An Automatic Early Risk Classification of Hard Coronary Heart Diseases using Framingham Scoring Model

Hoda Ahmed Galal Elsayed
Research Scholar, Department of Software Engineering,
Prince Sultan University
Riyadh, KSA
(+966) 0552902661
helsayed1993@gmail.com

Liyakathunisa Syed
Assistant Professor, College of Computer & Information Sciences,
Prince Sultan University
Riyadh, KSA
(+966) 0538124035
lsyed@psu.edu.sa

ABSTRACT

Coronary Heart disease is the global leading cause of death, accounting for 17.3 million deaths per year, and this number is expected to grow to more than 23.6 million by 2030 [20]. In healthcare, coronary artery diseases were found on the top of the healthcare problems, that many countries are facing nowadays. Data mining techniques have been widely used in many governmental sectors including healthcare to mine knowledgeable information from medical data. The current health care organizations use manual heart rate risk scoring models such as Framingham to calculate the early risk of coronary artery diseases. Due to the growing population and increase in the number of patients at health care, the manual process is becoming inefficient to treat the condition which may demand immediate treatment. In this research work, we are proposing an automated system for early risk classification of hard coronary heart diseases using Framingham scoring model. K-Nearest Neighbor and Random Forests algorithms were applied for heart rate risk prediction and the obtained results were compared to the results obtained through the manual process to measure the accuracy level. It was observed that, our proposed automated system for heart rate risk prediction using Framingham model was highly accurate when compared to the manual process. This work attempts to report the effectiveness of using K-Nearest Neighbor and Random Forests for Framingham heart and medical decision support in cardiology field.

CCS Concepts

• [Information Systems]: Information systems Applications—Data Mining—Medicine and Science

Keywords

Early risk detection; hard coronary heart diseases; K-Nearest Neighbor; Random Forests; Framingham scoring model; data mining

1. INTRODUCTION

There is a tremendous increase in medical information databases due to the rapid increase in new patient records and physician diagnosis information. Each record in these databases is indeed so useful not only for storage purposes but also in training medical decision-support systems to save many lives [27]. Data mining approaches can be adapted to support medical decisions. Data mining is an analysis process used to generate summary reports of large data from different perspectives and thus it converts these data to more reliable information [21]. The main goal behind it is to turn data that represents facts into knowledge-based data that can be further processed to detect possible observations much faster. Furthermore, it extracts the features that reflects common patterns that are hard to be tracked otherwise [7,14].

Data mining techniques are either predictive or descriptive techniques [10]. Data mining is highly related to machine learning process, which is part of the artificial intelligence, and thus it also involves data pre-processing and classification techniques that can be used to derive object classes with knowledge of its siblings [12]. Many researchers are utilizing data mining techniques in the medical domain to help professionals in heart diseases diagnosis [8,25]. The primary task of data mining is classification and predictions [2].

The world health organization (WHO) reported that cardiac diseases are the leading cause of sudden mortality in countries with different income levels [28] which reflects the existence of a wide dataset that can be used for knowledge discovery. On the other hand, the Framingham Heart Study (FHS) [4,9,11,15,26] identified the common characteristics that highly contribute to the development of cardiovascular diseases through close observation to different participants' generation for considerably long time. FHS has been designed using various Framingham scoring models to address different heart related diseases including: 1) atrial fibrillation, 2) cardiovascular diseases 3) congestive heart failure, 4) coronary heart diseases, 5) diabetes, 6) hypertension, 7) intermittent claudication, 8) strokes. By careful monitoring of the study, it has led to the identification of major risk factors that can lead to cardiac problems as well as their effects.

For coronary heart diseases early risk detection, the following four models were developed: 1) 10 years hard coronary heart diseases risk, 2) 10-years coronary heart disease risk, 3) recurrent coronary heart disease, 4) 2-years coronary heart disease risk. In this paper we will address the first type of these models which is the 10 years hard coronary heart diseases (HCHD) which is applicable for people with 20-79 age range. Hard coronary heart diseases are known to cause myocardial infarction or coronary death if not detected in the early stages [4, 9,11,15,26]. In this study, we focused on applying data mining for prediction purposes to indicate hard coronary heart diseases risks in advance, based on observations driven from dataset. In particular, we examine the accuracy level of both K-Nearest Neighbor and Random Forest in risk classification process using the Framingham risk factors for HCHD.

Organization of the paper is as follows: Section 2 presents related work that served as a reference for this study. Section 3 presents the

ICC '17, March 22 2017, Cambridge, United Kingdom
 ISBN 978-1-4503-4774-7/17/03$15.00
DOI: http://dx.doi.org/10.1145/3018896.3036384

proposed methodology. Section 4 presents the results obtained using R-studio and SPM. Finally, Section 5 presents conclusion of our study and findings.

2. RELATED WORK

KNN is considered one of the top classification approaches that is widely used in risk prediction not only in the medical field but also in different other life-related aspects. Shouman et al. [23] investigated the effectiveness of using KNN in diagnosing the different heart diseases. They also claimed that the integration of voting to the KNN can enhance the accuracy in the heart diseases' patient diagnosis. Their study reflected that KNN provided a higher accuracy compared to the decision tree, naïve bayed, fuzzy AIRS-KNN, bagging algorithm, neural networks and nine voting gain ratio decision tree. The authors reported that KNN's accuracy was (97.4%) compared to other data mining techniques that used a benchmarking dataset. However, it was also recorded that unlike for decision trees, voting wasn't effective in increasing the accuracy of KNN. Another study by Zhao et al [17], used KNN in predicting the highway traffic accidents risks to enhance the safety and efficiency of the transportation systems. In their study, the authors reported that KNN outperformed the other techniques by providing 80% accuracy rate.

Thamilselvan et al. [24] conducted a study to detect and classify lung cancer MRI images by using the Enhanced KNN approach (EKNN). In their paper, the authors highlighted that the accuracy of KNN, improved existing KNN and EKNN are 80%, 92.85% and 96.57% respectively. They reported that their approach was able to classify malignant cells exactly with low error signal. In 2016 [19], Mohanraj et al.'s study merged both KNN classification and k-mean clustering techniques to predict the heart diseases. The authors claimed that their approach is capable of enhancing the classification accuracy as well as the security level of the medical data. Jabbar et al [13] proposed a new algorithm that combines both KNN and genetic algorithm for classification. The accuracy results for KNN alone applied to their dataset and KNN with genetic algorithms are 95% and 100% respectively.

In [1], Abdullah A. used random forest classifier for coronary heart diseases events prediction (e.g. causes and symptoms). The result of applying that approach was higher than that of decision tree. Their study recorded that random forest's accuracy in coronary heart events' prediction is 63.33% whereas decision tree's accuracy was 50.67%. Ani et al. [6] used four classification data mining techniques for coronary heart diseases' prediction. The results of that paper reflect that the accuracy level of random forests, naïve bayes, decision tree- c4.5 and KNN are 89%, 84%, 81% and 77% respectively.

As observed in the previous studies, most of the researches applied in the cardiac domain focused on examining the effectiveness of data mining techniques in the prediction of different heart diseases for diagnosis purposes. Unlike these studies, ours focused on applying both KNN and random forests as they showed a high level of accuracy in the cardiac medical field. Furthermore, our study addresses the hard coronary heart diseases in specific that can lead to sudden death if not early detected. The biomarkers are collected based on the Framingham heart study's 10-years scoring model for HCHDs' early risk detection.

3. METHODOLOGY

In this research work we are using K-Nearest Neighbor and Random Forests techniques along with the Framingham risk scoring model for hard coronary heart diseases. Subsequent sections explain data elicitation, preprocessing and Framingham calculations for heart rate risk prediction.

3.1 Data Elicitation and Preprocessing

The data, elicited for the purpose of this study, were real data from patients, who had a cardiac checkup, taken from a hospital in Saudi Arabia. The data include the following attributes for each patient: age, gender, glucose level, total cholesterol, HDL, SBP and the family history. Few attributes had missing information (e.g. smoking status per each patient) which required preprocessing before applying the machine learning techniques. Moreover, the hypertension treatment was marked as "yes" for all the patients in our sample from the hospital records. Furthermore, as the ranges of numeric data varied widely, normalization was done to keep all the numeric data values within 0 and 1 range. Normalization using the equation (1) was required to apply to our dataset's numeric columns to make the numeric values more consistent in terms of format which will thus ease the process of learning. The sample size of our data include 60 patients originally but 52 currently after excluding six patients' data as they reflected a diabetic status which cannot be used by the Framingham's HCHD risk model as explained previously. Also, two more patient's results were excluded because they were 80 years old which is above the study's age range. The subjects involved in this study are 26 females (age 44-74) and 26 males (age 38-62). The total risk points and equivalent risk percentages were manually calculated using the Framingham's risk model for HCHD as explained previously. A summary of the process applied to the database can be shown in figure 1.

$$\textbf{\textit{E1}}: \frac{(x - min(x))}{(\max(x) - \min(x))} \quad (1)$$

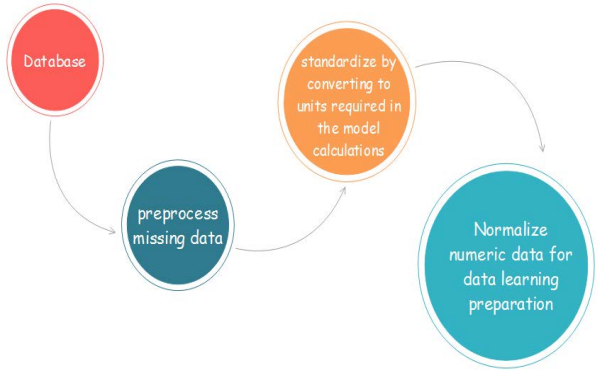

Figure1. Database Preparation Process to Machine Learning

3.2 Framingham Heart Study and HCHD's Biomarkers

The Framingham study has set the biomarkers, also known as risk factors, for the early risk detection of hard coronary heart diseases such as age, gender, total cholesterol, high density lipoproteins (HDL), systolic blood pressure (SBP), treatment for hypertension and smoking status. In our study we included the diabetic status as well as it was found that this model is applicable for individuals with no diabetes. Furthermore, the family history is considered in this study. The relation between these different biomarkers in the calculation process of the HCHDs' risks is illustrated in figure 2, which shows how the different risk factors are related in

determining the total points needed to estimate the 10-years risk percentage that is mapped later to the risk class with three levels: low, moderate and high.

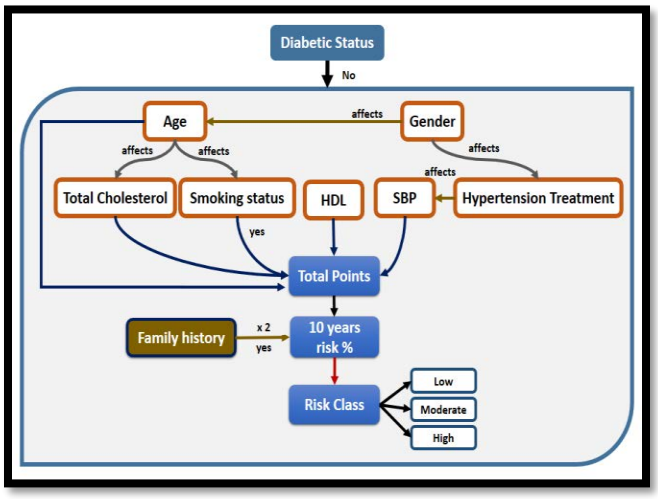

Figure 2. Risk Factors of Framingham Scoring Model

3.2.1 The Framingham calculation process of HCHDs' early risks

The Framingham's total points score is calculated by summing the values of these five risk factors which are age, HDL, SBP, total cholesterol and smoking status. The gender is also considered a risk factor as it participates in defining the age risk factor as well as the hypertension treatment attribute based on which SBP is calculated. Similarly, the age risk factor is not only used in calculating the total number points but also contributes in defining the points assigned to other risk factors, any change in it can easily affect the total number of risk points. The summed total of these points is then mapped to risk percentage. The role of family history risk factor takes place at this step to either duplicate the risk percentage or keep it as it is. If the patient has no family history to cardiac diseases, the risk percentage will remain the same otherwise, it will be duplicated. The risk is then defined based on the following criteria [16]: 1) low risk class if risk percent is less than or equals to ten, 2) moderate risk class if the risk percent is between 10 and 20 inclusive, 3) high risk class if the risk percent is over 20. A summary of the mapping process between total points and risk % is shown below in table 1.

Table 1. Total points to risk % to risk class mapping matrix

Females			Males		
Total points	Risk%	Risk Class	Total points	Risk%	Risk Class
<9	<1	*Low*	<0	<1	*Low*
9	1		0	1	
10	1		1	1	
11	1		2	1	
12	1		3	1	
13	2		4	1	
14	2		5	2	
15	3		6	2	
16	4		7	3	
17	5		8	4	
18	6		9	5	
19	8		10	6	
20	11	*Moderate*	11	8	
21	14		12	10	*Moderate*
22	17		13	12	
23	22	*High*	14	16	
24	27		15	20	
=>25	=>30		16	25	*High*
			=>17	=>30	

3.3 Data Classification Algorithms

3.3.1 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is one of the top 10 algorithms in data mining [29] that is widely used for classification in predictive modeling. By applying KNN, our dataset was split into two portions as shown in figure 3. 1) Training and 2) test using R-Programming [22]. The test data was predicted based on the patterns observed in the training data, by measuring the minimum distance between the new instance to the stored training data set. The distance was calculated using the *Euclidean distance* [23] applying equation (2)

$$\textbf{E2: } d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (2)$$

where q_i represents the attribute value of the new instance and p_i is the attribute value of the already stored data for training. The class to be predicted was defined afterwards according to training data that reflected the shortest distance. In case of equality of two or more data instances, the classification was done randomly which may increase the error rate produced by the built model classifier. One of the major challenges that is usually faced with KNN is that the higher the range of specific attribute, the higher its influence will be [23]. Thus, in order to avoid this problem, we normalized the data by applying equation (1) on all numeric attributes to keep their range between 0 and 1 to have same influence on the distance measure between the instances.

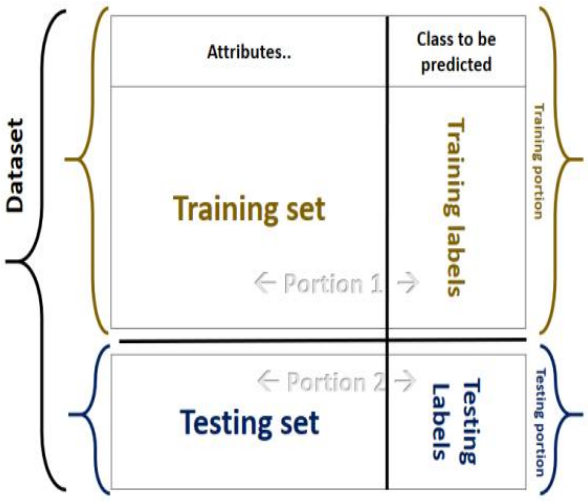

Figure 3. KNN Divisions of the dataset for predictive modeling

3.3.2 Random forest

Random forest is one of the most effective algorithms used in prediction [3]. It takes a training set as an input and produces multiple decision trees, hence called random forest. Random set is generated for each tree [5] wherein forecaster variables are used to split the training data into homogenous subsets [18]. New entry classification depends on combining the results from the n trees generated. In this study, we generated 200 decision trees on the Salford predictive modeler (SPM) for our dataset using random selected attributes. The risk is predicted accordingly and accuracy was recorded illustrated in the results section.

4. RESULTS AND DISCUSSION

In this section, we will summarize our findings on both KNN and random forests techniques after examining them on our dataset.

4.1 KNN

Using R- data mining, we have divided our data sample into 2 portions: 1) first randomly arranged 43 rows for training set (83%) and 2) last 9 randomly arranged test set (17%). The reason for randomly allocating the database rows is that the risk classes are not equally distributed among patients. In other words, the patients with low risk class are8 out of the 52 and those with high risk class are 7 in the whole sample while those with moderate risk class are the majority with a total of 37 as shown in figure 4 (results from R).

```
> table(nrisk$RiskClass)

    high      low moderate
       7        8       37
```

Figure 4. Risk Class Values Distribution

Thus, a well distributed mixture of these minimal data should be used in both the training set as well as the test set. KNN algorithm was applied to both the training and test data set to check for correctly classified and incorrect classified data samples.

Confusion matrix was generated using R programming to compare similarities and misclassifications. Figure 5 shows, the confusion matrix, which displays the classifier result using KNN. From the results, it is clear that we were able to detect six correct risk classes out of the 9 cases in the test set which reflects that our classifier's accuracy is 66.7%. However, the correct classifications all fall under the moderate risk class which is the highly existing risk class in our sample.

```
> table(nrisk_test_target, m1)
                  m1
nrisk_test_target high low moderate
         high        0   0        1
         low         0   0        2
         moderate    0   0        6
```

```
> library("gmodels")
> CrossTable(x = nrisk_test_target, y = m1, prop.chisq=FALSE)


   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  9


                 | m1
nrisk_test_target |  moderate | Row Total |
------------------|-----------|-----------|
             high |         1 |         1 |
                  |     0.111 |           |
------------------|-----------|-----------|
              low |         2 |         2 |
                  |     0.222 |           |
------------------|-----------|-----------|
         moderate |         6 |         6 |
                  |     0.667 |           |
------------------|-----------|-----------|
     Column Total |         9 |         9 |
------------------|-----------|-----------|
```

Figure 5. Confusion Matrix for KNN using R-Studio (k=1)

4.2 Random forests

Random forests classifier on SPM generated 200 random decision trees, where all of them were compared and the best outcome was extracted. In figure 6, a summary of these generated 200 trees and the balanced error rate of each is shown. In the graph, the circled tree represents the tree with the lowest error rate and thus it will be picked by the random forest classier for further analysis. A summary in terms of accuracy and error rate of this chosen decision tree can be seen in figure 7, which reflects that the best accuracy gained out of the 200 decision trees is 63.49% while the error rate is nearly 30.8%. The details behind this summary can be found in the confusion matrix, illustrated in figure 8. In that matrix, we show that a total of 33 risk classes were predicted correctly out of 52 sample patients' data. Although the accuracy rate generated by the random forest algorithm is considerably lower that of KNN, it was highly noticed that, unlike the KNN, the correct classification of the random forests covered all the risk classes. Figure 9, shows the misclassifications generated by the random forest classifier.

Name	OOB
Balanced Error Rate (Simple Average over classes)	0.30775
Class. Accuracy (Baseline threshold)	0.63462

Figure 7. Decision Tree's Accuracy and Error Rate

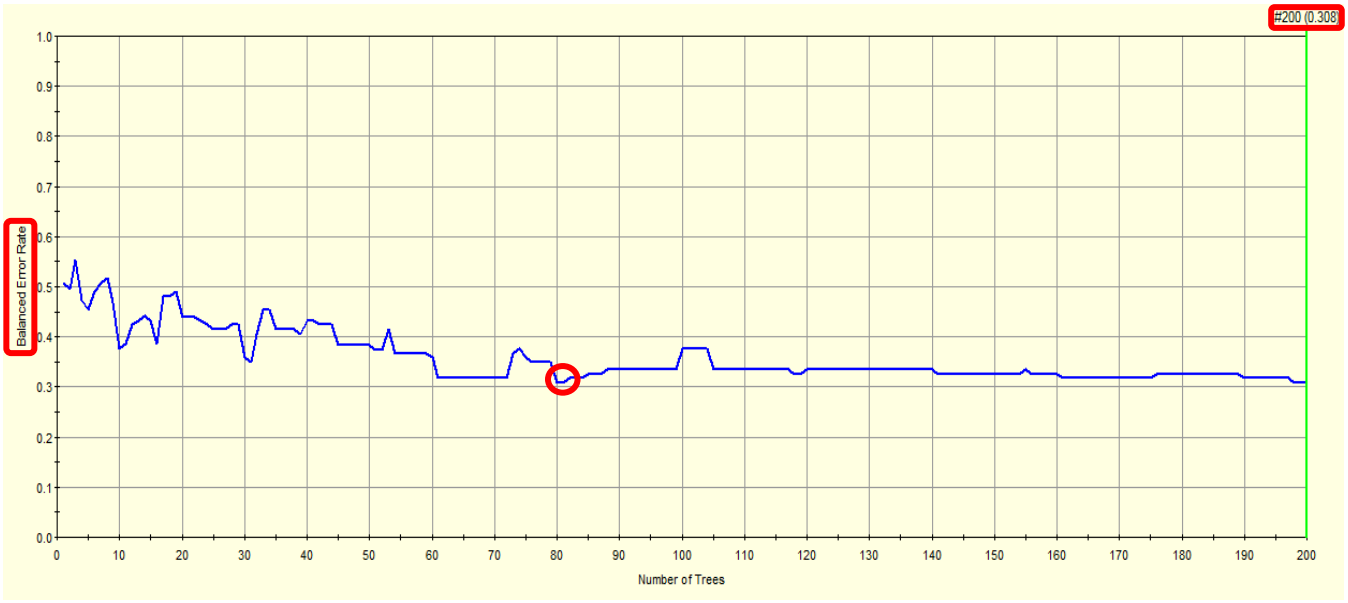

Figure 6. 200 Generated Decision Trees by the R

Actual Class	Total Class	Percent Correct	Predicted Classes		
			high N = 12	low N = 14	moderate N = 26
high	7	85.71%	**6**	0	1
low	8	62.50%	0	**5**	3
moderate	37	59.46%	6	9	**22**
Total:	52				
Average:		69.22%			
Overall % Correct:		63.46%			

Figure 8. Confusion matrix for Random Forest's

OOB Sample

Class	N Cases	N Mis-Classed	Pct. Error	Cost
high	7	1	14.29%	0.14286
low	8	3	37.50%	0.37500
moderate	37	15	40.54%	0.40541

Figure 9. Misclassifications by the Random Forests Classifier

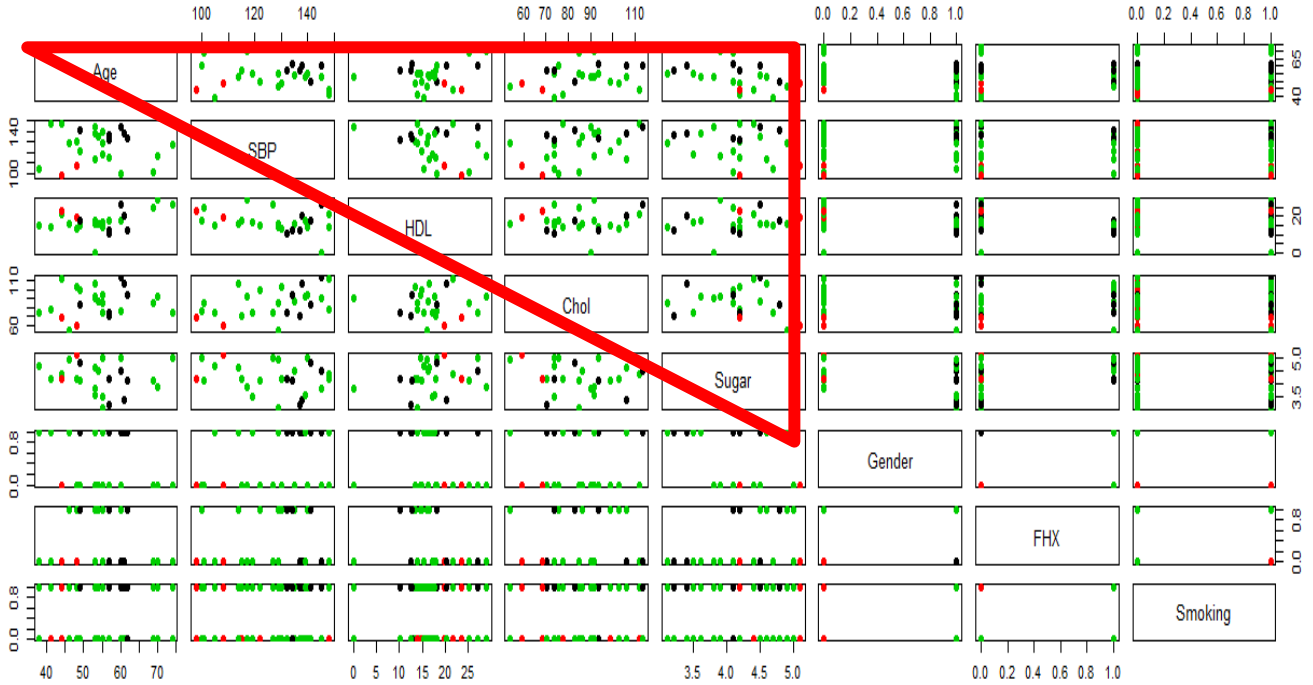

Figure 10. The Relationships among different Framingham Biomarkers and the Risk Classes

4.3 Impact of biomarkers on risk prediction

Using KNN, we plotted a scattered model that shows the impact of the 8 biomarkers provided by Framingham scoring model. This model is shown in figure 10 above. In this model, we highlighted the important relationship that highly contributes to the risk class prediction process. The triangular area reflects too many combinations that identify the risk category under which the patient is. These active attributes are age, SBP, HDL, cholesterol and glucose level. One thing to note here is that although the other attributes contribute little to the risk classification, they also can't be ignored. It was also found that patients with a family history and high blood pressure are most likely to be in the moderate or high risk categories results illustrated in figure 11 – part a, whereas the high risk possibility increases for males with older age than females who might be at moderate risk, results shown in figure 11- part b. Moreover, patients with high cholesterol and are smokers at the same time can develop a high risk possibility than non-smokers, as shown in figure 11 – part c.

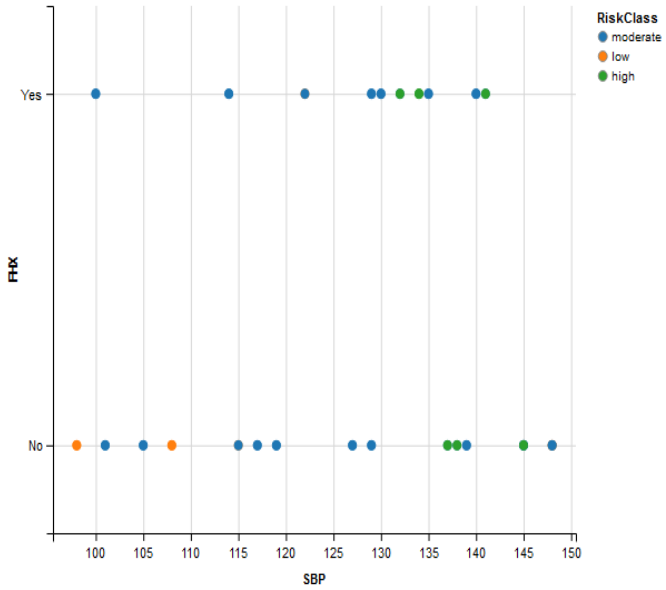

Figure 11(a). SBP vs. Family History in Risk Classification

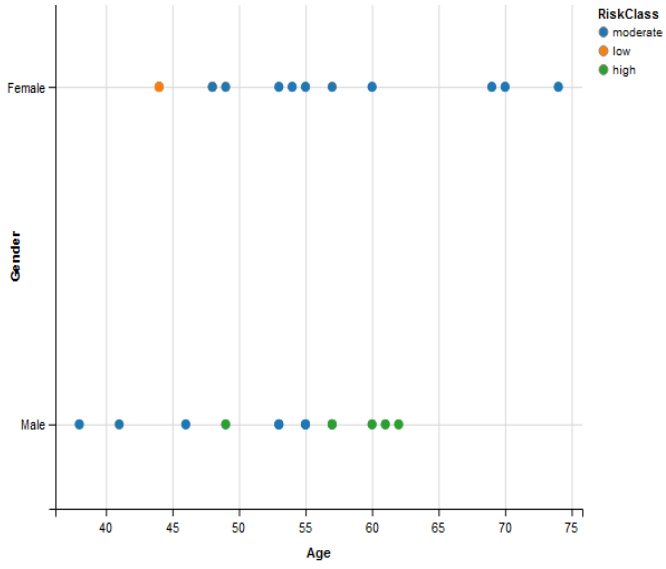

Figure 11(b). Age vs. Gender in Risk Classification

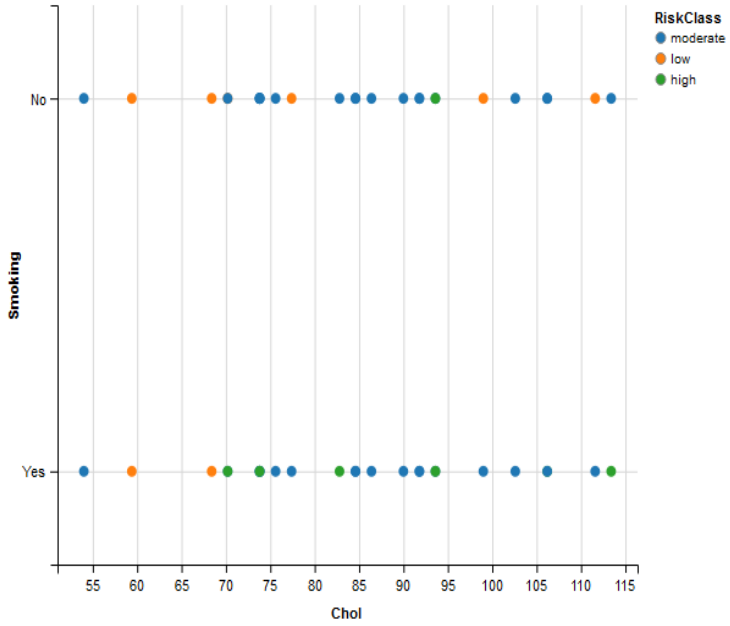

Figure 11(c). Cholesterol vs. Smoking Status in Risk Classification

Similarly, we reflected the attribute importance according to the random forests classifieras shownin figure 12. Both KNN and random forests agreed on the rank of age and SBP but some variances were spotted in the rest. Figure 13 shows the true positives vs the false positives being recorded based on the three classes under focus.The true positive rate in all three part of the figure 13(a) represnts represents the true positives versus false positives for high risk factor. Figure 13(b) represents represents the true positives versus false positives for low risk and figure 13(c) represents the true positives versus false positives for moderate risk . True positives are number of instances that are correctly identified whereas the false positive rate represents the falsely identified positives amongst all three risk classes (low, moderate and high). The rate of true positves is clearly higher in all three cases which reflects high accuracy of the automatic classification done through the random forest classifier.

Variable	Score
AGE	100.00
SBP	54.26
CHOL	48.08
HDL	37.92
GENDER	30.53
SMOKING	19.98
SUGAR	15.06
FHX	5.54

Figure 12. Variables' importance Report Generated by the Random Forest Classifier

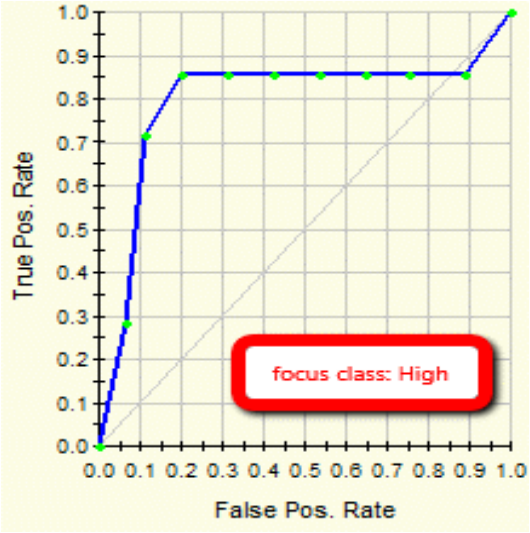

Figure 13(a). True Positives vs. False Positive class: High

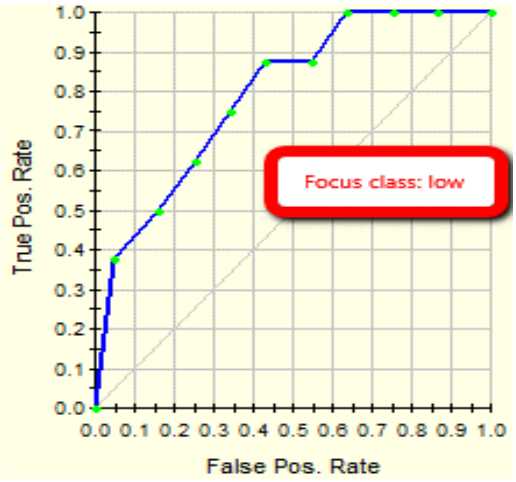

Figure 13 (b). True Positives vs. False Positive for class: Low

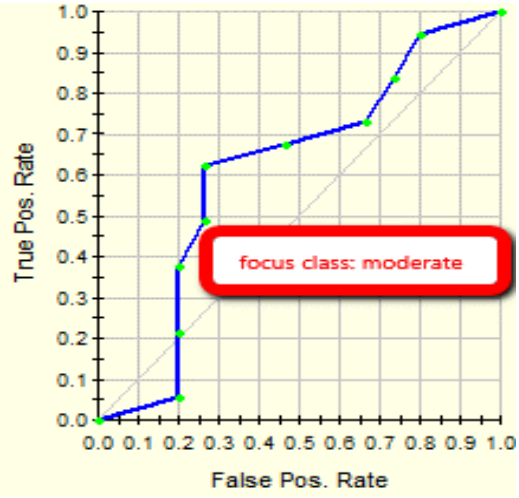

Figure 13 (c). True Positives vs. False Positive for class: Moderate

5. CONCLUSION

Data mining plays an inevitable role in the prediction process of many chronic diseases including deadly heart related diseases. In this study, we examined and revealed the results of applying both KNN and random forests to the Framingham scoring model designed for early risk prediction of HCHD. The results reflected that the accuracy of KNN was higher compared to random forests in identifying the risk classes among the test dataset compared to the training one. However, the accuracy rate is still to be improved. In our future work, we plan to use the enhanced KNN approach to enhance the classifier's performance [24]. Moreover, we will also apply clustering techniques along with the KNN. In addition, the sample size should increase to contain patients' cases with more low and high risk classes for better accuracy in the data.

6. ACKNOWLEDGMENTS

I would like to extend my thanks to Dr. Liyakathunisa Syed and Prince Sultan University for the continuous support and motivation to get this work done. Also, I would like to thank Dr. Mariam Ahmed Galal for the great medical guidance and for XYZ hospital for allowing access to the medical data required in the study.

7. REFERENCES

[1] Abdullah, A. S. (2012). A Data Mining Model for predicting the Coronary Heart Disease using Random Forest Classifier, (Icon3c), 22–25

[2] Adnan, M. H. M., Husain, W., & Rashid, N. A. (2012). Data Mining for Medical Systems: A Review. Acit 2012, (January 2012), 978–981. https://doi.org/10.3850/978-981-07-3161-8

[3] Akhil Jabbar M, Deekshatulua BL, Priti Chandra B (2013) Classification of heart disease using K-nearest neighbor and genetic algorithm. Int Conf Comput Intell: Model Tech Appl(CIMTA)

[4] Anderson M, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile: a statement for health professionals. Circulation. 1991;83:356-362.

[5] Ani, R., Augustine, A., Akhil N.C. and Deepa, O.S. (2016). Proceedings of the International Conference on Soft Computing Systems. Advances in Intelligent Systems and Computing, 397, 909–917. https://doi.org/10.1007/978-81-322-2671-0

[6] Ani, R., Augustine, A., Akhil, N. C., & Deepa, O. S. (2016). Random Forest Ensemble Classifier to Predict the Coronary Heart Disease Using Risk Factors. In Proceedings of the International Conference on Soft Computing Systems (pp. 701-710). Springer India.

[7] B. Thuraisingham, "A Primer for Understanding and Applying Data Mining," IT Professional IEEE, 2000.

[8] C. Helma, E. Gottmann, and S. Kramer, "Knowledge discovery and data mining in toxicology," Statistical Methods in Medical Research, 2000.

[9] Gordon T, Kannel WB. Multiple risk functions for predicting coronary heart disease: the concept, accuracy, and application. Am Heart J. 1982;103:1031-1039.

[10] H. Cheng, et al. (2010). Data mining for protein secondary structure prediction. 134.

[11] Coronary Risk Handbook: Estimating Risk of Coronary Heart Disease in Daily Practice. New York, NY: American Heart Association; 1973:1-35.

[12] I. H. Witten, et al., Data mining: practical machine learning tools and techniques, 3rd ed.: Morgan Kaufmann, 2011.

[13] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of Heart Disease Using K- Nearest Neighbor

and Genetic Algorithm. Procedia Technology, 10, 85–94. https://doi.org/10.1016/j.protcy.2013.12.340

[14] J. Han and M. Kamber, "Data Mining Concepts and Techniques," Morgan Kaufmann Publishers, 2006.

[15] Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol. 1976;38:46-51.

[16] Kim, H. C. (2012). Clinical utility of novel biomarkers in the prediction of coronary heart disease. Korean Circ J, 42(4), 223–228. https://doi.org/10.4070/kcj.2012.42.4.223

[17] Lv, Y., Tang, S., & Zhao, H. (2009, April). Real-time highway traffic accident prediction based on the k-nearest neighbor method. In 2009 International Conference on Measuring Technology and Mechatronics Automation (Vol. 3, pp. 547-550). IEEE.

[18] Mellor A, Haywood A, Stone C, Jones S. The performance of random forests in an operational setting for large area sclerophyll forest classification. Remote Sens 2013; 5:6 p. 2838–2856.

[19] Mohanraj, E., Subhasuryaa, K., Sudha, P., & K, S. K. (2016). Heart Disease Prediction using K Nearest Neighbour and K Means Clustering, (2).

[20] Mozaffarian D, Benjamin EJ**.(2015).** on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—update: a report from the American Heart Association. [published online ahead of print December 17, 2014]. Circulation. doi: 10.1161/CIR.0000000000000152.

[21] Q. Luo, "Advancing knowledge discovery and data mining," in WKDD '08 Proceedings of the First International Workshop on Knowledge Discovery and Data Mining, Washington, DC, USA, 2008.

[22] RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

[23] Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-nearest neighbour in diagnosing heart disease patients. International Journal of Information and Education Technology, 2(3), 220.

[24] Thamilselvan, P., & Sathiaseelan, J. G. R. (2016). Detection and Classification of Lung Cancer MRI Images by using Enhanced K Nearest Neighbor Algorithm, 9(November). https://doi.org/10.17485/ijst/2016/v9i43/104642

[25] V. Podgorelec et al., "Decision Trees: An Overview and Their Use in Medicine," Journal of Medical Systems, vol. 26, 2002.

[26] Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97:1837-1847.

[27] W. Lord and D. Wiggins, "Medical Decision Support Systems Advances in Health Care Technology Care Shaping the Future of Medical." vol. 6, G. Spekowius and T. Wendler, Eds., ed: Springer Netherlands, 2006, pp. 403-419.

[28] World Health Organization. (July 2007-Febuary 2011). [Online]. Available: http://www.who.int/mediacentre/factsheets/fs310.pdf

[29] Wu X, Kumar P, Quinlan JR, Ghosh J, Yang Q, Motoda H. Top 10 algorithm in data mining. Knowledge and Information Systems. 2008; 14(1):1–37.