Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach

Hossam Meshref
Associate Professor, Computer Science Department
College of Computers and Information Technology
Taif University, Taif, Saudi Arabia

***Abstract*—Research on heart diseases has always been the center of attention of the world health organization. More than 17.9 million people died from it in 2016, which represent 31% of the overall deaths globally. Machine learning techniques have been used extensively in that area to assist physicians to develop a firm opinion about the conditions of their heart disease patients. Some of the existing machine learning models still suffers from limited predication ability, and the chosen analysis approaches are not suitable. As well, it was noticed that the existing approaches pay more attention to building high accuracy models, while overlooking the ability to interpret and understand the recommendations of these models. In this research, different renowned machine learning techniques: Artificial Neural Networks, Support Vector Machines, Naïve Bayes, Decision Trees and Random Forests have been investigated to help in building, understanding and interpreting different heart disease diagnosing models. The Artificial Neural Networks model showed the best accuracy of 84.25% compared to the other models. In addition, it was found that despite some designed models have higher accuracies than others, it may be safer to choose a lower accuracy model as a final design of this study. This sacrifice was essential to make sure that a more transparent and trusted model is being used in the heart disease diagnosis process. This transparency validation was conducted using a newly suggested metric: the Feature Ranking Cost index. The use of that index showed promising results by making it clear as which machine learning model has a balance between accuracy and transparency. It is expected that following the detailed analyses and the use of this research findings will be useful to the machine learning community as it could be the basis for post-hoc prediction model interpretation of different clinical data sets.**

Keywords*—Heart diseases; machine learning; artificial neural networks; support vector machines; Naïve Bayes; decision trees; random forests; model interpretation; feature ranking cost index

I. Introduction

The field of machine learning has been progressing tremendously as its techniques became more popular and easily accessible. Applications ranged from face detection, system security, disease diagnosis, drug discovery, and many other revolutionary areas that impacted the lifestyle of many individuals. The basic idea behind building machine learning applications is different from most conventional programming methods. Basically, Machine learning models learn from patterns in the provided training examples without using explicit instructions, and then use inference to come up with useful predictions.

Some machine learning techniques, such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM), are very well known as successful prediction models, but sometimes they have problems. The main problem lies in the fact that they remain as black boxes after the model is built. In most of the cases, prediction models are built using historical data to make predictions about future situation that may take place. Understanding the reasoning behind the model prediction response could save organizations' stakeholders a lot of trouble as they may be carefully investigating different situations, while choosing the right medical treatment or assessing the risk of an investment plans for example. Some designed machine learning models play a very critical role in the health care system, and the designed system could recommend performing surgery on a patient. That decision should be extremely accurate to avoid life threatening situations. Making such a tough decision requires a thorough understanding of the reasons behind the model final recommendation before actually going on with the surgery.

In order to build machine learning models that could perform heart patient diagnosis, patients' data set examples need to be used. There are a few trusted websites that most researchers use when they collect data for analysis, such as UCI and Kaggle. The data set that has been used in this research is from the UCI Machine Learning Repository, and it is called the Cleveland Heart Disease data set, which consists originally of 76 features and has 303 instances. The data was originally collected from Cleveland Clinic Foundation, Cleveland, Ohio, and provided by Robert Detrano, M.D., Ph.D. of the V.A. Medical Center, Long Beach, CA [1].

At the UCI website, there were 4 heart diseases data sets to choose from: The Cleveland, Hungarian, Switzerland, and Long Beach, VA data sets. However, the Hungarian and the Cleveland data sets were the most promising in terms of the number of data instances and quality of data. Particularly, the Cleveland data set was almost complete with 303 instances, while the Hungarian data set had 294 instances, but some features were incomplete such as slope, ca and thal. On the other hand, the other two data sets, Switzerland and Long Beach, had 123 and 200 instances respectively, and they even have more incomplete features such as chol, exang and thalach.

Having introduced the available heart disease data sets, it is necessary to have a second look at the two major data sets that have more records than the rest: the Cleveland and the

Hungarian data sets. It was noticed that the distribution of heart disease categories in the Hungarian data set were proportional, whereas the in the Cleveland land data set, some disease categories such as Dis-Cat1 was more represented, see Fig. 1 and Fig. 2. In almost all the research work done on heart data sets, all the disease categories were grouped into one group. Therefore, that note was not important, and the vote for which data set to use was won by the Cleveland data. That led to having almost a balanced category of disease instances versus no-disease: 160 instances of people without heart disease versus 137 instances of people with chance for heart decease.

Most research papers that investigated the Cleveland data set have used only 13 features out of the original 76 features. The last column in the data set is num, which is a categorical variable having values: 0 for no disease, and 1, 2, 3 and 4 for variation of presence of heart disease, see Table I. However, as mentioned earlier, the heart disease variation levels, 1 to 4, have not been specified in the UCI available data sets description. Therefore, in this research, since it was logical to group all the disease levels 1, 2, 3 and 4 in one category, the num variable had only two categories: Dis, or NoDis.

The structure of this paper is organized as follows: Section two shed some light on the related research that has been done in heart disease diagnosis using machine learning. Section three focuses on the theory behind different methodologies used in this paper. Section four covers results and discussion, followed by a related section: models' interpretations, which complements the models' design processes as we try to achieve the best diagnosing model. Finally, section six concludes the presented work and provides a glimpse into possible future research avenues.

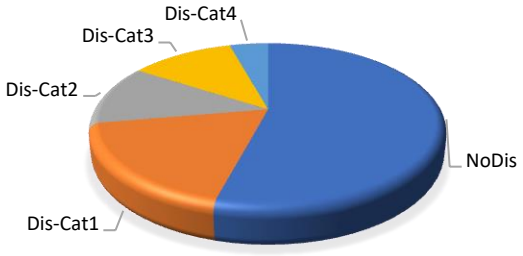

Fig. 1. Cleveland Data Categories.

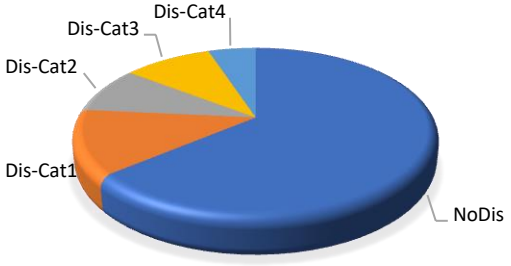

Fig. 2. Hungarian Data Categories.

TABLE. I. LIST OF FEATURES AND THEIR DESCRIPTIONS IN THE HEART DISEASE DATASET

Name Feature	Description
age	age in years
sex	patient sex
cp	chest pain type
trestbps	resting blood pressure
chol	serum cholesterol
fbs	fasting blood sugar
restecg	resting electrocardiographic result
thalach	maximum heart rate achieved
exang	exercise induced angina
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels (0-3) colored by fluoroscopy
thal	Exercise thallium scintigraphy
num	Response: diagnosis of heart disease

II. RELATED WORK

Different learning techniques have been used successfully in many medical applications to leverage human health conditions. For example, some applications addressed Liver Fibrosis prediction in Hepatitis patients as well as a decision support system for Diabetes diagnosis using soft computing fuzzy techniques [2,3]. Other applications focused on diagnostic systems for Heart Disease prediction for Coronary diseases using machine learning approaches. The machine learning methods used in these applications ranged between using a single machine learning technique such as hidden Naïve Bayes (NB), SVM, optimized ANN and Decision Tree (DT) classifiers [4,5,6,7], to using a collective or hybrid machine learning techniques [8,9,10]. Since the focus in this proposed research is on heart disease diagnosis, more attention will be devoted to its related literature.

The literature on using machine learning techniques to diagnose Heart Diseases were abundant. That was expected as the topic is very critical and is the center of attention of the World Health Organization as mentioned earlier. However, to build reliable machine learning models, rich data sets are needed. Unfortunately, most of the trusted data sources on heart diseases such as UCI or Kaggle have a relatively small number of instances when compared to Diabetes data sets for example [11]. Some machine learning techniques could be affected by that small number of instances such as ANN, which will eventually lead to building low accuracy models.

A few researchers have addressed the heart data instances sparsity issue and developed some techniques to handle it [12,13,14,15]. Some have combined two major heart data sets, the Cleveland and the Hungarian, to form a bigger set aiming to design better machine learning models and eventually achieve better results. Other researchers have used surrogate data sets that include synthetic observations in order to increase the number of instances in a heart disease data set. These models have done good efforts to improve the overall accuracy of the designed machine learning model despite data sparsity.

In general, most of the literature on using machine learning for heart disease diagnosis utilized two major techniques: the ANN [16,17,18,19] and the SVM [20,21,22,23]. They both have high classification accuracy, but they suffer from low learning speed when the number of instances or the number of features is huge. These facts make them great candidates for analyzing the Cleveland heart disease data at hand, since the number of instances and the features are relatively low. Most of the research focus in these two techniques tried improving the classification Accuracy and other validation metric scores such as the F-score.

In some applications that were based on ANN models research was focused on a serious pathophysiological heart condition, the Congestive Heart Failure (CHF), which is difficult to diagnose in some cases. Despite facing that difficulty, the combined design of convolutional neural network (CNN) and a distance distribution matrix (DDM) classification models was efficient enough to discriminate CHF patients from normal ones. The applications on deep learning continued to be used effectively in other research efforts focused on the same CHF heart condition. An important feature predictor, the Heart Rate Variability (HRV), which is an effective predictor, was used to analyze the designed model. Despite the challenges associated with the use of this predictor, the researchers have deployed an effective ensemble method for CHF detection using short-term HRV measured data and deep neural networks. They have added an extra analysis step, which is considered very important in model validation. They have conducted feature importance analysis to see if it agrees with their deep learning designed model chosen features. We Believe that the validation step is very important in model design, and we are going to use this step as well in our proposed research.

On the other hand, in some other applications that were based on SVM models, research was focused also on the CHF pathophysiological heart condition. One application deployed the same approach by using the HRV measures as input features to the support vector machine classifier. Their designed model was able to detect the CHF cases effectively, and therefore could be valuable if applied in other biomedical signal processing applications. One application using the support vector machine learning technique was focused on detecting patients with Heart Failure (HF). In it, researchers have utilized a hybrid grid search algorithm that can optimize multiple support vector machine models simultaneously. That algorithm showed an improvement over using the conventional stand-alone support vector machine models.

As a general comment on the literature, most of the applications found were focusing on getting the best performance model to perform predictions based on the available data. However, these efforts stopped at that point in most of the applications, and there were no further attempts of model interpretation. The focus in our research is not only to have the best performance model, but also to have a transparent model that could give interpretable trusted results. With interpretation, more useful information could be extracted from the data set in addition to prediction. This is a recent research trend in machine learning and the efforts in this area are growing in a promising direction.

This paper uses the UCI Machine Learning Repository Cleveland heart disease data set with the intention of performing the following main objectives: (1) to understand the heart disease data set at hand, to perform suitable data pre-processing, and to explore the best features to use during analysis (2) to build different machine learning classifiers and to use them to achieve the best prediction model. (3) to interpret the results achieved using these classification models and to provide suitable analysis about the transparency of these classifiers and the reasons to trust their resulted predictions. As much as the first two objective are considered as valuable contributions of this paper, the latter objective is considered to have more contribution by proposing the new metric: the Feature Ranking Cost index. In this research, that index was informative enough during the analysis phase to asses trust of different designed machine learning models based on their feature-sets used for prediction. We believe that it is a must-need step for every researcher who claim to have designed a robust machine learning model. We hope that it becomes a practice to adopt post-hoc validation techniques such as the one that we are proposing to guarantee the design of efficient and authentic machine learning models.

III. THEORITICAL BACKGROUND

In this research, a few supervised machine learning techniques have been chosen to build a diagnosing model for the Cleveland heart disease data set: MLP, NB, and a SVM, and Random Forests (RF) classifiers [24,25,26]. For experimentation, a 10-fold cross validation method was used to evaluate each model performance. During the validation process, the data set is divided into 10 folds. Each fold is held in turn for testing, while the other 9 folds are used for training. This validation process is repeated 10 times to guarantee that each data instance is used once for testing and 9 times for training. To further enhance the performance of the designed models in this research, stratified cross validation has been deployed where each fold used in the validation is balanced by having the right proportion of the class labels. In the following sections, a brief theoretical background about the deployed machine learning techniques will be introduced.

A. The ANN Model

ANN is designed based on the biological neural networks, which form the building blocks of the human nervous system. Multi-layer ANN consists of more than one processing layer of neurons, which represent the mathematical realization for the biological neural networks. During supervised learning, an ANN learns and gains experience from a set of predefined training examples. The error minimization process is supervised by a teacher. In this research use of the ANN model, a supervised training method is used to perform non-linear mapping in pattern classification based on back-propagation. During the training phase, the input examples are applied to the network, and the resulting, actual, response is compared with the desired response. If the actual response differs from the target response, an error signal is back propagated to adjust the network weights, see Fig. 3.

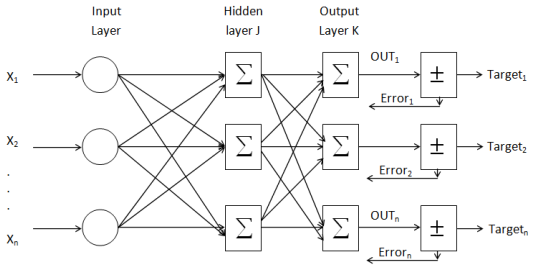

Fig. 3. The Back-Propagation ANN Structure.

The directions of two basic signal flows in a backpropagation network are: forward propagation of function signals and back-propagation error signals. For the forward propagation Pass:

1) Calculate $output_j$: actual-net-output
2) Calculate error ($target_j - output_j$) at the output units

The basic back-propagation algorithm is based on minimizing the error of the network using the derivatives of the error function. The most common measure of error is the mean-square error:

$$E = 1/2\ (target\ -\ actual)^2 \quad (1)$$

A small step, learning rate α, in the opposite direction will result in the maximum decrease of the local error function. Therefore, the new weight will be given by:

$$W_{new} = W_{old} - \propto \frac{\partial E}{\partial W_{old}} \quad (2)$$

For the backward propagation Pass:

1) Compute Δw for all weights of the output layer:

$\delta_j = f'(net_j)(target_j - output_j)$, therefore:

$$W_{new} = W_{old} + \alpha\, output_i f'(net_j)(target_j - output_j) \quad (3)$$

2) Compute Δw for all weights from hidden layer(s) back to input layer:

$\delta_j = f'(net_j)(\sum \delta_k W_{kj})$, therefore:

$$W_{new} = W_{old} + \alpha\, output_i f'(net_j)(\sum \delta_k W_{kj}) \quad (4)$$

B. The SVM Model

The key difference that discriminate SVM from other classifiers is that it focuses on the data points which are hard to classify, whereas in most other classifying techniques, the focus is on all the data points. For example, the basic Perceptron, in ANN, is searching for linear separability for each data point in the training set and stops when that condition is satisfied. However, these lines are not guaranteed to be the best separators. On the other hand, the SVM linear classification algorithm goal is to maximize the distance between the two hyper planes defined by: $wX - b = -1$ for the first class, and $wX - b = 1$ for the second class, see Fig. 4. The problem of finding that max-margin hyperplane, defined by $wX - b = 0$, could be solved by finding the distance:

$$\max_w \frac{2}{\|w\|} \quad (5)$$

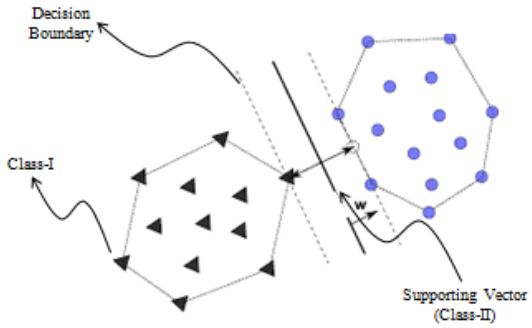

Fig. 4. The Concept behind the SVM Technique.

C. The NB Model

The Bayesian classifier is considered to be one of the most commonly used classification techniques in machine learning. The NB classifier, in particular, base its prediction on Bayes theorem, while assuming independence between the data set attributes, which makes its model easy to build. However, the assumption of independence is not accurate all the time and based on that NB classifiers may be considered less accurate than other more sophisticated machine learning algorithms. On the other hand, there are some advantages of its use such classification speed, tolerance to missing values and fewer model parameter handling. Therefore, when speed is needed during the analysis of big data sets, the NB classifier could be an appropriate choice. The Naive Bayes classification problem could be solved by estimating a classification ratio C, see equation (6). If C is greater than 1, then the first class is predicted, if not, then predict the second class.

$C = \frac{P(i|X)}{P(j|X)}$, therefore:

$$C = \frac{P(i)\prod P(X|i)}{P(j)\prod P(X|j)} \quad (6)$$

Where,

- P(i|X) is the posterior probability of the target class, given a predictor (i.e. attribute X).
- P(X|j), is the likelihood (i.e. the probability of predictor given class).
- P(i) and P(j) are the priori probability of first class i and second class j respectively.

D. The DT and RF Models

One major advantage of DT, unlike most other machine learning models, that it is transparent as you can follow its hierarchical structure to understand how the classification decision took place. In DT, Entropy measures disorder in the data, and can give an indication of how untidy the data is. For that reason, it is used as an algorithm to tidy the data by separating it and grouping the samples in the classes they belong to. A data set could be considered ordered, or tidy, when all the data items in it share the same label and is considered untidy if it has a blend of items with different labels. The DT algorithm uses the Entropy equation while looping around the training data set make sure that each sub data group is tidy and carries the same label, see equation (7).

In this research, J48 DT algorithm is used, which is an open source java implementation of the C4.5 algorithm. The information gain, Gain, is calculated based on testing the attributes, and the best attribute with the highest gain is used as a base for further branching, see equation (8). Given a node (attribute) split argument $\vec{S}$ by a certain value i, to calculate its Entropy we use the following equations:

$$\text{Entropy}(\vec{S}) = -\sum_{j=1}^{n} \frac{|S_i|}{|\vec{S}|} \log\left(\frac{|S_i|}{|\vec{S}|}\right) \quad (7)$$

And the overall Gain is calculated for an attribute j as:

$$\text{Gain}(\vec{S}, j) = \text{Entropy}(\vec{S}) - \text{Entropy}(j/\vec{S}) \quad (8)$$

The DT classifier that we have discussed so far is the basic building block of the RF classifier. In the RF classifier a subset of the original features is used when constructing a given tree. Then, the algorithm searches over sets of random features, e.g. N1 to N4, to choose the best one. Since there are different features for different DT, it is anticipated that it will form with different sizes, and eventually, the formed de-correlated DT are expected to produce different predictions. Based on that assumption, the RF classifier forms its final classification decision by majority voting, which is averaging all the predications of the formed sub trees #1-4, see Fig. 5.

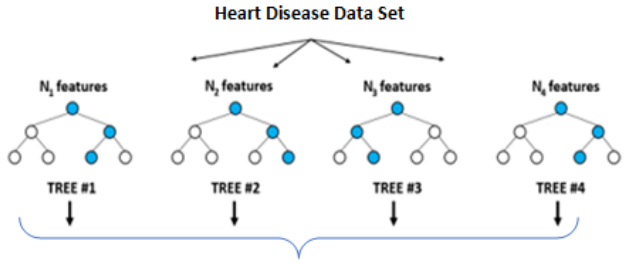

Fig. 5. The Concept behine the RF Technique.

E. Model Evaluation Metrics

During model evaluation, the confusion matrix played an important role in understanding the results obtained in this research, see Table II. True Positive value (TP) were those values that represent the number of patients who originally has heart disease and were actually predicted correctly. True Negative (TN), on the other hand, represented the number of patients who originally did not have heart disease and were actually predicted correctly. Conversely, False Positive (FP) were those patients, who originally did not have heart disease, but were predicted as positive. False negative (FN) on the other hand, were those patients predicted as negative, but originally did have a heart disease.

TABLE. II. THE CONFUSION MATRIX STRUCTURE

Actual			Predicted
		Negative	**Positive**
	Negative	TN	FP
	Positive	FN	TP

Recall Precision

Based on the previous definitions of TP, TN, FP and FN, the accuracy of the designed models in this research has been calculated using the following evaluation metric:

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (9)$$

It refers to the ratio between the sum of the total true-positives and true-negatives results to the number of examined instances of the training data. However, in most research papers, scholars agreed that accuracy should not be the end-all measure of model evaluation. Other metrics such as Precision, Recall, and F_1 score should also be considered:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (11)$$

$$F_1\text{Score} = 2\left(\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}\right) \quad (12)$$

Precision, or confidence, refers to the ratio between the positively predicated values and the total predicted positive values, whereas Recall, or sensitivity, is the ratio between the positively predicated values and the total actual positive values. The F_1 score on the other hand utilizes both Precision and Recall producing a new harmonic average that shows the balance between them.

IV. RESULTS AND DISCUSSION

A. Heart Data Set Pre-Processing

In this data set, there were 13 features and one target class value as a label. Missing values were not noticed in most of the attributes; however, only 6 missing values were found, 2 in thal, and 4 in ca. The instances that included these missing values were fully deleted, leaving 297 instances for further analysis. Since eventually we need to interpret the designed machine learning model, the idea of implementing feature reduction was not recommended. Instead, feature selection was implemented to identify those features that effectively contribute to the classification process.

It was necessary to remove outliers and extreme values to guarantee robustness of the designed machine learning models. The interquartile range (IQR) technique was used as a measure of the statistical dispersion for the data set features. There was one patient data instance with outliers in the chol feature with a value of 564, which deviates remarkably from the rest of the values in the data set. That instance was removed to avoid skewing in the result as it could have a significant effect on the mean and standard deviation.

As for the extreme values, it was found that there are 43 instances with extreme values, which constitutes 14.5% of the data records. However, for fear of falling into bias problems during models' design, the effect of the removal of these extreme records had to be checked. It was found that the 43 instances are divided almost equally between the two disease categories: 23 patients have a heart disease and 20 patients do not have a heart disease. That balance between the two class instances gave an indication that it is less likely to have class bias, and therefore, those 43 records were removed. Having

done this step, it was necessary to follow with a check for feature ranges to make sure that they are homogeneous.

It was noticed that the range of the features in the heart data set vary in a way that could affect the design of the machine learning models. For example, the maximum value for age and oldpeak are 77 and 6, while chol and thalach are 564 and 200 respectively, see Fig. 6. One effective method that was used in this research is to standardize all numeric attributes in the data set to have zero mean and unit variance. Overall, by conducting the previous pre-processing steps, the heart disease data set was ready for the models' design stage. All the models' performance evaluation results during the pre-processing stage are summarized in Table III.

As illustrated in Table III, the initial raw instances of the heart disease data set were used to build four different classifier models: MLP, NB, SVM, and RF. The SVM, NB, and the RF models showed the best performances compared to the MLP model, and significantly better than the base case classifier model: the ZeroR. Because there were a small number of instances with missing value and outliers, models' performances were not noticeably affected after their removal. However, removal of instances with extreme values as well as feature standardization led to an improvement in the MLP model accuracy reaching 81.88%, see also model building speeds in Fig. 7. The experiments were run on an Intel(R) Core (TM) i-2400S CPU @ 2.50GHZ processor, with 6.00 GB RAM, on a 64-bit operating system, x64-based processor system. Having dealt with the previous data processing steps, it was imperative to perform feature selection hoping to improve the performance of the models under investigation.

B. Feature Selection and Model Design

In this research, simple fast attribute selection methods have been studied such as single attribute evaluator with ranking and attribute subset selection methods. However, the single attribute method can allow redundancy, which is not recommended and may lead to inaccurate results. For example, problems such as redundancy have strong impact on the performance of the NB classifier, while overfitting could badly impact the MLP classifier. The attribute subset selection method, on the other hand, removes redundancy as well as irrelevant features, hence it was chosen in this research as a base method for feature selection.

Careful measures have been considered while applying that attribute selection method to the heart disease data set, with cross-validation, to have fair classification results. A problem could have happened if the entire data set is used to decide on the attribute subset. Therefore, in this research, attributes have been selected based on the training data only. Then, each designed classifier model has been trained on the training data as well with cross-validation in effect, followed by model evaluation using the test data.

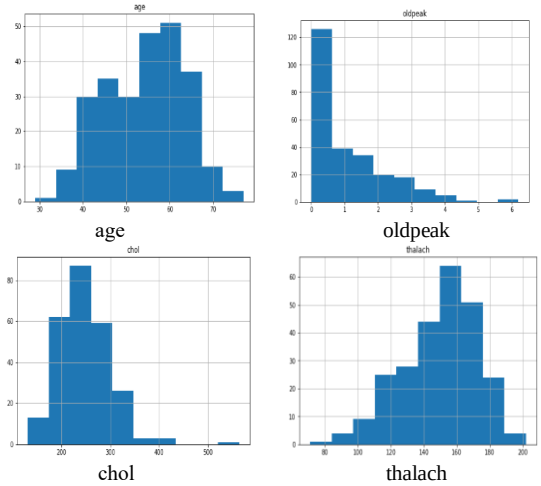

Fig. 6. Selected Features' Histograms.

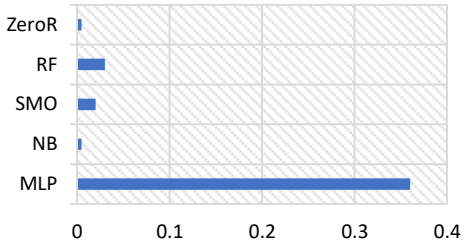

Fig. 7. Models' Building Speeds in Seconds.

TABLE. III. EFFECT OF PRE-PROCESSING STAGES

	Performance%	**MLP**	**NB**	**SVM**	**RF**	**ZeroR**
Original Data	Accuracy	78.88	83.5	83.82	81.85	54.12
	Precision	78.9	83.6	83.9	82.3	29.3
	Recall	78.9	83.5	83.8	81.8	54.1
	F_1	78.8	83.4	83.8	81.7	38.0
Missing & Outliers	Accuracy	78.37	83.11	83.44	82.43	53.72
	Precision	78.4	83.2	83.5	82.5	28.9
	Recall	78.4	83.1	83.4	82.4	53.7
	F_1	78.3	83.0	83.4	82.3	37.5
Extrems, Missing & Outlier	Accuracy	81.88	81.88	82.68	75.9	53.94
	Precision	81.9	81.9	82.7	76.2	29.1
	Recall	81.9	81.9	82.7	76.0	53.9
	F_1	81.9	81.9	82.6	76.0	37.8
-Standardization	Accuracy	81.88	82.28	82.67	79.13	53.93
	Precision	81.9	82.3	82.7	79.7	29.1
	Recall	81.9	82.3	82.7	79.1	53.9
	F_1	81.9	82.2	82.6	79.2	37.8

One good recommended practice, when performing attribute selection, is to use the same classification method as a wrapper substitute evaluator method. However, all possibilities have been tested in this research to find the best attribute selection method, and eventually come up with that specific set of features that lead to the best results. This feature-set is expected to be authentic in a sense that it actually affects the results at hand. In the following table, the row entries represent the used technique within the wrapper substitute evaluator method, while the column entries represent the classification technique used in building the models. Different model performance evaluation metrics such as accuracy, precision, recall, and F_1-score are presented in a separate column, see Table IV. It worth mentioning that the time taken for attribute selection and classifier training using MLP in the wrapping process was considerably long compared to other techniques, see a sample run in Fig. 8.

The best accuracy result obtained, 84.25%, after feature extraction, was for the MLP classifier with SVM as a feature selection wrapper substitute evaluator method. The NB and the SVM classifiers had a lower, but comparable, results at 83.07% and 82.28% respectively. The RF classifier came last with 78.35% accuracy despite the fact that the redundant features have been removed. Recalling an earlier comment in this paper about the practice of using the same classification method as a wrapper substitute evaluator method, it was noticed that SVM has the best accuracy performance at 82.28%, see the lightly shaded diagonal cells at Table IV.

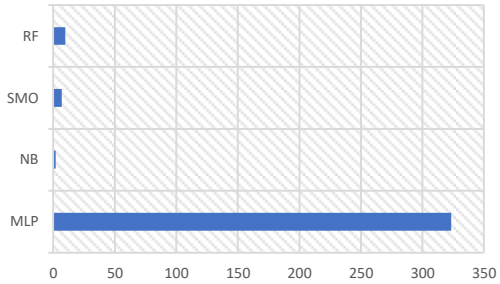

Fig. 8. Feature Extraction and Classification Model Building Speeds in Seconds.

During the feature selection analysis, different features have been selected by different wrapper substitute evaluator method. For example, MLP came up with 8 features in its recommended feature-set, followed by 5 features feature-sets for NB, SVM, and RF, see Table V. Up to this point of analysis, it could be fair to assume from Tables IV and Table V that the longer the run-time, while deciding on the best selected feature-set, the more the numbers of features selected. However, how authentic are these feature-sets required further analysis. The following section focuses on using one of the well-known model interpretation techniques, the DT, to find reasonable explanations for the resulted models' performances.

TABLE. IV. EFFECT OF FEATURE SELECTION METHODS

	Performance%	**MLP**	**NB**	**SVM**	**RF**
MLP	Accuracy Precision Recall F_1	79.13 79.1 79.1 79.1	79.92 80.2 79.9 80.0	79.53 79.5 79.5 79.5	80.32 80.4 80.3 80.3
NB	Accuracy Precision Recall F_1	81.1 81.1 81.1 81.1	81.1 81.1 81.0 81.0	82.7 82.8 82.7 82.6	78.35 78.3 78.3 78.3
SVM	Accuracy Precision Recall F_1	84.25 84.4. 84.3 84.2	83.07 83.2 83.1 83.0	82.28 82.3 82.3 82.2	78.35 78.4 78.3 78.4
RF	Accuracy Precision Recall F_1	78.3 78.3 78.3 78.3	80.71 80.9 80.7 80.7	77.56 77.7 77.6 77.6	79.92 79.9 79.9 79.9

TABLE. V. FREQUENCY OF SELECTED FEATURES

	age	**sex**	**cp**	**trestbps**	**chol**	**fbs**	**restecg**	**thalach**	**exang**	**oldpeak**	**slope**	**ca**	**thal**
MLP	√	√		√	√					√	√	√	√
NB			√					√		√		√	√
SVM		√	√						√			√	√
RF		√									√	√	√

V. Prediction Level Interpretation

All the machine learning models used in this research deploy supervised learning methods. These methods used the instances of heart disease data set to learn and to produce general hypotheses as predictions. DT is one of the supervised machine learning models that is frequently used to solve classification problems. One major advantage of DT models is that they could map non-linear relationships, while providing clear interpretation, and for that reason DT will have more focus in this section.

Further analysis was done using the J48 DT classifier, while performing attribute selection and using the same classification method, J48, as a wrapper substitute evaluator method, see Table VI. The DT designed model took 1.1 seconds to be build using the same earlier machine specs used for the other machine learning models: MLP, NB, SVM, and RF. The resultant model accuracy was 76.38%, which is considered low compared to those previous models except for the RF model. Most of those earlier models, despite having better accuracy, were not transparent, and therefore were hard to interpret.

Comparing Table V and Table VI, it was noticed that they are almost identical, except for one attribute difference between the RF and the DT models as wrapper substitute evaluator methods. Both models agreed on selecting attributes sex, ca and thal, but disagreed on slope and oldpeak. Focusing on table VI, one could conclude that the most frequently selected attributes for the J48 DT model were thal, ca, oldpeak, sex, and maybe cp as well. The following classification trees' samples were generated, while the designed DT models were evaluated, see Fig. 9 to Fig. 12. The root node in each illustrated classification tree is considered the attribute with the highest purity as it is more capable of discriminating between patients with and without heart disease and so forth down the tree.

From the model interpretation point of view, the purity of these features could be a reference point for measuring their contribution to the accuracies of their corresponding analyzed models. For example, if we look at classification Tree-1, in Fig. 9, it could be fair to deduce that in order to decide if a patient has heart disease or not, thal status need to be checked first. As well, the next feature to be checked in that tree is ca, whether the answer at the previous thal-node was Yes or No. One can comprehend such reasoning even at the third level of the tree while checking for sex and age. However, as we go deeper in that tree, we may get confused during analysis. That confusion could be more noticeable at classification Tree-4, which was build during the design of the MLP model.

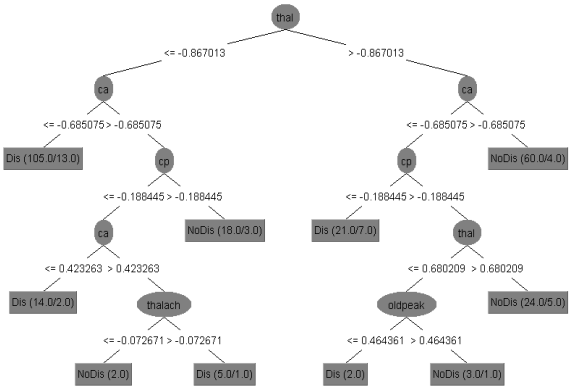

Fig. 9. Classification Tree-1.

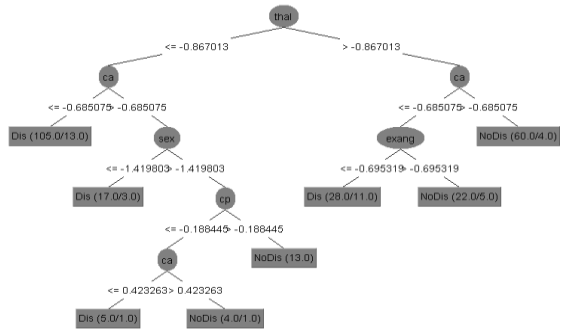

Fig. 10. Classification Tree-2.

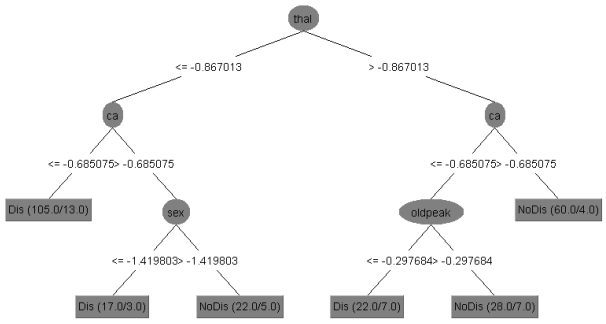

Fig. 11. Classification Tree-3.

TABLE. VI. Frequency of Selected Features-DT Only

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	Accuracy (%)
Tree1 NB			√					√		√		√	√	77.17
Tree2 SVM		√	√						√			√	√	79.53
Tree3 DT		√								√		√	√	76.38
Tree4 MLP	√	√		√	√					√	√	√	√	78.35

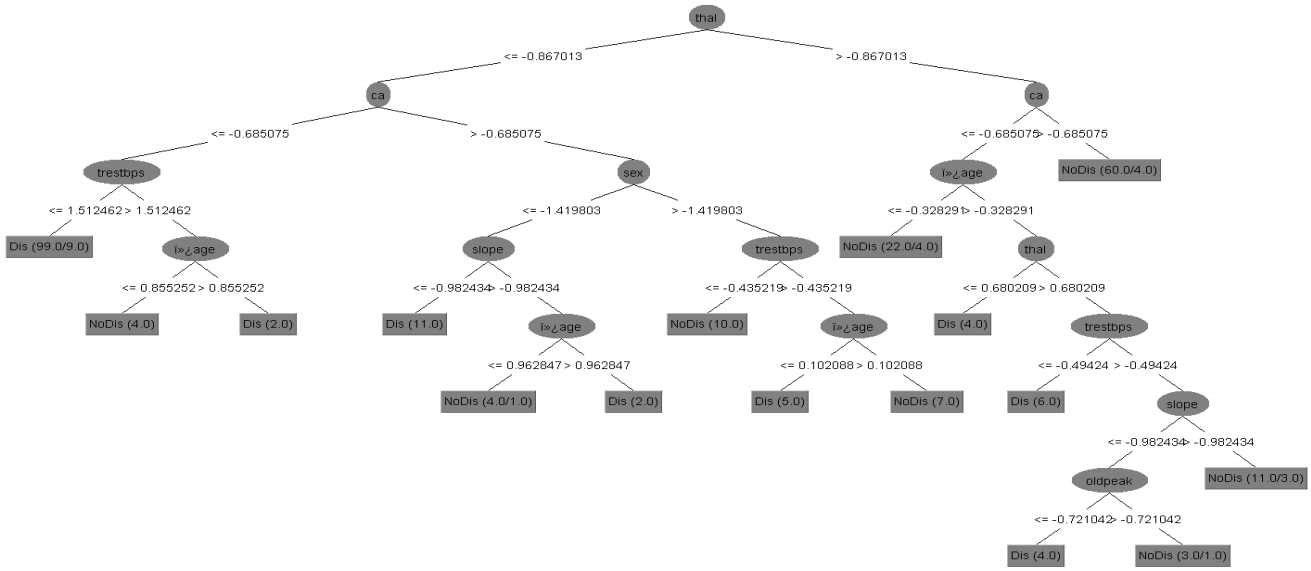

Fig. 12. Classification Tree-4.

A well-known method to handle the previous confusion problem is to use feature importance analysis using ensemble method of DT. The values representing feature importance are relative values, or scores, comparing the performance of the desired model with and without that specific feature. In this research, all features in the heart disease data set have been considered for feature analysis, see the chart shown in Fig. 13. Values of importance ranged between Zero for the fbs feature to 1.75 for the thal feature, which is considered for this model to be the most predictive value. Based on that concept, removing a feature such as thal is expected to considerably affect the designed model, while removing a feature such as fbs should not have an effect, and so forth for rest of the feature importance values.

Further analysis has been conducted using a new proposed technique, Feature Ranking Cost, to better understand and interpret the performances of the designed models: MLP, NB, SVM, and RF. The concept behind the creation of this new metric is to come up with a simple post-hoc technique that can help in evaluating the worthiness of a model performance based on the importance of its feature-set. After evaluating the designed models from that point of view, a simple corrective action could be taken by choosing the best final model that is capable of producing the best authentic result as much as possible.

Table VII illustrates how this index could be calculated with the support of the chart in Fig. 13 as well. The Feature Ranking Cost (FRC) index for a specific model M as well as the best model FRC could be calculated using Equation (13) and Equation (14) respectively.

$$M_{FRC_i} = \sum_{i=1}^{n} FIR_i \quad (13)$$

$$M_{Authentic_j} = \min_j M_{FRC_j},\ j{:}1 \rightarrow m \quad (14)$$

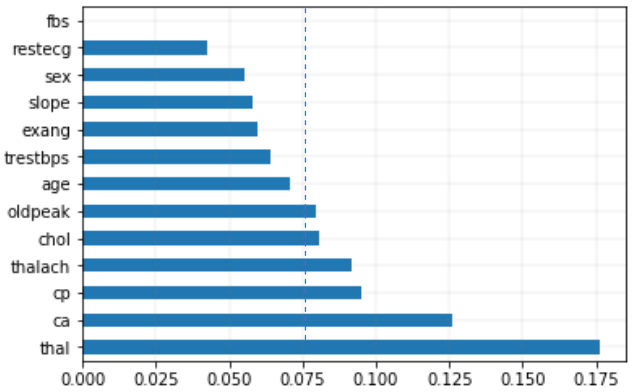

Fig. 13. Most Important Features based on DT– Entropy Function.

TABLE. VII. CREATION OF FEATURE RANKING COST INDEX

Feature Importance	Feature	MLP	NB	SVM	RF
0.07041383	age	7			
0.05492301	sex	11		11	11
0.09525694	cp		3	3	
0.06413777	trestbps	8			
0.08089077	chol	5			
0.0	fbs				
0.04242536	restecg				
0.09190829	thalach		4		
0.05973358	exang			9	
0.07978184	oldpeak	6	6		
0.05797686	slope	10			10
0.12622297	ca	2	2	2	2
0.17632879	thal	1	1	1	1
	Σ Cost @ 0.075	14	16	6	3

Where M_{FRC} is the sum of ranks for all the feature-set items of a model M given its Feature Importance Ranks (FIR's), and $M_{Authentic}$ is the minimum authentic value amongst all the m models (m=4 in this research). The FIRs could be found from Fig. 13 in an ascending order: 1 for thal, 2 for cp, and so forth until 13 for fbs. The reason behind choosing the minimum authentic value is that the most authentic model is expected to have the most important set of features and to show the best performance at the same time. It worth mentioning that the sum of all the feature importance values adds up to one. Therefore it is recommended when conducting research with a larger feature data set to use a Weighted Feature Ranking Cost to avoid computational problems. It could be calculated by multiplying each FIR by its corresponding Feature Importance Score (FIS), and then follow the previous calculation procedures:

$$M_{WFRC} = \sum_{i=1}^{n} FIR_i \times FIS_i \quad (15)$$

Deciding how many features should be in each feature-set is a challenging task. In other words, at what rank should we stop to perform the calculation of the M_{FRC} indices? In this research, the approach was to calculate the whole FIS range, and then use that as a reference as where to set the threshold value. Specifically, for Table VII, 50% of the whole FIS range was used, and every feature that has a lower value was not included in the features' analysis pool. Based on that range of choice, approximately 46% of the attributes in the feature importance chart were covered: thal, ca, cp, thalach, chol, and oldpeak.

The FIR values for each classification model were added to calculate its M_{FRC}. The lowest FRC index was for the RF model, using the RF wrapping attribute select method. It was chosen based on that criterion as it has the most relevant attribute-set than the other models: MLP, NB, SVM. As a final concluding point, recalling Table IV, the MLP model, at an accuracy of 84.25%, seemed to be the right choice for the heart disease final classification model design. However, after conducting the pos-hoc FRC analysis, it could be more accurate to resort to a safer lower accuracy model in our final design by choosing the RF model at an accuracy of 79.92%. By doing that the final designed model in this research will have a balance between accuracy and transparency.

It has been noticed that the RF model has around 40% of the most effective feature, while the MLP has around 65%. However their feature ranking cost (FRC) are 3 and 14, see Fig. 14. It is believed that the values of the feature ranking cost FRC should be proportional to the values of the feature count ratio, shown in Fig. 15, for each designed model. A few modification have been applied to equation (13) by introducing a new term, the feature count ratio C_i, see equations (16) and (17). To further enhance the proportional relation between FRC for each model and its C_i, it was necessary to factor it by the corresponding Feature Importance Ranks (FIR), see equation (18). Table VIII shows the new re-calculated FRC's for the four designed models, and it shows that the RF model still has the lowest cost, which means that the previous post processing analysis still holds. Fig. 16 combines the new values for feature ranking costs for each model compared to its feature count ratio based on the new equations. It has been noticed that the values of the feature ranking cost FRC have better proportion to the values of the feature count ratio. That result assures, as well, that despite the previous modifications, the choice of the RF model is still an authentic choice and more safer to rely on.

$$C_i = \frac{m_i}{n_\theta} \quad (16)$$

$$M_{FRC_i} = \frac{1}{C_i} \sum_{i=1}^{n} FIR_i \quad (17)$$

$$M_{WFRC_i} = \frac{1}{C_i} \sum_{i=1}^{n} FIR_i \times FIS_i \quad (18)$$

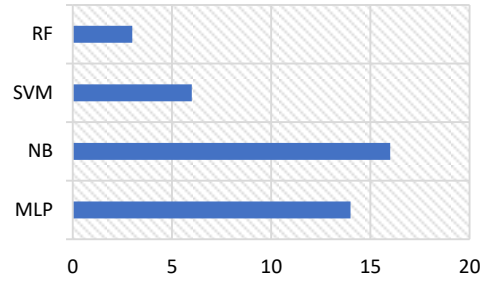

Fig. 14. Feature Ranking Costs based on Equation 13.

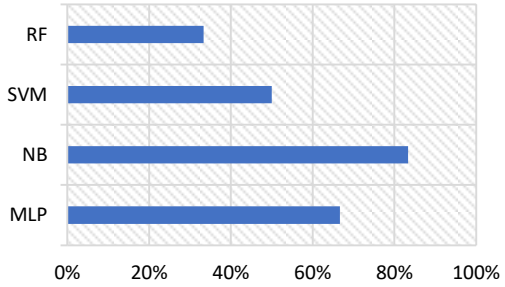

Fig. 15. Feature Count Ratio.

TABLE. VIII. CREATION OF WEIGHTD FEATURE RANKING COST INDEX

Feature Importance	Feature	MLP	NB	SVM	RF
0.07041383	age	0.4929			
0.05492301	sex	0.6042		0.6042	0.6042
0.09525694	cp		0.2858	0.2858	
0.06413777	trestbps	0.5131			
0.08089077	chol	0.4045			
0.0	fbs				
0.04242536	restecg				
0.09190829	thalach		0.3676		
0.05973358	exang			0.5376	
0.07978184	oldpeak	0.4787	0.4787		
0.05797686	slope	0.5798			0.5798
0.12622297	ca	0.2524	0.2524	0.2524	0.2524
0.17632879	thal	0.1763	0.1763	0.1763	0.1763
	Σ Cost @ 0.075	1.9679	1.8730	1.4291	1.2863

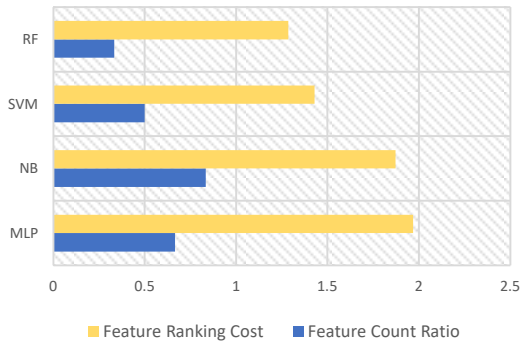

Fig. 16. Feature Ranking Costs Vs Feature Count Ratio.

VI. Conclusion

In this paper it was shown that we have conducted thorough analyses and understanding of the Cleveland heart data set. As well, different machine learning classifiers were designed and used to achieve the best diagnosing model. However, the previous discussion in the interpretation section highlight a few issues that need to be considered as we try to understand the machine learning designed models. If the design for the four models: MLP, NB, SVM, and RF was concluded based merely on calculating the initial used metrics: accuracy, precision, recall, and F1, there could have been a chance of ending up with an inaccurate model. For example, the MLP model, based on an SVM wrapping attribute select method, resulted in an 84.25% accuracy, but used an 8-features set to achieve that result. Based on the 50% threshold used in this research, its feature ranking score, M_{FRC}, was 15, which is a triple of the RF model score. This result indicated that it would not have been accurate to choose the MLP as a base for heart disease diagnosis model.

The investigation analysis done in this research have laid a reasonable foundation in exploring the nature of the heart disease data set. These efforts have been complemented by the interpretation analysis, which added more clarification of the designed models by the introducing the new FRC index. That index was an informative metric and led to a clear discrimination between the models based on their feature-set importance. The final chosen RF model, based on the post-hoc interpretation analysis, had a 79.92% accuracy, which was not a far compromise from the MLP model accuracy. In fact, it was a necessary step to choose the RF model instead of the MLP model to ensure that the final chosen model is authentic and has a balanced compromise between its transparency and its accuracy. It is anticipated that the use of the previous findings will be useful to the machine learning community as it could be the basis for post-hoc prediction model interpretation analysis on different clinical data sets.

For future work, a few main points could be considered. First, combing the Cleveland & Hungarian data sets and performing the required analysis may improve accuracy and give more insight into the transparency of each designed model. New challenges could arise such as missing data, but the 100% data instances increase may compensate for that problem. Second, performing association rule analysis could help in model interpretation and help in understanding the designed DT models, but rule post-processing may be needed to remove redundancy. Lastly, further in-depth post-hos prediction model interpretation analysis could be done to better understand and validate the designed models.

References

[1] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., and Froelicher, V., "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol. vol.64, pp.304–310, 1989.

[2] S. El-Sappagh, F. Ali, A. Ali, A. Hendawi, F. A. Badria and D. Y. Suh, "Clinical Decision Support System for Liver Fibrosis Prediction in Hepatitis Patients: A Case Comparison of Two Soft Computing Techniques," in IEEE Access, vol. 6, pp. 52911-52929, 2018.

[3] S. El-Sappagh, J. M. Alonso, F. Ali, A. Ali, J. Jang and K. Kwak, "An Ontology-Based Interpretable Fuzzy Decision Support System for Diabetes Diagnosis," in IEEE Access, vol. 6, pp. 37371-37394, 2018.

[4] M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier," 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), Bangalore, pp. 1-5, 2016.

[5] M. Ahmad, V. Tundjungsari, D. Widianti, P. Amalia and U. A. Rachmawati, "Diagnostic decision support system of chronic kidney disease using support vector machine," 2017 Second International Conference on Informatics and Computing (ICIC), pp. 1-4, Jayapura, 2017.

[6] M. Kumar, A. Sharma and S. Agarwal, "Clinical decision support system for diabetes disease diagnosis using optimized neural network," 2014 Students Conference on Engineering and Systems, Allahabad, pp. 1-6, 2014.

[7] Purushottam, K. Saxena and R. Sharma, "Efficient heart disease prediction system using decision tree," International Conference on Computing, Communication & Automation, Noida, pp. 72-77, 2015.

[8] M. Gudadhe, K. Wankhade and S. Dongre, "Decision support system for heart disease based on support vector machine and Artificial Neural Network," 2010 International Conference on Computer and Communication Technology (ICCCT), Allahabad, Uttar Pradesh, pp. 741-745, 2010.

[9] Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, pp. 704-706, 2015.

[10] Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), pp. 1-5, Srivilliputhur, 2017.

[11] J. Collins, J. Brown, C. Schammel, K. Hutson, and W. Edenfield, ”Meaningful Analysis of Small Data Sets: A Clinicians Guide,” Greenville Health System Proc., vol.2, no.1, pp. 16-19, June, 2017.

[12] Gárate-Escamilla, A.; El Hassani, A. and Andres, E., “Dimensionality Reduction in Supervised Models-based for Heart Failure Prediction,” In Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods – Volume 1: ICPRAM 2019, pp. 388-395, 2019.

[13] A. Sabay, L. Harris, V. Bejugama, and K. Jaceldo-Siegl, "Overcoming Small Data Limitations in Heart Disease Prediction by Using Surrogate Data," SMU Data Science Review, vol.1, no.3, 2018.

[14] S. Torgyn, and N. Khovanova, ”Handling limited datasets with neural networks in medical applications: A small-data approach,” Artificial intelligence in medicine, vol.75, pp. 51-63, 2017.

[15] L. Masitah, M. Azah, Y. Zeratul, M. Noor, and A. Mohd. "Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review," Journal of Physics: Conference Series, 2017.

[16] Y. Li et al., "Combining Convolutional Neural Network and Distance Distribution Matrix for Identification of Congestive Heart Failure," in IEEE Access, vol. 6, pp. 39734-39744, 2018.

[17] L. Wang, W. Zhou, Q. Chang, J. Chen and X. Zhou, "Deep Ensemble Detection of Congestive Heart Failure Using Short-Term RR Intervals," in IEEE Access, vol. 7, pp. 69559-69574, 2019.

[18] S. Rajamhoana, C. A. Devi, K. Umamaheswari, R. Kiruba, K. Karunya and R. Deepika, "Analysis of Neural Networks Based Heart Disease Prediction System," 2018 11th International Conference on Human System Interaction (HSI), Gdansk, pp. 233-239, 2018.

[19] S. Harjai and S. K. Khatri, "An Intelligent Clinical Decision Support System Based on Artificial Neural Network for Early Diagnosis of Cardiovascular Diseases in Rural Areas," 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, Unitsed Arab Emirates, pp. 729-736, 2019.

[20] B. Hu, S. Wei, D. Wei, L. Zhao, G. Zhu and C. Liu, "Multiple Time Scales Analysis for Identifying Congestive Heart Failure Based on Heart Rate Variability," in IEEE Access, vol. 7, pp. 17862-17871, 2019.

[21] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," in IEEE Access, vol. 7, pp. 54007-54014, 2019.

[22] C. Yang, B. An and S. Yin, "Heart-Disease Diagnosis via Support Vector Machine-Based Approaches," 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 3153-3158, Miyazaki, Japan, 2018.

[23] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," in IEEE Access, vol. 7, pp. 54007-54014, 2019.

[24] Witten, I.H., Frank, E. and Hall, M.A., Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publications, 2011.

[25] Michael Bowles, Machine Learning in Python: Essential Techniques for Predictive Analysis. John Wiley & Sons Publications, 2019.

[26] Simon Haykin, Neural Networks and Learning Machines. Pearson Prentice Hall Publications, 2009.