DATA MINING CLASSIFICATION TECHNIQUES FOR CARDIOVASCULAR DISEASE DIAGNOSIS

Hnin Ei Ei Cho$^{(1)}$, Nan Yu Hlaing$^{(2)}$
$^{(1)}$ Myanmar Institute of Information Technology, Mandalay, Myanmar
$^{(2)}$ Myanmar Institute of Information Technology, Mandalay, Myanmar

$^{(1)}$ Email: hnin_ei_ei_cho@miit.edu.mm
$^{(2)}$ Email: nan_yu_hlaing@miit.edu.mm

ABSTRACT

The huge amounts of data generated by healthcare transactions are complex and voluminous. We process and analyze them by using different traditional methods. The healthcare industry collects huge amounts of healthcare data, which, unfortunately, are not “mined” to discover hidden information. Mining Techniques offer a principled approach for developing sophisticated, automatic, and objective algorithms for analysis of high dimensional and multimodal biomedical data. Medical diagnosis is the process of determining which disease or condition explains a person’s symptoms and signs. In this study, we briefly examine the potential use of classification-based data mining techniques to massive volume of healthcare data. Aim of the paper is to propose a model for early detection and correct diagnosis of the disease, which will help the doctor in saving the life of the patient.

KEYWORDS: *Cardiovascular Disease*, *Classification, Random Forest (RF), Artificial Neural Network (ANN), Support Vector Machine (SVM)*

1. INTRODUCTION

Cardiovascular disease (CVDs) are a group of disorders of the heart and blood vessels. They include coronary heart disease, cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis and pulmonary embolism. This disease attacks a person so instantly that it hardly gets any time. [1].

CVD is the number one cause of death globally: more people die annually from this disease than from any other cause. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Of these deaths, 85% are due to heart attack and stroke. Over three quarters of CVDs, deaths take place in low- and middle- income countries. People with cardiovascular disease or with at high cardiovascular risk need early detection and management using counselling and medicines, as appropriate.

Recent advances in health related studies are concentrating on risk prediction of diseases. In the study of risk prediction from patients’ health records, different classification techniques are used. Classification algorithms take input data set and place given record in one of the pre-defined classes. Classification problems are prediction of class labels where number of classes is fixed and pre-defined. There is nothing like a particular classification method is accurate to classify the data in all situations. The accuracy of classification method is depends on the data we want to classify.

Effective and efficient automated cardiovascular disease prediction can benefit healthcare sector and this automation will save not only cost but also time [2]. This research paper highlights the utility and application of three different classification models of data mining techniques for prediction of cardiovascular disease to facilitate experts in the healthcare domain.

We divide this paper into the following sections: section two contains the related theory background; section three describes the methods and materials, which include a description of the datasets, data transformation techniques used and classification algorithms applied, and section four, which contains the conclusion and future scope.

2. CLASSIFICATION

Classification in data mining is a form of data analysis and can use in extracting models to describe important data classes or to predict future data trends. The classification process has two phases; the first phase is learning process, the classification algorithm analyze the training data. We shall represent the learned model or classifier in the form of classification rules. Next, the second phase is classification process where it uses the test data to estimate the accuracy of the classification model or classifier.

2.1. Random Forest (RF)

This algorithm considers numerous decision trees, thus forming a forest. Therefore, it is an ensemble of decision tree algorithms. The building of the random tree begins at the top of the tree with the dataset. The first step involves selecting a feature at the root node and then splitting the training data into subsets for every possible value of the feature. This makes a branch for each possible value of the attribute. Tree design requires choosing a suitable attribute selection measure for splitting and the selection of the root node to maximize dissimilarity between classes. If the information gain is positive; the node is split else the node will become a leaf node that would provide a decision of the most common target class in the training subset [3].

The steps for the Random Forest are as follows:

Step 1: From a total of *n* features, randomly *m* features are selected, *m* << *n*

Step 2: Calculate a node *d*, which belongs to the set of *m* nodes using the best split point

Step 3: Further, Split *d* into daughter nodes using the best split method

Step 4: Repeat Steps 1-3 until form a tree with a root node and having the target as the leaf node

Step 5: Steps 1-4 represent the creation of a tree. Repeat them the number of times to create a forest.

It has the following pros and cons:

- It can handle large set of data with high dimensionality.
- It is useful in the case of missing data.
- Fit for some datasets with noisy classification/regression tasks.
- Classifications made by random forests are difficult to interpret.

2.2. Artificial Neural Network (ANN)

Artificial Neural networks are those systems modeled based on the human brain working. As the human brain consists of synapses that interconnect with millions of neurons, a neural network is a set of connected input/output units in which each connection has a weight associated with it. Multi-Layer Perceptron (MLP) network models are the popular network architectures. It is used in most of the research applications in medicine, engineering, mathematical modeling, etc.

In MLP, we pass the weighted sum of the inputs and bias term to activation level through a transfer function to produce the output, and arrange the units in a layered feed-forward topology called Feed Forward Neural Network (FFNN) [4]. We represent unit along with the bias term of the input unit and hidden unit in Figure 1.

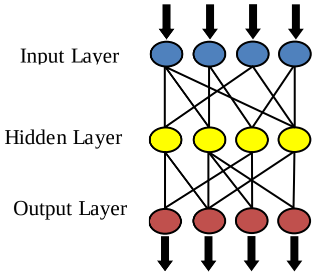

Figure 1. Feed forward neural network

It has these pros and cons:

1. Have the ability to work with inadequate knowledge and fault tolerance.
2. Have numerical strength that can perform more than one job at the same time.
3. It is nonlinear in nature; by this, it is suitable for all kinds of real-world problems.
4. The realization of the equipment is dependent.
5. Difficult to know how many neurons and layers we need to process and duration is also unknown.
6. When producing a solution, it does not give a clue as to why and how.

2.3. Support Vector Machine (SVM)

Support vector machine is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. These models closely relate to neural networks. They use a sigmoid kernel function, which is equivalent to a two-layer, perceptron neural network. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data [5]. It is a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. This classifier is very effective in high dimensional spaces.

Boser, Guyon and Vapnik invented SVM. The Support vector machine deals with pattern classification [6]. There are two types of patterns linear and non-linear. Linear patterns can be easily distinguishable and non-linear patterns are not easily distinguishable. The principal concept behind SVM is to develop optimal hyper-plane. We should use that hyper-plane for classification of linearly separable problems. The optimal hyper-plane means that the hyper-plane selected classifying patterns, which is having the maximum size [7]. It will be helpful to classify patterns correctly. If the margin size is large then there will be more correctly classified patterns.

Hyper-plane Equation:

$$\mathbf{X=AX+C} \quad (1)$$

The kernel function used to map given function is $\Phi(x)$.

$$\mathbf{X \rightarrow \Phi(x)} \quad (2)$$

The kernel functions used are SIGMOID, POLY, LINEAR and RBF.

The equation of Poly kernel function is:

$$\mathbf{K(x, y) =<x, y>\hat{}p} \quad (3)$$

The SVM gives the identically distributed, independent training samples. The following facts express SVM's advantages and disadvantages:

1. It scales well to high dimensional data.
2. The risk of overfitting is less.
3. It works well with unstructured and semi structured data.
4. Choosing a good kernel function is not easy.
5. Long training time for large datasets.
6. It is hard to visualize their impact.

3. PROPOSED SYSTEM

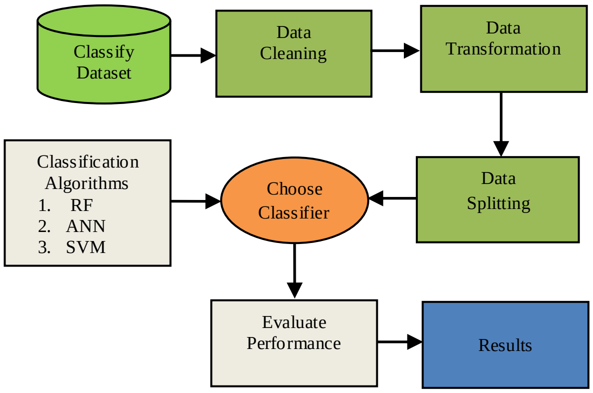

Figure 2. Cardiovascular disease diagnosis system

Proposed research work introduces a framework to develop a classifier based on data mining techniques. Another objective is to perform cross validation of different framework designed for different category of data. In this frameworks dataset is given to preprocessing stage which further classified by selected classifier. This approach involves:

1. Classify dataset
2. Data Pre-processing
3. Splitting training and testing data
4. Select classifier with the best performant
   i. RF
   ii. ANN
   iii. SVM
5. Interpret Results

3.1 Dataset Description

We design the system to integrate multiple indicators from many data sources to provide a comprehensive picture of the public health burden of CVDs and associated risk factors in the United States. There are two different data sets. Table 1 shows the details of the datasets.

Table 1. Dataset description

Database	No. of Attributes	No. of Instances
Cleveland	14	303
Cardio Train	12	700,000

The datasets used for experiments are:

1. Cleveland dataset provided by DHDSP, the National Cardiovascular Disease Surveillance System.
2. Cardio train dataset from Kaggle.

Cleveland data set has 75 attributes, but all published experiments refer to using a subset of 14 of them. This database have concentrated on simply attempting to distinguish presence or absence of cardiovascular disease. We are available the original dataset at [8]. By analyzing this dataset, 165 instances (54.46%) are having cardiovascular disease. We describe attributes descriptions in Table 2.

Table 2. Cleveland dataset attributes and their description

Attribute	Description
1. Age	Age in years
2. Sex	1=male;0=female
3. CP	Chest pain type(4 values)
4. trestbps	Resting blood pressure(mm Hg)
5. chol	Serum cholestoral in mg/dl
6. fbs	Fasting blood sugar(1=true;0=false)
7. restecg	Resting electrocardiographic results
8. thalach	Maximum heart rate achieved
9. exang	Exercise induced angina (1=yes;0=no)
10. oldpeak	ST depression induced by exercise relative to rest
11. slope	The slope of the peak exercise ST segment
12. ca	Number of major vessels(0-3)colored by flourosopy
13. thal	3=normal;6=fixed defect;7=reversible defect
target	0 or 1

The other one, cardio train dataset from Kaggle consists of 70,000 records of patients with 11 features plus target. We can download this dataset at [9].The amount of 34,979 instances (49.97%) have the disease. The attribute "cardio" describes the predictable attribute with value "1" for patients with cardiovascular disease and value "0" for patients with no disease. The attribute description for this dataset is as follows:

Table 3. Cardio train dataset attributes and their description

Attribute	Description
1. id	ID Number
2. age	Age in days
3. gender	1=women, 2=men
4. height	Height in cm
5. weight	Weight in kg
6. ap_hi	Systolic blood pressure
7. ap_lo	Distolic blood pressure
8. cholestrol	1=normal, 2=above normal, 3=well above normal
9. gluc	1=normal, 2=above normal, 3=well above normal
10. smoke	Whether patient smokes or not
11. alco	Binary Feature
12. active	Binary Feature
cardio	Target variable (0 or 1)

3.2 Data Preprocessing

Identification of the unnecessary attributes, which impedes the processing task, is crucial before the application of the classification technique. Besides acting as noise and disturbing the process, they also affect the classifier performance. To identify these, employ the statistical methods. We apply the data cleaning techniques first. Identify the missing values and replaced by the group median. Further, apply the min-max scaling technique to have the features value range between zero and one. Thus, the researchers have found that no classifier that generates the best result for each dataset.

3.2.1 Data Cleaning

Some instances have missing data for some of the features. Machine learning algorithms cannot work very well with missing data. To find a solution to "clean" the data, the easiest option is to eliminate all those records, but in this way, we would eliminate many important data. Another option is to calculate the median value for a specific column and substitute that value everywhere in the same column that have missing data.

3.2.2 Data Transformation

Most of the machine learning algorithms do not work very well if the features have a different set of values. The solution is to apply the feature scaling technique. Feature Selection Technique (FST) eliminates the less important features and reduces the time complexity of the machine learning technique. The type of scaling depends on the data fed to which model, so there is no universally best approach. In this paper, min-max normalization techniques is used. Min-max normalization preserves the relationships among the original data values. It always boosts the classification accuracy and minimizes the computational cost.

3.3 Splitting the Dataset

To check the performance of classifiers, part each dataset into two division – training and testing. Test a classifier using a testing dataset, is chosen based on its performance in comparison to other available classifiers. In this paper, we use the K-fold cross-validation method. It partitions the original data set into equal-sized sub-segments. The number of segments depends upon the value of *k* taken; in our case, we have taken *k* to be 3, 5, or 10. We use the first part to train the model ignoring the column with the pre assigned label. Then we use the trained model to make predictions on new data, which is the test dataset, not part of the training set, and compare the predicted value with the preassigned label.

The advantage of using this validation is that we can use every single data is for training as well as in testing the model and each entry in the dataset is used for validation of the result at least once. This helps to increase the accuracy of the model.

3.4. Comparison of different Algorithms

We compare the accuracy of multiple algorithms with two different datasets. To understand classifier's behavior, we should calculate metric Confusion Matrix. This matrix is a visualization tool that present the accuracy of the classifiers in classification [10]. Based on data mining techniques as explained above, evaluated all the developed models in terms of following error measures.

Table 4. Performance measures

Measures	Definitions	Formula
Accuracy (A)	Accuracy determines the accuracy of the algorithm in predicting instances.	A=(TP+TN) / (TN+TP+FP+FN)
Precision (P)	Precision measure the classifier's correctness/accuracy.	P = TP / (TP+ FP)
Recall (R)	To measure the classifiers' completeness or sensitivity, Recall is used.	R =TP / (TP+FN)

True positive (TP) denotes the number of identified positive samples in the positive set. True negative (TN) represents the number of classification negative samples in the negative set. False positive (FN) is the number of identified positive samples in the negative set. False negative (FN) means the number of identified negative samples in the positive set. The accuracy is as the ratio of the number of samples correctly classified by the classifier to the total number of samples.

In this research, we apply the K-Fold cross-validation technique by considering the different value of k to be 3, 5 and 10. In this work, we use Python programming language. Python provides a variety of efficient tools for data mining and data analysis. Among them, we used scikit-learn. It is a free software machine-learning library for the Python programming language. It features various classification, regression and clustering algorithms.

We present the resulted output of the three classifiers that predict the cardiovascular disease using small and large datasets in table 5.

Table 5. Performance comparison of three classifiers

Database	Kth Validation	Classification Model		
		RF	ANN	SVM
Cardio Train	10-fold	71.46	65.51	70.57
	5-fold	69.58	72.99	64.79
	3-fold	70	64.59	72.96
Cleveland	10-fold	83.87	87.10	87.10
	5-fold	88.52	85.25	86.86
	3-fold	84.62	81.32	81.32

By analyzing this result, we should apply K-fold cross-validation if the dataset is small because of getting more performant than larger dataset. According the comparison results, we consider that RF with 5-fold cross-validation gives the most performant algorithm for the Cleveland dataset. Then, using cardio train dataset, ANN with 5-fold have the highest accuracy in all of them. Therefore, we plot the confusion matrix of all classifiers based on accuracy measures using 5-fold cross-validation, which gives better accuracy.

Classification Matrix displays the frequency of correct and incorrect predictions [17]. It compares the actual values in the test dataset with the predicted values in the trained model. Table. 6 shows the results of the Classification Matrix for all the three algorithms. In Cleveland, we diagnosed that 85%, 85% and 91% patients have disease and can correctly classify 74%, 65%, 70% patients in cardio train. Then, we found that 15%, 15% and 19% patients for Cleveland and 26%, 35%, 30% for cardio train do not have cardiovascular disease but the model incorrectly classified that they had disease, it is very dangerous.

Table 6. Confusion Matrix for the two datasets

Model	**Actual Class**	**Cleveland**		**Cardio Train**	
		Predict Class		**Predict class**	
		Correct	Incorrect	Correct	Incorrect
RF	Yes	85	15	74	26
	No	91	9	65	35
ANN	Yes	85	15	65	35
	No	88	12	64	36
SVM	Yes	81	19	70	30
	No	88	12	70	30

In this study, we evaluate the performances of the models using the standard metrics of accuracy, precision, recall. We conducted three different experiments on the different size datasets using three algorithms: Random Forest, Neural Network and support vector machine as given in Table 7 and 8.

Table 7. Performance measures for Cleveland dataset

Model	**Cleveland**				
	Accuracy (%)	**Precision**		**Recall**	
		TP	TN	TP	TN
RF	89	0.89	0.88	0.91	0.85
ANN	85	0.86	0.85	0.88	0.81
SVM	87	0.88	0.85	0.88	0.85

According to Table.7, the True positive rate for Random Forest algorithm (0.89), Artificial Neural Network (0.86) and Support Vector Machine (0.88). Whereas Random Forest is best in True Positive Rate and Artificial Neural Network performed lowest in True Positive Rate. The True Negative Rate for Random Forest (0.91), Artificial Neural Network (0.88) and Support Vector Machine (0.88), we observed that all the three algorithms performed best in True Positive Rate.

For True positive, Support Vector Machine can predict with the lowest rate in Cardio Train dataset and Random Forest and Artificial Neural Network are better performance. By analyzing the result of performance measures for two dataset, we discover the models are best in identifying Negative cases for Cardio Train and best in Positive cases for Cleveland.

Table 8. Performance measures for Cardio Train dataset

Model	Cardio Train				
	Accuracy (%)	Precision		Recall	
		TP	TN	TP	TN
RF	70	0.72	0.68	0.65	0.74
ANN	73	0.72	0.70	0.70	0.72
SVM	65	0.65	0.65	0.64	0.65

To build the model, it took 2.21, 0.60 and 0.05 seconds time for Cleveland, and 2.16, 73.08 and 303.37 seconds for Cardio Train respectively. We described the comparisons of performance based on accuracy percentage with bar chart as in Figure 3 and 4.

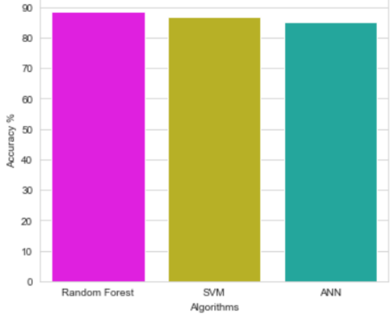

Figure 3. Performance evaluation of classifiers using Cleveland dataset

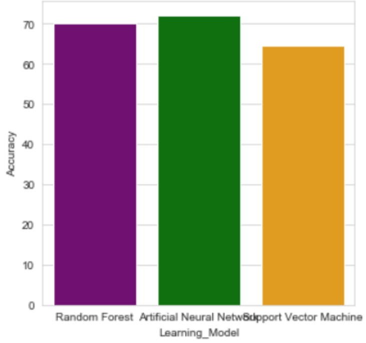

Figure 4. Performance evaluation of classifiers using Cardio Train dataset

In this paper, we used two different datasets, one consists of more than three hundred records and the other contains seventy thousands records. We found by the result with the highest accuracy of 88.52% was achieved by using the Random Forest with 5-fold cross-validation. From the results, we also observed that the Neural Network with 72.99% accuracy was performing better compared to all the other classification algorithms for larger dataset. The above results show that Random Forest is best in Cleveland and Artificial Neural Network is performing better amongst all the other algorithms for Cardio Train.

4. CONCLUSIONS

This research was using two different dataset. First, replace the group median values in all missing values. Further, apply the data transformation technique with a proper feature scaling method. We use three learning algorithms along with *k*-fold cross-validation with k=3, 5 and 10. This enabled to perform data analysis to obtain the optimal result. Every model can has best performance for specific dataset. The accuracy depend on the nature of dataset. The performance score are not ….We observed that Artificial Neural Network has much impressive power. It works best in large dataset and more robust when encountering with missing values. Future work will include trying a study with different data transformations or trying algorithms that we have not tested yet for further analysis of the dataset.

REFERENCES

[1]Dey M, Rautaray SS (2014) Study and analysis of data mining algorithms for healthcare decision support system. Int J Comput Sci Inf Technol 5: 470-477.

[2]Bhatla N, Jyoti K (2012) An analysis of heart disease prediction using different data mining techniques. Int J Adv Res Technol 1: 1-4.

[3]W. Almayyan, "Lymph Diseases Prediction Using Random Forest and Particle Swarm Optimization", J. of Intelligent Learning Systems and Applications, Vol. 8, No.3:51-62 (2016). DOI: 10.4236/jilsa.2016.83005.

[4]Padmavati J., "A Comparative study on Breast Cancer Prediction Using RBF and MLP," International Journal of Scientific & Engineering Research, vol. 2, Jan. 2011.

[5]XindongWu , Vipin Kumar , J. Ross Quinlan , Joydeep Ghosh , Qiang Yang , Hiroshi Motoda , Geoffrey J. McLachlan , Angus Ng , Bing Liu , Philip S. Yu , Zhi-Hua Zhou , Michael Steinbach , David J. Hand and Dan Steinberg, Top 10 algorithms in data mining, Knowl Inf Syst (2008) 14:1–37, Springer-Verlag London Limited 2007

[6]https://www.kaggle.com/sulianova/cardiovascular-disease dataset/download/IFCtlrBySQEGm0VBHLv%2Fversions%2F40V9qH7z2bTsP9V9cAsZ%2Ffiles%2Fcardio_train.csv?datasetVersionNumber=1

[7]Classification in Data Mining,https://www.tutorialspoint.com

[8]https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[9]Kim J, Lee J, Lee Y (2015) Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. Health Informatics J21: 167-174.