PRINT ISSN 1119-8362
Electronic ISSN 2659-1502

Full-text Available Online at
https://www.ajol.info/index.php/jasem
http://www.bioline.org.br/ja

J. Appl. Sci. Environ. Manage.
Vol. 26 (6) 1093-1098 June 2022

Enhancing the Performance of Heart Disease Prediction from Collecting Cleveland Heart Dataset using Bayesian Network

MUSA, UA; MUHAMMAD, SA

Department of Computer Science, Federal University Dutse, Dutse, Jigawa State, Nigeria

Corresponding Author Email:* **Usman.m@fud.edu.ng
Other Author Email: msirajoa@fud.edu.ng

ABSTRACT: Cardiovascular diseases are diseases affecting the general well-being of the heart. It is responsible for many deaths annually. Consequently, this paper focuses on improving the performance of heart disease prediction by collecting Cleveland heart datasets from the University of California Irvine machine learning repository. Different feature subset selection is performed on the dataset and modeled using machine learning models such as logistic regression, K-Nearest neighbor, Naïve Bayes and Bayesian Network. The proposed method achieved an accuracy of 88.53%. Based on the results obtained, we observed feature reduction on the Cleveland dataset could enhance the performance of the Bayesian network.

DOI: https://dx.doi.org/10.4314/jasem.v26i6.15

Impact factor: http://sjifactor.com/passport.php?id=21082

Google Analytics: https://www.ajol.info/stats/bdf07303d34706088ffffbc8a92c9c1491b12470

Dates: Received: 03 March 2022; Revised: 16 June 2022; Accepted: 20 June 2022

Keywords: Machine Learning, Bayesian Network, Naïve Bayes Logistic Regression, Heart Disease

Cardiovascular diseases are conditions that influence the construction or capacity of your heart, for example, Abnormal heart rhythms/arrhythmias, Aorta infection, Marfan disorder, Congenital coronary illness, Heart muscle sickness (cardiomyopathy), Stroke, and so on (Steinbaum 2019). These diseases share common risk factors namely; age, unhealthy diet sexual orientation, hypertension, diabetes mellitus, tobacco smoking, processed meat utilization, unnecessary liquor consumption, sugar consumption, family ancestry, weight, absence of exercise, psychosocial factors, and air contamination (WHO 2019). According to the Federal Ministry of Health (FMoH), "cardiovascular disease is a major public health concern, accounting for 11% of the over 2 million NCD deaths in Nigeria each year." (WHO 2019). The increase in the amount of health data gathered through the electronic health record (EHR) systems makes the use of strong analysis tools necessary. The need of making accurate predictions of heart disease made the use of machine learning algorithms to point out predictions based on many factors. Numerous machine learning algorithms such as Random Forest, Logistic Regression, ANN, K-Nearest Neighbor, SVM, etc. have been applied to Cleveland heart datasets however, not very much was done on modeling with a Bayesian Network (BN). Bayesian networks (BNs) have received increasing research attention as it possesses potential significant benefit to the healthcare system.

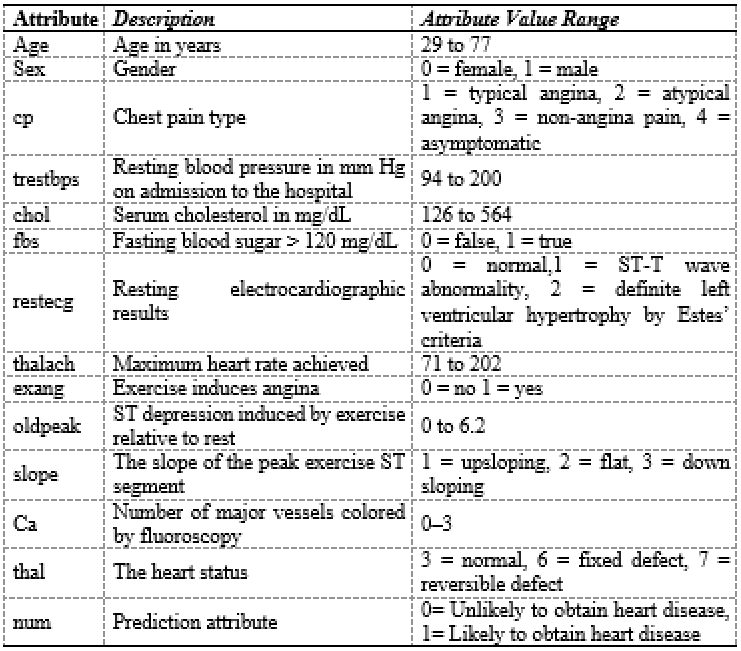
Many factors or input features even to machine learning often make a predictive modeling task more challenging to model thereby necessitating feature engineering. Feature engineering is important in increasing the efficiency and correctness of prediction on machine learning models. The study aimed to apply Wrappers for feature subset selection, Consistency Subset Evaluation, and Correlation-based Feature Subset Selection as means of feature reduction technique for extracting important features subset and the use of Naïve Bayes, Bayesian Network, KNN, and Logistic Regression to make predictions.

MATERIALS AND METHODS

Data Collection: To make a model for the prediction of heart disease and improvement of the model, the open-source program WEKA was used in the data mining procedure. WEKA is a tool for knowledge analysis that has multiple machine learning algorithms for data analysis (Srivastava 2014). The first step involves data collection followed by data preprocessing. After prepossessing the data, the full data set is put trained on the selected models and then evaluated. The preprocessed data is then put into a feature reduction technique and feature selection is done. The reduced dataset is then trained on the models and evaluated using the following steps:

Dataset Retrieval: This research uses the heart disease dataset downloaded from the University of California Irvine (UCI) machine learning repository called the Cleveland Heart Disease Data set. Cleveland Heart Disease Dataset is a publicly available supervised dataset provided by the Cleveland Clinic Foundation that was used for the ML model. This data set contains 14 total attributes of patient medical information for 303 patients. Table 1 shows the chosen attributes and their information.

Table 1: List of Attribure in Cleveland heart diseae Dataset



Attribute	*Description*	*Attribute Value Range*
Age	Age in years	29 to 77
Sex	Gender	0 = female, 1 = male
cp	Chest pain type	1 = typical angina, 2 = atypical angina, 3 = non-angina pain, 4 = asymptomatic
trestbps	Resting blood pressure in mm Hg on admission to the hospital	94 to 200
chol	Serum cholesterol in mg/dL	126 to 564
fbs	Fasting blood sugar > 120 mg/dL	0 = false, 1 = true
restecg	Resting electrocardiographic results	0 = normal,1 = ST-T wave abnormality, 2 = definite left ventricular hypertrophy by Estes' criteria
thalach	Maximum heart rate achieved	71 to 202
exang	Exercise induces angina	0 = no 1 = yes
oldpeak	ST depression induced by exercise relative to rest	0 to 6.2
slope	The slope of the peak exercise ST segment	1 = upsloping, 2 = flat, 3 = down sloping
Ca	Number of major vessels colored by fluoroscopy	0–3
thal	The heart status	3 = normal, 6 = fixed defect, 7 = reversible defect
num	Prediction attribute	0= Unlikely to obtain heart disease, 1= Likely to obtain heart disease

Data Preprocessing: Data preprocessing is also known as cleaning data. It is one of the most important steps to achieve the best from the dataset. This is a technique that removes data inconsistencies such as missing numbers, out-of-range values, unformatted data, and noise. Our preprocessing would involve data handling missing values, and data discretization. Handling Missing Values is a common problem faced by analysts. This occurs due to different reasons such as incomplete extraction, corrupt data, failure to load the information, etc. This is a great challenge that must be fixed because good models are generated when you make the right decisions on how to fix them. The missing numerical values were replaced with a mean value. Data Discretization is the technique of transforming continuous data attribute values into a finite set of intervals with minimal information loss known as data discretization. Discretization can help improve significantly the classification performance of some algorithms like Naïve Bayes that are sensitive to the dimensionality of the data (Lustgarten *et al.*, 2008).

Feature Selection: To select the best features for our model, Wrappers for feature subset selection, Consistency Subset Evaluation, and Correlation-based Feature Subset Selection as means of feature reduction were performed on the WEKA data mining tool and the result of each technique is presented in the result section.

Models Used In The Study: The proposed models comprise Naïve Bayes, Bayesian Network, KNN, and Logistic Regression. This is because the models proved to have high performance in predicting heart disease from previous studies.

Naive Bayes: Naïve Bayes or stupid Bayes is used to handling binary (two-class) and multiclass classification challenges, It has its name because the probabilities for each hypothesis are simplified to make its calculation tractable. (Jason 2014). In simple terms, a Naive Bayes classifier assumes that the presence of one attribute or feature in a class is unrelated to the presence of any other feature, i.e., predictor independence.

$$P(A|B) = \frac{P(B|A) \text{ x } P(A)}{P(B)}$$

Where P(A) is the prior distribution of parameter A; P(A|B) is the posterior distribution, the probability of A given new data B; and P(B|A) is the likelihood function, the probability of B given existing data.

Bayesian Network: A Bayesian network B = < N, A, θ > is a directed acyclic graph (DAG) <N, A> with a conditional probability distribution (CP) for each node, collectively represented by θ. Each arc a ∈ A between nodes represents a probabilistic dependency, and each node n ∈ N represents a domain variable. In general, a BN can be used to compute the conditional probability of one node, given values assigned to the other nodes; hence, a Bayesian Network can be used as a classifier to calculate the posterior probability distribution of a classification node given the values of other characteristics. (Cheng *et al.,* 2002). A Bayesian network, for example, could reflect the probability correlations between diseases and symptoms. Given a set of symptoms, the network may be used to calculate the likelihood of the presence of certain diseases. The Figure 1 below show a DAG representing Bayesian network P(x1) P(x2|x1) P(x3) P(x4|x1) P(x5|x2,x3,x4) P(x6|x3).

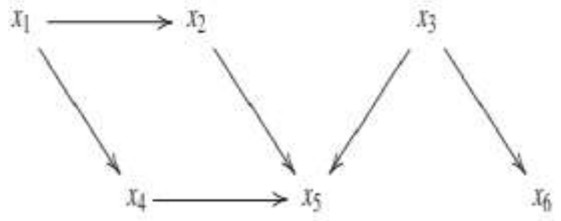


Figure 1**.** DAG representing a Bayesian

The main role of the network structure is to express the conditional independence relationships among the variables in the model through graphical separation, thus specifying the factorisation of the global distribution(Scutari, 2017). Bayesian networks provide a framework for presenting causal relationships and enable probabilistic inference among a set of variables(Bouchra *et al.,* 2019). They can be used to explore and display causal relationships between key factors and system outcomes. (Pollino and Henderson, 2010).

Logistic regression: The logistic function, also known as the sigmoid function, was created by statisticians to characterize the properties of population increase in ecology, such as how it rises swiftly and eventually reaches the environment's carrying capacity. It's an S-shaped curve that can transfer any real-valued integer to a value between 0 and 1 but never exactly between those two points.

$$\frac{1}{1+\ e^{-value}}$$

Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform.

Logistic regression has two alternative outcomes of a goal variable. This means that the input and output have a linear relationship and calculate the likelihood of the goal variable of the data. Logistic regression still considers the dependent variable to be bi-categorical. It is mostly used to forecast and calculate the likelihood of success. Molding the equation into the form of needed data entry is also part of Logistic regression. A basic equation is used here:

$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots\ldots\ \beta_nX_n$

The regression coefficients are estimated using the Maximum Likelihood Ratio (MLR). It aids in the calculation of statistical significance for dependent variables using independent variables. MLR tests and assesses the part of the independent variables.

K-Nearest Neighbour: KNN makes predictions using the training dataset directly. For each new data point, predictions are formed by exploring the whole training set for the k most similar examples (neighbors) and summing the output variable for those k instances. The most similar of the k instances in the training dataset to a new input is determined using a distance metric. The most frequent distance measure for real-valued input variables is Euclidean distance. The square root of the total of the squared discrepancies between two points a and b across all input qualities *i* is used to calculate the Euclidean distance.

$$\text{Euclidean distance}(\text{a}, \text{b}) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

KNN extracts data points from a dataset and calculates the closest output. Because there are various features in the heart disease dataset, this technique works well with pattern recognition. Along with the majority of KNN, it extracts logic and knowledge using the Euclidean distance Samples function d(Xi, Xj). Mathematically;

$$d(x_i, x_j) = \sqrt{(x_{i,1} - x_{j,1})^2 + \cdots + \ (x_{i,m} - x_{j,m})^2}$$

Performance Metrics Adopted For The Study: Performance metrics are used to evaluate how different algorithms perform based on various criteria The matrices used in the research include; Accuracy, Precision, Recall/ Sensitivity, F1 Score, MCC AUROC curve, and Kappa Statistics

RESULTS AND DISCUSSION

In total for machine learning classifiers, 303 records were used each with 14 total attributes. Three different attribute feature subset selections were made. Each feature selection technique's performance is evaluated. This list of the feature selection with selected attributes are shown as subscripts beneath Table 2. The result of the three different feature selection subset using Bayesian network is shown in the table 2:

Table 2. performance of the full dataset against that of the selected feature reduction techniques

Dataset	Precision	Recall	F-Measure	MCC	ROC Area	Accuracy
Dataset Before reduction	0.869	0.869	0.869	0.737	0.927	86.8852
Wrappers for feature subset selection	**0.888**	**0.885**	**0. 885**	**0. 77**	**0. 919**	**88.5246**
Consistency Subset Evaluation	0.869	0.869	0.869	0.737	0.927	86.8852
Correlation-based Feature Subset Selection	0.838	0.836	0.836	0.674	0.919	83.6066

Wrapper Subset Evaluation: : *age, sex, cp, exang, oldpeak, slope, ca, thal*
Correlation-based Feature Subset Selection: *sex, cp, restecg, thalach, exang, oldpeak, slope, ca, thal*
Consistency Subset Evaluation: *age, sex, cp, fbs, restecg, thalach, exang, oldpeak, slope, ca,thal*

As shown in Table 2, out of the 3 different feature selection techniques, the Wrapper feature subset proves to have better performance with the highest accuracy, Precision, Recall F-Measure, and MCC score. The wrapper feature subset had selected 8 out of the 14 features as it was shown to play a part in determining the best feature subset in heart disease diagnosis. The proportions of the feature set are also evaluated. The evaluation involved different classification methods, i.e., Naïve Bayes, Bayesian Network, KNN, and Logistic Regression. In this study, WEKA was used to classify the Cleveland dataset. This is done using different proportions of the dataset for training and testing and the accuracy measure is taken as shown in Table 3. In Table 3, Splitting the training and testing data into an 80:20 ratio has the highest accuracy. As shown in Table 3, 10 different proportions of training and testing ratios were split into the data with 4 different algorithms (ie. Naïve Bayes, Bayesian network, KNN, and Logistic regression) with accuracy as a performance measure. It is observed among all the results obtained that; the Bayesian network had the highest accuracy score on the 80:20 training/testing split with 88.553%. To have better insight into the performance of the split with the highest accuracy. The performance is shown on different other metrics and given in table 4.

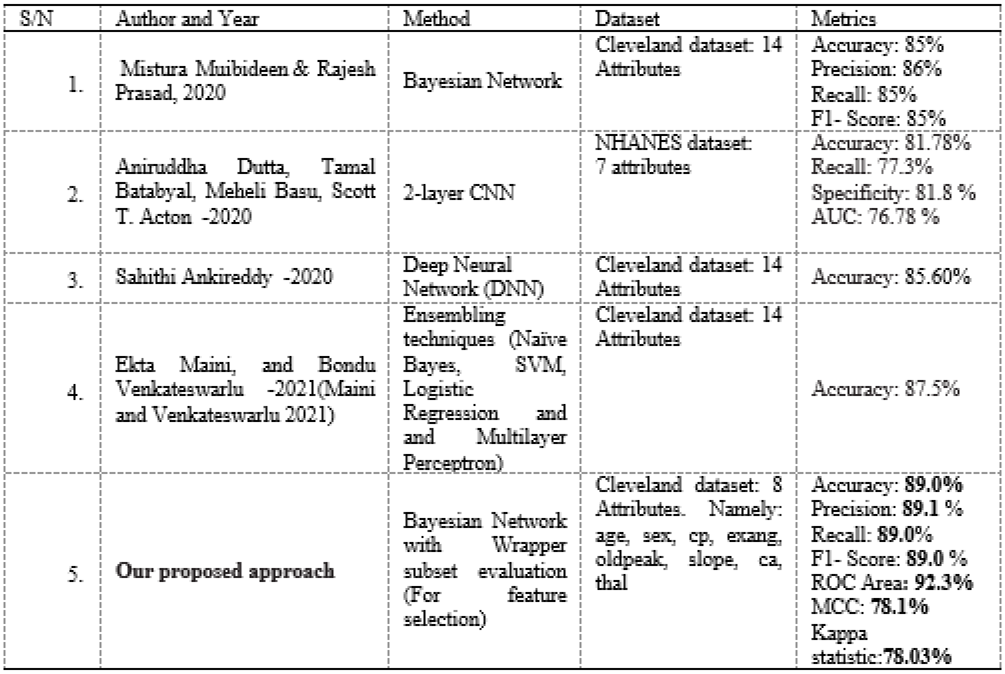
Table 3. Test proportion to determine the best ratio for the highest model accuracy (rounded to 4 digits)

ALGORITHMS	**Training/Testing Accuracy Ratio**								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Naïve Bayes	82.4176	83.8843	85.3774	84.0659	86.0927	87.6033	86.8132	86.8852	86.6667
Bayesian Network	81.685	83.4711	86.3208	85.1648	87.4172	87.6033	86.8132	88.5246	86.6667
KNN	79.8535	79.3388	79.717	80.2198	82.1192	80.1653	82.4176	80.3279	86.6667
Logistic Regression	68.8645	65.2893	75.000	82.4176	86.0927	83.4711	81.3187	85.2459	83.3333

Table 4: Matrices Comparison between Naive Bayes, Bayesian Network, Logistic Regression, and KNN

Model	Accuracy	Precision	Recall	F-Measure	MCC	ROC Area	Kappa statistic
Naïve Bayes	86.8852	0.870	0.869	0.868	0.738	0.920	0.7362
Logistic Regression	86.8852	0.870	0.869	0.868	0.738	0.908	0.7362
KNN	80.3279	0.804	0.803	0.803	0.606	0.873	0.6043
Model adopted by the study (Bayesian Network)	88.5246	0.888	0.885	0. 885	0. 77	0. 920	0.7803

Table 5: Comparison of Various Approaches with our Proposed Approach



S/N	Author and Year	Method	Dataset	Metrics
1.	Mistura Muibideen & Rajesh Prasad, 2020	Bayesian Network	Cleveland dataset: 14 Attributes	Accuracy: 85% Precision: 86% Recall: 85% F1- Score: 85%
2.	Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton -2020	2-layer CNN	NHANES dataset: 7 attributes	Accuracy: 81.78% Recall: 77.3% Specificity: 81.8 % AUC: 76.78 %
3.	Sahithi Ankireddy -2020	Deep Neural Network (DNN)	Cleveland dataset: 14 Attributes	Accuracy: 85.60%
4.	Ekta Maini, and Bondu Venkateswarlu -2021(Maini and Venkateswarlu 2021)	Ensembling techniques (Naïve Bayes, SVM, Logistic Regression and and Multilayer Perceptron)	Cleveland dataset: 14 Attributes	Accuracy: 87.5%
5.	**Our proposed approach**	Bayesian Network with Wrapper subset evaluation (For feature selection)	Cleveland dataset: 8 Attributes. Namely: age, sex, cp, exang, oldpeak, slope, ca, thal	Accuracy: **89.0%** Precision: **89.1** % Recall: **89.0%** F1- Score: **89.0** % ROC Area: **92.3%** MCC: **78.1%** Kappa statistic:**78.03%**

As shown in Table 4, the performance of the various models is given. Each model had proved to have good predictive power to heart disease. However, looking at the MCC score and accuracy score Bayesian Network, despite having a similar ROC Area score with Naive Bayes was the overall winner**.** Bayesian Network proved to be the better overall model as seen by its much higher performance when looking at accuracy score, precision, recall, F-Measure, Matthews Correlation Coefficient, and Kappa statistic. Thus, it can be deemed that it was the best algorithm out of the 4 tested.

Conclusion: The research method in the study reduces the dimensionality of the dataset using the WEKA wrapper method of data selection to select the best subset of Cleveland dataset features for better accuracy and efficiency in predicting heart disease. The selected features are had improve the prediction performce compared to the unreduced dataset. The method in the study has been evaluated with various metrics, and its performance results are compared with explores different machine learning algorithms. A very detailed, useful, and highly preferable Machine Learning-based model in this paper that helps medical practitioners diagnose heart diseases at an early stage to enable patients to take precautionary measures in a rectification window.

Acknowledgments: The author(s) wish to acknowledge the kind support of Dr. Muhammad Sirajo Aliyu for providing useful guidance in the research work.

REFERENCES

Cheng, J; Russell, G; Jonathan, K; David, B; Weiru, L. (2002). Learning Bayesian Networks from Data: An Information-Theory Based Approach. *Artificial Intelligence* 137 (1–2): 43–90.

Jason, B. (2014). "Naive Bayes for Machine Learning." *Machine Learning Algorithms*. https://machinelearningmastery.com/naive-bayes-for-machine-learning/.

Lustgarten, J L; Vanathi, G; Himanshu, G; Shyam, V. (2008). Improving Classification Performance with Discretization on Biomedical Datasets.

Maini, E; Bondu, V. (2021). Improving the Performance of Heart Disease Prediction System Using Ensembling Techniques. *AIP Conference Proceedings* 2316 (February).

Muibideen, M; Rajesh, P. (2020). A Fast Algorithm for Heart Disease Prediction Using Bayesian Network Model. 1–11.

Sai, S; Mani-Chand, MY; Mary, GL. (2021). Heart Disease Prediction Using Machine Learning. *Lecture Notes in Elect. Engineer.* 708 (Inccst 20): 603–9.

Srivastava, S. (2014). Weka: A Tool for Data Preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. *Inter. J. Comp. Applicad.* 88 (10): 26–29.

Steinbaum, SR. (2019). Cardiovascular (Heart) Diseases: Types and Treatments. 2019. https://www.webmd.com/heart-disease/guide/diseases-cardiovascular.

Webb, GI; Eamonn, K; Risto, M; Risto, M; Michele, S. (2011). "Naïve Bayes." *Ency. Mach. Learn.* 713–14.

WHO. (2019). WHO and Nigerian Government Move to Curb Cardiovascular Diseases | WHO | Regional Office for Africa. 2019. https://www.afro.who.int/news/who-and-nigerian-government-move-curb-cardiovascular-diseases.