Heart Disease Predictive Model Using Filter-Based Selection Techniques and Tree-Like Classifiers

Awe Oluwayomi[1], Aiyeniko Olukayode[2], Adedokun Olufemi Adewale[3], Funso Bukola Omolara[4], and Samuel Ruth Medinat[5]

[1]*Department of Computer Science, University of Lagos, Lagos State, Nigeria*
[2]*Department of Computer Science, Lagos State University, Lagos State, Nigeria*
[3]*Department of Computer Science, University of Ilorin, Ilorin, Nigeria*
[4,5]*Department of Computer Science Kogi State Polytechnic, Kogi State, Nigeria*

Abstract: **The attribute selection is considered a major phase that eliminates redundant attributes thereby improving the accuracy of the predictive or diagnostic model. Designing a model with unrelated attributes may influence the accuracy or result in more memory space used during diagnosis or prediction. This paper examined the impact of the filter-based attribute selection technique on the heart disease diagnostic model. Three filter-based techniques: Relief-F, Information Gain and Chi-square were applied to the heart disease dataset. Five tree-like learning algorithms:** ID3 **(Iterative Dichotomiser 3), C4.5 Decision Tree, Reptree (RP), Random Forest (RF), Classification and Regression Tree (CART) were applied to classify the reduced attributes. The experimental results in terms of accuracy, precision and recall showed that the relief-f attribute selection outperformed information gain and chi-square with the best predictive accuracy of 93.4983% in IDE, the precision value of 0.93500 in IDE and recall value of 0.93500 in IDE classifier.**

***Keywords*: Chi-square, Data Mining, Filter-based, Relief-F, Information Gain**

I. INTRODUCTION

Several deadly ailments affect humans, one among these diseases is heart disease [1]. Heart disease is common sickness in adults, this has currently increased the death rate over the globe [2]. Heart disease is an illness that affects the circulatory system of the heart [3]. The medical domain is surrounded by huge data but inadequate action has been given to this data in proffering answers to some life-threatening problems such as diagnosis of diseases. Available are many approaches to accomplish this task, but data mining remains the most significant method [4]. The approach of data mining uses data to methodically or logically discover inadequacies, ease costs and improve upon best practices in medicine [5]. Computational intelligent systems or data mining are tools that can be employed to perform predictions with medical datasets with many responses. Heart disease prediction is a difficult process that necessitates being achieved correctly and creditably [6]. Conclusions made by physicians may occasionally be centered on instinct rather than on the unknown information in a patient's data, this sometimes results in unwanted flaws and expensive costs in healthcare which also constitute adverse effects on the standard of service given to patients.

Many methods in data mining have been rendered to examine and infer unidentified relations that occur in features of clinical data to carry out some tasks such as prediction and diagnosis of diseases [7]. An evolving field with importance for the provision of diagnosis and comprehensive facts about medical data is data mining [8]. Prediction of heart disease by the application of techniques in data mining in hospitals is considered a suitable solution to confusing matters, which assists medical staff in arriving at intelligent conclusions to advance the standard of decisions in healthcare [9]. Reduction in the number of deaths from heart diseases requires quick and efficient prediction. Approaches to accomplish robust outputs in heart disease diagnosis such as combining the learning methods and attribute reduction techniques have been developed by researchers [10] [11]. One of the data pre-processing techniques is attributes reduction that is needed to incorporate a systematized structure into the information before its passage to the mining processor [12]. The selection of attributes is a dynamic domain of study in pattern recognition and data mining communities [13]. Selection of attributes is used to produce a subset of input variables by eliminating irrelevant attributes with little or no predictive information [2]. This technique can meaningfully enhance the unambiguousness of the resulting classifier models and often construct a model that simplifies better to hidden points. Attribute selection techniques can be grouped into the filter, wrapper and hybrid techniques.

Filter technique is generally used on data with huge dimensionality due to its generalization and complex computation [14]. The significance of the target variable is determined by the level of relationship between attributes and the target variable [15] The selection of attributes utilizing a filter is independent of classifiers, faster and more accessible than wrapper and hybrid techniques.

The paper applied three filter-based attribute selection methods to decrease the attributes of the heart disease dataset and also classified the decreased attributes app five learning algorithms: IDE, C4.5 Decision Tree, RP, CART and RF.

II. LITERATURE REVIEW

2.1 Attribute Selection

This is an approach to obtaining an important or possible attribute subset from the original attributes [16]. Different from attribute extraction, attribute selection is used on datasets with identified attributes. The techniques try to know the essential attributes and remove irrelevant or redundant attributes from the original attributes. There are several examples of attribute selection methods [17].

(i) Information Gain (IG)

The technique is based on entropy, generally applied in the field of machine learning [18]. The technique is employed as attribute selection, which explains the information level formed by the attribute items for the type of text [19]. Information gain aids the measurement of how mixed up the attributes are in text and data mining [20].

(ii) Relief-F

Relief-F calculates the value of an attribute by repetitively finding an instance and taking the attribute value given for the nearest instance of the equivalent and unlike class [21]. The method gives a value to each attribute based on the capability of the attribute to distinguish among the classes and then selects those attributes whose values are more than the user-defined threshold as significant attributes [22]. The value calculation is determined by the chance of the nearest neighbours from two different classes getting different values for an attribute and the chance of two nearest neighbours of the same class getting the same value of the attribute. The greater the difference between these two chances, the more relevant the attribute.

(iii) Chi-Squared

This approach is usually employed for the selection of features that evaluates the value of an attribute by calculating the chi-squared value concerning the class [23]. The first hypothesis H0 is the postulation that the two features are not the same, and it is tested by the chi-squared formula as shown in Equation (1).

$$X^2 = \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (1)$$

Where O_i the observed frequency and E_i the expected frequency is asserted by the null hypothesis. The greater the value X^2 the greater the evidence against the hypothesis H0.

2.2 Related Work

Several studies have been conducted employing attribute selection techniques in data mining before the classification process. These studies include:

A broad review was carried out to draw a summary of selected current and dominant machine learning approaches in all the contexts [24]. The highest number of papers reviewed in the study were from IEEE Explorer, Science Direct, PubMed, Springer, Hindawi, ACM digital library and MDPI libraries. It was shown that Support Vector Machines (SVM) and Artificial Neural Networks (ANN) were superseded in most of the studies in all the frameworks. A deep neural network was a comparatively newer machine learning method that is giving noticeable results in the classification of heart sound signals and cardiovascular images.

[25] experimentally assessed the effectiveness of models derived by techniques of learning algorithm applying appropriate attributes obtained by different attribute-selection techniques. Four heart disease datasets were tested using Principal Component Analysis, Chi-squared, Relief-f and Symmetrical uncertainty to produce unique attribute sets. Several classification algorithms were applied to produce models that were compared to find the optimal feature combinations to enhance the prediction rate of heart conditions. It was discovered that the advantages of using attribute selection are different subjects in the learning algorithm applied to the heart disease database. Results showed that the best model was obtained when a chi-squared attribute selection method was combined with the BayesNet algorithm in which an accuracy of 85.00% was achieved on the considered datasets.

[26] provided a survey of approaches on attribute selection and classification for the prediction of heart disease in the last ten years. Thus, this type of review is also applied as a reference in the selection of a technique for the prediction of heart disease in prospect.

[1] designed model by applying two supervised learning algorithms; random forest and logistic regression. Approach to heart disease prediction employing random forest with robust attributes predicts the heart failure stage of the patient. Logistic Regression Analysis was employed in the model to deduce how assertive can the predicted value be the real value when given as the patient's test data details in the prediction of the stages.

[27] presented a mixed method on the disease of heart dataset; the outputs showed the efficiency of the developed hybrid technique in dealing with different data types for the classification of heart disease. The study looked at several algorithms of machine learning and compared the outputs using diverse evaluation metrics, accuracy, precision, recall, f1-score and so on. The highest classification accuracy of 99.65% was recorded in optimized model proposed by FCBF, PSO and ACO. The results revealed that the effectiveness of the projected model outperformed that of the classification method.

[17] improved upon predictive model for heart attack by applying attribute selection methods. The main target of the study is to find out the best machine learning method and the best attribute selection algorithm to predict heart attacks. Many methods of machine learning with optimum parameters

and several techniques of attribute selection were used and evaluation was done using Starlog Heart disease dataset. Experimental results showed that SVM algorithm was the best machine learning with the linear kernel, while the relief-F method was identified as the best attribute selection algorithm. The two algorithms recorded the highest accuracy of 84.81%.

Development of a predictive model which contained two subsystems, the RFRS attribute selection and classification with an ensemble classifier [28]. The first system consists of three phases; discretization of data, extraction of feature using the Relief-F and reduction of feature with the heuristic Rough Set reduction algorithm developed. In the second system, an ensemble classifier was developed based on the C4.5 classifier. The Statlog Heart dataset, obtained from the UCI database was used for the simulation of the model. The highest classification accuracy of 92.59% was obtained.

A review was conducted by [3]. The study summarised selected current research on the predictive model for heart diseases using techniques of data mining, examined the different arrangements of mining algorithms and arrived at a conclusion on which techniques are effective and robust.

[29] a developed predictive model for heart disease applying feature selection methods with data mining techniques. Heart patient data from the UCI repository was used for experimentation. There were 689 unique instances. Different learning algorithms were applied to classify the data accurately. Certain significant points were reflected upon to obtain an appropriate tool for mining. Accuracy was considered to test the model performance. The results revealed that regression algorithms and SVM recorded the best accuracy of 99.7% compared with other approaches.

[29] applied data mining methods for heart disease diagnosis. Several classification techniques in heart disease diagnosis such as Decision Tree, K Nearest Neighbours (KNN), Naive Bayes (NB), and SMO were used to classify the dataset. Evaluation metrics like accuracy, precision, sensitivity, specificity, F-measure and area under the ROC curve were used to evaluate and compared the model. The comparative analysis outputs indicated that the Decision tree was identified as the best classifier for diagnosis of heart disease in the current dataset.

III. METHODOLOGY

The heart disease diagnostic model by applying selected techniques of filter-based selection approach and tree-like classifiers was conducted step wisely. Information Gain, Relief-F and Chi-square techniques were used to reduce attributes from the heart disease dataset. Five classifiers: C4.5 Decision Tree Artificial Neural Network, K-Nearest Neighhood, Naïve Bayes and Support Vector Machine were applied to classify the attributes. The evaluation of the model was achieved using a heart disease dataset. The dataset was used to diagnose the presence or absence of heart disease given the result of various medical tests carried out on the patients. Finally, an analysis was done comparatively on the attribute selection methods to determine the effect of the filter-based attribute selection techniques on the heart disease dataset.

3.1 Dataset Acquisition/Description

The collection of data was from four sources: Cleveland Clinic Foundation, Hungarian Institute of Cardiology in Budapest, V.A. Medical Center, Long Beach in CA and University Hospital, Zurich, Switzerland. The attributes in this dataset contain 14 attributes. The study used an online heart dataset that is publicly available in the UCI repository to evaluate the performance of the heart disease diagnosis.

3.2 Method of Analysis

The Weka data mining simulator tool was used to carry out the experimental analysis of heart datasets from the UCI repository. The heart disease dataset was properly pre-processed. Attributes from the dataset were reduced using filter techniques of attribute selection. The impact of the attribute selection techniques on the heart disease dataset was analysed. Finally, the individual classifier was evaluated in terms of accuracy, precision and recall.

IV. RESULTS AND DISCUSSION

4.1 Results of the Heart Disease Diagnostic Model

This section mentions and compares the various results obtained during performance evaluation based on different evaluation metrics such as accuracy, recall, and precision. The results of the analysis are shown in Tables 1, Table 2 and Table 3.

Table 1: Performance Evaluation of the Heart Predictive Model (Accuracy %)

Learning Algorithm	Attribute Selection Method (%)			
	No Attribute Selection	Relief-F	Information Gain	Chi-Square
IDE	92.8581	93.4983	91.5182	92.1782
C4.5	91.2284	92.4983	92.1584	92.8284
CART	91.5578	91.9770	92.6780	92.9076
RP	92.2370	93.3076	92.5810	92.2875
RF	91.3230	92.4900	92.4570	92.1880

From Table 1, the effect of attribute selection is noticeable in the Relief-F attribute selection method on the heart disease dataset with a predictive accuracy of 93.4983% obtained in IDE. The lowest predictive accuracy of 91.22845% was recorded in C4.5 when no attribute selection was applied, thus there is a clear observation that out of the three attribute selection techniques for the C4.5 classifier, chi-square outperformed others with a predictive accuracy of 92.8284%.

Table 2: Performance Evaluation of the Heart Disease Predictive Model (Precision)

Learning Algorithm	Attribute Selection Method (%)			
	No Attribute Selection	Relief-F	Information Gain	Chi-Square
IDE	0.92860	0.93500	0.91500	0.92180
C4.5	0.91230	0.92330	0.92200	0.92830
CART	0.91560	0.92000	0.92700	0.92910
RP	0.92300	0.93300	0.92580	0.92200
RF	0.91300	0.92500	0.92640	0.92190

From Table 2, the impact of attribute selection is seen in the Relief-F attribute selection method on the heart disease dataset with the precision value for the predictive model of 0.93500 obtained in IDE. The lowest precision value of 0.91230 was recorded in C4.5 when no attribute selection was applied, thus there is a clear observation that out of the three attribute selection techniques for the C4.5 classifier, chi-square outperformed others with a precision value of 0.92830.

Table 3: Performance Evaluation of the Heart Disease Predictive Model (Recall)

Learning Algorithm	Attribute Selection Method			
	No Attribute Selection	Relief-F	Information Gain	Chi-Square
IDE	0.92900	0.93500	0.91500	0.92190
C4.5	0.91200	0.92340	0.92210	0.92850
CART	0.91510	0.92100	0.92730	0.92940
RP	0.92310	0.93310	0.92600	0.92210
RF	0.91300	0.92520	0.92640	0.92190

From Table 3, the impact of attribute selection is shown in the Relief-F attribute selection method on the heart disease dataset with the recall value of 0.93500 recorded in the IDE classifier. The lowest recall value of 0.91200 was recorded in C4.5 when no attribute selection was applied, thus it was observed that out of the three attribute selection techniques for the C4.5 classifier, chi-square outperformed others with a recall value of 0.92850.

V. CONCLUSION

Attribute selection has been identified to be one of the significant stages in diverse fields such as data mining and pattern recognition, this process includes the reduction of input data into measurable attributes for the analysis and proper data processing. The importance of attribute selection is to obtain the subset of the original attribute which helps in building a robust learning model. The filter-based selection technique remains the widely employed approach due to its independence from the learning algorithms and less computational time among the techniques of attribute selection. This paper conducted a comparative analysis to reveal the effect of three filter-based attribute selection techniques on the prediction of heart disease. Relief-F, Information Gain and chi-square were applied to select attributes of the heart disease dataset. Results revealed that Relief-F performed effectively compared with other filter-based techniques.

REFERENCES

[1] G. C. Anusha, M. S. Apoorva, N. Deepthi, V. Dhanushree, and R. Firdaus, "Heart Disease Diagnosis Using Machine Learning," Int. J. Eng. Res. Technol., vol. 7, no. 10, pp. 1–7, 2020.

[2] N. Boyko and I. Dosiak, "Analysis of Machine Learning Algorithms for Classification and Prediction of Heart Disease," in 4th International Conference on Informatics & Data-Driven Medicine, 2021, vol. 9363, pp. 0–2.

[3] S. K. Mandal, A. Gupta, A. Mukherjee, and A. Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review," Adv. Comput. Sci. Technol., vol. 10, no. 7, pp. 2137–2159, 2017.

[4] M. M. Altaf, N. F. Idris, and M. Arfian, "Machine Learning Classification Techniques for Breast Cancer Diagnosis Machine Learning Classification Techniques for Breast Cancer Diagnosis," in OP Conference Series: Materials Science and Engineering, 2019, pp. 1–17.

[5] T. Karthikeyan, B. Ragavan, and V. A. Kanimozhi, "A Study on Data mining Classification Algorithms in Heart Disease Prediction," Int. J. Adv. Res. Comput. Eng. Technol., vol. 5, no. 4, pp. 1076–1081, 2016.

[6] V. V Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: a survey," Int. J. Eng. Technol., vol. 7, no. 2, pp. 684–687, 2018.

[7] M. S. Gharibdousti, K. Azimi, S. Hathikal, and D. H. Won, "Prediction of Chronic Kidney Disease using data mining techniques," 67th Annu. Conf. Expo Inst. Ind. Eng. 2017, no. February 2018, pp. 2135–2140, 2017.

[8] S. A. Patil, P. R., & Kinariwala, "Automated Diagnosis of Heart Disease using Data Mining Techniques," Int. J. Adv. Res. Ideas Innov. Technol., vol. 3, no. 2, pp. 560–567, 2017.

[9] S. S. Mirzajani and siamak Salimi, "Prediction and Diagnosis of Diabetes by Using Data Mining Techniques," Avicenna J. Med. Biochem., vol. 6, no. 1, pp. 3–7, 2018.

[10] L. Duan, C. Aggarwal, M. Shuai, M. Tiejun, and J. Huai, "An Ensemble Approach to Link Prediction," IEEE Trans. Knowl. Data Eng., vol. 10, no. 10, pp. 1–14, 2017.

[11] S. Vijayarani and S. M. Sylvia, "Comparative Analysis of Dimensionality Reduction Techniques," Int. J. Innov. Res. Comput. Commun. Eng., vol. 4, no. 1, pp. 23–29, 2016.

[12] N. Varghese and V. Verghese, "a A Survey of Dimensionality Reduction and Classification," Int. J. Comput. Sci. Eng. Surv., vol. 3, no. 3, pp. 45–54, 2012.

[13] V. Kumar and S. Minz, "Feature Selection: A Literature Review," Smart Comput. Rev., vol. 4, no. 3, 2014.

[14] M. Kayaalp, F., & Basaralan, "Performance Analysis of Filter Based Feature Selection Methods On Diagnosis Of Breast Cancer And Orthopedics," in 6th International Congress on Fundamental and Applied Sciences, 2019, pp. 1–12.

[15] D. Ashraf, M., Chetty, G. & Tran, "Feature Selection Techniques on Thyroid, Hepatitis and Breast Cancer," Int. J. Data Min. Intell. Inf. Technol. Appl., vol. 3, no. 1, p. 2013, 2013.

[16] S. Patra, B., & Dash, "A FRGSNN Hybrid Feature Selection Combining FRGS filter and GSNN wrapper," Int. J. Latest Trends Eng. Technol., vol. 7, no. 2, pp. 8–15, 2016.

[17] H. Takci, "Improvement of heart attack prediction by the feature selection methods," Turkish J. Electr. Eng. Comput. Sci., vol. 2, no. 1, pp. 1–10, 2018.

[18] S. iyaswamy V. Hariharan, K., Vigneshwar, W.S., Sivaramakrishn an, N., "A Comparative Study on Heart Disease Analysis Using," Int. J. Pure Appl. Math., vol. 119, no. 12, pp. 13357–13366, 2018.

[19] J. Novaković, P. Strbac, and D. Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," Yugosl. J. Oper. Res., vol. 21, no. 1, pp. 119–135, 2011.

[20] A. Sonam Ni., & Karandikar, "Prediction of Heart Disease Using Machine Learning Algorithms," Int. J. Adv. Eng. Manag. Sci., vol. 2, no. 6, pp. 617–620, 2016.
[21] S. K. Prabhakar, H. Rajaguru, and D. O. Won, "A Holistic Performance Comparison for Lung Cancer Classification Using Swarm Intelligence Techniques," J. Healthc. Eng., pp. 1–13, 2021.
[22] K. B. Al Janabi and R. Kadhim, "Data Reduction Techniques: A Comparative Study for Attribute Selection Methods," Int. J. Adv. Comput. Sci. Technol., vol. 8, no. 1, pp. 1–13, 2018.
[23] O. Olatubosun, I. Gabriel, O. Samuel, and A. Tunde, "Hough Transform and Chi-Square-Based Iris Recognition," International J. Comput., vol. 21, no. 1, pp. 1–15, 2016.
[24] A. O. Balogun et al., "A novel rank aggregation-based hybrid multifilter wrapper feature selection method in software defect prediction," Entropy, vol. 23, pp. 1–28, 2021.
[25] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," Digit. Heal., vol. 6, pp. 1–10, 2020.
[26] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," Int. J. Pharm. Res., vol. 12, no. 4, pp. 56–66, 2020.
[27] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," Int. J. Intell. Eng. Syst., vol. 12, no. 1, pp. 242–252, 2019.
[28] X. Liu et al., "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method," Comput. Math. Methods Med., vol. 2017, pp. 1–12, 2017.
[29] K. Uma and M. Hanumanthappa, "Feature Selection Methods for Heart Disease Prediction with Data Mining Techniques," in Seventh International on Advances in Computer Engineering, 2016.