A Classification Model for Predicting Heart Failure in Cardiac Patients

Muhammad Saqlain(✉), Rao Muzamal Liaqat, Nazar A. Saqib, and Mazhar Hameed

College of Electrical and Mechanical Engineering (E&ME), National University of Sciences and Technology (NUST), Islamabad, Pakistan
m.saqlainl240@yahoo.com, muzammilliaqat@gmail.com, nazar.abbas@ce.ceme.edu.pk, mazharhameedsw@gmail.com

Abstract. Today the most significant public health problem is Heart Failure (HF). There are a lot of raw medical data available to healthcare organizations in the form of structured and unstructured datasets, but the need is to analyze this data to get information and to make intelligent decisions. By using data mining, classification tool on a real dataset of cardiac patients we propose a model which classified these patients into four major classes. This model will help to identify the risk of HF and patients who have no HF signs but structural irregularities. We can also identify the patients having HF signs and irregularities and those having the critical stage of HF. This paper provides a detailed summary of modern strategies for management and analysis of HF patients by classes (1 to 4) that have appeared in the past few years.

Keywords: Data mining · Classification techniques · Heart Failure · Predictive model · Support Vector Machine

1 Introduction

Heart failure (HF) has become a foremost reason for cardiovascular morbidity and mortality [1], and its occurrence is increasing day by day [2]. In common population, the chance of getting HF for a healthy person at 40 years of age is 1 in 5 [3]. It has become the key public health care precedence to control high HF patient's mortality rate [4]. It is the major goal for healthcare organizations to identify the cost-effective techniques to minimize the occurrence of hospitalization. An accurate prediction model can be very useful for physicians as well as for patients. Using this model a physician can recommend new insistent treatment plan and the patient can follow this plan more confidently [7].

Raw data collected from the patient's history can be very helpful for healthcare organizations if they can get the meaningful hidden patterns from it [5], and these hidden patterns are used to build predictive models for medical practitioners to control diseases and to making intelligent decisions before actual diseases occur. Data mining is one of the most important techniques for knowledge discovery in the dataset (KDD) and it can be used for disease prediction and for extracting hidden patterns [6]. There are a lot of databases available for healthcare organizations in the form of

M.U. Ahmed et al. (Eds.): HealthyIoT 2016, LNICST 187, pp. 36–43, 2016.
DOI: 10.1007/978-3-319-51234-1_6

radiology reports, images, medication profiles, treatment records, signals, patient history, and pathology report. This type of data can be very complex, heterogeneous, noisy and uncertain [8].

In this research study, we take a real dataset of cardiac patient's by Armed Forces Institute of Cardiology (AFIC), Pakistan. We manually extract the important attribute of the unstructured dataset and propose a classification model using data mining, classification algorithms Support Vector Machine (SVM). We classify cardiac patients according to their conditions into four important classes as given below.

- Class 1: Patients with risk of HF
- Class 2: Patients having no HF symptoms, but structural heart irregularities
- Class 3: Patients having HF symptoms and structural heart irregularities
- Class 4: Patients with critical stage of HF

This study will present a detailed update on modern techniques in the management and diagnosis of heart failure by classes 1 to 4 that have to appear in the past few years. On behalf of various cardiac studies, we also present a treatment plan for patients belonging to different classes of our proposed model. This treatment plan will be very helpful for patients as well as for medical researchers and cardiologist to overcome the problem of each class separately. It will also focus on recent research results and strategies that may give the positive impact on clinical practice.

Section 2 contains the related research studies by different researchers of the same domain. In Sect. 3, we discuss our proposed classification model in detail. Section 4 contains the conclusion, which provides the overall summary of our research work.

2 Literature Review

Predictive modeling of cardiac disease using electronic health record (EHR) data has become a very broad research area. The reason behind this is that HF has become the main cause of death for adults [9]. There are many machine learning strategies available for classification, such as Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), Naïve Bayes (NB), Decision Trees (DT) and much more. In [10], the authors take data from the National Health and Nutrition Examination Survey (NHANES) and applied SVM for classification of diabetes patients and find Area under the Curve (AUC) of 83%. RF was applied by [11] for prediction the chances of depression due to Traumatic Brain Injury (TBI) identification. Authors of [12] take a dataset from EHR propose a model for detection of HF within the time period of 6 months before the real heart failure occurrence. They also provide the performance comparison of SVM, LR, and Boosting.

Data mining, classification strategies have being used for identification and prevention of cardiac diseases. In [13], the authors present a performance comparison on behalf of accuracy for ANN, SVM, DT, and RIPPER techniques. The results show that SVM with an accuracy of 84% was the best technique with all of these. An isolated cardiac detecting system was introduced for prevention of HF by [14], by using mobile gateways. This system extracted the highly related features and then applied SVM classifier and finds the accuracy of 87.5%. Authors of [15] take a real dataset from VA

Medical Center, California and provide the performance comparison of DT, ANN, and LR for prediction of cardiac disease and ANN show the highest accuracy. In [16] authors proposed a model to accept different strategies of machine learning to handle concealed dataset. They applied their model for predicting the repetition of cancer and give the performance comparison of DT, Cox regression, and NB. [17] Provided a prediction model to show HF patient's survival risk by using some common classification algorithms such as SVM, RF, DT, and LR. The results of their study show that LR provides the highest accuracy.

The authors of [21] present a classification system called Clinical Decision Support System (CDSS) for diagnosis purpose of cardiovascular disease by using four special classification methodologies (ANN, BN, SVM, and DT). Their system checks the disease level with an accuracy of more than 94%. An Intelligent Heart Disease Predication System (IHDPS) was presented by [22], used to extract hidden information and their relation with HF from a huge dataset of cardiac patients. This is a hybrid system created by three common data mining strategies: NB, DT, and ANN. They concluded that NB creates a more effective prediction. Another HF predictive system called "Intelligent and Effective Heart Attack Prediction System" (IEHAPS) was created by [23]. They use different methodologies of prediction: ANN, frequent pattern mining, and clustering. Maximum Frequent Itemset Algorithm (MAFIA) technique was used to filter most significant patterns and finally, ANN was trained by these patterns for prediction of heart failure in a very efficient manner.

3 HF Diagnosis and Prediction Model

There are many diseases and several other interrelated factors that may cause of HF for a normal person. That's why HF is a very heterogeneous disease and detection and prediction of this disease also a very tough job. So, data mining has introduced many algorithms that are being used to develop intelligent prediction models for physicians and medical practitioners which increase the accuracy of diagnosis of HF. In this study, we propose a data mining, classification model using real data of cardiac patients. Figure 1 shows the architectural view of our proposed model. We describe our model in different phases in detail as discussed below.

3.1 Data Preparation

We take raw data of 500 cardiac patients in the form of their medical reports from AFIC, Pakistan. We manually extract useful 32 features from these reports with the close collaboration of cardiologists and medical practitioners. We create a better understanding of these patient's medical reports with the help of cardiac specialists and make sure that these features are enough to get valuable results for our model. This approach provides a deep knowledge of cardiology and helps to understand the domain of the problem. These extracted features were stored in MS Excel to create a database. To make our data structured, we applied machine learning algorithms. Such as, we applied mapping table for transforming textual data into numeric data.

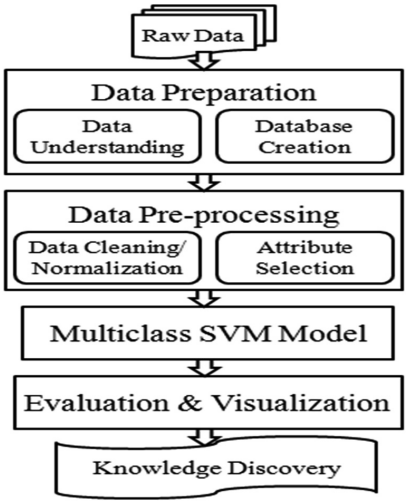

Fig. 1. Basic flow of proposed approach

3.2 Data Pre-processing

This phase includes several sub-processes such as data cleaning, data reduction, and data transformation. First, we handle identical, missing and inconsistent values in the dataset for cleaning purpose by removing and replacing with correct values. Finally, when we fed this dataset into database of Rapid Miner tool, it again cleans the dataset and replaced missing values with average value of that attribute by using an operator called "Replace Missing Value". "Normalize" operator was applied to cleaned dataset for standardization of data. This operator normalizes the attribute values of the selected attributes. Some important selected features are given in Fig. 2. By applying this strategy, we reduced the complexity of our dataset and it helps to develop a classification model with the highest accuracy [18].

3.3 Multi-class SVM Classification Modeling

Data was uploaded in Rapid Miner to develop SVM model. As SVM deal only with binary data, where our dataset contains multi-class data, so we applied "Class-Binarization" techniques to transform multi-class data into binary class data [19]. The dataset was managed into four subsets having individual class labels. Now we have four different classes' having the same number of attributes, but the diverse number of patients. SVM operator randomly takes 70% of the dataset for training. Now put the unseen 30% testing data into a trained model by "Apply Model" operator and find some valuable performance measures by applying "Performance" operator. It shows the results for each class and four separate models were created respectively. Our resulting attribute was "Result Class" and its categories are; Class 1, Class 2, Class 3 and Class 4.

Patient-ID	Diabetes	HeartRate-BI_BPM
Age	Atrial Fibrillation	HeartRate-MA_BPM
Gender	Hyperlipidemia	BP-BI_mmHg-uppLim
Smoking History	Chronic Kidney Disease	BP-BI_mmHg-lowLim
Family History	Ischemic Heart Disease	ACE inhibitor
BMI	Pulmonary	Beta-Blocker
Hypertension	Hemoglobin (g/dL)	STATIN user
Cataract	Sodium(mEq/L)	Diuretic user
Anemia	Cholestrol(mg/dL)	LVEF
Rheumatoid Arthritis	Lymphocytes(*10 (9)/L)	Reported Class

Fig. 2. Selected attributes in proposed model

3.4 Result Analysis

Different classification measures were used such as precision, classification error, sensitivity, specificity, F-measure, AUC, and accuracy to get the overall result of our model. Each class was evaluated individually and finally got overall results as shown in Table 1. All these resulting attributes are independent of each other, so higher the value of these attributes give the best performance of our prediction model. The result shows that a class having the highest value of accuracy will be the least value of precision and vice versa, as explained in [20]. By calculating the overall average result of all four classes we find the accuracy of the SVM model of 82%. As our dataset is very heterogeneous and have higher dimensional space, so we prefer SVM to other state-of-the-art classification models. SVM also gives better results for text classification.

Table 1. Accuracy measures for SVM models

SVM model	Precision (%)	F-Measure (%)	Sensitivity (%)	Classifi. error (%)	AUC (%)	Accuracy (%)
Class 1	81.97	88.5	96.15	8.67	96.8	91.33
Class 2	100	3.17	1.61	40.67	51.6	59.33
Class 3	100	8.7	4.55	14	73.8	86
Class 4	91.28	95.44	100	8.67	84.2	91.33
Average	93.3	68.53	50.58	18	76.6	81.99

3.5 Knowledge Discovery

On behalf of various cardiac studies and results of our model, we create an important treatment plan, explain in Fig. 3. Patients of class 1 are very common in our society because they ignore the risk of HF even they are already under attack of hypertension,

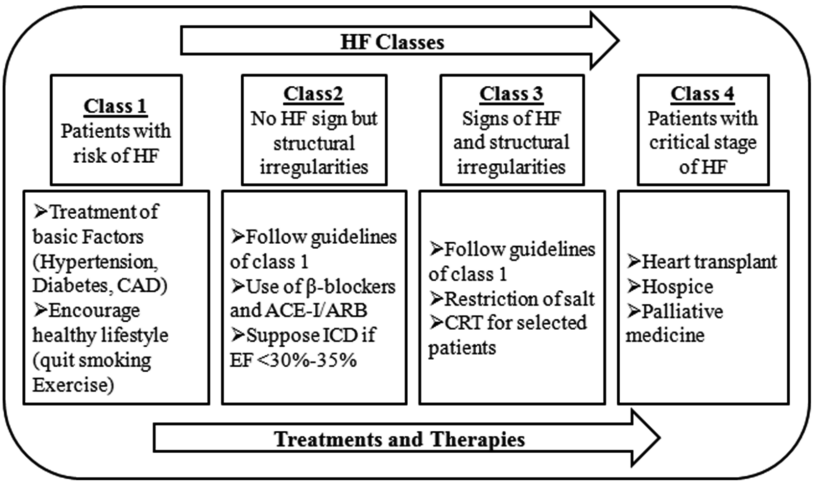

Fig. 3. Classes in development of HF and their suggested treatment

and diabetes. So patients with these diseases should never forget the risk of HF. They should promote their lifestyle to lose their weight and to quit smoking. Class 2 contains the patients that have some structural irregularities in the cardiac system and have more chances of HF, so they should be more careful about it. They should follow all instructions for class 1 patients and also should use β-blockers if they have reduced EF.

Patients of class 3 have symptoms of HF and they should prefer the use of diuretics. Patients of class 4 are in the most critical form of HF. These patients have preferred to use of palliative medicine or they are treated with heart transplants. So, this research study can be very helpful for cardiologists to treat the cardiac disease very efficiently.

Some important result of our study can be defined as:

- *Our dataset has 69% male patients, which conclude that the chances of HF are more in males than females.*
- *Dataset contains 73% of patients having more than 50 years of age, which means adults are more affected by this disease.*
- *We used SVM, a data mining algorithm to propose a classification model that classifies our data into 4 classes, which are very important for treatment of HF.*

4 Conclusion

This research study proposes a framework, in which we used a data set from AFIC, Pakistan. After applying all the necessary preprocessing steps of data mining we applied SVM, to classify our dataset into 4 important classes. Our proposed classification model gave the accuracy and AUC of 82% and 77% respectively. Our model with its excellent result is very helpful for medical practitioners to understand the causes of HF and they can make intelligent decisions to control the conditions of

patients on behalf of these results. We also propose a treatment plan of cardiac patients belonging to different classes of our proposed model. This flowchart is very helpful for medical practitioners as well as for patients, because by following this plan they can treat cardiac disease in the best way. Patients of each class should make sure that they are not moving toward the next more dangerous class and this study will help to do this.

References

1. McCullough, P.A., Philbin, E.F., Spertus, J.A.: Confirmation of a heart failure epidemic: findings from the resource utilization among congestive heart failure (REACH) study. JACC **39**, 60–69 (2002)
2. McMurray, J.J., Adamopoulos, S., Anker, S.D.: ESC guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: the task force for the diagnosis and treatment of acute and chronic heart failure 2012 of the European society of cardiology. EURJHF **14**, 803–869 (2012)
3. Ouwerkerk, W., Adriaan, A.V., Zwinderman, A.H.: Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. JACC. Heart Fail **2**, 429–436 (2014)
4. Gerber, Y., Weston, S.A., Redfiled, M.M., Chamberlain, A.M., Manemann, S.M., Killian, J. M., Roger, V.L.: A contemporary appraisal of the heart failure epidemic in Olmsted county, Minnesota, 2000 to 2010. JAMA Intern. Med. **175**, 996 (2015)
5. Tan, G., Cbye, H.: Data mining applications in healthcare. J. Healthcare Inf. Manage. **19**, 64 (2004)
6. Giudici, P.: Applied Data Mining: Statistical Methods for Business and Industry. Wiley, New York (2003)
7. Taslimitehrani, V., Dong, G.: Developing EHR-driven heart failure risk prediction models using CPXR (Log) with the probabilistic loss function. J. Biomed. Inform. **60**, 260–269 (2016)
8. Cios, K.J., Moore, G.W.: Uniqueness of medical data mining. Intell Med. **26**, 1–24 (2002)
9. Health World Organization: The top 10 causes of death. http://www.who.int/mediacentre/factsheets/fs310/en/index.html
10. Wei, Y., Liu, T., Valdez, R., Gwinn, M., Khoury, M.J.: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med. Inform. Decis. Mak. **10**, 16 (2010)
11. Kennedy, R.E., Livingston, L., Riddick, A., Marwitz, J.H., Kreutzer, J.S., Zasler, N.D.: Evaluation of the neurobehavioral functioning inventory as a depression screening tool after traumatic brain injury. Jorn. Head Trauma Rehab. **20**, 512–526 (2005)
12. Wu, J., Roy, J., Stewart, W.F.: Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med. Care **48**, 106–113 (2010)
13. Kumari, M., Godara, S.: Comparative study of data mining classification methods in cardiovascular disease prediction. IJCST **2** (2009)
14. Kwon, K., Hwang, H., Kang, H., Woo, K.G., Shim, K.: A remote cardiac monitoring system for preventive care. In: Proceedings of ICCE, pp. 197–200. IEEE (2013)
15. Kurt, I., Ture, M., Kurum, A.T.: Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst. Appl. **34**, 366–374 (2008)

16. Zupan, B., Demsar, J., Kattan, M.W., Beck, J.R., Bratko, I.: Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artif. Intell. Med. **20**, 59–75 (2000)
17. Panahiazar, M., Taslimitehrani, V., Pereira, N., Pathak, J.: Using EHRs and machine learning for heart failure survival analysis. Study Health Technol. Inform. Med. **216**, 40–44 (2015)
18. Deekshatulu, B.L., Chandra, P.: Classification of heart disease using artificial neural network and feature subset selection. Global J. Comput. Sci. Technol. **13** (2013)
19. Fürnkranz, J.: Round Robin classification. J. Mac. Learn. Res. **2**, 721–747 (2002)
20. http://www.ncsu.edu/labwrite/Experimental%20Design/accuracyprecision.htm
21. Hong, J., Kim, S., Zhang, B.: AptaCDSS-E: a classifier ensemble-based clinical decision support system for cardiovascular disease level prediction. Expert Syst. Appl. **34**, 2465 (2008)
22. Awang, R., Palaniappan, S.: Intelligent heart disease predication system using data mining technique. Int. J. Comput. Sci. Netw. Secur. **8** (2008)
23. Patil, S., Kumaraswamy, Y.: Intelligent and effective heart attack prediction system using data mining and artificial neural network. Eur. J. Sci. Res. **31** (2009)