International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Feature Selection and Classification Algorithms for Chronic Disease Prediction Using Machine Learning Techniques

Saranya K R

mailme270914[at]gmail.com
https://orcid.org/0000-0003-3461-0168

Abstract: ***This paper deals the various feature selection and classification algorithms for the prediction of chronic diseases such as diabetes, cardiovascular, kidney hepatitis, hypothyroid, obesity and cancer using the machine learning techniques. The effects of feature selection and the inclusion of the clinical data on chronic disease prediction accuracy are additionally examined. Feature selection is one of the main issues in machine learning algorithms. In high-dimensional data sets, several features are all related, and a few are zero-importance or irrelevant; understanding both of these types of higher dimensional data has become a struggle and also an important issue.***

Keywords: Feature Selection, Data Classification, Disease Prediction, Chronic Disease

1. Introduction

Machine Learning is a division of Artificial Intelligence (AI) that enables the Computer to learn by itself from the readily available data without any programming. ML acts as an individual brain that helps to recognize patterns and decision making. The main aim of machine learning is that it studies the structure of the datasets and creates a model which will be easy for the people to understand. The traditional computational approaches will depend on any explicit programming but machine learning does not require any explicit programming, it learns from the input and find out patterns from them. It is widely used in lots of real time applications including virtual personal assistants, self-diving cars, Google interpret, fraud detection and mainly in healthcare industries. Machine learning in healthcare was initially used to maintain the electronic health record of the patients and automatic medical billing, but nowadays it is widely used in the diagnosis of chronic life threatening diseases like diabetes, cardiovascular diseases, kidney related diseases, hepatitis, obesity, hypothyroid and cancer.

1.1 Machine Learning and its types

The machine learning has four different learning techniques they are, supervised, unsupervised, semi-supervised and reinforcement learning.

1.1.1Supervised learning

As the name implies, it is like learning under the guidance of a teacher or a supervisor. It basically divides the dataset into training (labeled data) and testing data. Using the training data, the system will classify the test data.

1.1.2Unsupervised learning

It does not contain any labeled data; it classifies the data according to their similarities and difference, without any guidance or supervision, this method try to find out some useful patterns from the available data.

1.1.3Semi-Supervised learning

It lies between supervised and unsupervised learning and it is similar to supervised learning, because it also contains labeled data, but the main difference between the two leaning methods is the ratio between the labeled and unlabeled data. In supervised learning method the labeled data will be more compared to the unlabeled data and it is vice versa in semi-supervised learning method.

1.1.4Reinforcement learning

This method is similar to unsupervised learning because it does not contain any labeled data. But it will give award and penalty over performing the machine learning tasks correctly or incorrectly. If the model works correctly then it will get award or else penalty. Basically it learns from the previous history. Figure 1.1 depicts the different types of machine learning techniques.

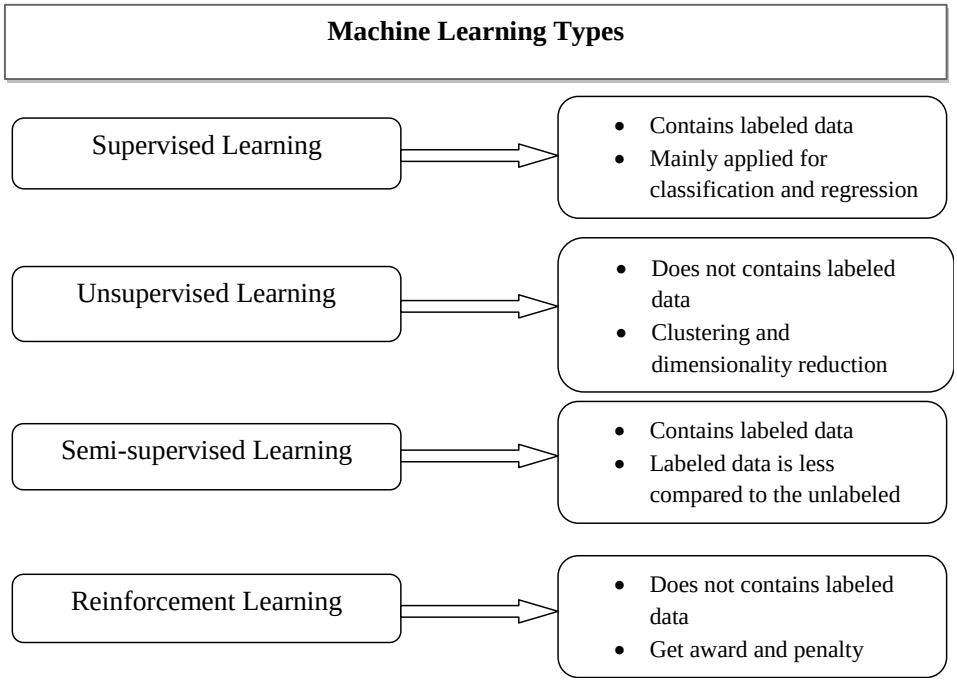

Figure 1.1: Different machine learning techniques

1.2 Feature Selection

Selecting the important features is known as feature selection or attributes selection or variable selection. The most important advantage of this feature selection is it not just improves the accuracy, but additionally, it boosts the efficiency of the classification. It can be done either manually or automatically, but selecting manually with a massive dataset is time consuming and complicated, for that reason various feature selection algorithms have been presented in machine learning. Feature selection in machine learning has different types like filter, wrapper and embedded methods. Figure 1.2 represents the different feature selection types in machine learning.

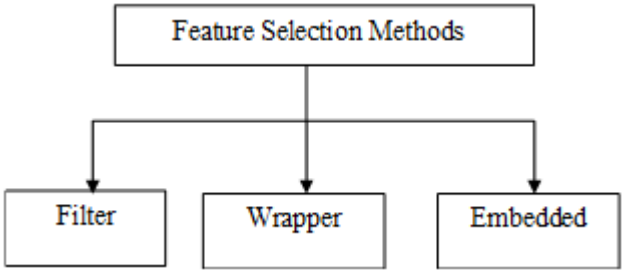

Figure 1.2: Feature selection types

1.2.1Filter method

This method makes use of statistical approaches to find the correlation or relationship between the attributes; so that the attribute with least correlation can be removed easily which in turn generates a subset of features. It does not depend on any specific machine learning algorithms. It not only reduces the time consumption but also improves the accuracy of the classification model. The filter method is faster compared to the wrapper method.

1.2.2Wrapper method

Unlike filter, it does depend on specific machine learning algorithms. It creates several models which have different feature subset. The subset which gives better performance for a specific model will be selected. The computational time will be high when the model is dealing with many features. Since it creates several models with different subset of feature it may lead to over fitting.

1.2.3Embedded method

It contains the quality of both filter and wrapper method. The feature is selected during the model building process that is the features are selected during the iteration of the model training. Thus selecting the relevant features will definitely reduce the computational time and improves the accuracy of the classification model.

1.3 Classification

Classification is a data mining process that mainly helps to assign a data item to a target class and it is also a supervised learning approach. It predicts for the target class from the observatory data or history, just such as the individual brain identifying the exact color of a thing from previously known colors. Example, classifying a person as if he/she's having or may be not having diabetes, for that number of attributes must have been taken into account for example age, insulin level, BMI, blood glucose level and blood pressure of the patients. The classifier is trained with the labeled data; once the training is done it's about to classify the test data. The three main types of classification are binary, multi-class and multi-label classification. In binary classification the final outcome will be two. E.g.: diagnosis of disease as whether the test result is positive or negative. Regarding the multi-class classification, more than two classes will be present. E.g.: classify a set of images of animal which may be dog, cat or lion, but it assigns each sample to one and only class i.e an image can be either dog or cat but not the both. In multi-labeled classification each sample is labeled to more than one target classes. E.g.: classifying the genre of various movies. For performing classification a classifier or a model has to be developed. The system will develop a classifier

using various classification algorithms like Decision Tree, Naïve Bayes, Artificial Neural Network, Support Vector Machine and K- Nearest Neighbor etc. Figure 1.3 represents the classification process.

Steps to build a classification model

Step1: Once the preprocessing work is done, the data is divided into training and test data.

Step2: Choose any classification algorithm to build the classifier

Step3: Classification algorithms are applied over the training data and a classifier is built.

Step4: Once the classifier is built using the training data, test data is given as the input for classification.

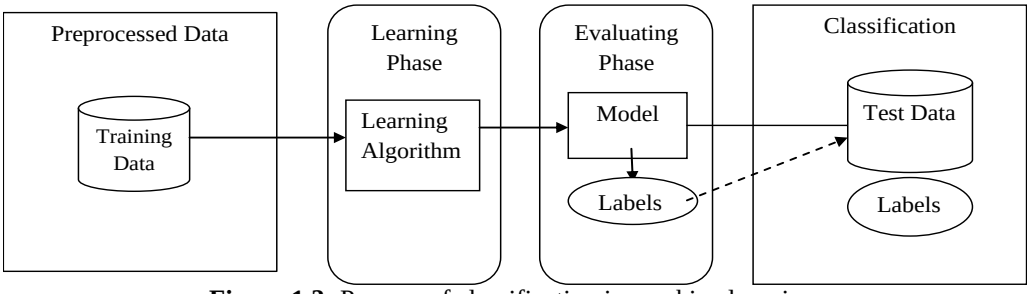

Figure 1.3: Process of classification in machine learning

It is clear from Figure 1.3 that the classification process starts with the preprocessed data and dividing it into training and test data. Since classification is a supervised learning method, the training data (labeled data) will be more than the test data (unlabeled). Then the training data is sent into learning phase which includes the classification algorithms like DT, NB, ANN, SVM and K-NN. The classification algorithm is used to build the classifier or model which is evaluated by giving the test data. To classify any data set the attributes included may be large; it may take time as well as it reduces the accuracy. To overcome that, removing the irrelevant features or selecting the relevant features is very essential.

2. Literature Review

This part covers the existing works of the applications of machine learning in chronic disease prediction such as prediction of diabetes, cardiovascular problem, kidney related disease, hepatitis, obesity, hypothyroid, lung cancer and breast cancer.

Tarun Jhaldiyal & Pawan Kumar Mishra (2014) proposed a model to predict Diabetes Mellitus using the feature selection algorithm called Principal Component Analysis. It also makes uses of Reduced Error Pruning Tree (REP) and Support Vector Machine (SVM). REP can reduce over fitting. Sneha&TarunGangil (2019) proposed a method for predicting Diabetes Mellitus at an early stage by optimal feature selection. The main aim of this method is to use the necessary features using machine learning techniques and to train various classifiers and it is found that Decision Tree algorithm and Random Forest provided the accuracy of 98.20% and 98% respectively in analyzing the diabetic datasets.

Anna Karen Garate- Escamila *et al.* (2020) developed a different classification models to predict the heart disease using feature selection methods. The datasets used for this work were Cleveland, Hungarian and Cleveland- Hungarian Heart disease dataset from UCI repository. They have implemented Chi Square test for feature selection and Principal Component Analysis (PCA) for feature extraction. For the classification purpose they make use of decision tree, gradient-boosted tree, logistic regression, multilayer Perceptron, Naïve Bayes, and random forest and for the valuation process accuracy, recall, F1 score and Matthews Correlation Coefficient (MCC) were calculated. The final results were high for Chi Square and PCA with Random forest were 98.7% for Cleveland dataset, 99% for Hungarian dataset and 99.4% for combined Cleveland and Hungarian dataset. Amin UlHaq *et al.* (2018) developed an intelligent system for the prediction of Heart disease using Machine learning algorithms. This system can classify the people with heart disease from healthy people and the dataset used for this work is Cleveland heart disease dataset from UCI repository. Relief, MRMR and Least Absolute Shrinkage Selector Operator (LASSO) were the feature selection method used to select the relevant features and the selected features are fed into the classifiers such as logistic regression, K-NN, ANN, SVM, NB and DT. The performances of the classifiers are evaluated using accuracy, sensitivity, specificity, precision, MCC, Receiver Operating Characteristic Curve (ROC) and Area Under the Curve (AUC). The classifier logistic regression with FS algorithm Relief gave the better performance in terms of accuracy with 89%.

Revathy *et al.* (2019) developed a model to predict the chronic kidney disease using machine learning techniques. Three classification algorithms have been used such as DT, SVM and Random Forest and their accuracies were evaluated. Out of the three, Random Forest achieved the highest accuracy of 99.16%. Sirage Zeynu & Shruti Patil (2018) carried out a survey on the prediction of chronic kidney disease (CKD) using classification techniques and feature selection. Before the classification process, the dataset has to be preprocessed by replacing or removing the missing values. After preprocessing relevant feature has to be selected either using wrapper or filter method. Finally the model was created using classification algorithms like K-NN, DT, ANN and J48.

Xiaolu Tian *et al.* (2019) developed a model to predict hepatitis B using machine learning techniques. The results showed the efficiency of machine learning algorithm in

predicting hepatitis B by utilizing the clinical data. EbruAydindagBayrak*et al.* (2019) carried out a study on classifying the hepatitis dataset using the machine learning classification techniques such as NB, logistic regression and DT. To improve the classification, feature selection was done by applying filter-based feature selection such as CFS subset evaluation, Information Gain attribute evaluation, and Principal Component Analysis (PCA). The performances of the classifiers were evaluated by the metrics including precision, recall, F-measure and ROC. The hepatitis dataset used for this study was obtained from UCI repository which contains 155 instances with 19 attributes and the outcome was whether the patient affected by hepatitis virus will survive or not. After the preprocessing work, the feature selection step was performed. This was mainly done to improve the performance of the classifier. CFS subset evaluation mainly evaluates the subset of attributes and also considers the degree of redundancies in the dataset and it uses best first search method. IG attribute evaluation evaluates the attributes by calculating the information gain in comparison with the class variable which make use of ranker search method. Finally the work performs the principal component analysis (PCA) which also makes use of ranker as a search method. Then the performances of the classifiers were evaluated, Naïve Bayes classifiers achieved a highest accuracy of 84.51% compared to other two classifiers.

Ankita Tyagi *et al.* (2018) developed an Interactive System using machine learning for thyroid disease prediction. The different machine learning techniques used for the disease prediction were ANN, SVM, K-NN and DT. The dataset used for this research were obtained from UCI repository and SVM achieved the maximum accuracy of 99.63%. Xiaolu Cheng *et al.* (2021) developed a machine learning and statistical method to predict relationship between obesity and physical activities. Eleven classification algorithms such as Logistic Regression, Naïve Bayes (NB), Radial Basis Function (RBF), k-nearest neighbors (KNN), classification via regression (CVR), random subspace, decision table, multiobjective evolutionary fuzzy classifier, random tree, J48, and multilayer Perceptron were implemented and their performances were compared. The datasets used for this study was obtained from National Health and Nutrition Examination Survey (NHANES, 2003 to 2006). Senthil & Ayshwarya (2018) developed lung cancer prediction methods using Feed Forward Back Propagation Neural Networks with Optimal Features. The developed system performs the classification using Neural Network and before performing the classification, feature extraction is done using Particle Swarm Optimization (PSO). The different classification algorithms such as K-NN, Bayes Network, SVM, and Neural Network were implemented and their results were compared with the proposed NN-PSO algorithm. Among all the algorithms NN-PSO achieved the maximum accuracy of 97. 8%. Nikita Rane*et al.* (2020) developed a breast cancer classification and prediction system using machine learning methods. Six different machine learning algorithms were used such as Naive Bayes (NB), Random Forest (RT), Artificial Neural Networks (ANN), Nearest Neighbour (KNN), Support Vector Machine (SVM) and Decision Tree (DT). The dataset used for this study was Wisconsin Diagnostic Breast Cancer (WDBC) dataset obtained from UCI repository.

3. Conventional Feature Selection and Classification Algorithms for Chronic Disease Prediction

This part focuses on the various conventional feature selection and classification algorithmsfor the prediction of chronic diseases such as diabetes, cardiovascular, kidney hepatitis, hypothyroid, obesity and cancer, also it contrasts the prediction accuracies obtained by traditional machine learning techniques. The effects of feature selection and the inclusion of the clinical data in chronic disease prediction accuracy are additionally examined. Feature selection is one of the main issues in machine learning algorithms. In high-dimensional data sets, several features are all related, and a few are zero-importance or irrelevant; understanding both of these types in higher dimensional data have become a struggle and also an important issue.

3.1 Feature Selection Methods and Datasets used for theExperimentation

3.1.1 Correlation based Feature Selection (CFS)

It is a filter-based feature selection method. It selects the features according to the correlation between the attributes. CFS is an automatic algorithm which does not ask the user to specify the number of features to be selected, it will select on its own. It evaluates a heuristic function based on the correlation of the attributes. The heuristic function generates a subset of features which includes attributes which are highly correlated with the class label but least or not correlated with each other. The attributes that are least correlated with the class label should be ignored and highly correlated were selected. The generated subset of features can be evaluated using the Equation (3.1):

$$M_s = \frac{a\overline{t_{cf}}}{a+a(a-1)\overline{t_{ff}}} \quad (3.1)$$

$M_s$is the evaluation function of the feature subset which contains ‘a’ number of features which are highly correlated with the class label. $\overline{t_{cf}}$is the average correlation between the feature and class label, $\overline{t_{ff}}$ is the average correlation between any two features.

3.1.2 Chi Square

It is a useful machine learning methods that can be used to find out the relationship between the features. It is a statistical procedure which helps to find out the independence of two events. The formula for calculating the Chi Square is given in the Equation (3.2).

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (3.2)$$

where, O- is the observed value, E is the expected value and c is the degree of freedom. Consider a dataset which deals with the prediction of diabetes disease which may include number of attributes like age, gender, insulin level, blood pressure, smoking habit, life style, family history and finally the class label positive or negative with the disease. Here the Chi Square is calculated for each attribute to rank them. If

we take the attribute gender with values male and female is compared with the class label. Hence the Chi Square is calculated to find out the relationship between the attribute gender and the class label. The different steps that are required to calculate the Chi Square are given below:

1) Define hypothesis
2) Contingency table
3) Need to find out the expected value
4) Next Chi Square statistic has to be calculated
5) Finally accept or reject the hypothesis

The first step deals with the hypothesis which includes two types null and alternate hypothesis. Null hypothesis means there is no relationship between two variables and on the other hand alternate hypothesis tells the relationships between the two variables. Step two deals with the contingency table which represents the relationship between two variable in rows and columns. Here the distribution of one variable will be in row and other in column. From the contingency table the degrees of freedom are calculated.

3.1.3 Information Gain (IG)

It helps to find the attribute which contains more information about the class label. The information gain of any attribute can be measured using a technique called Entropy. The entropy mainly predicts the degree of impurity of an attribute. If an attribute that highly related to the class label then it is called as 'pure' attribute or else it will be called as 'impure'. If the value of entropy is small then the attributes of the dataset are more pure and vice versa. This can be done by splitting the dataset using a random variable and the formula in Equation (3.3) helps to calculate the information gain is:

$$IG(C,A) = E(C) - E(C,A) \quad (3.3)$$

where, IG(C,A) is the information of the class variable C and the random variable A, E(C)is the entropy of the class variable and E(C,A) is the conditional entropy of both class variable and the random variable. The entropy of the class variable is calculated by the Equation (3.4).

$$E(C) = \sum_{i=1}^{C} -p_i(C) log_2 p_i(C) \quad (3.4)$$

Here p_i is the probability of class *i* in the dataset i.e. if the dataset is to predict diabetes then it may contain two classes like positive and negative, so the *i* can be either positive or negative. The conditional entropy of the class variable and the random variable can be calculated using the Equation (3.5):

$$E(C,A) = \sum_{C \in X} P(C)E(C) \quad (3.5)$$

IG is used to determine the best feature in the dataset which should contain non-zero IG value. Information Gain tells how essential a given feature is. It places the features according to their importance.

3.1.4 Gain Ratio (GR)

It is an extension of the information gain and used to normalize the IG of any particular attribute according to its entropy values. IG has problem with multi-values attributes, to overcome that Gain Ratio is used.The formula to calculate the GR is given in the Equation (3.6):

$$GainRatio = \frac{InformationGain}{Entropy} \quad (3.6)$$

From the above formula it is clear that the information gain value will be high provided if the entropy value is low and vice versa. The information gain of all the attributes have to be found and then the average of the information gain has to be computed. Then the Gain ratio has to be calculated for all the attributes, among them attribute with higher Gain ratio values has been selected to split the dataset.

3.1.5Relief

This algorithm can help to identify the quality of the attributes based upon the values that differentiate the instances that are near to each other. Basically it will rank the attributes according to their weight. Initially the weights of all the features are assigned as zero W[A] =0. This algorithm calculates two values: one is called nearest hit and another is nearest miss. First, a random instance has to be selected for example$x_i = \{x_{1i}, x_{2i}, \ldots x_{ni}\}$, then two neighbors nearest hit H, one from the same class and the nearest miss M, one from the different class has to be selected. Based on the values ofx_i, H and M it will update the weight of the features.

Relief Algorithm:

Step 1: Set all weights W[A] = 0.0;
Step 2: For i:=1 to m do begin;
Step 3: Randomly select an instance R_i;
Step 4: Find the nearest hit H and nearest miss M;
Step 5: For A:=1 to a do;
Step 6: W[A] := W[A]-diff(A,R_i,H)/m + diff(A,R_i,M)/m;
Step 7: End;

Figure 3.1 Pseudo-code for the relief algorithm

Figure 3.1 tells about the pseudo-code for the Relief algorithm, initially all the weights of the features were set as zero W [A]=0, m is the training instances that has to be selected randomly and the value of m usually given by the user. R_i is the target instance, once the random instance has been selected the nearest hit H and nearest miss M were calculated. Finally the weight of the feature has to be updated by the difference between the R_i and either nearest hit H or nearest miss M. Once the weight of all the features has been found, it is ranked according to that.

3.1.6 Symmetrical Uncertainty (SU)

It is used to measure the redundancy of the features. The dataset may have redundant features and using those features will definitely reduce the accuracy of the classification model. To overcome this issue, Symmetrical Uncertainty method is used to remove the redundant information. It makes use of the concept of both information gain and entropy. The symmetrical uncertainty is calculated using the Equations (3.7) and (3.8):

$$IG(C,A) = E(C) - E(C,A) \quad (3.7)$$

$$SU(C,A) = 2 * \frac{IG(C,A)}{[E(C)+E(A)]} \quad (3.8)$$

where, $E(C) + E(A) and IG(C, A)$ are the entropy and the information gain of the class variable and the random variable respectively.

Correlation-based Feature Selection (CFS), Chi Square, Information Gaing (IG), Gain Ratio (GR), Relief and Symmetrical Uncertainty are some of the important traditional feature selection techniques which are employed on diabetes, cardio vascular, kidney and hepatitis diseases data set in this study which is represented in Table 3.1 and the results are compared and also the performance of the classifiers are evaluated.

Table 3.1: Chronic disease datasets used for the experimentation

S. No	Dataset Name	Sample	Features
1.	PIMA Indian Diabetes	768	9
2.	Questionnaire Diabetes	952	18
3.	Hospital Frankfrut (Germany) Diabetes	2000	9
4.	SPECT Heart Data	268	23
5.	Cleveland Heart Disease	303	14
6.	Hypothyroid	1993	25
7.	Kidney Disease Detection	153	25
8.	Hepatitis Disease	80	18
9.	Obesity	2112	17
10	Lung Cancer	310	16
11.	Breast Cancer	570	32

4. Result and Analysis

This part of the paper deals with the results and analysis of the chronic disease datasets. All the chronic disease datasets mentioned in the Table 3.1 were fed to the traditional feature selection methods and their performance were evaluated using the benchmark classifiers such as DT, KNN, SVM, ANN and NB using WEKA tool. The detailed results of PIMA Indian Diabetes Dataset and the final results of remaining datasets were given below.

4.1PIMA Indian Diabetes Dataset

It was obtained from the National Institute of Diabetes and Digestive and Kidney Disease available in UCI repository. The main aim to use this dataset is to find out whether a patient is having diabetes or not. It contains 768 instances and 9 features including the class variable.

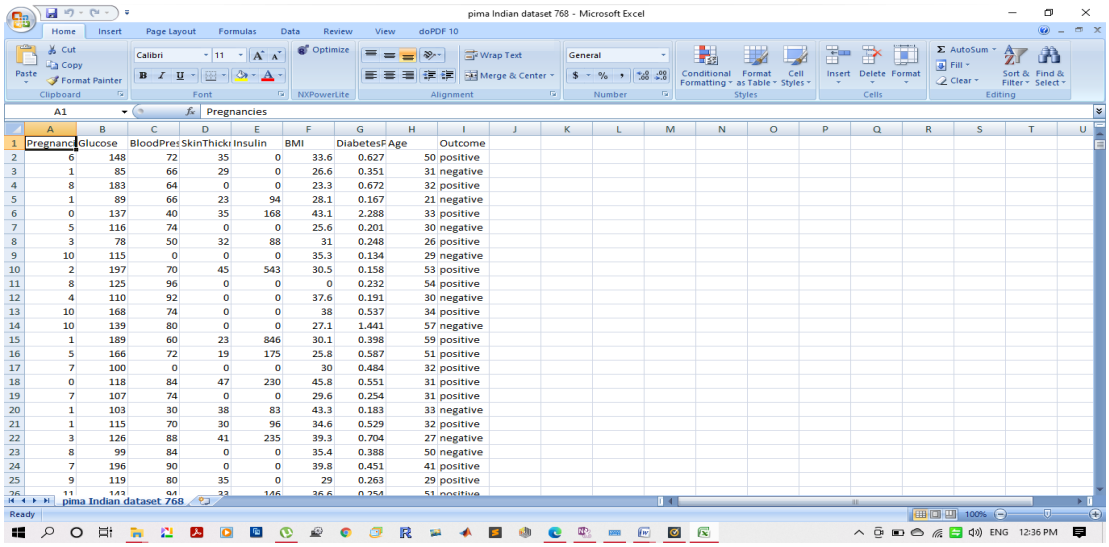

Figure 4.1: Sample PIMA Indian diabetic dataset

Figure 4.1 represents the sample PIMA Indian Diabetic Dataset and it was fed to the traditional feature selection methods and their performance were evaluated using the benchmark classifiers such as DT, KNN, SVM, ANN and NB using WEKA tool and their results are given below.

Table 4.1: Classifier accuracy on original features for PIMA Indian diabetes dataset

Classification Algorithms used	Accuracy in %	Precision	Recall	F-Score	Kappa	ROC	Time Taken in seconds
Decision Tree	73.8281	0.632	0.597	0.614	0.4164	0.751	0.21
K- Nearest Neighbor	69.1406	0.571	0.463	0.511	0.2895	0.714	0
Support Vector Machine	77.3438	0.740	0.541	0.625	0.4682	0.720	0.4
Artificial Neural Network	75.1302	0.653	0.612	0.632	0.4445	0.791	2.94
Naïve Bayes	76.3021	0.678	0.612	0.643	0.4664	0.819	0.01
Average	**74.3489**	**0.6548**	**0.565**	**0.605**	**0.417**	**0.759**	**0.7**

Table 4.2: Classifier accuracy on reduced features by CFS subset evaluation for PIMA Indian diabetes dataset
Selected Features: (2, 6, 7, 8)

Classification Algorithms used	Accuracy in %	Precision	Recall	F-Score	Kappa	ROC	Time Taken in seconds
Decision Tree	74.8698	0.668	0.556	0.607	0.445	0.791	0.02
K- Nearest Neighbor	69.9219	0.580	0.500	0.537	0.3161	0.764	0
Support Vector Machine	77.0833	0.740	0.530	0.617	0.4601	0.715	0.1
Artificial Neural Network	75.2604	0.659	0.604	0.630	0.445	0.805	1.03
Naïve Bayes	77.474	0.717	0.586	0.645	0.4823	0.829	0.01
Average	**74.9218**	**0.6728**	**0.5552**	**0.6072**	**0.4297**	**0.7808**	**0.23**

Table 4.3: Classifier accuracy on reduced features by chi square attribute evaluation for PIMA Indian diabetes dataset
Selected Features: (2, 8, 6, 5)

Classification Algorithms used	Accuracy in %	Precision	Recall	F-Score	Kappa	ROC	Time Taken in seconds
Decision Tree	74.349	0.645	0.590	0.616	0.424	0.766	0.01
K- Nearest Neighbors	71.6146	0.605	0.537	0.569	0.3586	0.783	0
Support Vector Machine	76.0417	0.712	0.526	0.605	0.4387	0.706	0.1
Artificial Neural Network	76.9531	0.702	0.590	0.641	0.4732	0.808	1.08
Naïve Bayes	75.3906	0.689	0.537	0.604	0.4292	0.813	0
Average	**74.8698**	**0.6706**	**0.556**	**0.607**	**0.4247**	**0.775**	**0.24**

Table 4.4: Classifier accuracy on reduced features by gain ratio for PIMA Indian diabetes dataset
Selected Features: (2, 6, 8, 1)

Classification Algorithms used	**Accuracy in %**	**Precision**	**Recall**	**F-Score**	**Kappa**	**ROC**	**Time Taken in seconds**
Decision Tree	74.8698	0.652	0.601	0.625	0.4367	0.753	0.01
K- Nearest Neighbors	71.3542	0.608	0.504	0.551	0.3434	0.779	0
Support Vector Machine	76.1719	0.711	0.534	0.610	0.4433	0.709	0.09
Artificial Neural Network	76.0417	0.684	0.582	0.629	0.4538	0.815	1.04
Naïve Bayes	75.5208	0.674	0.578	0.622	0.4429	0.823	0
Average	**74.7917**	**0.6658**	**0.5598**	**0.6074**	**0.4240**	**0.7758**	**0.23**

Table 4.5: Classifier accuracy on reduced features by information gain for PIMA Indian diabetes dataset
Selected Features: (2, 6, 8, 5)

Classification Algorithms used	Accuracy in %	Precision	Recall	F-Score	Kappa	ROC	Time Taken in seconds
Decision Tree	74.349	0.645	0.590	0.6161	0.424	0.766	0.02
K- Nearest Neighbors	71.6146	0.605	0.537	0.569	0.3586	0.783	0
Support Vector Machine	76.0417	0.712	0.526	0.605	0.4387	0.706	0.07
Artificial Neural Network	76.9531	0.702	0.590	0.641	0.4732	0.808	1.05
Naïve Bayes	75.3906	0.689	0.537	0.604	0.4292	0.813	0
Average	**74.8708**	**0.6706**	**0.556**	**0.6070**	**0.4247**	**0.7752**	**0.25**

Table 4.6: Classifier accuracy on reduced features by relief for PIMA Indian diabetes Dataset
Selected Features: (2, 6, 4, 1)

Classification Algorithms used	Accuracy in %	Precision	Recall	F-Score	Kappa	ROC	Time Taken in seconds
Decision Tree	74.349	0.634	0.627	0.630	0.434	0.782	0.01
K- Nearest Neighbors	69.401	0.577	0.463	0.513	0.2942	0.747	0
Support Vector Machine	76.3021	0.719	0.526	0.608	0.4438	0.708	0.07
Artificial Neural Network	75	0.678	0.541	0.602	0.4228	0.810	1.02
Naïve Bayes	75.3906	0.679	0.560	0.613	0.4354	0.821	0
Average	**74.0885**	**0.6574**	**0.5434**	**0.5932**	**0.4060**	**0.7736**	**0.22**

Table 4.7: Classifier accuracy on reduced features by symmetrical uncertainty for PIMA Indian diabetes dataset
Selected Features: (2, 6, 8, 5)

Classification Algorithms used	Accuracy in %	Precision	Recall	F-Score	Kappa	ROC	Time Taken in seconds
Decision Tree	74.349	0.645	0.590	0.6161	0.424	0.766	0.01
K- Nearest Neighbors	71.6146	0.605	0.537	0.569	0.3586	0.783	0
Support Vector Machine	76.0417	0.712	0.526	0.605	0.4387	0.706	0.06
Artificial Neural Network	76.9531	0.702	0.590	0.641	0.4732	0.808	1.02
Naïve Bayes	75.3906	0.689	0.537	0.604	0.4292	0.813	0
Average	**74.8708**	**0.6706**	**0.556**	**0.6070**	**0.4247**	**0.7752**	**0.22**

Table 4.1 represents the classifiers performance on the full dataset i.e. before performing feature selection. Table 4.2 to 4.7 represents the classifiers performance after performing feature selection using the traditional feature selection methods like CFS, IG, GR, Chi Square, Relief and SU.

The following observations were made from Table 4.1 to 4.7:

- Considering the Full dataset, the accuracy of DT,SVM,NB, K-NN and ANN are said to be 73.82%,

77.34%, 76.30%, 69.14% and 75.13% respectively are shown in Table 4.1.

- Table 4.2 shows the Dataset with Reduced Features by CFS are fed to DT,SVM,NB, K-NN and ANN classifiers and their accuracies are 74.86 %, 77.08%, 77.47%, 70% and 75.26% respectively.
- Table 4.3 shows the Dataset with Reduced Features by Chi Square are fed to DT,SVM,NB, K-NN and ANN classifiers and their accuracies are 74.34 %, 76.04%, 75.39%, 72% and 77% respectively.
- Table 4.4 shows the Dataset with Reduced Features by Gain Ratio are fed to DT,SVM,NB, K-NN and ANN classifiers and their accuracies are 74.86 %, 76.17%, 75.52%, 71.35% and 76.04% respectively.
- Table 4.5 shows the Dataset with Reduced Features by Information Gain are fed to DT,SVM,NB, K-NN and ANN classifiers and their accuracies are 74.35 %, 76.04%, 75.39%, 71.61% and 77% respectively.
- Table 4.6 shows the Dataset with Reduced Features by Relief are fed to DT,SVM,NB, K-NN and ANN classifiers and their accuracies are 74.35 %, 76.30%, 75.4%, 69.40% and 75% respectively.
- Table 4.7 shows the Dataset with Reduced Features by Symmetrical Uncertainty are fed to DT,SVM,NB, K-NN and ANN classifiers and their accuracies are 74.35 %, 76.04%, 75.4%, 72% and 77% respectively.

Table 4.8: Average accuracy and time taken by conventional FS for PIMA Indian diabetes dataset

Feature Selection Method	Average Accuracy in %	Time Taken in Seconds
CFS Subset Evaluation	**74.9218**	0.23
Chi Square	74.8698	0.24
Gain Ratio	74.795	0.23
Information Gain	74.8708	0.25
Relief	74.0885	**0.22**
Symmetrical Uncertainty	74.8708	**0.22**

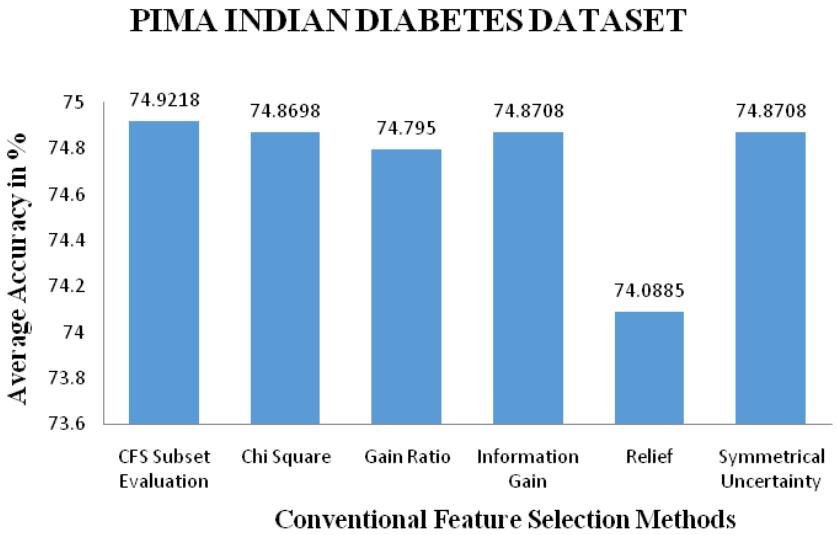

Figure 4.2: Average accuracy of the conventional FS methods for PIMA Indian diabetes dataset

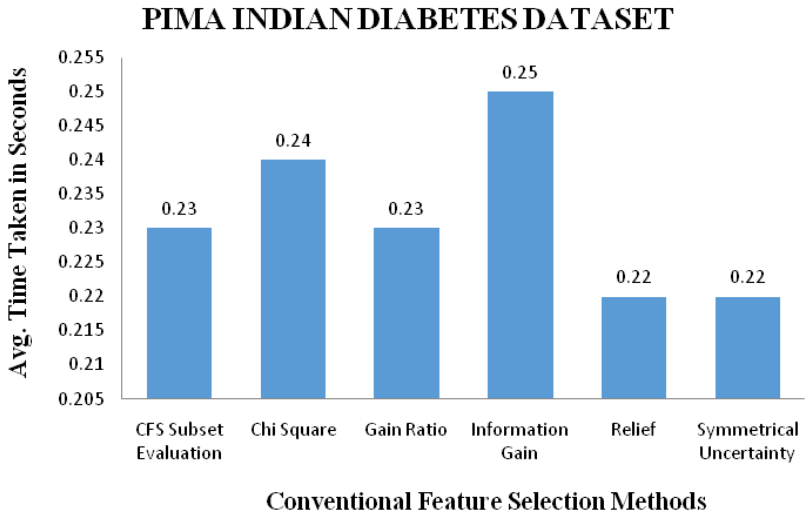

Figure 4.3: Average time taken by the conventional FS methods for PIMA Indian diabetes dataset

Table 4.8 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **CFS Subset Evaluation** achieved the maximum accuracy of **74.9218%** which is comparatively higher than the other conventional methods and **Relief and Symmetrical Uncertainty** had taken less time i.e. **2.22 seconds** to build the model and it is represented in Figure 4.2 and 4.3.

4.2 Questionnaire Diabetes Dataset (Kaggle):

This datasetcontains 952 instances with 17 attributes including the class variable. Out of the 952 participants 580 were male and 372 were female. They were asked to answer a questionnaire type of test and their response was collected.

Table 4.9: Average accuracy and time taken by conventional FS for questionnaire diabetes dataset

Feature Selection Method	Average Accuracy in %	Average Time Taken in Seconds
CFS Subset Evaluation	87.7007	**1.51**
Chi Square	91.1208	5.3
Gain Ratio	91.1217	5.23
Information Gain	91.1217	5.16
Relief	**91.1868**	5.82
Symmetrical Uncertainty	91.1208	5.24

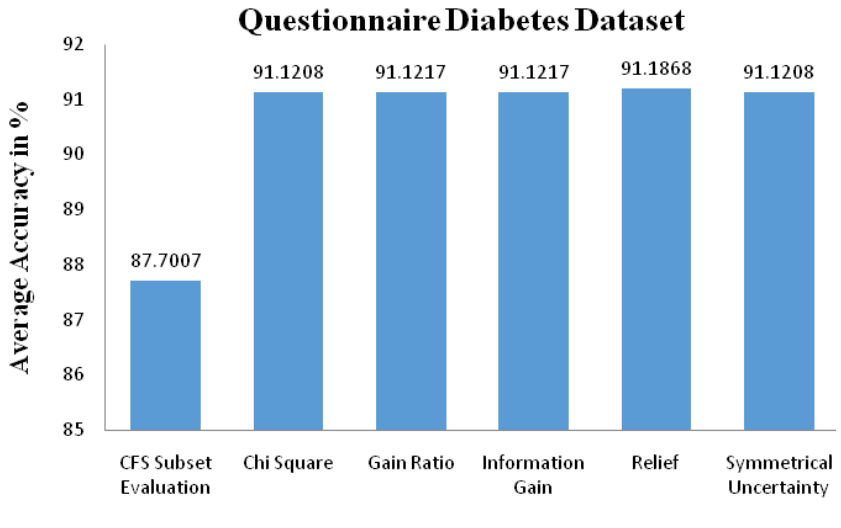

Figure 4.4: Average accuracy of the conventional FS methods for questionnaire diabetes dataset

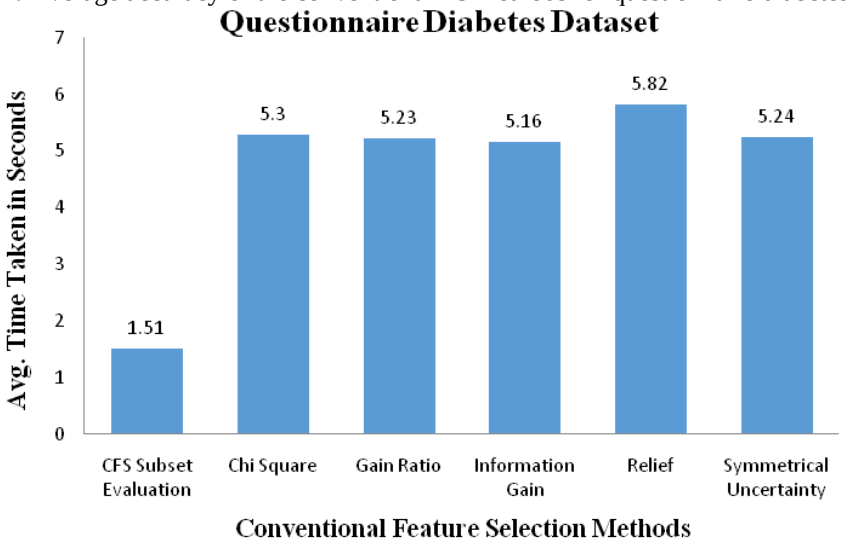

Figure 4.5: Average time taken by the conventional FS methods for questionnaire diabetes dataset

Table 4.9 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **Relief** methodachieved the maximum accuracy of **91.1868%** which is comparatively higher than the other conventional methods and **CFS Subset Evaluation** had taken less time i.e. **1.51 seconds** to build the model and it is represented in Figure 4.4 and 4.5.

4.3 Hospital Frankfrut (Germany) Diabetes Dataset

This datasetis taken from Hospital Frankfrut, Germany and it is available in Kaggle Dataset repository. It contains 2000 instances and 9 attributes.

Table 4.10: Average accuracy and time taken by conventional FS for hospital Frankfrut diabetes dataset

Feature Selection Method	Average Accuracy in %	Average Time Taken in Seconds
CFS Subset Evaluation	**84.78**	0.8
Chi Square	83.79	**0.6**
Gain Ratio	83.79	**0.6**
Information Gain	83.79	**0.8**
Relief	83.61	**0.6**
Symmetrical Uncertainty	83.79	**0.6**

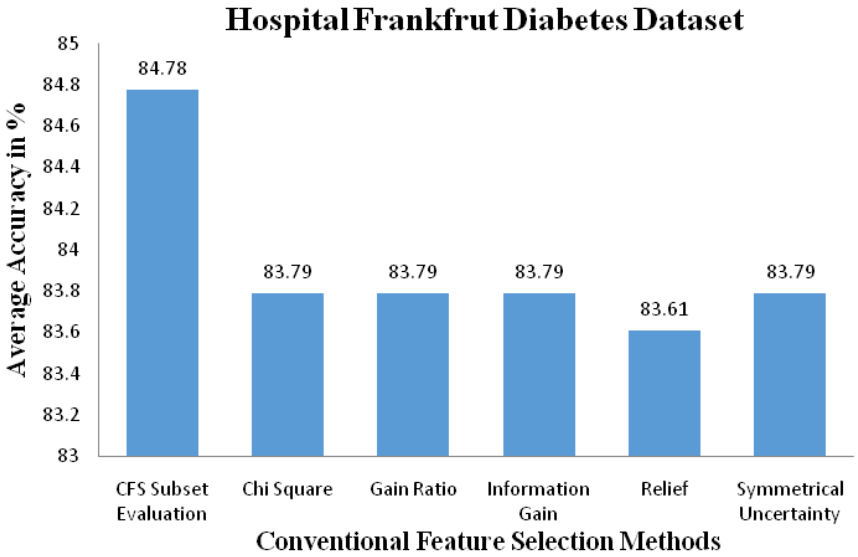

Figure 4.6: Average accuracy of the conventional FS methods for hospital Frankfrut diabetes dataset

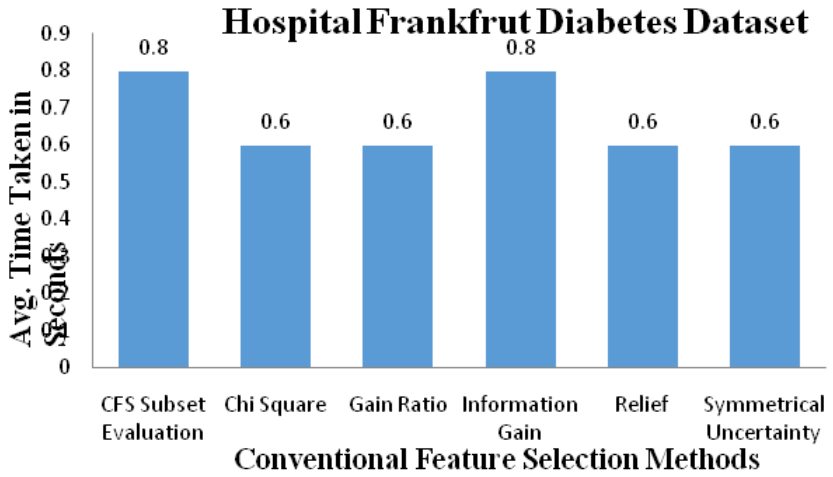

Figure 4.7: Average time taken by the conventional FS methods for hospital Frankfrut diabetes dataset

Table 4.10 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **CFS Subset Evaluation** methodachieved the maximum accuracy of **84.78%** which is comparatively higher than the other conventional methods and except CFS Subset Evaluation all other method taken less time of **0.6 seconds** to build the model and it is represented in Figure 4.6 and 4.7.

4.4 Single Proton Emission Computed Tomography (SPECT) Heart Disease Dataset

It consists of diagnosis of cardiac Single Proton Emission Computed Tomography (SPECT) images and available in UCI repository. It contains two categories which classifies the patients as normal and abnormal. Initially it contained 44 feature patterns with continuous values, which was further processed to obtain 22 feature patterns with binary values. The SPECT Heart Data contains 268 instances and 23 attributes (F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11, F12, F13, F14, F15, F16, F17, F18, F19, F20, F21, F22, Diagnosis (Class)) including the class variable.

Table 4.11: Average accuracy and time taken by conventional FS for SPECT heart disease dataset

Feature Selection Method	Average Accuracy in %	Average Time Taken in Seconds
CFS Subset Evaluation	81.11	0.3
Chi Square	**81.81**	0.3
Gain Ratio	78.72	0.3
Information Gain	78.64	0.3
Relief	80.51	0.3
Symmetrical Uncertainty	78.57	0.3

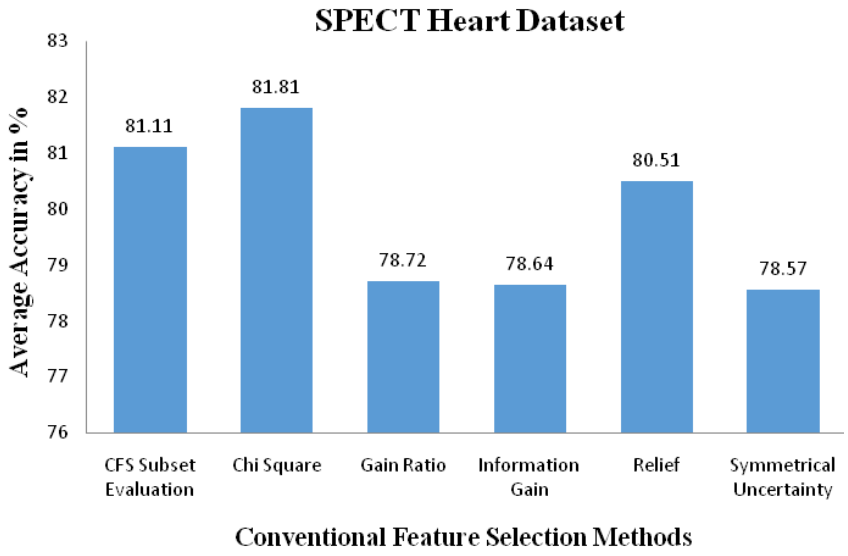

Figure 4.8: Average accuracy of the conventional FS methods for SPECT heart disease dataset

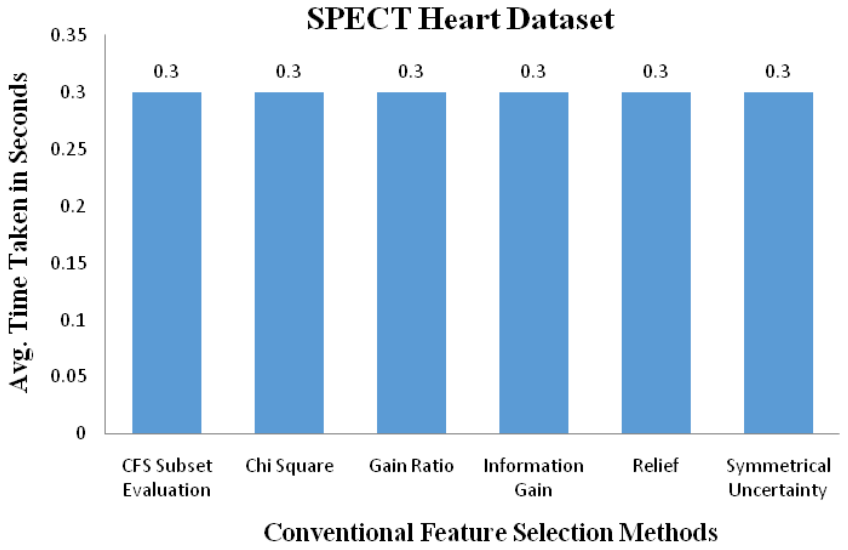

Figure 4.9: Average time taken by the conventional FS methods for SPECT heart disease dataset

Table 4.11 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **Chi Square** methodachieved the maximum accuracy of **81.81%** which is comparatively higher than the otherconventionalmethods and all methods took **0.3 seconds** to build the model and it is represented in Figure 4.8 and 4.9.

4.5 Cleveland Heart Disease Dataset

This datasetis obtained from UCI repository which consists of 14 attributes including a class attribute which classify the patient is having heart problem or not. It contains a total of 303 samples.

Table 4.12: Average accuracy and time taken by conventional FS for Cleveland heart disease dataset

Feature Selection Method	Average Accuracy in %	Average Time Taken in Seconds
CFS Subset Evaluation	81.26	0.4
Chi Square	81.1	0.47
Gain Ratio	81.07	0.4
Information Gain	81.1	**0.39**
Relief	**82.42**	0.5
Symmetrical Uncertainty	81.1	0.47

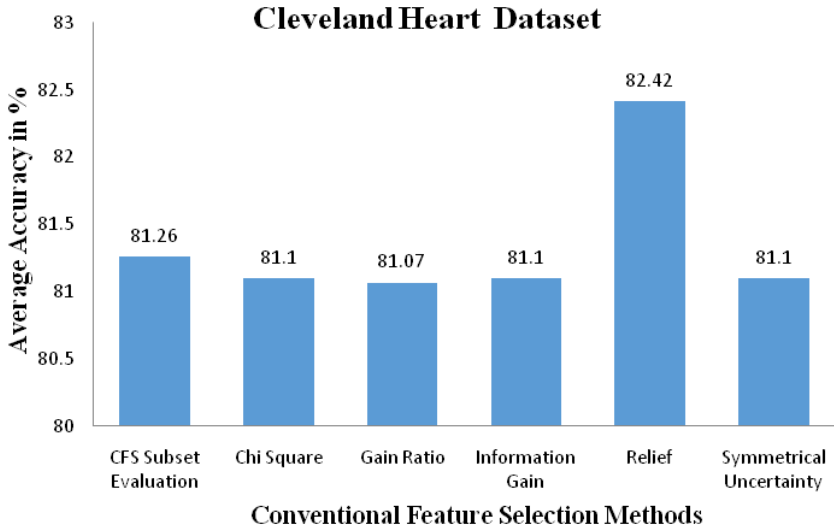

Figure 4.10: Average accuracy of the conventional FS methods for Cleveland heart disease dataset

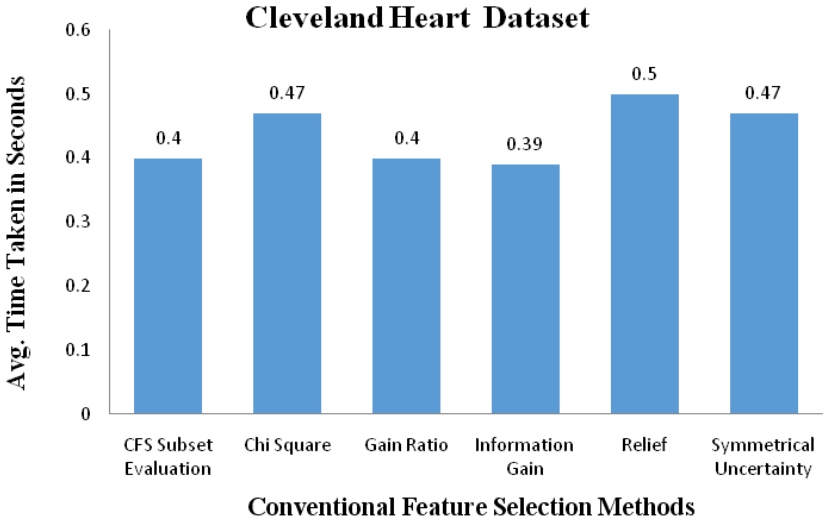

Figure 4.11: Average time taken by the conventional FS methods for Cleveland heart disease dataset

Table 4.12 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **Relief** methodachieved the maximum accuracy of **82.42%** which is comparatively higher than the other conventional methods and Information Gain took **0.3 seconds** to build the model and it is represented in Figure 4.10 and 4.11.

4.6 Hypothyroid

It was obtained from Kaggle data repository and it contains 25 attributes including a class attribute and 1993 instances. The class attribute classify whether the patients having hypothyroid or not.

Table 4.13: Average accuracy and time taken by conventional FS for hypothyroid dataset

Feature Selection Method	Average Accuracy in %	Average Time Taken in Seconds
CFS Subset Evaluation	**97.7209**	**0.7**
Chi Square	97.6004	1.2
Gain Ratio	97.7108	1.4
Information Gain	97.6004	1.1
Relief	97.5703	1.1
Symmetrical Uncertainty	97.7108	1.1

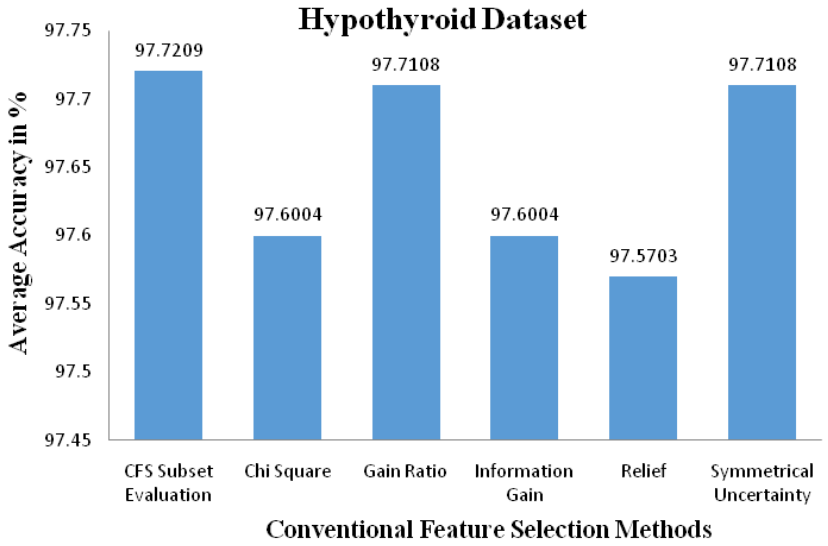

Figure 4.12: Average accuracy of the conventional FS methods for hypothyroid dataset

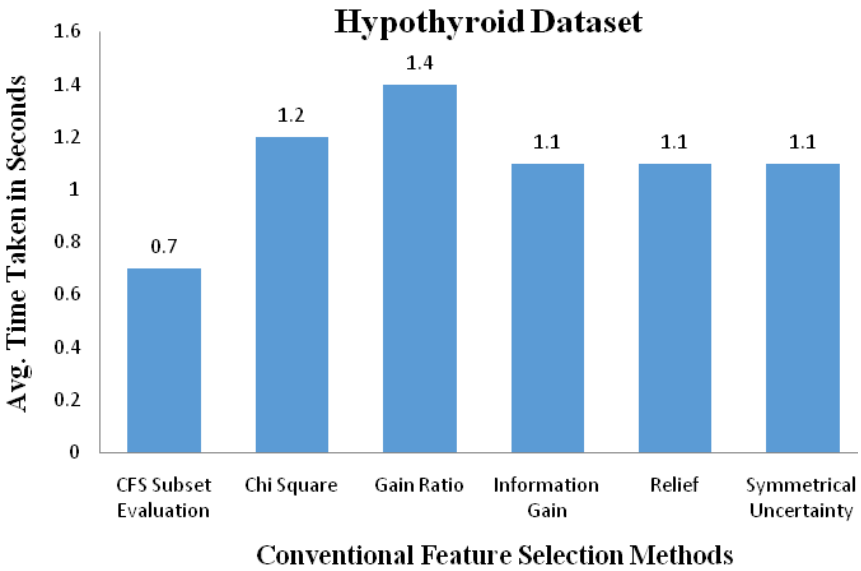

Figure 4.13: Average time taken by the conventional FS methods for hypothyroid dataset

Table 4.13 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **CFS Subset Evaluation** methodachieved the maximum accuracy of **97.7209%** which is comparatively higher than the other conventional methods and also took **0.11 seconds** to build the model and it is represented in Figure 4.12 and 4.13.

4.7 Kidney Disease Detection Dataset

It can be used to diagnose the patients with chronic kidney disease and it is obtained from Kaggle Dataset repository. It contains 153 instances with 24 attributes and one class variable.

Table 4.14: Average accuracy and time taken by conventional FS for kidney disease dataset

Feature Selection Method	Average Accuracy in %	Average Time Taken in Seconds
CFS Subset Evaluation	**99.34**	**0.1**
Chi Square	98.8	0.2
Gain Ratio	98.9	0.2
Information Gain	98.8	0.2
Relief	99.20	0.2
Symmetrical Uncertainty	98.9	0.2

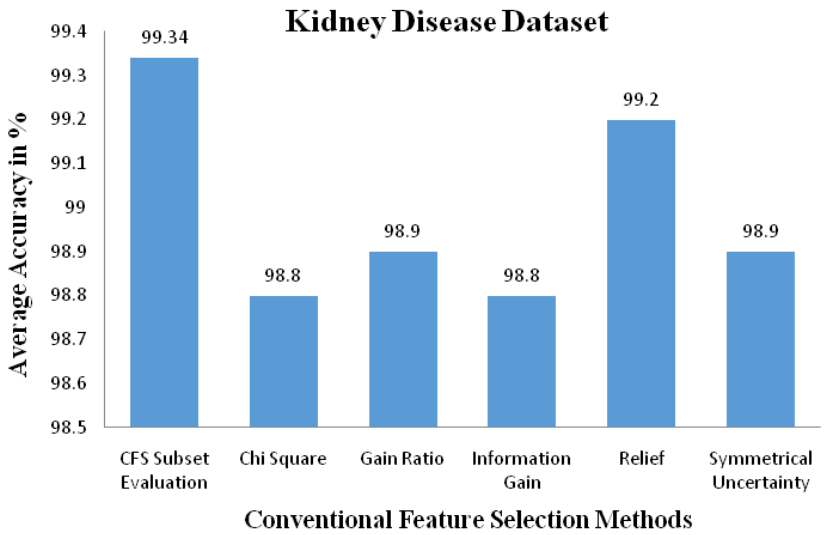

Figure 4.14: Average accuracy of the conventional FS methods for kidney disease dataset

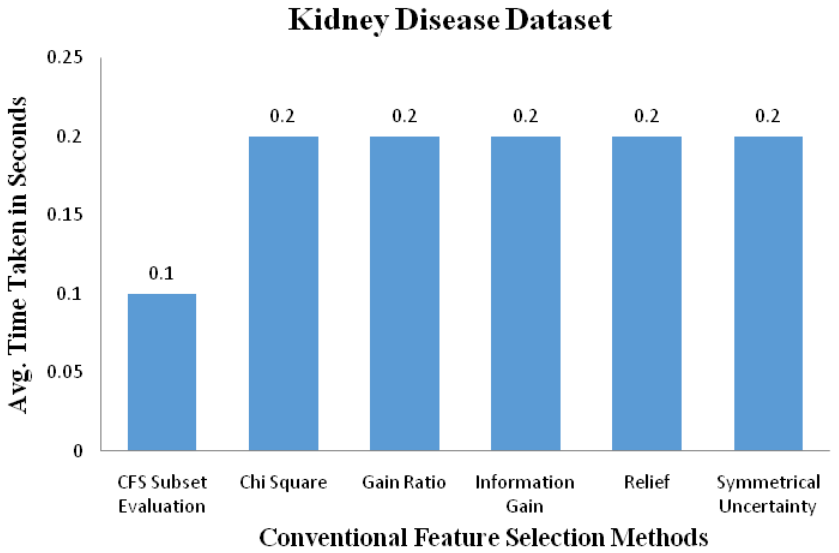

Figure 4.15: Average time taken by the conventional FS methods for kidney disease dataset

Table 4.14 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **CFS Subset Evaluation** method achieved the maximum accuracy of **99.34%** which is comparatively higher than the other conventional methods and it took **0.1 seconds** to build the model and it is represented in Figure 4.14 and 4.15.

4.8 Hepatitis Disease Dataset: This dataset is obtained from Kaggle Dataset repository which includes 80 instances and 18 attributes.

Table 4.15: Average accuracy and time taken by conventional FS for hepatitis disease dataset

Feature Selection Method	Average Accuracy in %	Average Time Taken in Seconds
CFS Subset Evaluation	85.56	0.2
Chi Square	84.03	0.2
Gain Ratio	82.02	0.1
Information Gain	84.03	0.2
Relief	79.49	0.1
Symmetrical Uncertainty	84.03	0.2

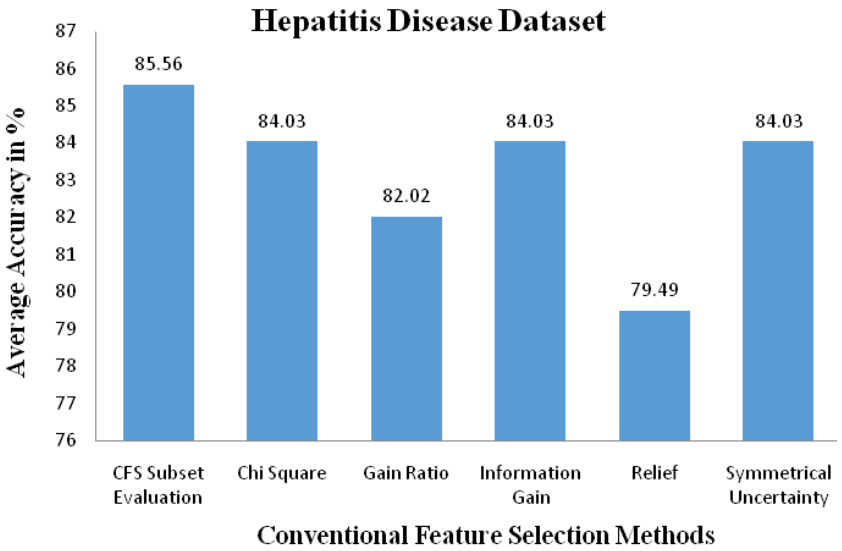

Figure 4.16: Average accuracy of the conventional FS methods for hepatitis disease dataset

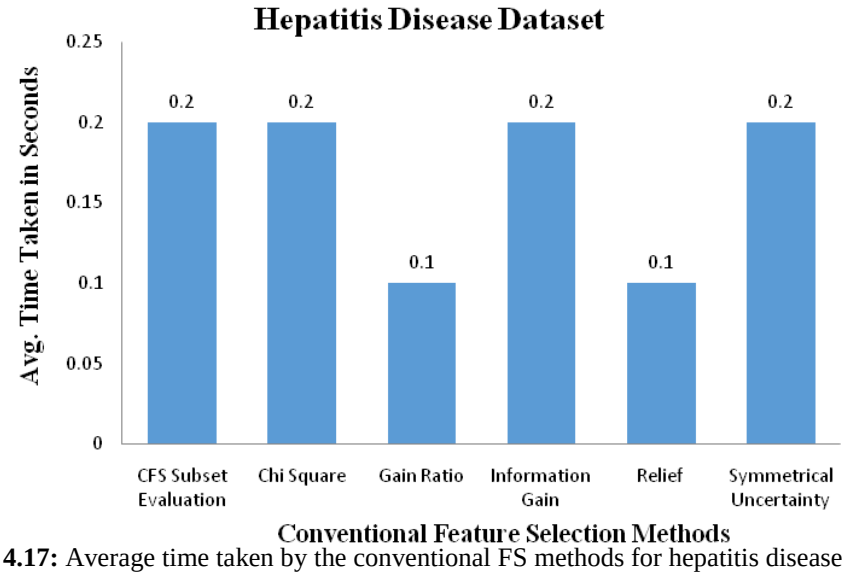

Figure 4.17: Average time taken by the conventional FS methods for hepatitis disease dataset

Table 4.15 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **CFS Subset Evaluation** methodachieved the maximum accuracy of **85.56%** which is comparatively higher than the other conventional methods and **Gain Ratio and Relief** has taken less time of **0.1 seconds** to build the model and it is represented in Figure 4.16 and 4.17.

4.9 Obesity Dataset

This dataset was obtained from Kaggle Dataset repository and this dataset is based upon the eating habits and physical conditions of the individual from Colombia, Peru and Mexico. It contains 17 attributes with 2112 samples. The class attribute classify the patients as Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

Table 4.16: Average accuracy and time taken by conventional FS for obesity dataset

Feature Selection Method	Average Accuracy in %	Average Time Taken in Seconds
CFS Subset Evaluation	**80.82**	5.2
Chi Square	75.22	3.09
Gain Ratio	77.22	3.56
Information Gain	75.22	**2.7**
Relief	76.83	4.33
Symmetrical Uncertainty	77.94	2.9

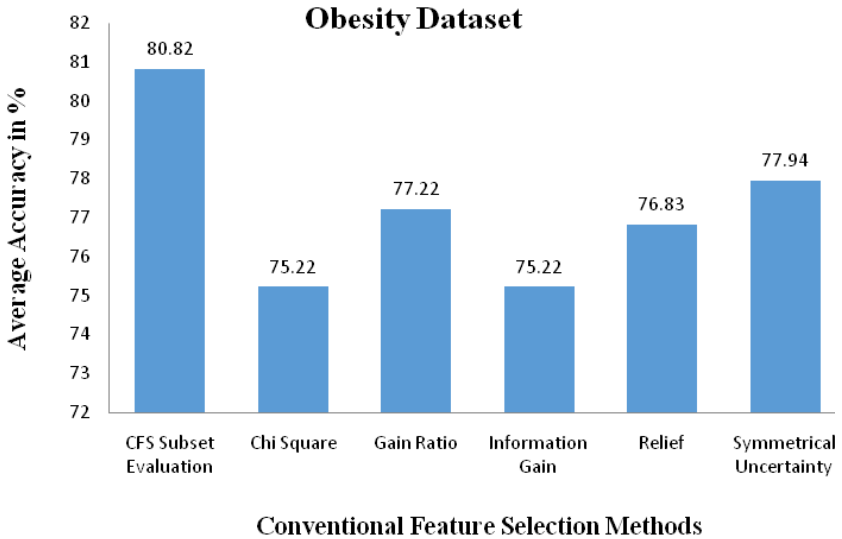

Figure 4.18: Average accuracy of the conventional FS methods for obesity dataset

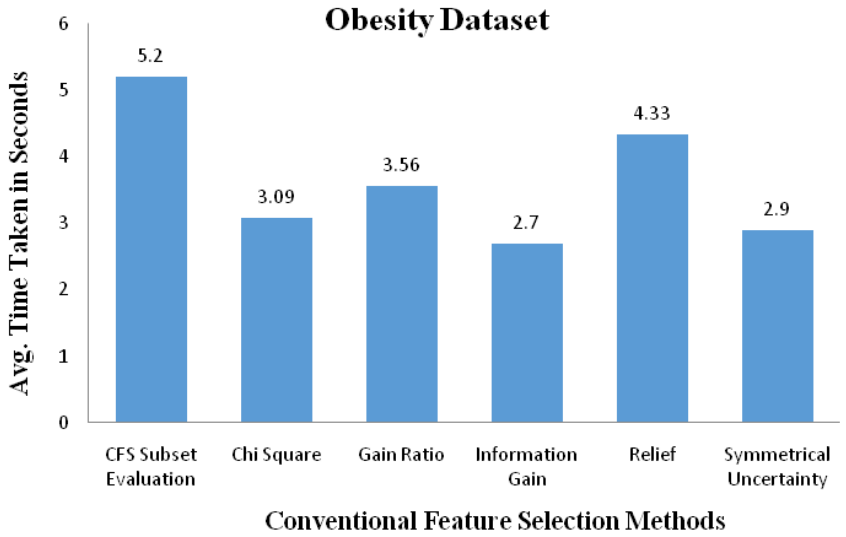

Figure 4.19: Average time taken by the conventional FS methods for obesity dataset

Table 4.16 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **CFS Subset Evaluation** methodachieved the maximum accuracy of **80.82%** which is comparatively higher than the other conventional methods and **Information Gain** has taken less time of **2.7 seconds** to build the model and it is represented in Figure 4.18 and 4.19.

4.10 Lung Cancer Dataset: Ithas been obtained from Kaggle dataset repository which includes 16 attributes and 310 samples. This dataset was collected as a survey from different patients by answering different questions.

Table 4.17: Average accuracy and time taken by conventional FS for lung cancer dataset

Feature Selection Method	Average Accuracy in %	Average Time Taken in Seconds
CFS Subset Evaluation	87.63	**0.11**
Chi Square	86.86	0.4
Gain Ratio	86.86	0.4
Information Gain	86.86	0.4
Relief	**88.03**	0.2
Symmetrical Uncertainty	86.86	0.4

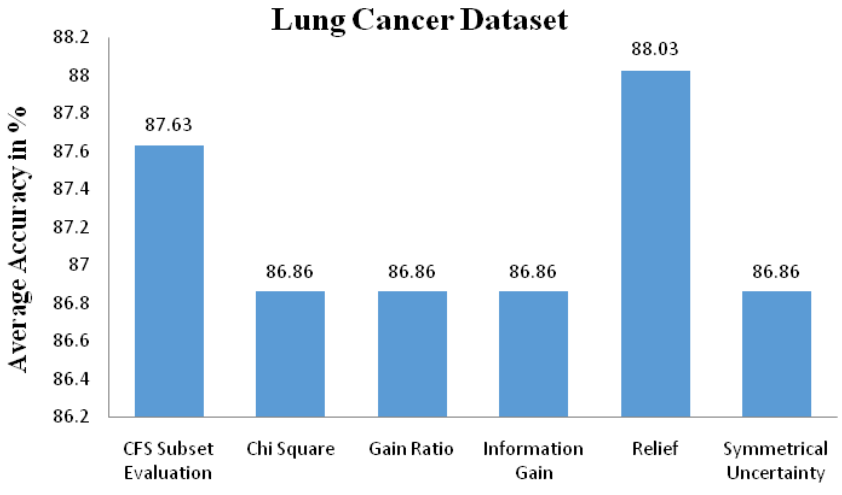

Figure 4.20: Average accuracy of the conventional FS methods for lung cancer dataset

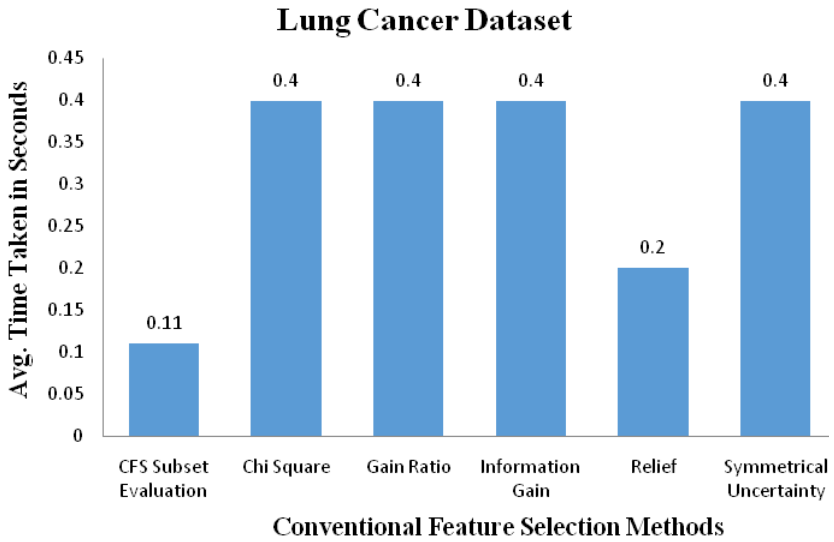

Figure 4.21: Average time taken by the conventional FS methods for lung cancer dataset

Table 4.17 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **Relief** methodachieved the maximum accuracy of **88.03%** which is comparatively higher than the other conventional methods and **CFS Subset Evaluation** has taken less time of **0.11 seconds** to build the model and it is represented in Figure 4.20 and 4.21.

4.11 Breast Cancer Dataset: Ithas been obtained from UCI repository which includes 32 attributes with a class attribute and 570 instances.

Table 4.18: Average accuracy and time taken by conventional FS for breast cancer dataset

Feature Selection Method	Average Accuracy in %	Average Time Taken in Seconds
CFS Subset Evaluation	94.2706	**0.51**
Chi Square	93.7434	0.8
Gain Ratio	93.7785	0.64
Information Gain	93.7434	0.6
Relief	**95.5008**	0.61
Symmetrical Uncertainty	93.7434	0.61

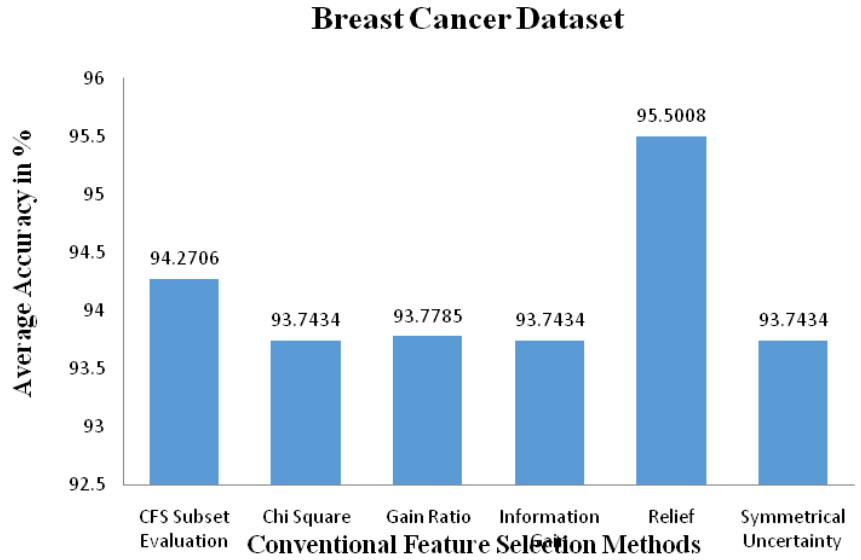

Figure 4.22: Average accuracy of the conventional FS methods for breast cancer dataset

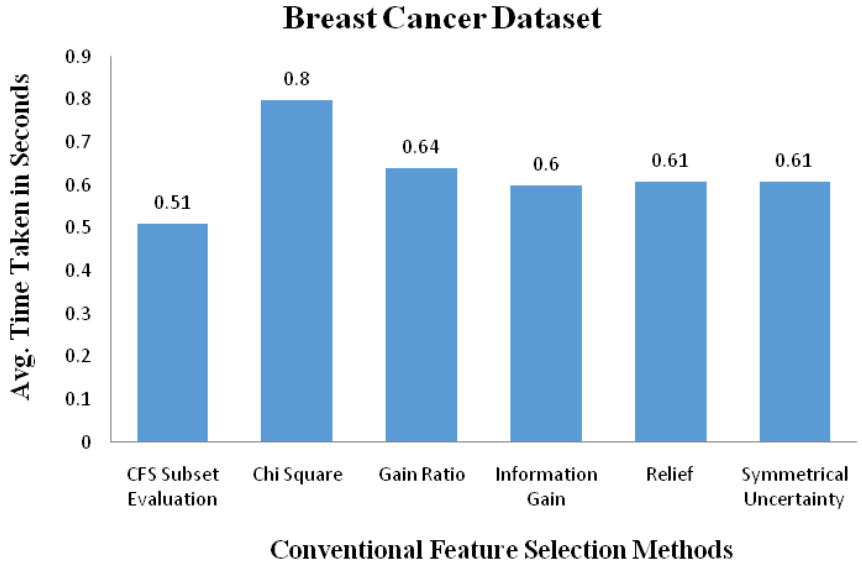

Figure 4.23: Average time taken by the conventional FS methods for breast cancer dataset

Table 4.18 represents the relationship between the conventional feature selection methods with its average accuracy and time taken to build the model. It is clear that **Relief** methodachieved the maximum accuracy of **95.5008%** which is comparatively higher than the other conventional methods and **CFS Subset Evaluation** has taken less time of **0.51 seconds** to build the model and it is represented in Figure 4.22 and 4.23.

Table 4.19: Maximum accuracy achieved by different FS methods for various datasets

S. No	Name of the Datasets	Conventional Feature Selection Methods	Maximum Accuracy in %
1.	PIMA Indian Dataset	**CFS Subset Evaluation**	74.92
2.	Questionnaire Diabetes Dataset	Relief	91.18
3.	Hospital Frankfrut Diabetes Dataset	**CFS Subset Evaluation**	84.78
4.	SPECT Heart Data	Chi Square	81.81
5.	Cleveland Heart Disease	Relief	82.42
6.	Hypothyroid	**CFS Subset Evaluation**	97.72
7.	Kidney Disease Detection Dataset	**CFS Subset Evaluation**	99.34
8.	Hepatitis Disease Dataset	**CFS Subset Evaluation**	85.56
9.	Obesity Dataset	**CFS Subset Evaluation**	80.82
10.	Lung Cancer Dataset	Relief	88.03
11.	Breast Cancer Dataset	Relief	95.50

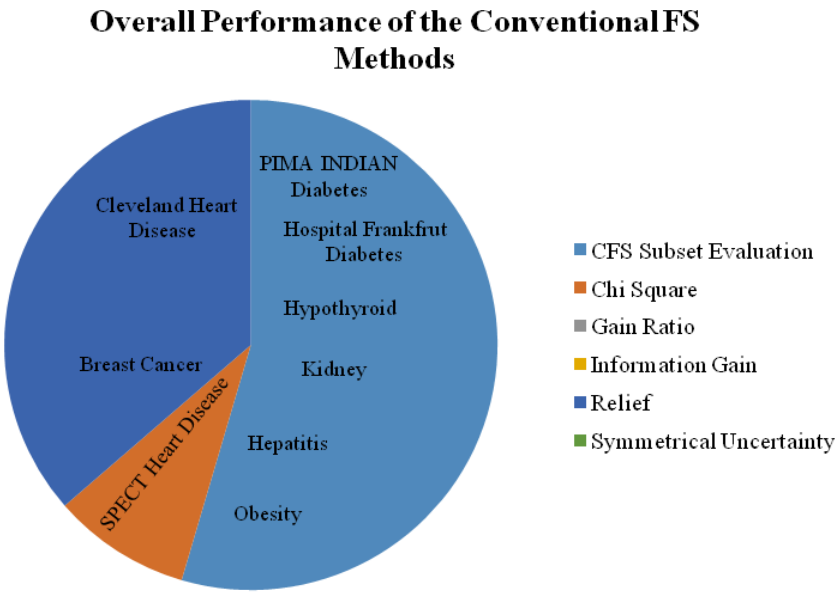

Figure 4.24: Graphical representation of the maximum accuracy achieved by different FS methods for various datasets

Table 4.20: Average time taken in seconds by the classifiers after performing feature selection using the traditional feature selection methods

Average Time Taken in Seconds							
S. No	Dataset Name	CFS	Chi Square	Gain Ratio	Information Gain	Relief	Symmetrical Uncertainty
1.	PIMA Indian Diabetes	0.23	0.24	0.23	0.25	0.22	0.22
2.	Questionnaire Diabetes	1.51	5.3	5.23	5.16	5.82	5.24
3.	Hospital Frankfrut (Germany) Diabetes	0.8	0.6	0.6	0.8	0.6	0.6
4.	SPECT Heart Data	0.3	0.3	0.3	0.3	0.3	0.3
5.	Cleveland Heart Disease	0.4	0.47	0.4	0.39	0.5	0.47
6.	Hypothyroid	0.7	1.2	1.4	1.1	1.1	1.1
7.	Kidney Disease Detection	0.1	0.2	0.2	0.2	0.2	0.2
8.	Hepatitis Disease	0.2	0.2	0.1	0.2	0.1	0.2
9.	Obesity	5.1	3.1	3.5	2.7	4.3	2.9
10	Lung Cancer	0.1	0.4	0.4	0.4	0.2	0.4
11.	Breast Cancer	0.5	0.8	0.6	0.6	0.6	0.6
Average Time Taken		**0.9**	**1.2**	**1.2**	**1.1**	**1.3**	**1.1**

5. Conclusion

Datasets used in this study may contain vast information that may not be necessary for the disease classification, using all the information in the dataset may lead to more time consumption to avoid that Feature Selection is used. Feature Selection is either selecting relevant data or removing irrelevant data. This section of the thesis deals with the study of important conventional feature selection methods like Correlation based Feature Selection (CFS), Chi Square, Information Gain (IG), Gain Ratio (GR), Relief and Symmetrical Uncertainty. The results were computed using WEKA tool, both the results with full dataset i.e without performing the feature selection and results after performing feature selection were compared. It is clear from those tables that CFS Subset Evaluation achieved maximum accuracy for most of the datasets, which is represented in the Table 4.19 and Figure 4.24. Also after selecting the relevant features using CFS Subset Evaluation, the classification algorithms taken less time to build the classifiers and to perform classification which is represented in Table 4.20.

References

[1] TarunJhaldiyal&Pawan Kumar Mishra 2014, 'Analysis and prediction of diabetes mellitus using PCA, REP and SVM', International Journal of Engineering and Technical Research (IJETR), ISSN: 2321-0869, vol. 2, no. 8, pp. 164-166.

[2] Sneha, N &TarunGangil 2019, 'Analysis of diabetes mellitus for early prediction using optimal features selection', Journal of Big Data, vol. 6, no. 13, pp. 1-19.

[3] Anna Karen Garate-Escamila, Amir Hajjam El Hassani& Emmanuel Andres 2020, 'Classification models for heart disease prediction using feature selection and PCA', Informatics in Medicine Unlocked, vol. 19, pp. 1-11.

[4] Amin UlHaq, Jian Ping Li, Muhammad HammadMemon, Shah Nazir&Ruinan Sun 2018, 'A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms', Mobile Information Systems, vol. 1, pp. 1-21.

[5] Revathy, Bharathi, Jeyanthi& Ramesh 2019, 'Chronic kidney disease prediction using machine learning model', International Journal of Engineering and

Advanced Technology (IJEAT), vol. 9, no. 1, pp. 6364-6367.

[6] SirageZeynu&ShrutiPatil 2018, 'Survey on prediction of chronic kidney disease using data mining classification techniques and feature selection', International Journal of Pure and Applied Mathematics, vol. 118, no. 8, pp. 149-156.

[7] XiaoluTian, Yutian Chong, Yutao Huang, Pi Guo, Mengjie Li, Wangjian Zhang, Zhicheng Du, Xiangyong Li &YuantaoHao 2019, 'Using machine learning algorithms to predict hepatitis B surface antigen seroclearance', Hindawi Computational and Mathematical Methods in Medicine, vol. 1, pp. 1-7.

[8] EbruAydindagBayrak, Pinar Kirci&TolgaEnsari 2019, 'Performance analysis of machine learning algorithms and feature selection methods on hepatitis disease', International Journal of Multidisciplinary Studies and Innovative Technologies, vol. 3, no. 2, pp. 135-138.

[9] AnkitaTyagi, RitikaMehra&AdityaSaxena 2018, 'Interactive thyroid disease prediction system using machine learning technique', 5th IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC-2018), pp. 689-693.

[10] Xiaolu Cheng, Shuo-yu Lin, Jin Liu, Shiyong Liu, Jun Zhang, PengNie, Bernard, F, Fuemmeler, Youfa Wang & Hong Xue 2021, 'Does physical activity predict obesity-A machine learning and statistical method-based analysis', International Journal of Environmental Research and Public Health, vol. 18, no. 8, pp. 1-11.

[11] Senthil&Ayshwarya 2018, 'Lung cancer prediction using feed forward back propagation neural networks with optimal features', International Journal of Applied Engineering Research, vol. 13, no. 1, pp. 318-325.

[12] Nikita Rane, Jean Sunny, RuchaKanad&Sulochana Devi 2020, 'Breast cancer classification and prediction using machine learning', International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 2, pp. 576-580.