Repository. Numbers of Experiments are conducted on Data Sets to verify the proposed approach. Summary of the data sets presented in the Table 1.

TABLE 3.1: DATA SETS

| **Dataset** | **#Instances** | **#Attributes** |
|---|---|---|
| Heart Disease | 303 | 75 |
| Chronic Kidney Disease | 400 | 25 |
| Breast Cancer | 699 | 10 |
| Lung Cancer Data Set | 32 | 56 |

### IV. Experimental Results

All the features in data set may not relevant to the prediction variable. Having irrelevant features in data set can decrease the accuracy of a model. The performance of proposed algorithm has been tested with 4 data sets from medical data base. Table 4.1 shows the Accuracies obtained by using proposed algorithm and also comparison with other methods and the Comparison of proposed method with existing method is shown in the Fi.4.1,Fig.4.2,Fig.4.3 and Fig 4.4

**Table 4.1 Accuracies of Different Classification Algorithms**

| Data Set | ID3+Genetic Algorithm | | Multi Layer Perceptron Neural Network | | Logistic Regression | | Naive Bayes | |
|---|---|---|---|---|---|---|---|---|
| | Without Feature Subset Selection | With Feature Subset Selection | Without Feature Subset Selection | With Feature Subset Selection | Without Feature Subset Selection | With Feature Subset Selection | Without Feature Subset Selection | With Feature Subset Selection |
| Heart Disease | 91 | 96 | 78.14 | 85.67 | 85.5 | 90.21 | 85.18 | 89.82 |
| Chronic Kidney Disease | 86 | 95 | 93.15 | 97.85 | 94 | 99 | 93.5 | 96.2 |
| Breast Cancer | 90 | 93 | 94.28 | 95.72 | 92 | 98.7 | 94.56 | 96.28 |
| LungCancer Data Set | 78 | 93 | 65.62 | 78.38 | 81.25 | 99 | 78.12 | 90 |

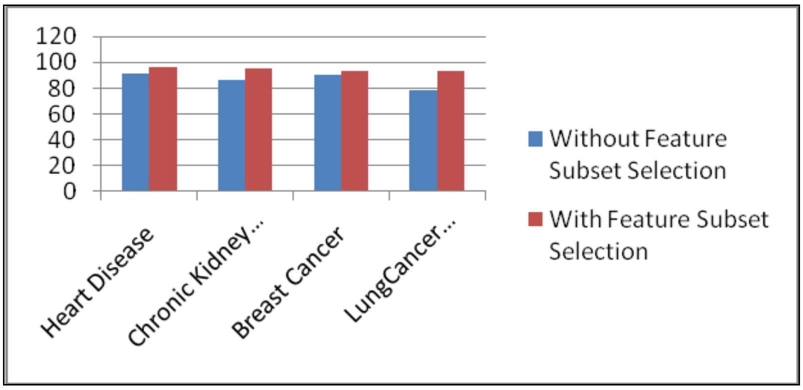

**Figure 4.1: Accuracy of ID3+GA for Different Datasets**

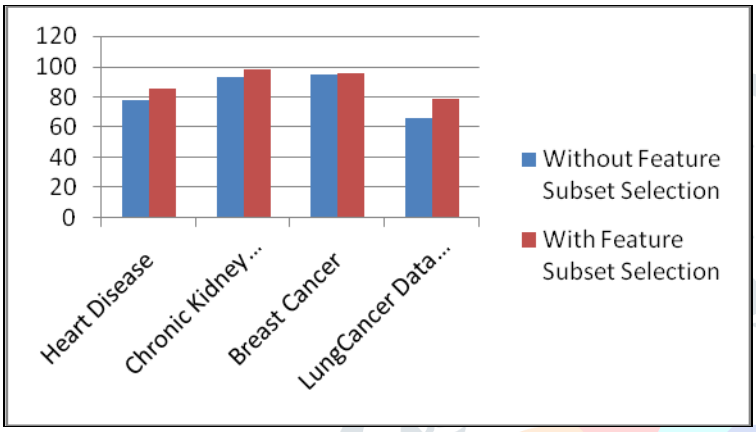

**Figure 4.2: Accuracy of Multi-Layer Perceptron Neural Network for different data sets**

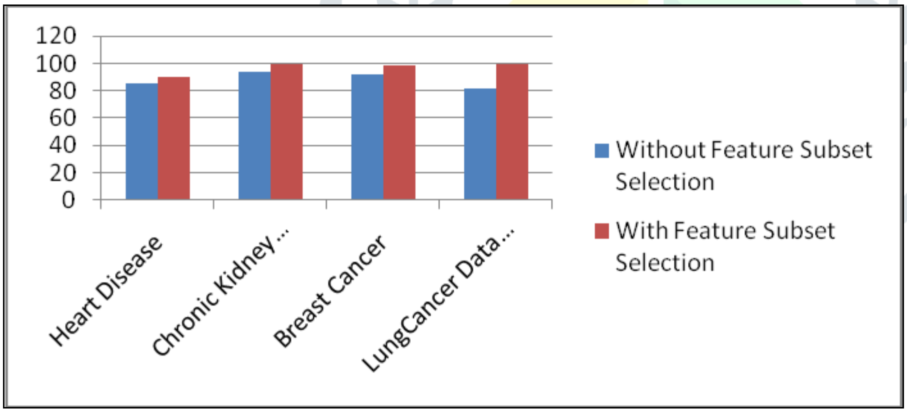

**Figure 4.3: Accuracy of Logistic Regression for different data sets**

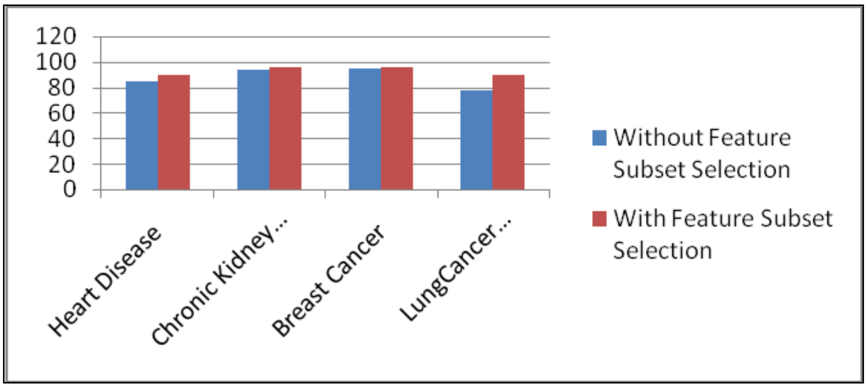

**Figure 4.4: Accuracy of Naïve Bayes for different data sets**

## V. CONCLUSIONS

Health Care Industry producing more data and data rate is growing exponentially. All the features in the data may not be relevant to predict to whether patient is having disease or not. Feature Subset Selection is pre-processing technique to identify the significant attributes from original data set. A new Feature Selection Algorithm is proposed and using the proposed Algorithm is efficient and effective. The performance of proposed algorithm is tested with four medical data sets and results shows that disease prediction is done better with Feature Selection Algorithm. And also prediction model construction time is reduced with selected features.

## REFERENCES

[1]. Anu Chaudhary, PuneetGarg,(2014) Detecting and Diagnosing a Disease by Patient Monitoring System, International Journal of Mechanical Engineering And Information Technology, Vol. 2 Issue 6 //June //Page No: 493-499.

[2]. Daniyal, Wei-Jen Wang, Mu-Chun Su , Si-Huei Lee , Ching-Sui Hung ,Chun-Chuan Chen, "A guideline to determine the training sample size when applying big data mining methods in clinical decision making", Proceedings of IEEE International Conference on Applied System Innovation 2018, ISBN 978-1-5386-4342-6, pp:678-681

[3]. DoronShalvi and Nicholas DeClaris, "An Unsupervised Neural Network Approach to Medical Data Mining Techniques", In IEEE proc of International Joint Conference on Neural Networks, pp. 171-176, Vol.1, ISBN: 0-7803-4859-1, 1998

[4]. D. S. Vijayarani, Mr. S. Dhayanand, "Data Mining Classification Algorithms for Kidney Disease Prediction", International journal of Cybernetics and informatics (IJCI), pp 13-25 Vol. 4, No. 4, 2015.

[5]. Hai Wang et.al,"Medical Knowledge Acquisition through Data Mining", Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education 978-1-4244- 2511-2/08©2008 Crown.

[6]. Harsh Vaziraniet. al.," Use of Modular Neural Network for Heart Disease", Special Issue of IJCCT Vol.1 Issue 2, 3, 4; 2010 for International Conference [ACCTA-2010], 3-5 August 2010, page no. 88-93.

[7]. I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, L. Hua, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", J. Med. Syst., vol. 36, no. 4, pp. 2431-2448, May 2011.

[8]. My ChauTu AND Dongil Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms" In IEEE proc of Eighth International Conference on Dependable, Autonomic and Secure Computing, pp. 183-187, ISBN: 978-1-4244-5421-1, 2009

[9]. Narander Kumar, SabitaKhatri, "Implementing WEKA for medical data classification and early disease prediction", 3rd IEEE International Conference on "Computational Intelligence and Communication Technology" (IEEE-CICT 2017), 978-1-5090-6218-8, pp:1-6.

[10]. Priyanka N, Dr.PushpaRaviKumar, "Usage of Data mining techniques in predicting the Heart diseases – Naïve Bayes & Decision tree", International Conference on circuits Power and Computing Technologies 978-1- 5090-4967- 7/17/$31.00 © 2017 IEEE.

[11]. Ruey Key, "Constructing Models for Chronic Kidney Disease Detection and Risk Estimation", IEEE International Symposium on Intelligent Control.