

*Hypothesis*

# Performance Evaluation of a Proposed Machine Learning Model for Chronic Disease Datasets Using an Integrated Attribute Evaluator and an Improved Decision Tree Classifier

Sushruta Mishra <sup>1,\*</sup>, Pradeep Kumar Mallick <sup>1</sup>, Hrudaya Kumar Tripathy <sup>1</sup>,  
Akash Kumar Bhoi <sup>2</sup> and Alfonso González-Briones <sup>3,4,5,\*</sup>

<sup>1</sup> Kalinga Institute of Industrial Technology, School of Computer Engineering, Deemed to Be University, Bhubaneswar 751024, India; pradeep.mallickfcs@kiit.ac.in (P.K.M.); hktripathyfcs@kiit.ac.in (H.K.T.)

<sup>2</sup> Department of Electrical and Electronics Engineering, Sikkim Manipal Institute of Technology, Sikkim Manipal University, Majitar 737136, Sikkim, India; akash.b@smit.smu.edu.in

<sup>3</sup> Research Group on Agent-Based, Social and Interdisciplinary Applications (GRASIA), Complutense University of Madrid, 28040 Madrid, Spain

<sup>4</sup> BISITE Research Group, University of Salamanca, Calle Espejo s/n, Edificio Multiusos I + D + i, 37007 Salamanca, Spain

<sup>5</sup> Air Institute, IoT Digital Innovation Hub, Calle Segunda 4, 37188 Salamanca, Spain

\* Correspondence: sushruta.mishrafcs@kiit.ac.in (S.M.); alfonsogb@ucm.es (A.G.-B.)

Received: 23 September 2020; Accepted: 4 November 2020; Published: 17 November 2020



**Abstract:** There is a consistent rise in chronic diseases worldwide. These diseases decrease immunity and the quality of daily life. The treatment of these disorders is a challenging task for medical professionals. Dimensionality reduction techniques make it possible to handle big data samples, providing decision support in relation to chronic diseases. These datasets contain a series of symptoms that are used in disease prediction. The presence of redundant and irrelevant symptoms in the datasets should be identified and removed using feature selection techniques to improve classification accuracy. Therefore, the main contribution of this paper is a comparative analysis of the impact of wrapper and filter selection methods on classification performance. The filter methods that have been considered include the Correlation Feature Selection (CFS) method, the Information Gain (IG) method and the Chi-Square (CS) method. The wrapper methods that have been considered include the Best First Search (BFS) method, the Linear Forward Selection (LFS) method and the Greedy Step Wise Search (GSS) method. A Decision Tree algorithm has been used as a classifier for this analysis and is implemented through the WEKA tool. An attribute significance analysis has been performed on the diabetes, breast cancer and heart disease datasets used in the study. It was observed that the CFS method outperformed other filter methods concerning the accuracy rate and execution time. The accuracy rate using the CFS method on the datasets for heart disease, diabetes, breast cancer was 93.8%, 89.5% and 96.8% respectively. Moreover, latency delays of 1.08 s, 1.02 s and 1.01 s were noted using the same method for the respective datasets. Among wrapper methods, BFS' performance was impressive in comparison to other methods. Maximum accuracy of 94.7%, 95.8% and 96.8% were achieved on the datasets for heart disease, diabetes and breast cancer respectively. Latency delays of 1.42 s, 1.44 s and 132 s were recorded using the same method for the respective datasets. On the basis of the obtained result, a new hybrid Attribute Evaluator method has been proposed which effectively integrates enhanced K-Means clustering with the CFS filter method and the BFS wrapper method. Furthermore, the hybrid method was evaluated with an improved decision tree classifier. The improved decision tree classifier combined clustering with classification. It was validated on 14 different chronic disease datasets and its performance was recorded. A very optimal and consistent classification performance was observed. The mean values for accuracy, specificity, sensitivity and f-score metrics were 96.7%, 96.5%, 95.6% and 96.2% respectively.

**Keywords:** attribute selection; wrapper; filter; classification; regression; chronic diseases; diabetes; heart disease; breast cancer; decision tree; K-means clustering

---

## 1. Introduction

Chronic related disorders tend to create long term health issues in patients where the immune system and related body functionalities are weakened, and it usually ends in restricting their autonomy and freedom [1]. Normal lifestyle gets severely affected and the cost of healthcare steadily rises. Apart from this, it also affects the emotional stability of the patients. Provided a proper diagnosis, some chronic disorders are curable, but the treatment is usually very lengthy and painful [1]. Nevertheless, the majority of these diseases are not yet curable. The available drugs help palliate the symptoms. Consequently, the patient's lifespan can be increased [2]. Effective diagnosis and treatment is required for better public health. Among all the chronic diseases found in India, diabetes, breast cancer and heart disease are relatively common [3]. However, the treatment of these diseases is a challenging task due to the exponential rise in population. Moreover, the regular manual analysis of these diseases has led to inefficient medical decisions, given the large number of patients. Diagnosing a large number of patients with a potential chronic disease can be extremely hectic and time consuming. Further, the manual detection and diagnosis of symptoms may lead to errors. Consequently, machine learning technologies are critical to the successful mitigation of this problem [4]. The automatic classification of patients on the basis of symptoms is being gradually explored using machine learning [3,4]. Disease-related datasets can be analysed and the main symptoms that underlie a specific chronic disease can be uncovered through machine learning [4]. However, extracting the most common symptoms from raw unstructured data records is a significant challenge since this type of data reduces the efficiency of machine learning algorithms.

Attribute selection is a reliable pre-processing technique where less relevant data in an unstructured raw dataset is eliminated [5]. It acts as an optimizing tool that reduces the data size by selecting the relevant attributes [6]. This improves classification performance, reducing latency and increasing accuracy [7]. The main reasons for applying attribute selection techniques are that they:

- Enable faster execution of data mining algorithms,
- minimize the model complexity and enhance its interpretability,
- enhance the classification/prediction accuracy rate,
- reduce over-fitting,

In general, an attribute selection process is divided into four phases which include the following [8].

- Generation of subsets,
- evaluation of subsets,
- termination condition,
- validation of results.

In the first phase, a candidate attribute subset is generated using an effective search algorithm. In the next phase, every subset is evaluated on the basis of a predefined evaluation condition and verified with the previous optimal subset. It is replaced with the previous subset if it is found to be a better subset. This subset generation and evaluation procedure continues until a termination factor is reached. Finally, the selected attribute subset is verified and validated with predetermined testing data samples. The overall process is illustrated in Figure 1.

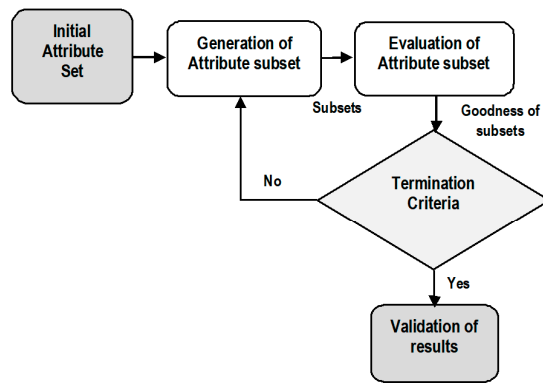


Figure 1. Attribute Selection Process.

After the generation of attribute subsets, the evaluation of these subsets is conducted on the basis of predefined criteria to compute their performance. Usually, the performance of an attribute set is determined by its ability to differentiate the distinct classes. In terms of dependency on learning classification algorithms, attribute selection is categorized into two distinct types which include wrappers and filters.

The two categories of attribute selection are Wrappers and Filters based on dependency and interrelation with the inductive learning algorithms. On the one hand, Wrapper methods compute the attributes on the basis of the accuracy of the target classification algorithm [9]. On the other hand, Filter methods calculate attributes using statistical interrelationship between an attribute subset and the target class [10]. Every attribute is assigned a rank by a scoring value. Using the ranking of attributes, the ones with the lowest ranking are eliminated from the dataset. The prediction accuracy of an induced data mining algorithm is utilized to evaluate an attribute subset in case of Wrappers. The performance of every subset of attributes is evaluated by the use of an induced classifier to the dataset with all attributes. Wrappers are implemented to generate a subset of attributes and to obtain high prediction accuracy rate. This is achieved by considering the most relevant attributes. A system model is trained by using an attribute subset that are inserted or deleted from the subset on the basis of evaluation of previous model. Therefore, it is a search problem that is computationally expensive. Figure 2 presents the wrapper model where a predictive approach is used to compute the attribute subsets.

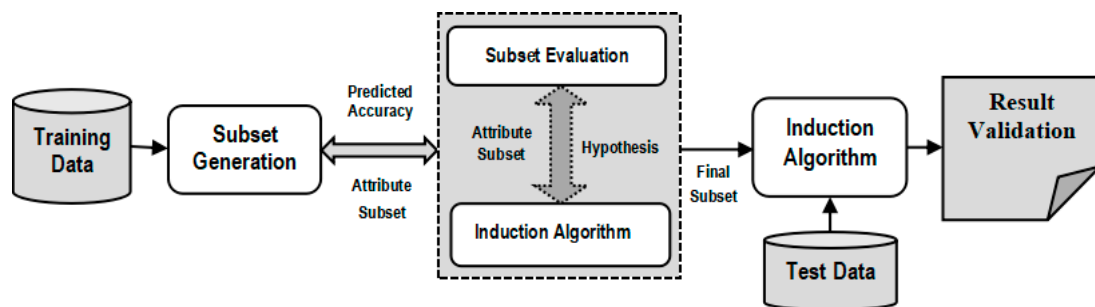


Figure 2. Wrapper Approach for Attribute Selection.

Unlike Wrappers, Filter techniques for attribute selection do not depend on an induction classifier. These techniques do not make an attribute subset. They simply used to evaluate each attribute independently and choose the best attribute set.

The intrinsic properties of data samples are used for the evaluation of attribute subsets. Figure 3 shows the filter approach for attribute selection.

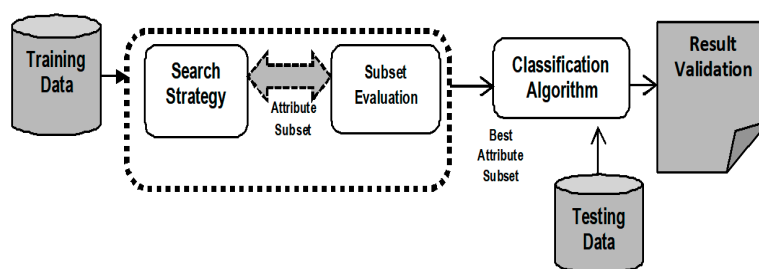


Figure 3. Filter Method for Attribute Selection.

Attributes are characterized by their relevance with respect to target class. Since induction is not applied to determine the attribute subset, the prediction accuracy rate is comparatively lower than that of Wrappers. Filters are more generalized than Wrappers since they do not depend on any induction technique for classification. In Filters, the attributes are selected on the basis of their ranking achieved in several test conditions for their association with the resultant labelled class.

This attribute selection technique can be implemented for effective disease diagnosis [11]. Inconsistencies are usually detected in the collected chronic disease datasets. Datasets of diseases such as hepatitis, lung- and heart-related diseases, breast cancer and diabetes are highly unstructured. This means that the data require pre-processing which will remove redundant and noisy features [11]. The presence of noisy features and outliers hampers the classification mining task. Sometimes, not all the symptoms present in the datasets are required for the disease classification task. Thus, the less relevant symptoms should be excluded from the final datasets. By selecting the most relevant attributes, it is possible to improve accuracy and reduce processing time [12]. In addition, in some situations, the structure of attributes is difficult to interpret. A suitable attribute optimization approach is needed for accurate the accurate treatment of diseases.

The main objective of this work is to analyse popular attribute selection techniques and the impact of filter and wrapper approaches on the evaluation of the classification of chronic diseases from datasets. Common chronic disease datasets like diabetes, heart disease and breast cancer are used to determine the impact of filters and wrappers on classification performance. The benefits and limitations of wrappers and filters are presented through experimental demonstration using the WEKA software [13]. Due to its ease of use and versatile nature, decision tree is used as classifier in the research. On the one hand, correlation feature selection (CFS), information gain (IG) and chi-square are used as filters in analysis. On the other hand, the best first search (BFS), greedy step wise search (GSS) and linear forward selection (LFS) methods are the wrappers used. Ranker algorithm is the predefined search scheme applied to filter approach while WrapperSubsetEval is the predefined search method used in combination with wrapper methods. Furthermore, on the basis of the outcome of this analysis, it was observed that CFS method provided better outcome than others while among wrapper methods, BFS performed better results compared to others. Based on this, a new hybrid three step attribute evaluator was proposed. It integrated CFS as the filter approach, BFS as the wrapper and it used a new variant of K-Means named Integrated Supervised K-Means for outliers detection. Based on the evaluated attributes, a classification model using an improved decision tree is presented. The implemented results are recorded and further analysed.

This study is divided into specific sections. This section introduces the overall theme, motivation and objective of the research work. It presents the current scenario of chronic disease treatment in India and emphasizes the importance of attribute selection in the diagnosis of chronic disorders. Section 2 presents relevant research work that has been carried out by several renowned researchers and academicians in this domain. Some widely spread chronic disease dataset details are presented in Section 3. Information on the diabetes, heart disease and breast cancer datasets is depicted in this section. The computational methods and software packages that have been used are discussed. A graphical illustration of a classification model using the filter and wrapper approach is highlighted.

The results obtained from implementing filters and wrappers are also described in detail. Section 4 presents the proposal of a new hybrid and integrated Attribute Evaluator method that can be used with an improved clustering-based decision tree for the classification of chronic disease risk disorders. The impact of the proposed attribute selection method is analysed through its implementation. Various performance parameters such as accuracy rate, latency and f-score are used to evaluate the effectiveness of classification using wrapper and filter techniques. Finally, Section 5 concludes the research by summarizing the obtained results and highlighting the most significant findings of the study.

## 2. Literature Survey

Chronic disease datasets are collected in numerous ways. Medical experts save massive quantities of patient-related data in unstructured samples [13]. Qualitative Research is conducted using various analyses based on statistics and machine learning methods. The objective of such research is to support medical decision-making. Several possible symptoms that are responsible for chronic diseases are identified. Vital factors include age, hypertension [14], total cholesterol [15], diabetes [16], hypertension, inherent heart risk factors [17], obesity, and lack of physical activity [18]. It is critical to have a practical idea of the risk parameters associated with the specific disease to support decision-making. These symptoms are extracted from datasets using attribute selection techniques such wrappers and filters. Different analyses require specific attribute selection approaches. There are numerous studies in the scope of disease diagnosis and symptom analysis using attribute selection. Different filter and wrapper methods are used to rank and select relevant attributes and eliminate less relevant attributes from disease datasets, such as lung cancer, thyroid, diabetes, kidney disease and heart disorders. In this section, the recent and relevant research applying the filter and wrapper approach to disease datasets are discussed and presented. Primarily, the classification accuracy is the criterion used for the detailed survey of different works. This literature survey analyses several relevant studies focusing on the use of filters and wrappers for the selection of relevant symptoms from chronic disease datasets. Special attention has been paid to evaluation metrics, such as accuracy rate, precision and execution time.

L.A. Simons et al. [19], applied various machine learning techniques such as decision table, RBF classifier, Multi-layer perceptron, support vector machine to coronary heart disease dataset and was able to successfully predict a Framingham risk disorder in heart disease for senior citizens in Australia. Surekha S. et al. [20], discussed the impact of wrapper methods on thyroid disease diagnosis. They compared the performance of genetic algorithm and particle swarm optimization algorithm with that of quick reduction and Johnson's reduction methods of rough set theory. Classifiers like naive Bayes and decision tree were used for thyroid classification. Results demonstrated that evolutionary approaches outperformed rough set theory methods in terms of attribute selection accuracy and minimum attribute subsets. Moreover, Rahul Deo Sah [21], evaluated the performance of vital computational techniques like support vector machine (SVM) and K Nearest neighbours (KNN) in the prediction of liver, cancer and heart disorders from healthcare disease datasets. It has been observed that SVM achieved better classification performance than the KNN algorithm. Li-Yeh Chuang et al. [22], integrated an information gain method for feature ranking with binary PSO for the generation of feature subsets and selection of gene subsets. The evaluation of the performance indicated that the developed approach selected minimal subsets of genes thereby obtaining a good classification accuracy rate. In [23], Sushruta et al., have discussed and applied vital bio-inspired optimization methods to classify various types of tumours. It used a genetic algorithm and a PSO algorithm as wrappers and implemented an MLP classifier for the categorization of tumours.

Félix Fernando González-Navarro [24], discussed an attribute selection approach to be applied to a brain tumour dataset. When evaluated according to a series of performance metrics, it was noted that a quantity of repetitive metabolisms helped in the enhancement of class label separability and its classification accuracy. Anurag Kumar Verma et al. [25], presented a new approach to computational

learning for effective prediction of skin diseases. The classification models used were Adaboost, bagging and gradient boosting methods. A new feature selection method that selects 15 most relevant attributes was used to determine an optimal feature subset. The result showed an improvement in the prediction accuracy of dermatological testing data samples. Ibrahim M. El-Hasnony [26], introduced a comparison among different data reduction methods. These data reduction methods were tested against classification algorithms like MLP and KNN algorithms for accuracy. The observation noted that the fuzzy rough feature selection method performed much better than correlation feature selection, gain ratio and principal component analysis methods. B.M. Patil [27], developed a hybrid machine learning model to predict diabetes risk disorders. It used K-Means clustering, followed by the C4.5 algorithm for classification. A very high 92.38% accuracy was obtained with this hybrid model. A predictive model for knee joint disorder detection using VAC signals was proposed by [28]. The a priori algorithm and the genetic algorithm were used as feature evaluators while LS-SVM were the classifiers used for the study. A classification accuracy of 94.31% was the outcome when LS-SVM had been integrated with the a priori algorithm and the genetic approach. Selwyn Piramuthu [29], demonstrated various inter class and probabilistic attribute selection methods based on distance computation. It was used as a pre-processing approach to induce decision trees. The results indicated that the inter class distance parameter generated optimum performance in comparison to probabilistic metrics.

Karegowda AG et al. [30], developed a hybrid classification model using the genetic algorithm as the wrapper method and the back-propagation algorithm as the classifier in detecting the presence of PIMA Indian diabetes. The hybrid model outperformed the back-propagation algorithm in classification. In [31], a cardiac arrhythmia dataset was used with Information Gain and Relief-F as filter methods. SVM and regression were the classifiers used. The performance was evaluated with accuracy, precision and recall metrics. It was observed that the Relief-F filter method combined with the SVM algorithm produced optimum result. Saeid Fallahpour et al. [32], presented a sequential floating forward selection (SFFS) wrapper method to identify an optimal feature set in mental stress dataset. SVM was used for classification and it was observed that the SFFS method combined with the SVM achieved very good accuracy when compared to other algorithms such as the genetic algorithm and information gain. The combination of Principal component analysis (PCA) with the Relief-F method was applied to a breast cancer and diabetes dataset in [33] and evaluated with the KNN algorithm. The classification performance was compared with other filter methods like correlation-based feature selection and information gain. The hybrid combination of PCA and Relief-F methods achieved a higher accuracy than others. Burak Kolukisa et al. [34], used a coronary artery disease dataset and applied an embedded model of gain ratio and chi-square filter approaches to rank attributes and remove the irrelevant features. Then, the attributes are classified with Random Forest, SVM and bagging algorithms. It was noted that the hybrid model produced an impressive result while classification with random forest achieved the highest accuracy with lowest execution time. The authors of [35], presented four different filter approaches such as gain ratio, Relief-F, information gain and chi-square methods on a tumour dataset, to select the most relevant attributes for the SVM algorithm. Classification with SVM and gain ratio gave the best performance in terms of accuracy rate and precision value.

Jianli Ding et al. [36], proposed a new embedded model that integrates information gain and a wrapper approach which used a decision tree classifier on various disease datasets like diabetes, lung disease, hepatitis and heart disease. A classification accuracy of 92.86% was achieved with the reduced attribute set. A hybrid combination of the PCA-LDA model was presented in [37], for the classification of neuron-generative diseases. Statistical attributes were ranked and optimized, and it achieved a credible accuracy rate. The authors in [38], discussed a Sequential Forward Selection (SFS) approach using 10-fold cross validation on a lung disease dataset. It was validated with SVM for classifying and it showed an impressive result with the lowest execution time on an even larger set of data samples. An extensive literature review has been carried out to analyse the existing studies associated with the scope of this research. Different authors have presented numerous attribute selection techniques for disease diagnosis. Similarly, common wrapper methods such as BFS, GSS,

Particle swarm optimization and genetic algorithm have also been discussed. The relevant studies on the use of wrappers in the classification of diseases are summarized in Table 1.

**Table 1.** An outline of relevant study on wrapper based attribute selection.

| Valuable Research Contribution  | Wrapper Method Used  | Datasets Used                          | Year |
|---------------------------------|--|--|------|
| El Akadi [39]                   | Genetic algorithm  | Dengue datasets                        | 2011 |
| Qaunz et al. [40]               | Sparse coding approach                                       | Knowledge based datasets               | 2012 |
| Hafizah et al. [41]             | GLCM and intensity histogram                                 | Kidney ultrasound images               | 2012 |
| Vieira, S.M [42]                | Modified Binary Particle Swarm Optimization (MBPSO) method   | Several disease datasets               | 2013 |
| Thananan Prasartvit [43]        | Artificial Bee Colony algorithm                              | Gene expression data                   | 2013 |
| Mokeddem S [44]                 | Genetic algorithm  | Coronary artery disease                | 2013 |
| Chen et al. [45]                | Rough set theory   | Heterogeneous medical datasets         | 2014 |
| Kora P. et al. [46]             | Improved version of bat algorithm                            | ECG signals data                       | 2015 |
| Priya M. [47]                   | Particle Swarm Optimization (PSO)                            | Liver disease                          | 2018 |
| Keerthi Priya et al. [48]       | Particle Swarm Optimization and Whale Optimization Algorithm | Breast cancer and hepatitis            | 2018 |
| Kanan et al. [49]               | ant colony optimization                                      | Face recognition dataset               | 2008 |
| Erguzel et al. [50]             | Genetic algorithm  | EEG Signal                             | 2015 |
| Guyon et al. [51]               | Support vector machine                                       | Cancer dataset                         | 2002 |
| Mustafa Serter Uzer et al. [52] | Artificial Bee Colony Algorithm                              | Liver, diabetes and hepatitis datasets | 2013 |
| E. Dogantekin et al. [53]       | linear discriminant analysis                                 | Hepatitis datasets                     | 2009 |
| L.F. Chen et al. [54]           | Particle swarm optimization                                  | obstructive sleep apnea                | 2012 |
| D. Karaboga et al. [55]         | Artificial bee colony  | Heterogeneous medical datasets         | 2011 |
| Mishra S. [56]                  | Enhanced Genetic Algorithm                                   | Diabetes datasets                      | 2020 |
| Mishra S. [57]                  | Best first search method                                     | Hepatitis datasets                     | 2020 |

Furthermore, numerous filter methods such as One-R, Relief-F, Chi-square, CFS and Gain ratio have been used in different studies. Based on different datasets, appropriate filter techniques have been employed to evaluate the goodness of attribute set. Table 2 highlights some vital research works using filter approach of attribute selection process.

**Table 2.** An outline of relevant study on wrapper based attribute selection.

| Valuable Research Contribution | Filter Method Used   | Datasets Used                                 | Year |
|--------------------------------|--|---|------|
| Kohavi et al. [58]             | Sequential forward selection                               | Thyroid dataset                               | 1997 |
| Uguz et al. [59]               | First ranker algorithm                                     | Clinical datasets for text categorization     | 2011 |
| K.K. Gandhi et al. [60]        | Correlation feature Selection                              | PIMA Indian diabetes                          | 2014 |
| Jabbar et al. [61]             | Chi-square error detection                                 | Corel images dataset                          | 2015 |
| Peker et al. [62]              | Relief-F and Sequential Forward Selection                  | Healthcare datasets                           | 2015 |
| Attia M.W. [63]                | Principal Component Analysis                               | Ultrasound kidney images                      | 2015 |
| Kavitha et al. [64]            | Principal Component Analysis                               | Heart disease dataset                         | 2016 |
| Xiao Liu et al. [65]           | ReliefF and Rough Set (RFRS) method                        | Statlog (Heart) dataset                       | 2017 |
| Haq A.U. [66]                  | Relief, minimal-Redundancy-Maximal-Relevance (mRMR)        | Heart Cleveland Dataset                       | 2018 |
| Gutlein et al. [67]            | Linear Forward Selection technique                         | High dimensional datasets from various fields | 2019 |
| Battiti R. et al. [68]         | Mutual Information   | Chronic datasets                              | 1994 |
| Verma L et al. [69]            | Correlation based feature selection                        | Coronary Artery                               | 2016 |
| Yu L. et al. [70]              | Fast correlation-based feature selection (FCBF) algorithm. | High dimensional medical datasets             | 2003 |
| Maldonado et al. [71]          | Sequential backward elimination                            | Heterogeneous datasets                        | 2009 |
| Pinar Yildirim [72]            | One-R and Consistency Based Subset Evaluation              | Hepatitis dataset                             | 2015 |
| G. Jothi et al. [73]           | Soft Set Based Feature Selection Approach                  | Lung Cancer dataset                           | 2010 |
| A. Heshmati et al. [74]        | Relief-F method  | Tumour dataset                                | 2011 |
| Ding J. et al. [75]            | Information Gain   | Breast cancer and diabetes dataset            | 2018 |

Our extensive literature survey highlights some important research in the field of disease diagnosis using machine learning. Several researchers and academicians used different computational techniques

to tackle the complexity of disease datasets. Among the different studies discussed in this section, various filter, wrapper and hybrid attribute selection methods were implemented for dimensionality reduction in chronic disease datasets. Figure 4 summarizes the overall use of these attribute selection methods discussed in our analysis.

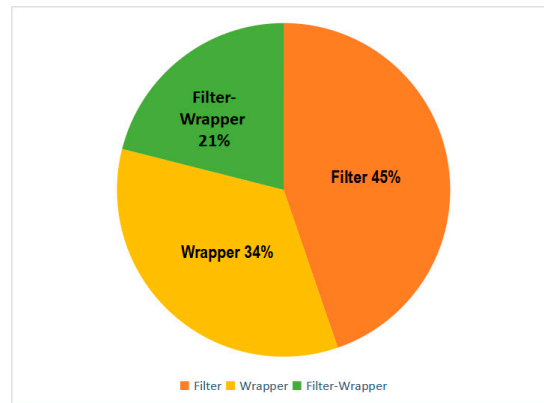


Figure 4. Distribution graph of the Filter and Wrapper approach in our literature survey.

### 3. Demonstration of Filter-Wrapper Approach in Evaluating Chronic Disease Datasets

The objective of this study is to determine the impact of the filter and wrapper approach for attribute selection on the classification performance. In this case, the focus was on the classification of disease datasets. This section presents popular chronic disease datasets used in research. System configuration requirements are provided in this section. The implemented computational techniques are presented along with the software packages. Graphs illustrating the impact of filters and wrappers on classification performance are presented. The results obtained upon the implementation of the proposed graphical data flow model are discussed.

Chronic diseases are a critical worldwide public health challenge [76]. Diabetes, hepatitis, liver disorders, lung cancer, breast cancer and heart disease are the most common chronic diseases [77]. Therefore, there is a significant number of patients suffering from these chronic diseases. Consequently, it is crucial to have a faster mechanism for effective patient support. In this study, three chronic disease datasets are used. These include diabetes, breast cancer and heart disease datasets. These are the most widely observed chronic diseases [78]. The datasets have been collected, extracted and aggregated from the UCI repository, the University of California. The PIMA Indian diabetes dataset utilized in our research constitutes a total of 8 distinct attributes along with 768 instances [79], as shown in Table 3; Table 4 depicts the breast cancer data used in this study. There was a total of 10 distinct attributes with 286 unique records [79].

Table 3. PIMA Indian dataset details [79].

| Name of Attribute | Description                           | Domain Range                |
|-------------------|---------------------------------------|-----------------------------|
| preg              | Frequency of Pregnancy                | 0–15                        |
| plas              | Concentration of Plasma glucose level | 0–199                       |
| pres              | Diastolic blood pressure              | 0–122 (mm Hg)               |
| skin              | The thickness of Triceps skin (mm)    | 0–99 (mm)                   |
| insu              | Serum insulin (2-h)                   | 0–846 (mu U/mL)             |
| mass              | Body mass index                       | 0–67.1 (kg/m <sup>2</sup> ) |
| Pedi              | Diabetes pedigree function            | 0.08–2.42                   |
| age               | Age of person                         | 21–81 years                 |
| class             | Class label                           | 0 = absence; 1 = presence   |



**Table 4.** Breast Cancer dataset details [79].

| Name of Attribute | Description  | Domain Range   |
|-------------------|--|--|
| Class             | Class label  | Non-recurrence, Recurrence   |
| age               | Age in years   | 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99                      |
| menopause         | whether the patient is pre- or post-menopausal during treatment      | ge40 or lt40 or premeno  |
| tumor-size        | Tumour size (in mm)  | 0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59.    |
| inv-nodes         | Total axillary lymph nodes that contain metastatic breast cancer     | 0–2, 3–5, 6–8, 9–11, 12–14, 15–17, 18–20, 21–23, 24–26, 27–29, 30–32, 33–35, 36–39 |
| node-caps         | If tumour penetrated in lymph node capsule                           | Yes or no  |
| deg-malig         | Histological level of the tumour                                     | 1, 2 or 3  |
| breast            | Which side of breast is affected                                     | Right or Left  |
| breast-quad       | Breast is partitioned into four quadrants with nipple as the centre. | right-up, left-up, right-low, left-low, central                                    |
| irradiat          | Patient’s radiation (X-rays) therapy history                         | Yes or no  |

The heart disease dataset, as shown in Table 5, is implemented in this analysis. It consists of 270 instances with two different class labels to detect the presence and absence of heart disease in patients [79]. The data record contains 13 unique symptom attributes.

**Table 5.** Heart Disease dataset details [79].

| Name of Attribute | Description                                  | Domain Range  |
|-------------------|--|---|
| age               | Age  | 1–100 years old   |
| sex               | The gender of a person                       | 1 = Male; 0 = Female  |
| cp                | Uncomfortable chest pain                     | General angina/non-anginal pain/asymptomatic/atypical angina/ |
| Trestbps          | blood pressure at rest                       | Measured in mm Hg after admitted to medical centre            |
| chol              | Serum Cholesterol level                      | Measured in mg/dL   |
| fbs               | Fasting period blood sugar level             | 0: <120 mg/dL; 1: >120 mg/dL                                  |
| Restecg           | Electrocardiography outcome at rest time     | Values of 0,1 or 2  |
| Oldpeak           | Exercise induced ST depression prior to rest | 3.05–3.81   |
| exang             | exercise induced angina                      | 1 = yes; 0 = no   |
| smoke             | Smoker or not                                | Value: 1 = yes; 0 = no  |
| Slope             | ST segment peak exercise slope               | 1: Upsloping; 2: Flat; 3: Downsloping                         |
| Ca                | Major vessels count                          | 0–3   |
| thal              | maximum heart rate achieved                  | 3 = normal; 6 = fixed defect; 7 = reversible defect           |

Our experiment has been conducted with 10-fold cross-validation using the attribute selection function of the Waikato Environment for Knowledge Analysis (WEKA) tool [80]. This software tool supports distinct functionalities like data visualization, data pre-processing, data classification, cluster analysis and pattern mining. The dataset which has been input in the WEKA software is in Attribute-Relation File Format (.arff) format [81,82]. A 64-bit Windows Operating system with a Quad-core processor and a minimum of 8 GB RAM is a key requirement. Moreover, Java 1.7 version or higher is required to install WEKA software. Several algorithms and methods have been used in this analysis. The algorithms used in the study have been made available in the WEKA software in the form of different packages and libraries. These packages are shown in Table 6 to give a brief introduction to the computational techniques and methods used in the study. Table 7 highlights the important packages of the WEKA software used in this analysis.

**Table 6.** Computational Methods used in the study.

| Computational Techniques                      | A Brief Description of the Methods  |
|---|---|
| Ranker Algorithm                              | It generates a ranked attribute set thereby specifying the number of attributes chosen among the ranked attribute list.   |
| Best First Search Algorithm (BFS)             | It uses a priority queue and a heuristic to find the most promising attribute and later analyse that attribute.   |
| Correlation Feature Selection Algorithm (CFS) | It determines the importance of prediction nature of an attribute and the redundancy level between a pair of attributes. Attribute sets with high class correlation and low inter-correlation are suitable.   |
| Chi-square Method                             | It demonstrates the significance of an attribute by calculating the value of the chi-squared metric in relation to the class label.   |
| Information Gain Method (IG)                  | It computes the value of an attribute by determining the net gain in information in association with the class.<br>$\text{InfoGain}(\text{Class}, \text{Attribute}) = H(\text{Class}) - H(\text{Class} \text{Attribute}).$  |
| Linear Forward Selection (LFS) Method         | It evaluates the impact of attribute selection by taking a limited attribute set. This limited attribute set is selected either by initial ordering or a ranking process. The direction of search can be forward or floating forward with selective backward steps.   |
| Greedy Step-wise Algorithm                    | It performs a greedy forward or backward search through an attribute subset space. It starts with all or no attributes and terminates when further addition or removal of attributes reduces the evaluation. It may also generate a ranked attribute set through traversal of attribute subsets and reordering the selected attributes. |
| Wrapper Subset Eval Algorithm                 | It evaluates attribute sets by using a learning approach and later a Cross validation technique to compute the classification accuracy of the learning method for a certain set of attributes.  |

**Table 7.** Algorithms and their packages in WEKA tool.

| Algorithms Used                 | Library Package Used in WEKA                    |
|---------------------------------|---|
| Ranker Algorithm                | weka.attributeSelection.Ranker                  |
| Best First Search Algorithm     | weka.attributeSelection.BestFirst               |
| CFS Algorithm                   | weka.attributeSelection.CfsSubsetEval           |
| Chi-square Method               | weka.attributeSelection.ChiSquaredAttributeEval |
| Information Gain Method         | weka.attributeSelection.InfoGainAttributeEval   |
| Decision Tree Classifier        | weka.classifiers.trees.J48                      |
| Greedy Step wise Algorithm      | weka.attributeSelection.GreedyStepwise          |
| WrapperSubsetEval Method        | weka.attributeSelection.WrapperSubsetEval       |
| Linear Forward Selection Method | weka.attributeSelection.LinearForwardSelection  |

Figure 5 demonstrates the general workflow model of the framework proposed in this study. Chronic disease datasets are the input file to which pre-processing is applied to detect and eliminate irrelevant features. Moreover, the missing values in datasets or occurrence of repetitive values are handled through data pre-processing. Missing and repetitive values are substituted with the mean value from that column. Finally, the data is ready to be processed using the attribute selection tool.

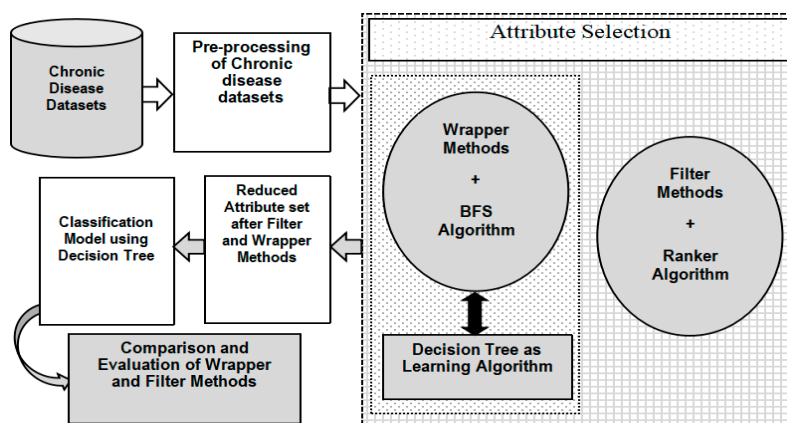


Figure 5. The Work-flow Model of Wrapper-Filter Methodology.

In this study the procedure is partitioned into two different stages. The first part is the attribute evaluator through which attribute subsets are evaluated. The second part includes the search scheme that enables the attributes to search the space of possible subsets and find the most relevant attribute subset. The attribute evaluator section is further partitioned into two different steps: the filter approach method and the wrapper approach. On the one hand, in the filter approach, attribute selection does not depend on classification algorithms. The attributes are selected on the basis of their ranking in statistical tests. This study includes three popular filter methods, namely, CFS, IG and Chi-square methods. These filter methods are used to select relevant attributes from pre-processed chronic disease datasets. Then, the attributes are ranked by the Ranker algorithm in WEKA. The Ranker method is responsible for generating a ranked attribute list for attribute evaluators. The output of this method is the reduced and optimized attribute set which is used for classification using the Decision Tree algorithm.

The second part of the attribute evaluator is the wrapper approach where a wrapper method in combination with a suitable search algorithm generates an optimal set of attributes. In wrapper methods, an attribute subset is employed in the training of a system prototype that will use those attributes. Once the predecessor model has obtained a result, further decisions are made regarding the inclusion or exclusion of attributes from the candidate subset. BFS, LFS and GSS are the three popular wrapper methods used in this analysis. These methods have been combined with WrapperSubsetEval which acts as a predefined search algorithm in WEKA. The combination of these three components generates a reduced attribute set. Once the reduced attribute set is generated through the application of the filter and wrapper approach, the next phase is the classification process where the reduced chronic disease datasets are subjected to classification using the Decision Tree algorithm. The Decision Tree algorithm has been tested on diabetes, breast cancer, and heart disease datasets.

The authors have analyzed the impact of the filter wrapper approach on the attribute selection process and its impact on the efficiency of chronic disease classification. Three filter methods have been used in this study, including the Information gain method, the Correlation feature selection method and the Chi-square method. The ranker algorithm was the predefined search algorithm in this approach. In the case of the wrapper approach, WrapperSubsetEval was the used attribute evaluator while the best first search, the linear forward selection method and the greedy step wise search method were used as search algorithms. This section presents the results obtained after implementing the proposed model in the WEKA software. This section is divided into two subsections. Firstly, the attribute significance graph has been presented to highlight the importance of individual symptoms upon the application of both filters and wrappers. The attributes are ranked on the basis of their relevance and priority. On the basis of the importance of the symptoms, the attribute significance graphs have been developed to identify the relevant and irrelevant symptoms in the dataset. Every attribute is assigned a score level out of 10 on the attribute significance graph. The X-axis represents the attributes of a chronic disease dataset while the Y-axis denotes the score level of the individual attributes that have

been obtained after the application of an attribute evaluator. According to the results shown in the graphs, the less relevant symptoms have been eliminated from the dataset. The obtained results are discussed throughout this section.

### 3.1. Comparison of Filter Approach

This subsection highlights the importance of the filter approach for chronic disease datasets. The filter approach was used to generate a reduced attribute set for different chronic disease datasets, where the attributes were the relevant symptoms. 270 heart disease dataset instances have been collected with 13 distinct attributes initially present. These samples were subjected to the filter approach. Individual attribute score can be seen from the attribute significance graph and as seen ‘ca’ and ‘Exang’ are the least relevant attributes according to the filter approach. Figure 6 highlights the attribute significance analysis concerning heart disease data. Figure 7 depicts the attribute significance graph of the Diabetes dataset that has a total 768 instances with 8 attributes. The attribute ‘skin’ generated a comparatively lower score of 0.3 with filter approach. Similarly, the filter approach was evaluated on a breast cancer dataset. The attribute significance graph is presented below. The significance score of the attribute ‘irradiat’ was the lowest, with only 0.3, in comparison to other attributes. It is highlighted in Figure 8.

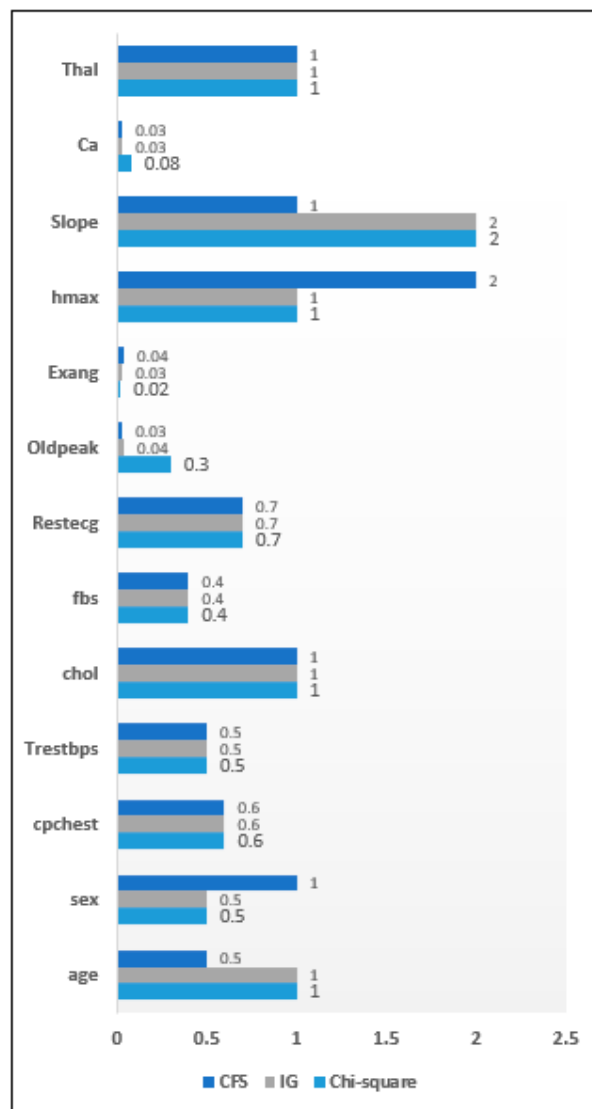


Figure 6. Attribute Significance Graph for the Heart disease dataset with Filters.

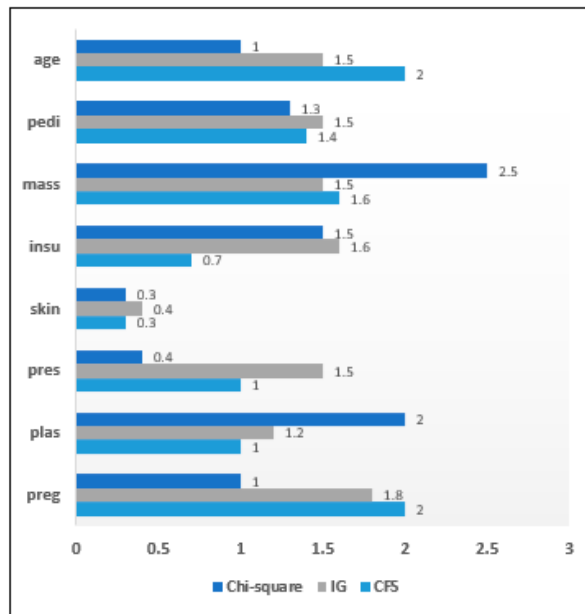


Figure 7. Attribute Significance Graph for the Diabetes dataset with Filters.

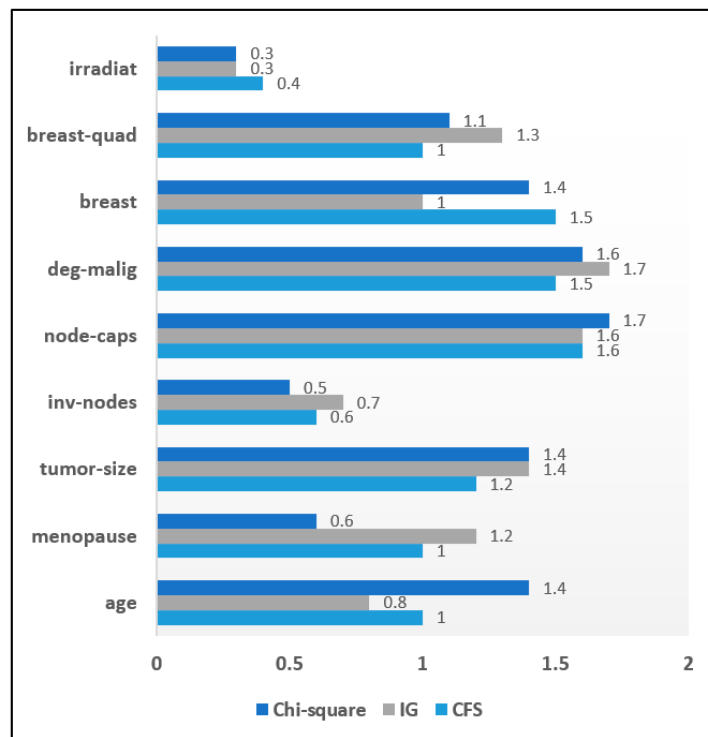


Figure 8. Attribute Significance Graph for the Breast Cancer dataset with Filters.

An attribute significance analysis was performed on all three datasets based on ranking of attributes. The low ranked attributes are given less priority. The set of relevant attributes obtained by the CFS method from the heart disease dataset was 11 while in the IG method and the chi-square method it has been reduced to 10 attributes. When filter methods were applied to diabetes data, it was seen that both the CFS and the IG methods included 7 attributes in the optimized attribute set on the basis of a priority ranking. The chi-square method included 6 attributes in the resulting set. It is observed that the CFS and chi-square methods included 7 attributes in the reduced sets of attributes

while the IG filter method of evaluation included 8 attributes in the resulting attribute set. When filter methods were applied to the breast cancer dataset, it was found that 2 attributes had less importance for both the CFS and Chi-square methods, while only 1 attribute was irrelevant for the IG method. Table 8 shows an overall analysis of the impact of filter methods on disease datasets and the sets of selected attributes.

**Table 8.** Impact of filter methods on disease datasets.

| Chronic Disease Dataset | Dataset Details       | CFS | IG  | Chi-Square |
|-------------------------|-----------------------|-----|-----|------------|
| Heart Disease Dataset   | Number of instances   | 270 | 270 | 270        |
|                         | Initial Attribute set | 13  | 13  | 13         |
|                         | Reduced Attribute set | 11  | 10  | 10         |
| Diabetes Dataset        | Number of instances   | 768 | 768 | 768        |
|                         | Initial Attribute set | 8   | 8   | 8          |
|                         | Reduced Attribute set | 7   | 7   | 6          |
| Breast Cancer Dataset   | Number of instances   | 286 | 286 | 286        |
|                         | Initial Attribute set | 9   | 9   | 9          |
|                         | Reduced Attribute set | 7   | 8   | 7          |

After determining the relevant score level of all attributes of the chronic disease datasets, the reduced and optimized attribute set is classified using the Decision Tree algorithm. The classification process has been conducted to detect the presence or absence of the chronic disorder in patients. The filter methods have been applied to classify the chronic disease datasets using the Decision Tree classifier; the obtained results are interesting. It was observed that the CFS method outperformed other filter methods concerning the accuracy rate and execution time. In this work, accuracy rate is selected as the heuristic function to rank attributes. CFS method performed better than others as it evaluates the attribute subsets which exhibit strong association with class label but least dependence on each other. Thus, weekly correlated attributes are less significant and thus dropped. With heart disease dataset, the CFS method achieved an accuracy of 93.8%. Furthermore, this method produces an 89.5% accuracy rate and only a 1.02 s delay in execution with the diabetes dataset. The CFS method also provided 96.8% accuracy and a minimum execution time delay of only 1.01 s when classified with breast cancer data. Table 9 presents a tabular view of classification accuracy and latency analysis using filter methods.

**Table 9.** Analysis of classification accuracy and latency with filter methods on disease datasets.

| Chronic Disease Dataset | Filter Method | Search Method | Classifier    | Accuracy | Latency |
|-------------------------|---------------|---------------|---------------|----------|---------|
| Heart Disease Dataset   | CFS           | Ranker        | Decision tree | 93.8%    | 1.08 s  |
|                         | IG            | Ranker        | Decision tree | 89.6%    | 1.34 s  |
|                         | Chi-square    | Ranker        | Decision tree | 91.2%    | 1.16 s  |
| Diabetes Dataset        | CFS           | Ranker        | Decision tree | 89.5%    | 1.02 s  |
|                         | IG            | Ranker        | Decision tree | 88.2%    | 1.14 s  |
|                         | Chi-square    | Ranker        | Decision tree | 88.8%    | 1.05 s  |
| Breast Cancer Dataset   | CFS           | Ranker        | Decision tree | 96.8%    | 1.01 s  |
|                         | IG            | Ranker        | Decision tree | 89.2%    | 1.09 s  |
|                         | Chi-square    | Ranker        | Decision tree | 86.8%    | 1.21 s  |

### 3.2. Comparison of the Wrapper Approach

The result of the application of the wrapper approach to the chronic disease datasets is summarized in this section. The attribute significance graph was developed for all three chronic datasets. The evaluation of wrapper methods was carried out on the heart disease data instances. The selected attribute list is shown in Figure 9 in the form of an attribute significance graph. In total 3 attributes

which include 'Ca', 'Exang', 'Oldpeak' were found to be least relevant. The attribute 'skin' was recorded with a significance score of 0.3 in the diabetes dataset and hence eliminated. Figure 10 shows an attribute significance graph for diabetes data. The wrapper approach was evaluated on the breast cancer dataset and its attribute significance graph is depicted in Figure 11. Every method achieved a score level of around 0.3, 'irradiat' has been found to be the least relevant attribute. It can be clearly seen in the graphical analysis that three attributes were relatively less relevant than the others in case of the heart disease dataset. In the diabetes dataset, only one attribute was less important with both diabetes and breast cancer dataset samples.

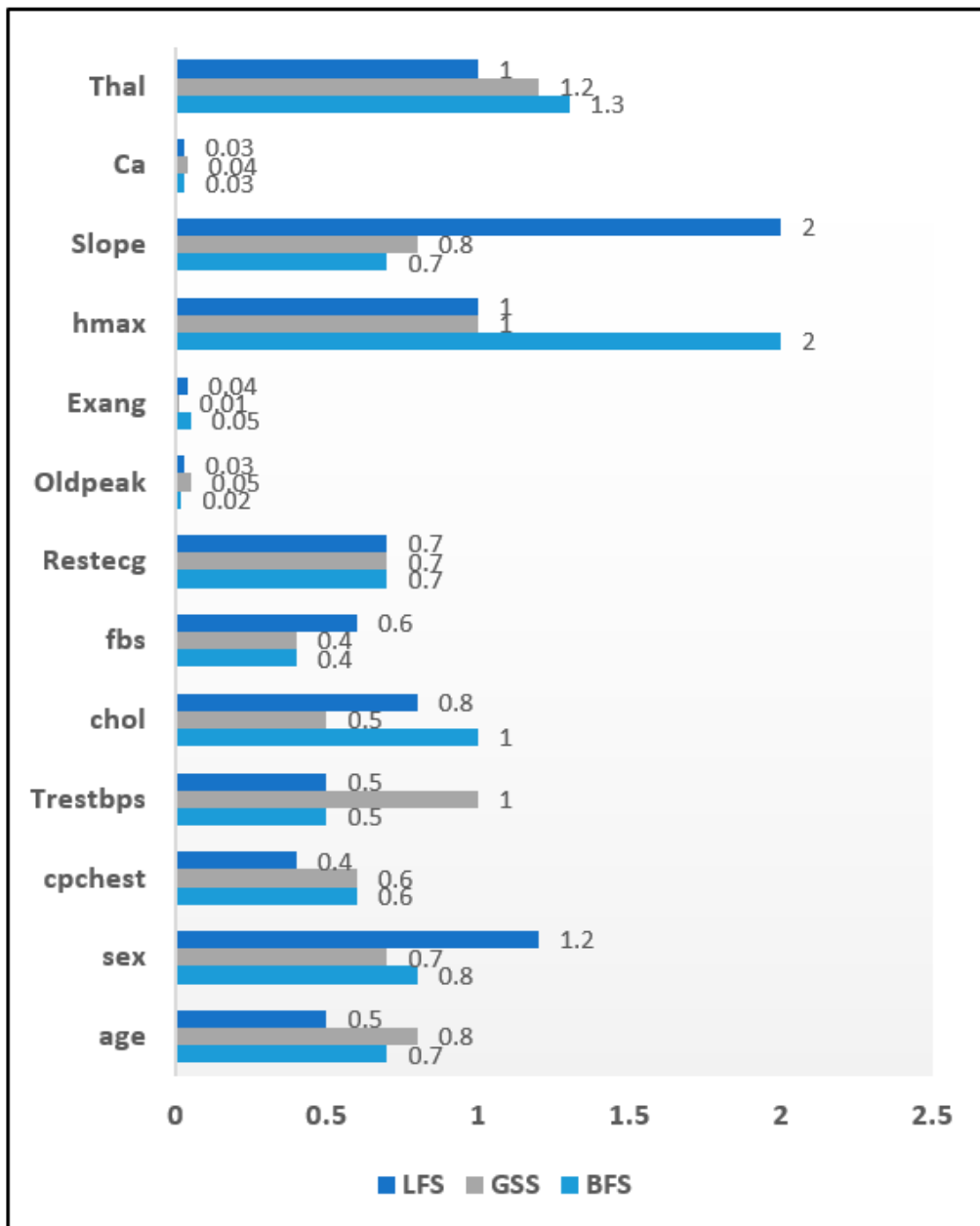


Figure 9. Attribute Significance Graph for the Heart disease dataset with Wrappers.

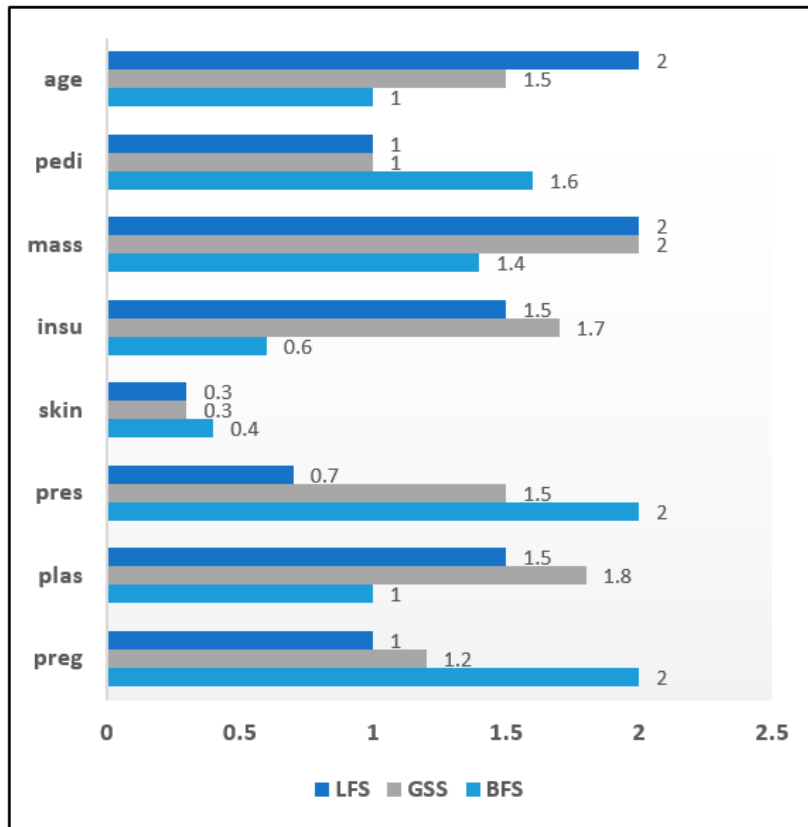


Figure 10. Attribute Significance Graph for the Diabetes dataset with Wrappers.

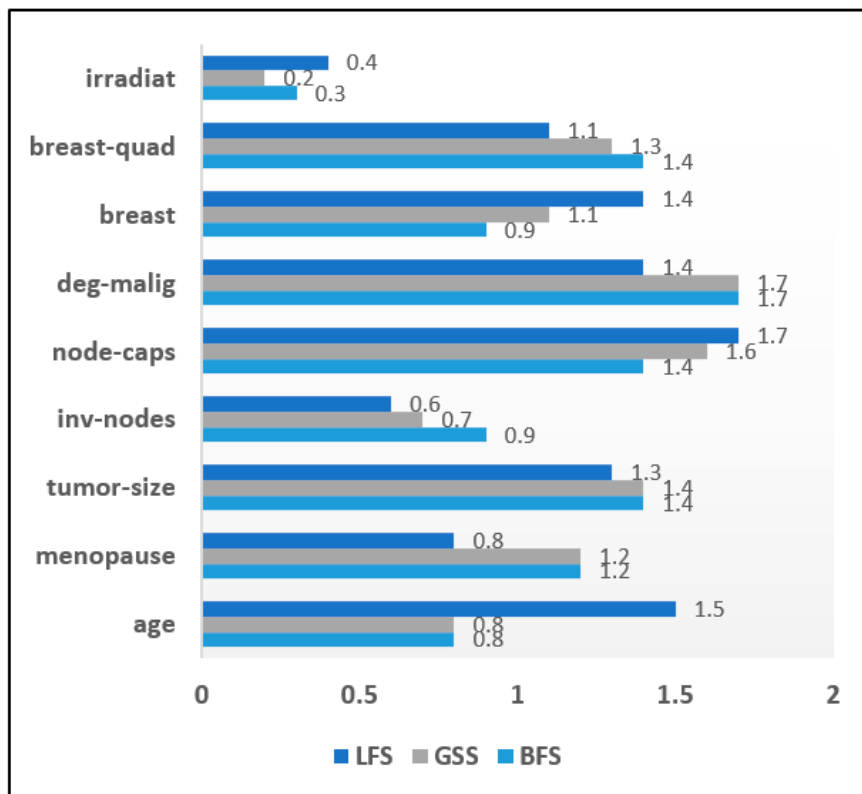


Figure 11. Attribute Significance Graph for the Breast cancer dataset with Wrappers.



The reduced sets of attributes obtained by the BFS, LFS and GSS methods were identical, which constitutes 10 attributes while eliminating 03 irrelevant attributes. Wrapper methods were evaluated with the diabetes dataset to get an optimized attribute set. It has been observed that among eight attributes, only one attribute was identified as irrelevant. Consequently, it was eliminated from the resulting set in all three wrapper methods. The BFS and LFS methods obtained reduced sets of attributes for breast cancer data that contained eight attributes in total. The GSS method included only seven attributes in its reduced set. Table 10 depicts the tabular analysis of the reduced attribute sets after using the wrapper approach.

**Table 10.** The impact of the wrapper methods on disease datasets.

| Chronic Disease Dataset | Dataset Details       | BFS | LFS | GSS |
|-------------------------|-----------------------|-----|-----|-----|
| Heart Disease Dataset   | Number of instances   | 270 | 270 | 270 |
|                         | Initial Attribute set | 13  | 13  | 13  |
|                         | Reduced Attribute set | 10  | 10  | 10  |
| Diabetes Dataset        | Number of instances   | 768 | 768 | 768 |
|                         | Initial Attribute set | 8   | 8   | 8   |
|                         | Reduced Attribute set | 7   | 7   | 7   |
| Breast Cancer Dataset   | Number of instances   | 286 | 286 | 286 |
|                         | Initial Attribute set | 9   | 9   | 9   |
|                         | Reduced Attribute set | 8   | 8   | 7   |

The reduced attribute set is further classified using the decision tree algorithm. In heart disease prediction, it was observed that the best first search method generates an optimal accuracy rate of 94.7%. The greedy step wise search achieves a 94.2% accuracy and linear forward selection method lies in between these two methods with an accuracy of 92.6%. When the wrapper approach was applied to diabetes datasets, an accuracy of 95.6% was achieved with the greedy step wise search and the overall time delay was 1.97 s. Linear forward selection search had an accuracy of 90.6% with a latency delay of 1.56 s. The best first search method was relatively faster than the linear forward selection search method with only a 1.56 s execution delay, thereby achieving a 90.6% accuracy rate. In case of breast cancer classification, the greedy step wise method shows an impressive accuracy rate of 96.8% while the best first search method produces a lower accuracy of 92.8%. The latency delay with the greedy step wise search is also much lower, with 1.08 s, in comparison to the other two wrapper methods. In general, it was noted that BFS method offered better result than others. Reason for this good performance by BFS method was due to the fact that it is able to switch among different paths to solution with ease. It is able to find a good solution without figuring out all nodes and not being trapped at dead ends. Moreover, the time complexity of BFS method is quite less compared to others. The results are summarized in Table 11.

**Table 11.** An analysis of the classification accuracy and latency of the wrapper methods on disease datasets.

| Chronic Disease Dataset | Wrapper Method    | Search Method | Classifier    | Accuracy | Latency |
|-------------------------|-------------------|---------------|---------------|----------|---------|
| Heart Disease Dataset   | WrapperSubsetEval | BFS           | Decision tree | 94.7%    | 1.42 s  |
|                         | WrapperSubsetEval | LFS           | Decision tree | 92.6%    | 1.44 s  |
|                         | WrapperSubsetEval | GSS           | Decision tree | 94.2%    | 1.48 s  |
| Diabetes Dataset        | WrapperSubsetEval | BFS           | Decision tree | 95.8%    | 1.44 s  |
|                         | WrapperSubsetEval | LFS           | Decision tree | 90.6%    | 1.56 s  |
|                         | WrapperSubsetEval | GSS           | Decision tree | 95.6%    | 1.66 s  |
| Breast Cancer Dataset   | WrapperSubsetEval | BFS           | Decision tree | 96.8%    | 1.32 s  |
|                         | WrapperSubsetEval | LFS           | Decision tree | 93.6%    | 1.55 s  |
|                         | WrapperSubsetEval | GSS           | Decision tree | 96.4%    | 1.28 s  |

The experimental evaluation was successfully conducted on three chronic disease datasets. Both filter and wrapper methods for attribute selection were implemented in the analysis. The results

show that in wrapper approach, interaction with the classifier is high. However, it is more computationally intensive. The filter approach is independent of the classification algorithm. Consequently, it requires less computational cost. Furthermore, the classification accuracy rate of the wrapper approach was better than that of the filter approach.

However, the latency rate is minimal with the filter approach. Table 10 summarizes the differentiating elements between both wrapper and filter approaches for attribute selection. It is worth noting that CFS method emerged as the best filter method while BFS clearly outperformed other wrapper methods.

#### **4. A Hybrid Classification Model Using 3-Step Attribute Evaluator and Improved Decision Tree to Enhance Classification Performance of Chronic Disease Datasets**

Today, the clinical sector faces many obstacles in analyzing heaps of the medical data of different patients. The detailed processing of such medical records is a big challenge due to the presence of unstructured and heterogeneous data in the records. There is an urgent need for the evolution of the clinical industry so as to address the issue of data processing and management. Identifying an appropriate attribute set from raw and unstructured data records is the main challenge. Any good machine learning model depends on the attributes that it is fed as input to the system. If the selected attribute set is more relevant, then the subsequent classification and prediction task will also be more efficient and accurate. Identifying the characteristics of attributes to discover their relevance in a dataset is a very critical issue. The majority of existing attribute selection methods overlook the structure of attributes. Recently some studies applied clustering methods to attribute sets, to enhance the overall performance of a computational model and demonstrate it is more effective than conventional attribute selection approaches. Krier et al. [83], successfully developed a hybrid model using hierarchical clustering on spectral features and then selecting an optimal attribute set by means of the mutual information method. Van Dijck and Van Hulle [84], worked on the same model as developed in [83], with the exception that the former model forced each cluster to select successive attributes. Song et al. [85], presented a Fast clustering-based attribute reduction method (FAST) which generated a more useful and independent set of attributes that enhanced the performance of machine learning algorithms. Xu et al. [86], discussed the importance of the clustering approach in simplifying the computation in a machine learning task.

Moreover, recent studies have concluded that the structure of attributes can be effectively interpreted, and noisy variables can be removed from data records by means of the clustering approach [87,88]. In the previous section, it has been observed that the CFS filter method and the BFS wrapper method generated most optimal results. The proposed attribute evaluator makes use of BFS and CFS method for generating reduced attribute set.

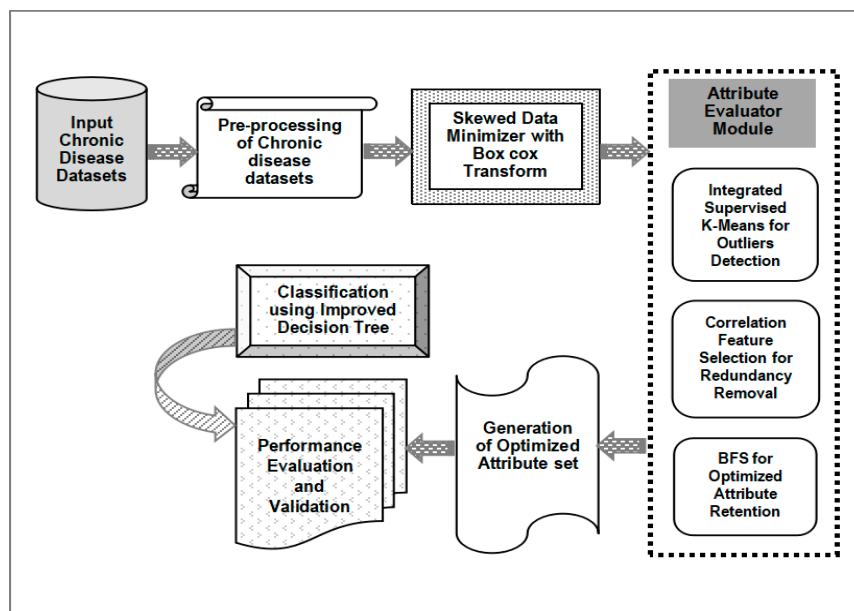
In this era of advanced technology, users face the hurdle of datasets especially if they comprise of large number of attributes and instances. Though decision tree is quite efficient in classification tasks, still it is susceptible to noisy datasets. Certain datasets are tedious to handle since sometime decision trees is dependent on various metrics such as topological configuration of decision tree itself and the type of dataset used. Cluster analysis is widely applied in various domains but is rarely been used with decision tree. In this proposed work of chronic disease diagnosis, clustering approach is integrated with decision tree classifier to enhance the accuracy of classification. A suitable clustering technique like K-Means algorithm can be used to filter the data samples of training set which can help in reducing noise from decision tree thereby elevating the classification accuracy rate.

##### *4.1. Description of the Proposed Hybrid Classification Model*

In this section, a new hybrid classification model for chronic disease datasets is presented. Figure 12 illustrates the proposed machine learning model using the integrated 3-step attribute evaluator method and an improved decision tree classifier. A chronic disease dataset collected from a suitable and reliable source is the input to the model. Many times, these raw datasets are in unstructured form and thus

it needs to be filtered and pre-processed to get it in proper format. The pre-processing of chronic disease data samples is the next step. The data records are scanned and the presence of unintended delimiters, commas or any other symbols is identified and removed. Inconsistencies like missing values or repetitive instances are detected and are replaced with the mean value in the corresponding column. Once the dataset is made free from any ambiguity and thoroughly pre-processed, it is pushed into the skewed data minimizer. The pre-processed data samples are usually unbalanced and uneven. Due to this uneven distribution of data instances the data visualization does not follow a normal distribution thereby its classification performance gets hampered. In this model this skewed data is effectively handled by using box cox transformation. It is a transformation that transforms a random dependent features into normal forms. It enables the predictors to execute the datasets with a wide range of use cases. The core part of box cox transformation is an exponential value which ranges in between  $-5$  to  $5$ . values within the specified range are taken into consideration and its optimum value is chosen. The optimum value denotes the value that creates the best normal distribution curve. It can be represented as in Equation (1).

$$\begin{aligned}
 Y, \lambda &= (y + \lambda_2)^{\lambda_1} - 1/\lambda_1, \lambda_1 \neq 0 \\
 Y, \lambda &= \log(y + \lambda_2), \lambda_1 = 0
 \end{aligned}
 \tag{1}$$



**Figure 12.** Illustration of the Hybrid Classification Model based on 3-Step Attribute Evaluator.

The normalized data obtained is made input to Attribute Evaluator Module. It takes advantage of a clustering approach, which helps it obtain a more accurate selection of attributes form a chronic disease dataset. This ‘Attribute Evaluator’ module integrates K-Means clustering with the CFS method used as a filter method and the BFS method as a wrapper. Clustering with K-Means is the first step in this module. The attribute set is the input and multiple clusters are formed on the basis of similarity among attributes. K-Means clustering selects ‘K’ attributes as its cluster centers from the initial disease dataset. On the basis of the similarity measure, distances between the cluster mean and the attributes are computed. Updated mean value is calculated for every cluster. The most similar attribute is allotted to the corresponding cluster. Euclidean distance is used in our proposed model to determine the attributes that exhibit similarity, as shown in Equation (2). The squared distance between two vectors  $p = (p_1, p_2)$  and  $q = (q_1, q_2)$  is denoted as the sum of squared differences in their coordinates in Equation (3).

$$d^2_{pq} = (p_1 - q_1)^2 + (p_2 - q_2)^2
 \tag{2}$$

$$d_{pq} = \sqrt{(p1 - q1)^2 + (p2 - q2)^2} \tag{3}$$

A new variation of K-Means algorithm referred to as ‘Integrated Supervised K-Means’ is presented and implemented. The main idea behind this new variant is to perform logistic regression in every iteration. Since regression approach is combined with unsupervised clustering hence it is named as Integrated Supervised K-Means. Regression generates a cluster member probability and it is used to eliminate appropriate samples from participating in finding cluster mean process. It is useful in dealing with real time datasets which comprises of heterogeneous instances with overlapping samples. Here the K-Means algorithm initializes with K initial cluster means. Then the process is iterated thereby allotting samples to its nearest cluster and the cluster means is recomputed. Basic notations used in the algorithm are listed in Table 12.

**Table 12.** Parameters used in Integrated Supervised K-Means.

| yi      | It Denotes the i-th Sample Vector (i = 1, ... n)                               |
|---------|--|
| zi      | Arrangement of cluster for i-th sample, zi ∈ {1, ... M}, i = 1, ... n          |
| I (I,m) | Returns 1 for zi = m else returns 0  |
| Cm      | Mean vector corresponding to the mth cluster (m = 1, ... K)                    |
| CS      | Cumulative sum of squared distance   |
| β       | Converging factor for linear difference in cumulative sum of squared distances |
| CMP     | Cluster Member Probability of a sample under consideration                     |

Squared Euclidean distance to the mean is calculated for every sample observation. Every sample is assigned to the nearest cluster. It is followed by solving logistic regression equations represented in Equations (4)–(7).

$$\text{Prob}(z_i = M) = \frac{1}{1 + \sum_{m=1}^{M-1} e^{\alpha_m \cdot y_i}} \tag{4}$$

$$\text{Prob}(z_i = 1) = \frac{e^{\alpha_1 \cdot y_i}}{1 + \sum_{m=1}^{M-1} e^{\alpha_m \cdot y_i}} \tag{5}$$

$$\text{Prob}(z_i = 2) = \frac{e^{\alpha_2 \cdot y_i}}{1 + \sum_{m=1}^{M-1} e^{\alpha_m \cdot y_i}} \tag{6}$$

$$\text{Prob}(z_i = M - 1) = \frac{e^{\alpha_{M-1} \cdot y_i}}{1 + \sum_{m=1}^{M-1} e^{\alpha_m \cdot y_i}} \tag{7}$$

Here the predicted value of regression is used to predict cluster member probability for every sample in all clusters. This probability value is reordered in descending order to determine the proportion of the two highest probabilities. Cluster means is recomputed, and cumulative summation of total distance is determined. Variation in cumulative sum of square distances from predecessor round is measured and compared with the converging factor.

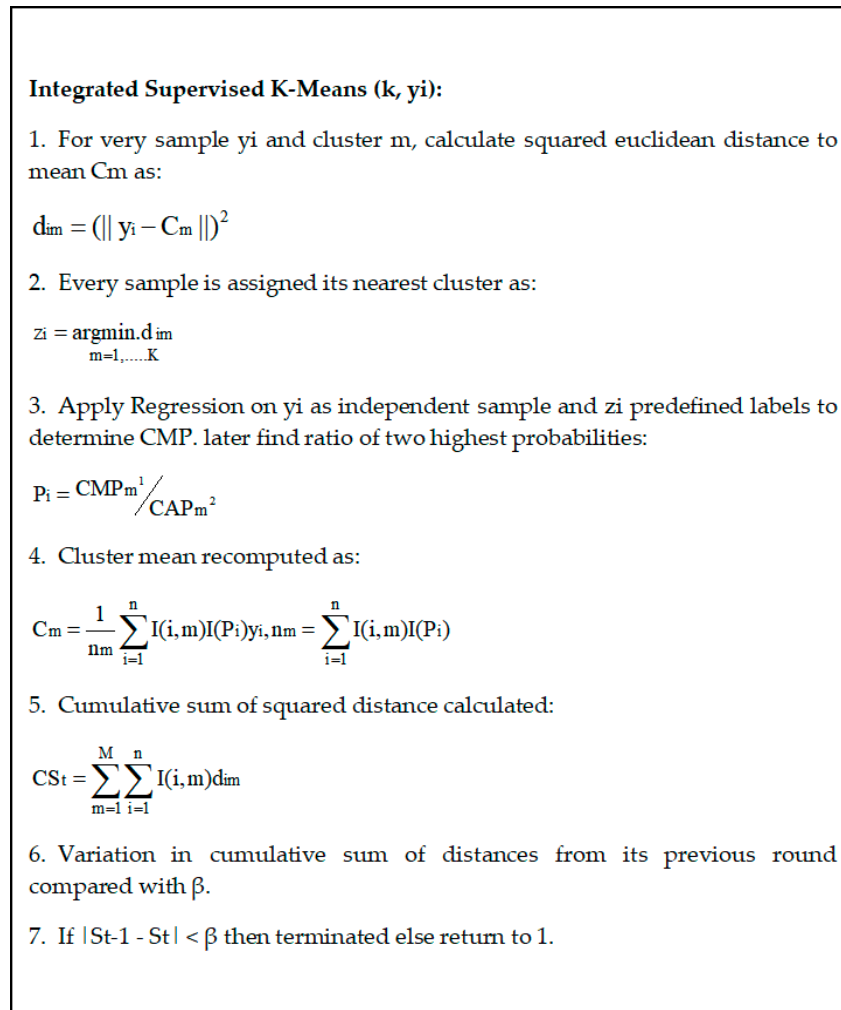
After cluster formation, the left-over attributes which do not belong to any cluster are identified as outliers and are eliminated. Then, the Correlation Feature Selection (CFS) method is applied. It computes correlation among a set of attributes by ranking all attributes in descending order. Correlation among attributes is computed using Pearson’s linear correlation coefficient, as in Equation (8). A has a values and class labels R with r values where A and R are assumed to be random variables.

$$\varphi(A, R) = \frac{E(A, R) - E(A)E(R)}{\sqrt{\sigma^2(A)\sigma^2(R)}} \tag{8}$$

$\varphi(A, R)$  is regarded as 0 when there is no correlation or as 1 when A and R exhibit linear dependency. An error function is used to estimate the correlation between the two attributes as shown in Equation (9).

$$P(A \approx R) = \text{err}(|\varphi(A, R)| \sqrt{M/2}) \quad (9)$$

The pseudo code highlighting integrated supervised k-means is presented in Figure 13.



**Figure 13.** Pseudo code of Integrated Supervised K-Means.

The list of attributes ranked in terms of decreasing order of  $P(A \approx R)$  serves as an attribute ranking order. Highly correlated and redundant attributes have been detected and eliminated. Furthermore, the non-redundant attributes have been input to the BFS method which explored all attributes and found a reduced optimal attribute subset. The output of the 'Attribute Evaluator' led to the generation of a better and reduced attribute set which constitutes only the relevant attributes that can be considered in the diagnosis process. Subsequently, the classification has been performed using an improved decision tree on the attribute set to determine the risk level of a person being affected by a chronic disease. Finally, performance of the developed model has been assessed according to vital performance parameters, determining the effectiveness of the classification.

A new improved decision tree-based model is developed and implemented. It combines both clustering and classification techniques. In this work a novel hybrid model is proposed to improve accuracy of decision tree using clustering approach. This integrated system is formed by a clustering algorithm, a decision tree and a parameter generation module to identify and select suitable and best parameters for the clustering algorithm. These three modules working together are capable to increase the accuracy of the solutions. Clustering acts both as a data instance selection method as well as for classification at front end. As a result, data instances which are not classified by clustering are input

to the decision tree for further classification. This approach is helpful while dealing with complex datasets with noisy environments.

The proposed classification model makes use of the cluster analysis in decision tree classifier as presented in Figure 14. The hybrid model presented here comprises of three units which include the following.

- ❖ Cluster Analysis Unit (CAU): It is used to develop a clustering method with K-Means algorithm.
- ❖ Decision Tree Unit (DTU): It is to create two separate decision tree modules which include complete decision tree and non-clustered decision tree. Complete decision tree utilizes the entire dataset while the non-clustered decision tree operates with only the data samples which are not classified to any cluster.
- ❖ Parameter Check Unit (PCU): It functions as an automated parameter generator which provides relevant parameters for cluster analysis unit without the need to develop the complete model. Dataset is the input too this unit and an appropriate parameter set for cluster analysis unit is the output produced.

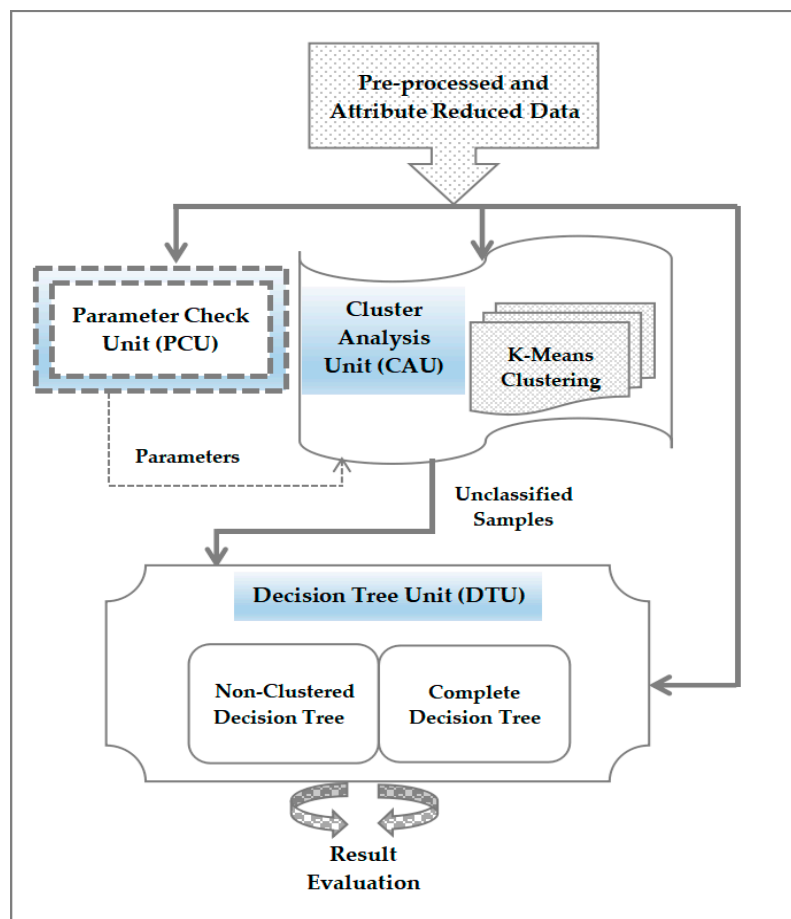


Figure 14. Functioning workflow of Improved decision Tree Model.

#### 4.1.1. Functioning of Cluster Analysis Unit (CAU)

Training data samples are used to build a cluster-based model with removal of the class label from dataset. Every cluster is verified as to determine if it constitutes a larger instances of a class label. It is done by counting the data instances for every class which is allotted to a cluster. If it is beyond the threshold value, then it is inferred that the specific cluster is sorted and classified. A particular cluster is tagged as non-classified if no class is exceeding the threshold value. So, an instance is classified

by passing it through the CAU where that instance is assigned a cluster using K-Means algorithm. Then that instance is tagged as classified or non-classified based on the value in the table built during model creation. Accordingly, the instance is categorized with the same class label of cluster else it is identified as unclassified instance. If it is found to be unclassified then it is left out without any class label.

#### 4.1.2. Functioning of Decision Tree Unit (DTU)

DTU is responsible to build the decision tree classifier. Here two distinct models are built which is being used to function four classifiers. The first model is called complete decision tree which is used to train the complete training set of data samples. The second model called non-clustered decision tree is applicable on the training samples which failed to be classified by CAU. Both decision tree models are subsequently validated with both testing dataset as well as the unclassified data samples of CAU. Thus, the four classifiers created in the process are listed here.

- Decision tree classifier which forms the base of comparative analysis.
- Decision tree which functions on only unclassified data of CAU but is used to categorize all data samples.
- Decision tree which functions on all data samples but used to categorize only unclassified CAU data.
- Decision tree which functions on unclassified CAU data samples and is also used to classify unclassified data from CAU.

Among these four models used in the research work, first model is the usual decision tree classifier. Second model is mainly utilized to minimize the number of data samples which are used for training the classifier. It is the third and fourth model which are the newly developed ones of our study which intermix clustering with classification to provide decision tree optimum classification capability. Use of clustering algorithm in CAU to categorize data samples and assign them into classified and unclassified data is highlighted in Figure 15.

**Step 1:** Dataset split into train set (TN) and test set (TT) for every cross-validation fold;

**Step 2:** DC = Develop\_Cluster (TT, Parameters);

**Step 3:** For every cluster of DC:

**Step 3.1:** Cluster.Class = -1;

**Step 4:** For every dataset class:

**Step 4.1:** m = Frequency\_count(Class samples of cluster);

**Step 5:** if (m/length(Cluster.samples) > threshold):

**Figure 15.** Pseudo code for Clustering process in Improved decision tree model.

#### 4.1.3. Functioning of Parameter Check Unit (PCU)

This unit is responsible for validating the parameters which are significant and adequate for a dataset without executing the entire decision tree classifier. A simple yet effective heuristic is

used on the training set result which takes into account correctly and incorrectly classified data to achieve the task. PCU utilizes a parameter list like number of clusters created and employs K-Means algorithm. Then a clustering model is developed with this data values for every testing values. This unit then computes the classified clusters for every model built. It is followed by evaluation of training data for all resultant models is performed. Eventually a heuristic is used with the results obtained. Parameters having maximum heuristic value is chosen as a potential candidate to be used for classification. This approach generates parameters for the clustering process and the threshold determined as the deserving candidate.

Heuristic value (h) is calculated by considering the number of correctly classified samples (SC), incorrectly classified samples (SIC), total data samples (S), parameter (P) and cluster count (MC) as denoted in Equation (10).

$$h = \frac{SC - SIC}{S.(P.Mc)^{0.2}} \quad (10)$$

Prime purpose of using this heuristic is to enhance the correctly classified data and at the same time to reduce the error while classification. The best parameter generation procedure is depicted in Figure 16.

```

Step 1: Good = 0;

Step 2: For every par in parameter:

    Step 2.1: For every th in thresholds:

        Step 2.1.1: DC = Develop_Cluster (th, par);

Step 3: For every cluster in DC:

    Step 3.1: Cluster.class = -1;

Step 4: For every dataset class:

    Step 4.1: m = Frequency_count(Class samples of cluster)

Step 5: if (m/length(Cluster.samples) > threshold):

    Step 5.1: Cluster.class = class;

Step 6: if Cluster.class != NULL:

    Step 6.1: For every sample in cluster:

        Step 6.1.1: if Sample.class = Cluster.class:

            Step 6.1.1.1: SC = SC + 1;

```

**Figure 16.** Pseudo code for parameter adjustment and generation in Improved decision tree model.



#### 4.2. Results and Analysis

The effectiveness of a machine learning model is determined in terms of its uniformity and scalability. Various performance metrics are available to figure out the efficiency of any newly developed predictive model. In this subsection, our proposed model using ‘Attribute Evaluator’ was evaluated on several chronic disease datasets and its performance has been assessed according to a series of vital performance parameters. The obtained Results are discussed in detail further on in this subsection.

The proposed ‘Attribute Evaluator’ comprising K-Means clustering, the CFS method and the BFS method, was used to generate an optimized attribute subset. It was used as a dimension reduction approach on 15 disease datasets collected from the UCI data repository. It was further compared with the CFS filter method and the BFS wrapper method in selecting the reduced attribute set. The result is shown in the table. The difference in performance is quite evident, especially on datasets with a higher number of attributes. The proposed model (K-Means, CFS and BFS) performs very well in generating a reduced attribute set for almost all disease dataset. With a relatively lower attribute set, the proposed model performs marginally better. The Arrhythmia data samples initially contained 279 attributes, the proposed model removed outliers and redundancies, reducing the number of attributes to 262. However, when optimizing diabetes data which contained only 8 attributes, the distinction was not as evident, as the attribute set was reduced to 6. This set is identical to the reduced set using wrapper and filter approach taken alone. Hence, it can be inferred that the dimensionality reduction of the proposed ‘Attribute Evaluator’ is optimal and beneficial when the analysed disease data samples are large and have a relatively higher number of attributes. On the contrary, when smaller datasets are evaluated the model’s performance is equivalent to the wrapper or filter approach. Table 13 gives a summary of the results.

**Table 13.** The Impact of Proposed Attribute Evaluator Method on disease datasets.

| Chronic Disease Dataset | Instances | Initial Attribute Set | K-Means | CFS | BFS | CFS + BFS | Proposed Attribute Evaluator |
|-------------------------|-----------|-----------------------|---------|-----|-----|-----------|------------------------------|
| Pima Indians Diabetes   | 768       | 8                     | 8       | 7   | 7   | 6         | 6                            |
| Chronic Kidney Disease  | 400       | 25                    | 23      | 22  | 22  | 21        | 20                           |
| Statlog (Heart)         | 270       | 13                    | 12      | 11  | 10  | 9         | 9                            |
| Breast Cancer Wisconsin | 699       | 10                    | 10      | 9   | 9   | 8         | 8                            |
| Arrhythmia              | 452       | 279                   | 274     | 271 | 268 | 264       | 262                          |
| Hepatitis               | 155       | 19                    | 19      | 18  | 17  | 16        | 15                           |
| Lung Cancer             | 32        | 56                    | 54      | 53  | 51  | 50        | 49                           |
| Parkinson’s             | 197       | 23                    | 22      | 21  | 20  | 19        | 18                           |
| Liver Disorder          | 345       | 7                     | 7       | 7   | 6   | 6         | 6                            |
| Primary Tumour          | 339       | 17                    | 17      | 16  | 15  | 14        | 13                           |
| Thyroid Disease         | 7200      | 21                    | 20      | 18  | 17  | 16        | 16                           |
| Cervical Cancer         | 858       | 36                    | 35      | 34  | 33  | 32        | 30                           |
| Breast Cancer           | 286       | 9                     | 9       | 8   | 8   | 8         | 8                            |
| Autistic Disorder       | 292       | 21                    | 20      | 20  | 19  | 19        | 18                           |

The proposed ‘Attribute Evaluator’ approach is combined with decision tree for the classification of datasets. Its performance is assessed according to performance indices, such as accuracy rate, specificity, sensitivity and F-Score. Figure 17 shows the evaluation result. It has been observed that classification with decision tree yielded a very impressive result and the performance was very consistent on almost all disease datasets. The classification of cervical cancer data generated the highest accuracy of 98.4% while the lowest accuracy was 94.8% with the Parkinson disease dataset.

A maximum specificity value of 98.2% was achieved on cervical cancer data and a minimum specificity value of 93.6% was achieved on autism disorder dataset. Similarly, a very high sensitivity rate of 97.6% was achieved on thyroid data and the lowest sensitivity of 92.9% on autism data. Harmonic mean of specificity and sensitivity called the F-Score metric was also evaluated. Both thyroid and cervical cancer recorded the highest F-Score value of 97.8% while autism gave a comparatively low value of 93%. In general, the generated mean values for accuracy, specificity, sensitivity and f-score were 96.8%, 96.5%, 95.6% and 96.2%, respectively using the proposed attribute evaluator approach.

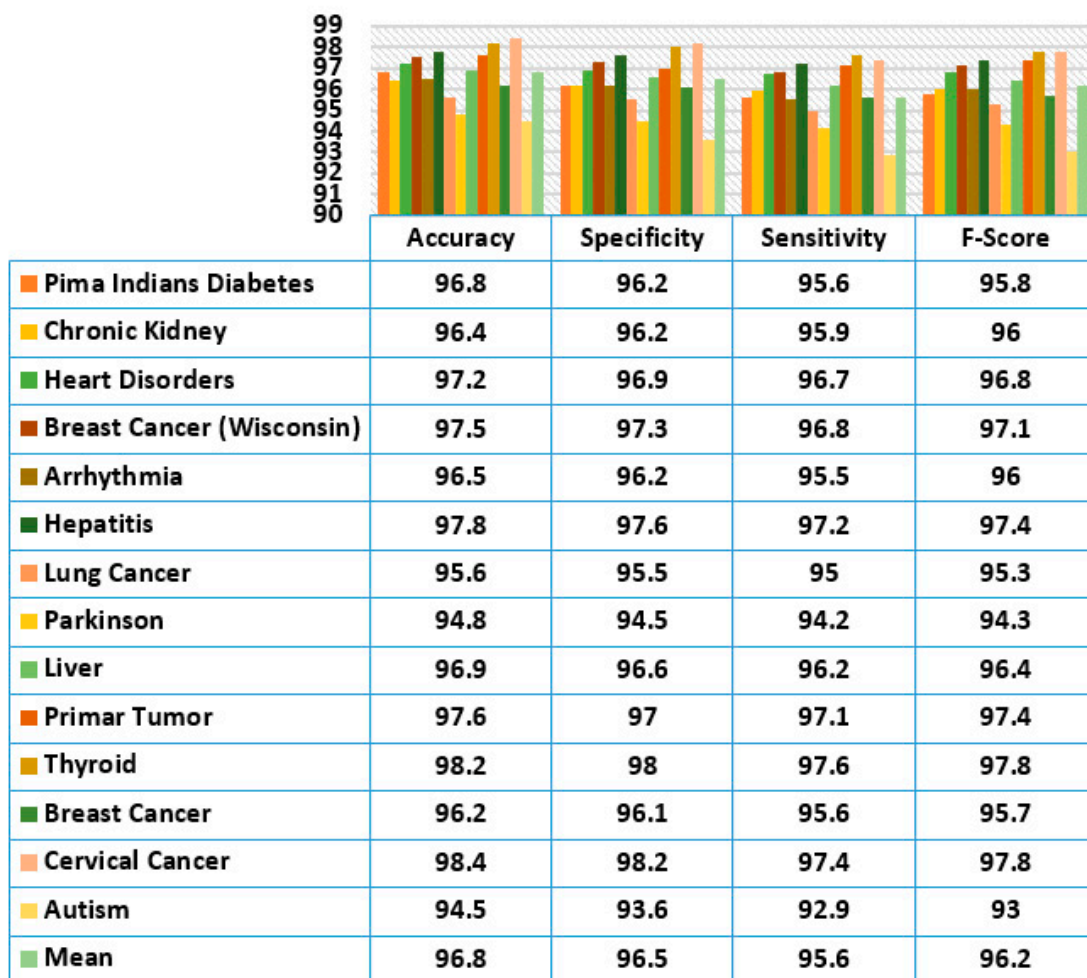


Figure 17. Evaluation of the Classification Performance of the Proposed Attribute Evaluator Method.

The classification accuracy of the decision tree with proposed attribute optimization model was evaluated with some popular filter approaches to determine the preciseness of the generated attribute set. The result is illustrated in Figure 18. CFS, IG and Chi-square methods were used as filter methods. Highly consistent performance was observed with the proposed model on almost all disease datasets with the exception of lung cancer and Parkinson disease. While the accuracy of the CFS filter method for lung cancer and Parkinson was noted to be 95.7% and 95.2%, slightly lower values of 95.6% and 94.8% were obtained using the proposed model. Overall, the proposed classification model outperformed other filter approaches.

A comparative analysis of the accuracy achieved by the proposed model and the wrapper approach has also been carried out, using decision tree as classifier. The wrapper methods considered in the comparison were: BFS, LFS and GSS. The result is analysed in Figure 19. A highly optimal classification performance was recorded. Except the classification of breast cancer data, other disease datasets generated a very high accuracy rate has been achieved by compared to its wrapper approaches. Only in breast cancer data, the BFS method 96.8% of accuracy while the proposed model slightly lagged behind with 96.2%. Overall, the use of the proposed model for classification has yielded optimal performance; superior to that of wrappers.

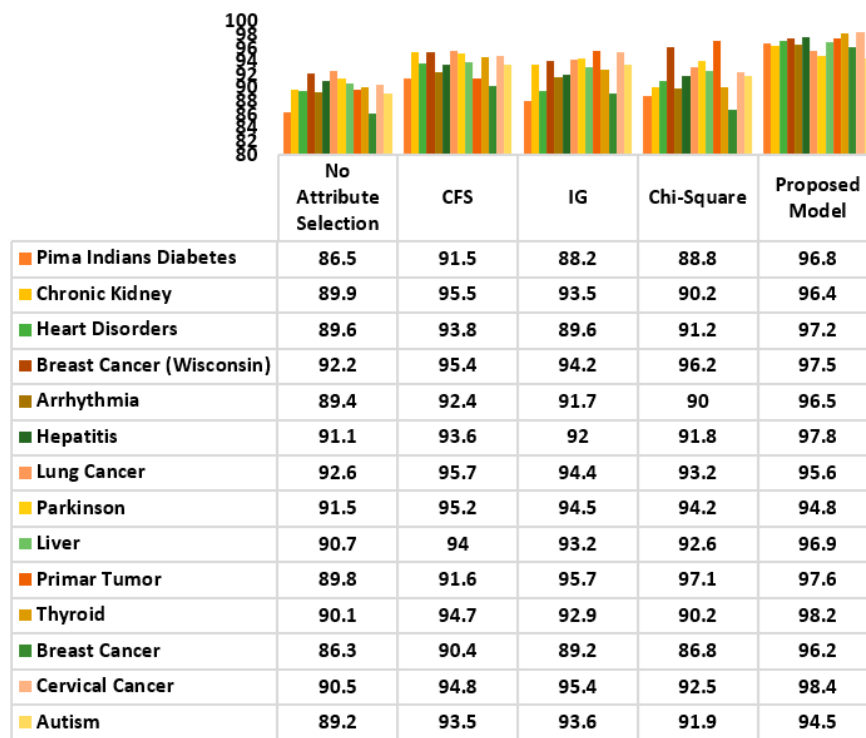


Figure 18. Comparative Study of the Classification Accuracy of the Proposed Attribute Evaluator Method and Other Filter Methods.

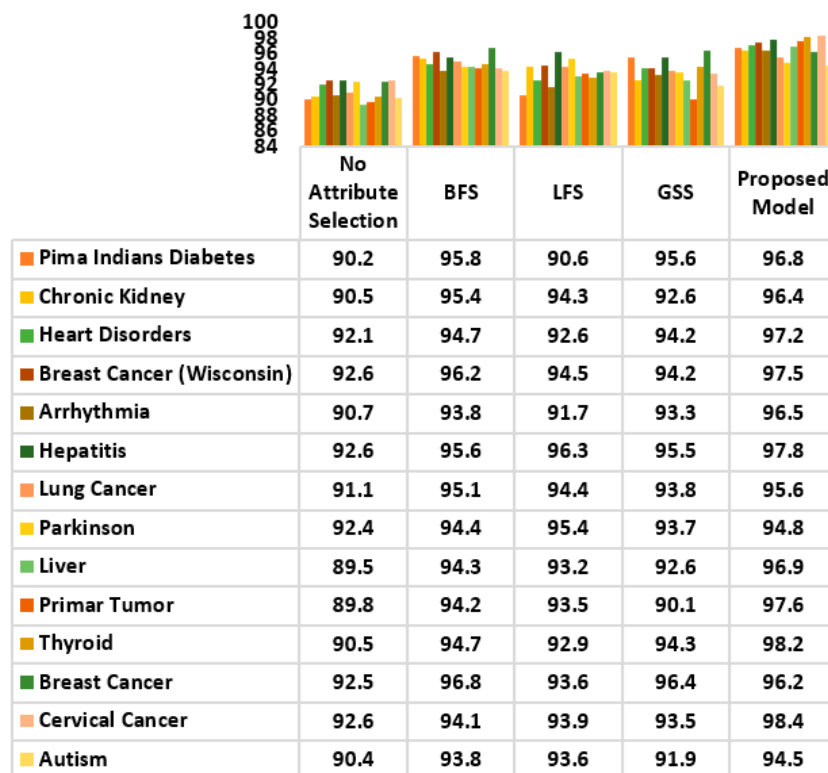
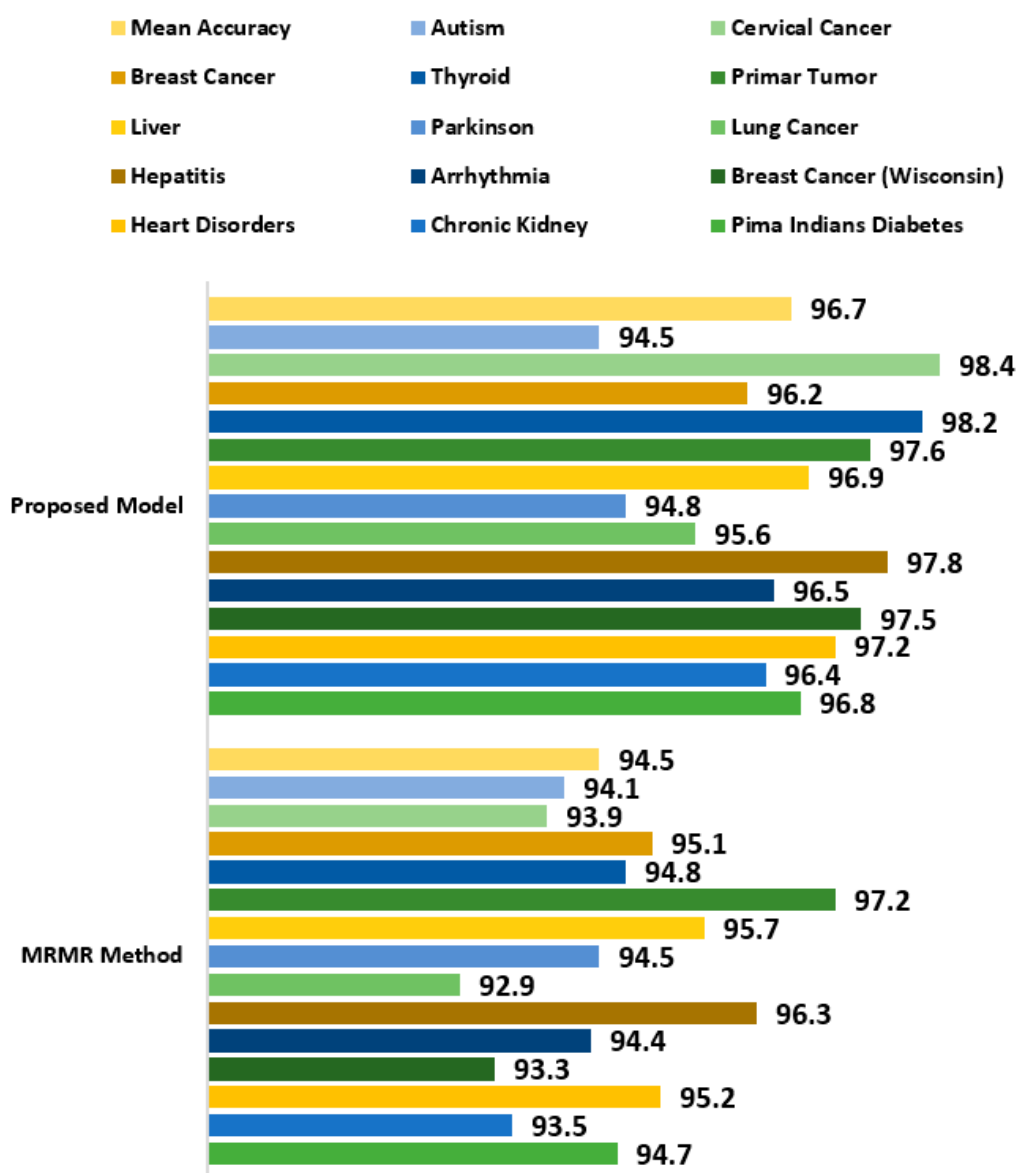


Figure 19. Comparative Study of the Classification Accuracy of the Proposed Attribute Evaluator Method and Other Wrapper Methods.

The proposed attribute evaluator approach was further compared with the Minimum Redundancy Maximum Relevance (MRMR) method which is a widely popular attribute selection technique being applied to several datasets in different domains [89]. It is used as a heuristic method and is applicable for both discrete as well as continuous data samples. It computes the most promising attributes by measuring the relevance and redundant attributes. The MRMR method was evaluated against the proposed attribute evaluator using decision tree as classifier. 14 chronic disease datasets as mentioned in Table 13 were used for analysis. MRMR method gave a very good performance but the proposed attribute evaluation approach generated a better outcome with almost all datasets. A mean accuracy rate of 96.7% was recorded with the proposed approach while a slightly less accuracy of 94.5% was observed with MRMR method as observed in Figure 20.

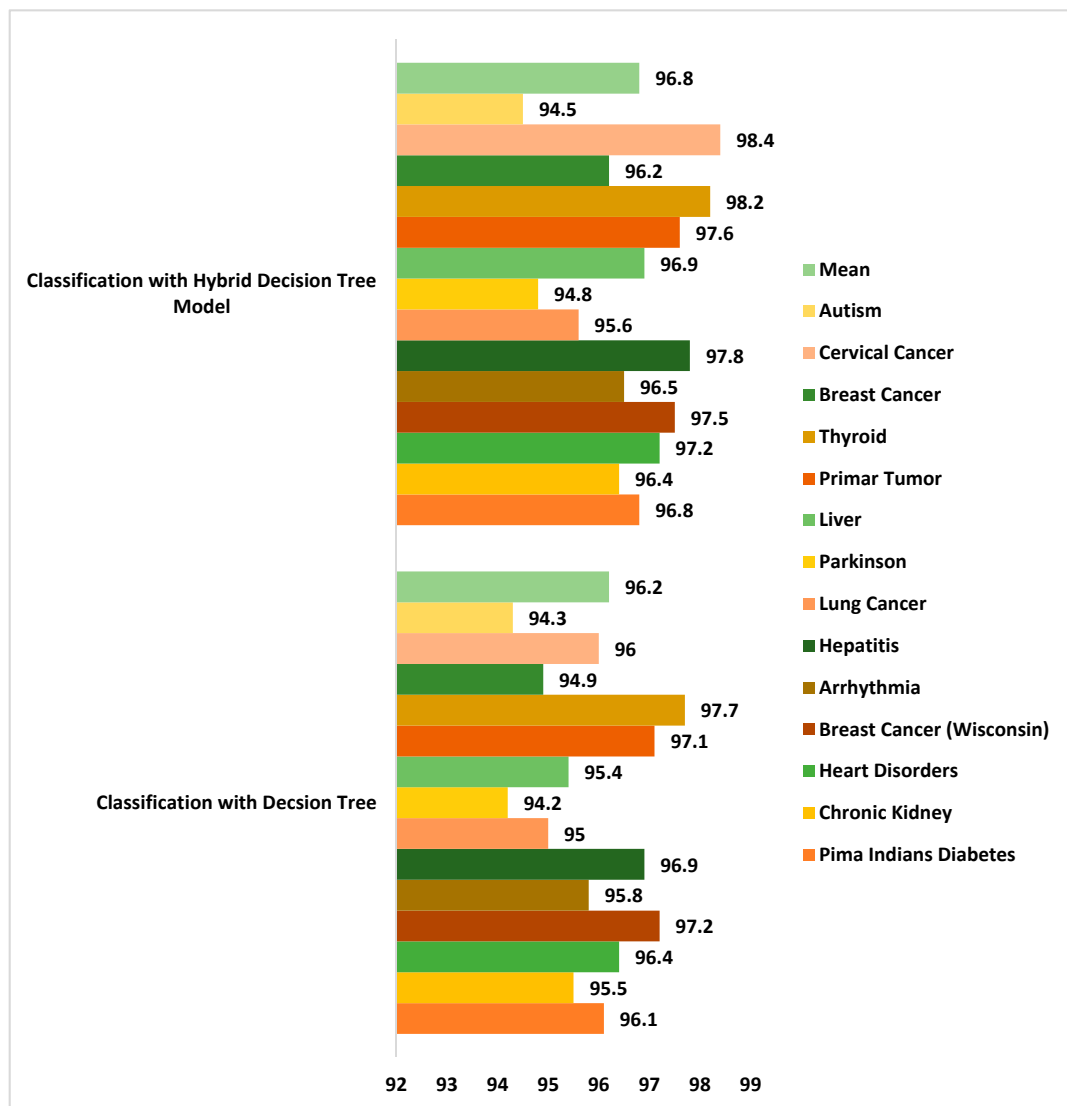


**Figure 20.** Comparative Study of the Classification Accuracy of the Proposed Attribute Evaluator Method and MRMR method.

The data from the datasets is partitioned before being used, using a 10% of the data for adjusting the parameter, and the rest for a 10-fold cross-validation [90], which implies that each fold uses 81% of the total data of the data set for training and 9% for testing. The chronic disease datasets under use is

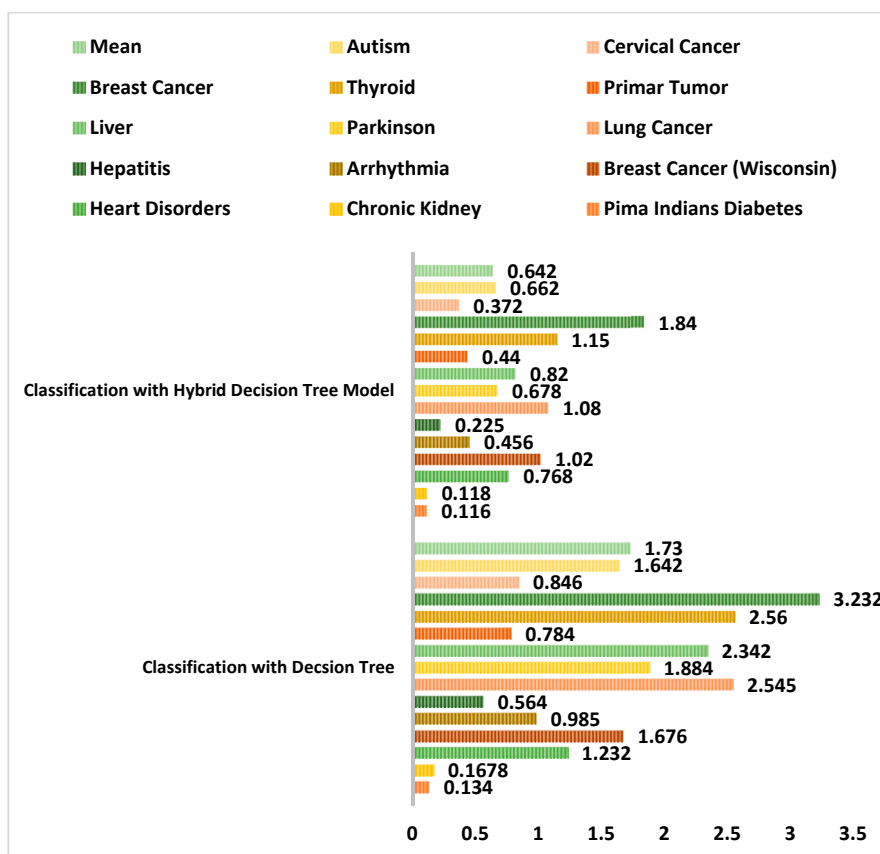
divided before being implemented. Among the data samples, 15% instances were utilized to adjust and generate best parameters. Remaining 85% data were used for cross-validation where 75% data samples were used as training set while the rest 10% instances were used for testing.

Classification accuracy analysis was done between the newly developed hybrid classifier model and decision tree alone. A maximum 98.4% accuracy was generated using cervical cancer dataset while the mean accuracy recorded was 96.8% with the hybrid classification model [91,92]. Decision tree too gave a good accuracy performance but slightly less than the hybrid model as seen in Figure 21.



**Figure 21.** Comparison of accuracy rate between the Proposed Hybrid Decision Tree and the classical decision tree method on chronic disease datasets.

The performance of the proposed hybrid clustering based decision tree classification model was compared with the general decision tree classifier with respect to the model prone to error in Figure 22. It was observed that a least error rate generated was 0.116 with diabetes dataset when used with the new hybrid classification model. An error rate of 1.84 was noted with breast cancer data samples. Similarly, the hybrid decision tree model generated less error rate when compared to the decision tree classifier taken alone. The mean error rate recorded with the new improved hybrid model was just 0.642 as compared to 1.73 with decision tree classifier.



**Figure 22.** Comparison of error rate between the Proposed Hybrid Decision Tree and the classical decision tree method on chronic disease datasets.

Thus, in this section, a new integrated attribute evaluator approach using an enhanced K-Means clustering, the CFS filter method and the BFS wrapper method has been presented and evaluated on 14 different chronic disease datasets. The impact of the proposed model was clearly visible on datasets with higher numbers of attributes. As the number of attributes increased, the relevance of attributes was more evident, and outliers were eliminated. Various performance indicators have been used to determine its classification performance. Impressive maximum metric values were recorded. 96.7%, 96.5%, 95.6% and 96.2% were the mean percentages for accuracy, specificity, sensitivity and f-score, respectively. Performance of improved decision tree model was also compared with decision tree classifier taken alone and it was noted that the improved decision tree performed much better in terms of accuracy rate and error rate. It is due to the fact that clustering approach in improved decision tree classifies majority of data samples. Only the left-over unclassified instances were fed into decision tree unit for further classification. Hence it is more effective and less prone to error.

### 5. Conclusions and Future Scope

This paper has presented a comparative analysis of wrapper and filter selection while determining its impact on prediction performance in different chronic disease datasets. The Decision Tree algorithm has been used as a classifier for this analysis and is implemented through the WEKA tool. The overall research analysis can be bifurcated into two parts. In the first part, a comparative analysis of wrappers and filters has been carried out and its impact on three chronic disease datasets was analysed. Heart disease, diabetes and breast cancer data were contained in the datasets. CFS, IG and Chi-Square were the filter methods used while BFS, LFS and GSS were the wrapper methods used in the research. An optimum accuracy of 93.8%, 89.5% and 96.8% was recorded for heart disease, diabetes and breast cancer, respectively, using the CFS method. Latency delays of 1.08 s, 1.02 s and 1.01 s were observed

using the same method on the respective datasets. Among wrapper methods, BFS gave the best classification performance. Maximum accuracy of 94.7%, 95.8% and 96.8% were achieved for heart disease, diabetes and breast cancer, respectively. Latency delays of 1.42 s, 1.44 s and 132 s were observed using the same method on the respective datasets. Moreover, it was noted that though wrappers generated maximum classification accuracy, they were compute-intensive. Classifications with filter methods incurred less computational cost and their execution time was lower than that of wrapper methods. It was inferred that many attribute selection methods did not give much importance to the structure of attributes and the possible outliers were ignored during the generation of an optimized attribute set.

In the second part of the paper, an integrated hybrid attribute evaluation method has been proposed. It combined a new variant of K-Means cluster analysis named Integrated Supervised K-Means along with the CFS and BFS methods. This new variant of k-means algorithm used regression approach to generate cluster member probabilities which is used to detect less relevant instances from participating in finding cluster mean. Classification with the proposed method using an improved decision tree maximise the performance. The improve decision tree used clustering approach which classifies majority instances. The remaining instances were input to the decision tree unit which classifies them into respective classes. It was evaluated on 14 different chronic disease datasets using various performance metrics. Optimum mean values of 96.8%, 96.5%, 95.6% and 96.2% were recorded for accuracy, specificity, sensitivity and f-score. The use of the newly proposed integrated attribute evaluator model along with improve decision tree classifier generated a mean accuracy of 96.85 which is more than 96.1% accuracy rate generated by using decision tree alone. The mean error rate recorded with the hybrid classification model is as low as 0.642. Thus, it can be suggested for chronic disease assessment which can assist medical experts, acting as a decision support system. Furthermore, this analysis can support future research initiatives using machine learning for chronic disease classification. Future lines of research involve the application of the proposed model to image related datasets. Moreover, assessing impact of ensemble learning and image datasets on attribute evaluators will be an important element of future research.

**Author Contributions:** All authors have designed the study, developed the methodology, performed the analysis, and written the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Jain, D.; Singh, V. Feature selection and classification systems for chronic disease prediction: A review. *Egypt. Inform. J.* **2018**, *19*, 179–189. [CrossRef]
2. Ganiger, S.; Rajashekharaiyah, K. Chronic Diseases Diagnosis using Machine Learning. In Proceedings of the International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), Kottayam, India, 21–22 December 2018; pp. 1–6.
3. Apoorva, R. All about the Top 5 Chronic Diseases in India, Medlife Blog: Health and Wellness Tips. 2018. Available online: <https://www.medlife.com/blog/top-5-chronic-diseases-india/> (accessed on 4 November 2020).
4. Canlas, R.D. Data Mining in Healthcare: Current Applications and Issues. Master's Thesis, School of Information Systems & Management, Carnegie Mellon University, Australia, 2009; pp. 1–11.
5. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
6. Sarojini, B.; Ramaraj, N. Enhancing Medical Prediction using Feature Selection. *Int. J. Artif. Intell. Expert Syst.* **2011**, *1*, 1–9.
7. Novakovic, J.; Strbac, P.; Bulatovic, D. Toward optimal feature selection using ranking methods and classification algorithms. *Yugosl. J. Oper. Res.* **2011**, *21*, 119–135. [CrossRef]
8. Dash, M.; Liu, H. Feature Selection for Classification. *Intell. Data Anal.* **1997**, *1*, 131–156. [CrossRef]

9. Mishra, S.; Mishra, B.K.; Sahoo, S.; Panda, B. Impact of swarm intelligence techniques in diabetes disease risk prediction. *Int. J. Knowl. Discov. Bioinform. (IJKDB)* **2016**, *6*, 29–43. [[CrossRef](#)]
10. Zhao, Z.; Wang, L.; Liu, H.; Ye, J. On Similarity Preserving Feature Selection. *IEEE Trans. Knowl. Data Eng.* **2011**, *25*, 619–632. [[CrossRef](#)]
11. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
12. Top5 Advantages and Disadvantages of Decision Tree Algorithm. Available online: <https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a> (accessed on 4 November 2020).
13. Awang, R.; Palaniappan, S. Web-based Heart Disease Decision Support System using Data Mining Classification Modeling techniques. In Proceedings of the 9th International Conference on Information Integration and Web-Based Application & Services, Jakarta, Indonesia, 3–5 December 2007; pp. 177–187.
14. Tu, M.C.; Shin, D.; Shin, D. Effective Diagnosis of Heart Disease through Bagging Approach. In Proceedings of the 2nd International Conference on Biomedical Engineering and Informatics, Tianjin, China, 17–19 October 2009; pp. 1–4.
15. Benkaci, M.; Jammes, B.; Doncescu, A. Feature Selection for Medical Diagnosis Using Fuzzy Artmap Classification and Intersection Conflict. In Proceedings of the 24th International Conference on Advanced Information Networking and Applications Workshops, Perth, Australia, 20–23 April 2010; pp. 790–795.
16. Heller, R.F.; Chinn, S.; Pedoe, H.D.; Rose, G. How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. *BMJ* **1984**, *288*, 1409–1411. [[CrossRef](#)]
17. Wilson, P.W.F.; D’Agostino, R.B.; Levy, D.; Belanger, A.M.; Silbershatz, H.; Kannel, W.B. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* **1998**, *97*, 1837–1847. [[CrossRef](#)]
18. Wang, H.; Khoshgoftaar, T.M.; Van Hulse, J. A Comparative Study of Threshold-Based Feature Selection Techniques. In Proceedings of the 2010 IEEE International Conference on Granular Computing, San Jose, CA, USA, 14–16 August 2010; pp. 499–504.
19. Simons, L.; Simons, J.; Palaniappan, L.; Friedlander, Y.; McCallum, J. Risk functions for prediction of cardiovascular disease in elderly Australians: The Dubbo Study. *Med. J. Aust.* **2003**, *178*, 113–116. [[CrossRef](#)] [[PubMed](#)]
20. Surekha, S.; JayaSuma, G. Comparison of Feature Selection Techniques for Thyroid Disease. In Proceedings of the International Conference on Intelligent Systems, Control & Manufacturing Technology (ICICMT 2015), Abu Dhabi, UAE, 16–17 March 2015; pp. 20–26.
21. Sah, R.D.; Sheetalani, J. Review of Medical Disease Symptoms Prediction Using Data Mining Technique. *IOSR J. Comput. Eng.* **2017**, *19*, 59–70. [[CrossRef](#)]
22. Chuang, L.-Y.; Ke, C.-H.; Yang, C.-H. A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification. In Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, China, 19–21 March 2008; pp. 146–150.
23. Mishra, S.; Tripathy, H.K.; Mishra, B.K. Implementation of biologically motivated optimisation approach for tumour categorisation. *Int. J. Comput. Aided Eng. Technol.* **2018**, *10*, 244–256. [[CrossRef](#)]
24. González-Navarro, F.F.; Munoz, L.A.B. Feature Selection for the Prediction and Visualization of Brain Tumor Types Using Proton Magnetic Resonance Spectroscopy Data. *Comput. Vis.* **2012**, 83–97. [[CrossRef](#)]
25. Verma, A.K.; Pal, S.; Kumar, S. Comparison of skin disease prediction by feature selection using ensemble data mining techniques. *Inform. Med. Unlocked* **2019**, *16*, 100202. [[CrossRef](#)]
26. El-Hasnony, M.I.; El Bakry, H.M.; Saleh, A.A. Comparative Study among Data Reduction Techniques over Classification Accuracy. *Int. J. Comput. Appl.* **2015**, *122*, 9–15. [[CrossRef](#)]
27. Patil, B.; Joshi, R.; Toshniwal, D. Hybrid prediction model for Type-2 diabetic patients. *Expert Syst. Appl.* **2010**, *37*, 8102–8108. [[CrossRef](#)]
28. Nalband, S.; Sundar, A.; Prince, A.A.; Agarwal, A. Feature selection and classification methodology for the detection of knee-joint disorders. *Comput. Methods Progr. Biomed.* **2016**, *127*, 94–104. [[CrossRef](#)]
29. Piramuthu, S. Evaluating feature selection methods for learning in data mining applications. *Eur. J. Oper. Res.* **2004**, *156*, 483–494. [[CrossRef](#)]
30. Karegowda, A.G.; Manjunath, A.; Jayaram, M. Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of PIMA Indians Diabetes. *Int. J. Soft Comput.* **2011**, *2*, 15–23. [[CrossRef](#)]



31. Elsayyad, A.; Al-Dhaifallah, M.; Nassef, A.M. Features selection for arrhythmia diagnosis using Relief-F algorithm and support vector machine. In Proceedings of the 2017 14th International Multi-Conference on Systems, Signals & Devices (SSD), Marrakech, Morocco, 28–31 March 2017; pp. 461–468.
32. Fallahpour, S.; Lakvan, E.N.; Zadeh, M.H. Using an ensemble classifier based on sequential floating forward selection for financial distress prediction problem. *J. Retail. Consum. Serv.* **2017**, *34*, 159–167. [[CrossRef](#)]
33. D Jain, D.; Singh, V. Diagnosis of Breast Cancer and Diabetes using Hybrid Feature Selection Method. In Proceedings of the 2018 5th International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India, 20–22 December 2018; pp. 64–69.
34. Kolukisa, B.; Hacilar, H.; Goy, G.; Kus, M.; Bakir-Gungor, B.; Aral, A.; Gungor, V.C. Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 2232–2238.
35. Tengku Mazlin, T.A.H.; Sallehuddin, R.; Zuriahati, M.Y. Utilization of Filter Feature Selection with Support Vector Machine for Tumours Classification. In Proceedings of the Joint Conference on Green Engineering Technology & Applied Computing, Bangkok, Thailand, 4–5 February 2019.
36. Mishra, S.; Tripathy, H.K.; Mishra, B.; Sahoo, S. Implementation of Classification Rule Mining to minimize Liver Disorder risks. *Int. J. Control Theory Appl.* **2017**, *10*, 117–124.
37. Amin, S.; Singhal, A. Identification and classification of neuro-degenerative diseases using feature selection through PCA-LD. In Proceedings of the 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, India, 26–28 October 2017; pp. 578–586.
38. Narayanan, B.N.; Hardie, R.C.; Kebede, T.M. Performance Analysis of Feature Selection Techniques for Support Vector Machine and its Application for Lung Nodule Detection. In Proceedings of the NAECON 2018 IEEE National Aerospace and Electronics Conference, Dayton, OH, USA, 26–26 July 2018; pp. 262–266.
39. El Akadi, A.; Amine, A.; El Ouardighi, A.; Aboutajdine, D. A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowl. Inf. Syst.* **2010**, *26*, 487–500. [[CrossRef](#)]
40. Quanz, B.; Huan, J.; Mishra, M. Knowledge Transfer with Low-Quality Data: A Feature Extraction Issue. *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 1789–1802. [[CrossRef](#)]
41. Hafizah, W.M.; Supriyanto, E.; Yunus, J. Feature Extraction of Kidney Ultrasound Images Based on Intensity Histogram and Gray Level Co-occurrence Matrix. In Proceedings of the 2012 6th Asia Modelling Symposium, Bali, Indonesia, 29–31 March 2012; pp. 115–120.
42. Vieira, S.M.; Mendonça, L.F.; Farinha, G.J.; Sousa, J.M. Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. *Appl. Soft Comput.* **2013**, *13*, 3494–3504. [[CrossRef](#)]
43. Prasartvit, T.; Banharnsakun, A.; Kaewkamnerdpong, B.; Achalakul, T. Reducing bioinformatics data dimension with ABC-kNN. *Neurocomputing* **2013**, *116*, 367–381. [[CrossRef](#)]
44. Mokeddem, S.A.; Atmani, B.; Mokaddem, M. Supervised Feature Selection for Diagnosis of Coronary Artery Disease Based on Genetic Algorithm. In Proceedings of the 1st International Conference on Computational Science and Engineering, Dubai, UAE, 18–19 May 2013; pp. 41–51.
45. Chen, C.P.; Zhang, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf. Sci.* **2014**, *275*, 314–347. [[CrossRef](#)]
46. Kora, P.; Krishna, K.S.R. Improved Bat algorithm for the detection of myocardial infarction. *SpringerPlus* **2015**, *4*, 666. [[CrossRef](#)]
47. Priya, M.; Juliet, P.L.; Tamilselvi, P.R. Performance analysis of liver disease prediction using machine learning algorithms. *Int. Res. J. Eng. Technol.* **2018**, *5*, 206–211.
48. Keerthi, P.; Hema Latha, N.; Hari Gokul, K.; Prasad, V.; Arun Kumar, T. Wrapper Based Feature Selection for Disease Diagnosis using Optimization Algorithms. *Int. J. Eng. Res. Technol.* **2018**, *6*, 1–10.
49. Kanan, H.R.; Faez, K. An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system. *Appl. Math. Comput.* **2008**, *205*, 716–725. [[CrossRef](#)]
50. Erguzel, T.T.; Ozekes, S.; Tan, O.; Gultekin, S. Feature Selection and Classification of Electroencephalographic Signals. *Clin. EEG Neurosci.* **2014**, *46*, 321–326. [[CrossRef](#)] [[PubMed](#)]
51. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389–422. [[CrossRef](#)]

52. Uzer, M.S.; Yilmaz, N.; Inan, O. Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification. *Sci. World J.* **2013**, *2013*, 1–10. [[CrossRef](#)] [[PubMed](#)]
53. Dogantekin, E.; Dogantekin, A.; Avci, D. Automatic hepatitis diagnosis system based on Linear Discriminant Analysis and Adaptive Network based on Fuzzy Inference System. *Expert Syst. Appl.* **2009**, *36*, 11282–11286. [[CrossRef](#)]
54. Chen, L.-F.; Su, C.-T.; Chen, K.-H.; Wang, P.-C. Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis. *Neural Comput. Appl.* **2011**, *21*, 2087–2096. [[CrossRef](#)]
55. Karaboga, D.; Ozturk, C. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Appl. Soft Comput.* **2011**, *11*, 652–657. [[CrossRef](#)]
56. Mishra, S.; Tripathy, H.K.; Mallick, P.; Bhoi, A.K.; Barsocchi, P. EAGA-MLP—An Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis. *Sensors* **2020**, *20*, 4036. [[CrossRef](#)]
57. Mishra, S.; Mallick, P.K.; Jena, L.; Chae, G.-S. Optimization of Skewed Data Using Sampling-Based Preprocessing Approach. *Front. Public Heal.* **2020**, *8*, 274. [[CrossRef](#)]
58. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [[CrossRef](#)]
59. Uğuz, H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowl. Based Syst.* **2011**, *24*, 1024–1032. [[CrossRef](#)]
60. Gandhi, K.K.; Prajapati, N.B. Study of Diabetes Prediction using Feature Selection and Classification. *Int. J. Eng. Res. Technol.* **2014**, *3*, 1–7.
61. Jabbar, M.A.; Deekshatulu, B.L.; Chandra, P. Prediction of Heart Disease Using Random Forest and Feature Subset Selection. *Adv. Intell. Syst. Comput.* **2015**, 187–196.
62. Peker, M.; Arslan, A.; Sen, B.; Celebi, F.V.; But, A. A novel hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection and random forest algorithm (ReliefF+RF). In Proceedings of the 2015 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Madrid, Spain, 2–4 September 2015; pp. 1–8.
63. Wagih, M.; Abou-Chadi, F.; El-Din, H.; Mekky, N. Classification of Ultrasound Kidney Images using PCA and Neural Networks. *Int. J. Adv. Comput. Sci. Appl.* **2015**, *6*, 53–57. [[CrossRef](#)]
64. Kavitha, R.; Kannan, E. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. In Proceedings of the 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, India, 24–26 February 2016; pp. 1–5.
65. Tolkacheva, E.G.; Zhao, X.; Zlochiver, S.; Mori, Y. Computational and Mathematical Methods in Cardiovascular Diseases. *Comput. Math. Methods Med.* **2017**, *2017*, 1–2. [[CrossRef](#)] [[PubMed](#)]
66. Haq, A.U.; Li, J.P.; Memon, M.H.; Nazir, S.; Sun, R. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Mob. Inf. Syst.* **2018**, *2018*, 1–21. [[CrossRef](#)]
67. Gutlein, M.; Frank, E.; Hall, M.A.; Karwath, A. Large-scale attribute selection using wrappers. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA, 30 March–2 April 2009; pp. 332–339.
68. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550. [[CrossRef](#)]
69. Verma, L.; Srivastava, S.; Negi, P.C. A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. *J. Med. Syst.* **2016**, *40*, 1–7. [[CrossRef](#)]
70. Yu, L.; Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning, Washington, DC, USA, 21–24 August 2003; Volume 3, pp. 856–863.
71. Maldonado, S.; Weber, R. A wrapper method for feature selection using Support Vector Machines. *Inf. Sci.* **2009**, *179*, 2208–2217. [[CrossRef](#)]
72. Yildirim, P. Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. *Int. J. Mach. Learn. Comput.* **2015**, *5*, 258–263. [[CrossRef](#)]
73. Jothi, G.; Inbarani, H. Soft Set Based Feature Selection Approach for Lung Cancer Images. *Int. J. Sci. Eng. Res.* **2010**, *3*, 1–7.
74. Heshmati, A.; Amjadifard, R.; Shanbehzadeh, J. ReliefF-Based Feature Selection for Automatic Tumor Classification of Mammogram Images. In Proceedings of the 2011 7th Iranian Conference on Machine Vision and Image Processing, Teheran, Iran, 16–17 November 2011; pp. 1–5.

75. Ding, J.; Fu, L. A Hybrid Feature Selection Algorithm Based on Information Gain and Sequential Forward Floating Search. *J. Intell. Comput.* **2018**, *9*, 93–101. [[CrossRef](#)]
76. Jena, L.; Patra, B.; Nayak, S.; Mishra, S.; Tripathy, S. Risk Prediction of Kidney Disease Using Machine Learning Strategies. In *Intelligent and Cloud Computing*; Springer: Singapore, 2019; pp. 485–494.
77. Ray, C.; Tripathy, H.K.; Mishra, S. Assessment of Autistic Disorder Using Machine Learning Approach. In Proceedings of the International Conference on Intelligent Computing and Communication, Hyderabad, India, 9–11 January 2019; pp. 209–219.
78. Sahoo, S.; Mishra, S.; Mishra, B.K.K.; Mishra, M. Analysis and Implementation of Artificial Bee Colony Optimization in Constrained Optimization Problems. In *Handbook of Research on Modeling, Analysis, and Application of Nature-Inspired Metaheuristic Algorithms*; IGI Global: Pennsylvania, PA, USA, 2018; pp. 413–432.
79. Dua, D.; Graff, C. UCI Machine Learning Repository. 2019. Available online: <http://archive.ics.uci.edu/ml> (accessed on 4 November 2020).
80. Panda, B.; Mishra, S.; Mishra, B.K. A Meta-Model Implementation with Tabu Search Technique to Determine the Buying Pattern of Online Customers. *Indian J. Sci. Technol.* **2016**, *9*, 1. [[CrossRef](#)]
81. Mishra, S.; Dash, A.; Jena, L. Use of Deep Learning for Disease Detection and Diagnosis. In *Bio-Inspired Neurocomputing*; Springer: Singapore, 2021; pp. 181–201.
82. Mallick, P.K.; Mishra, S.; Chae, G.-S. Digital media news categorization using Bernoulli document model for web content convergence. *Pers. Ubiquitous Comput.* **2020**, 1–16. [[CrossRef](#)]
83. Krier, D.; Rossi, F.; Verleysen, M. Feature clustering and mutual information for the selection of variables in spectral data. In Proceedings of the European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning, Bruges, Belgium, 25–27 April 2007; pp. 157–162.
84. Van Dijck, G.; Van Hulle, M.M. Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis. In Proceedings of the 16th International Conference on Artificial Neural Networks, Athens, Greece, 10–14 September 2006; pp. 31–40.
85. Song, Q.; Ni, J.; Wang, G. A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. *IEEE Trans. Knowl. Data Eng.* **2011**, *25*, 1–14. [[CrossRef](#)]
86. Xu, Y.-M.; Wang, C.-D.; Lai, J.-H. Weighted Multi-view Clustering with Feature Selection. *Pattern Recognit.* **2016**, *53*, 25–35. [[CrossRef](#)]
87. Chaudhury, P.; Mishra, S.; Tripathy, H.K.; Kishore, B. Enhancing the capabilities of Student Result Prediction System. In Proceedings of the 2nd International Conference on Information and Communication Technology for Competitive Strategies, Udaipur, India, 4–5 March 2016; pp. 1–6.
88. Mishra, S.; Tripathy, H.K.; Panda, A.R. An Improved and Adaptive Attribute Selection Technique to Optimize Dengue Fever Prediction. *Int. J. Eng. Technol.* **2018**, *7*, 480–486. [[CrossRef](#)]
89. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [[CrossRef](#)]
90. Sushruta, M.; Hrudaya, K.T.; Brojo, K.M. Filter Based Attribute Optimization: A Performance Enhancement Technique for Healthcare Experts. *Int. J. Control Theory Appl.* **2017**, *10*, 295–310.
91. Mishra, S.; Tadesse, Y.; Dash, A.; Jena, L.; Ranjan, P. Thyroid Disorder Analysis Using Random Forest Classifier. In *Intelligent and Cloud Computing*; Springer: Singapore, 2019; pp. 385–390.
92. Mishra, S.; Chaudhury, P.; Mishra, B.K.; Tripathy, H.K. An implementation of Feature ranking using Machine learning techniques for Diabetes disease prediction. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, Udaipur India, 4–5 March 2016; pp. 1–3.

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).