A Framework for Classifying Unstructured Data of Cardiac Patients: A Supervised Learning Approach

Iqra Basharat
Department of Computer Engineering
National University of Sciences and Technology
Islamabad, Pakistan

Ali Raza Anjum
Former Specialist - Business Analysis and Planning
Mobilink
Islamabad, Pakistan

Mamuna Fatima, Usman Qamar, Shoab Ahmed Khan
Department of Computer Engineering
National University of Sciences and Technology
Islamabad, Pakistan

***Abstract*—Data mining has recently emerged as an important field that helps in extracting useful knowledge from the huge amount of unstructured and apparently un-useful data. Data mining in health organization has highest potential in this area for mining the unknown patterns in the datasets and disease prediction. The amount of work done for cardiovascular patients in Pakistan is scarcely very less. In this research study, using classification approach of machine learning, we have proposed a framework to classify unstructured data of cardiac patients of the Armed Forces Institute of Cardiology (AFIC), Pakistan to four important classes. The focus of this study is to structure the unstructured medical data/reports manually, as there was no structured database available for the specific data under study. Multi-nominal Logistic Regression (LR) is used to perform multi-class classification and 10-fold cross validation is used to validate the classification models, in order to analyze the results and the performance of Logistic Regression models. The performance-measuring criterion that is used includes precision, f-measure, sensitivity, specificity, classification error, area under the curve and accuracy. This study will provide a road map for future research in the field of Bioinformatics in Pakistan.**

Keywords—bioinformatics; classification techniques; heart disease in Pakistan; heart disease prediction; multinomial classification; logistic regression

I. INTRODUCTION

Cardiac disease is one of the most serious and death causing health problem. This disease has its bad effects not only on older people but affected severely younger generation also. There are some significant risk elements of cardiac disorder like excessive amount of cholesterol, high blood pressure, hypertension, smoking and sometimes family history. Various precautionary mediations for individuals are available, involving either prescription or change in daily life routine. There exists raw data in the form of patients' history and complex reports. All these resources are key factors to extract meaningful results, for better medical diagnosis. This data can be processed and analyzed to extract valuable in-formation that helps practitioner in decision-making and cost saving. Medicine or Bioinformatics technology has flourished a lot in the past few years that we have applied in healthcare industry such as; treatment effectiveness, customer relationship management, healthcare information management fraud and abuse. However, the significant applications in data processing perhaps implicate predictive modeling [1].

Data mining enacts a substantial part in various applications of different domains such as business corporation's, e-commerce, healthcare industry, science and engineering. In the health care industry, primarily it is used to predict disease. Various diseases like Diabetes, Hepatitis, Cancer and are diagnosed using data mining or machine learning techniques [2]. According to Benko and Wilson in healthcare organizations where data mining is being practiced are performing better in meeting their long term needs. Benko and Wilson argue [3] "In healthcare, data mining is becoming increasingly popular, if not increasingly essential."

To assist the medical practitioners, intelligent information system, knowledge based system and prediction systems are being developed. Healthcare organizations have kept voluminous data of patients in the form of medical reports, patients' history, electronic test results etc. [4]. This data in its present unstructured form is complex, noisy, high dimensional and discrete [5]. Considerably there is a lot of useful knowledge buried in those records. However, the question arises that how can we mine and transform those unstructured and complex reports into practically useful information, that could assist the doctors to draw knowledgeable medical conclusions.

A. Research Motivation

The main motivation of this research is to propose a framework to extract the hidden valuable information from the unstructured records of patients in the form of medical reports and to classify the data into important classes or patients' impressions that could assist the healthcare experts to make intelligent decisions and for predictive analysis.

B. Objectives and Contribution

This study is designed at predictive analysis/classification of cardiac patients by proposing a classification framework for un-structured data. We propose to use a classification

framework that emphasize on pre-processing of unstructured data of healthcare organization in the form of patients' reports. Hence, contributions of this research study are as follows:

1) We present a manual approach that extracts attributes (like age, sex, blood pressure, LVEF value, BMI, defected areas, etc.) from patients' reports in order to classify the patient's condition. This study aims to provide a framework that uses supervised learning techniques of data mining.

2) We have used the multinomial class label (namely, fair, moderate, risk and critical).

3) A classification framework is proposed that uses supervised learning techniques to classify unstructured of data patients into four classes mentioned above.

4) We use best-known machine learning technique, Logistic Regression, which is most widely being used in prediction of heart disease.

5) Comparison performance evaluation is presented based on some performance measures are explained later section 3.

II. Literature Review

Cardiac syndrome is a serious and a death-causing syndrome [6]. However, the science and Bioinformatics has developed a lot and treatment for this disease is possible and available to almost every person. The increase in no of deaths occurring because of cardiac disease all around the world has focused the attention of medical practitioners and researcher on this serious issue. There is quite a comprehensive literature available on this topic. Various useful medical applications and decision support system have been advanced to aid the medical practitioner in better medical treatment of their patients. . These systems predict the likelihood of patients getting heart disease or heart attack, etc. data mining has played an important role in this field. Here, patients historical data is used to make and develop such system where artificial intelligence and machine learning techniques are used widely. The ongoing research in this field has provided much success and opened up the doors for further improvements.

It has been noticed from the detailed survey of the literature that SVM, Logistic regression, Neural Nets and Naïve Bayes, are most widely use algorithms for heart disease prediction. To predict the survival of cardiac patients, three prediction models were built on 1000 cases of cardiac heart disease patients. Using a binary categorical variable (1 for survival and 0 for non-survival), 10-fold cross validation procedure was performed on SVM, ANN and Decision trees. This gives less biased prediction and highest classification accuracy in SVM [6].

To identify and prevent the cardiovascular disease, classification techniques are used. Authors of [7] present comparison of Artificial Neural Network, Decision Tress and RIPPER and SVM techniques. Based on accuracy measures these techniques were compared. The results of this study show that the Support Vector Machine model is the best giving 84% accuracy. Ripper, a classification algorithm based on association rules with reduced error pruning algorithm gives 81% accuracy, whereas Decision tree gives the least accuracy and sensitivity ratio among all three-classification models.

Literature reports various comparative studies on data mining, classification algorithms for predicting heart disease. One such comparative analysis was carried out using the Cleveland cardiovascular disease dataset from UCI repository with 13 attributes and 303 instances [8]. On this dataset three classification models were developed, namely, Sequential Minimal Optimization, Logistic Function and Multilayer Perceptron Function. The Accuracy of these classification techniques was determined through kappa statistics, ROC, True positive rate and F measures. All these accuracy measures show that logistic regression gives better results than other techniques. The different error rates calculated shows that the Logistic Function algorithm performs much better than other two classification algorithms. The rate of true positives and ROC Area of the point reaches the maximum accurateness in the logistic function algorithm. Even Kappa statistics and F measures give better results in the logistic function than SMO and Multilayer perceptron.

Using physiological measuring devices like Point-of-Care devices (PoC), mobile gateways and monitoring server, a remote cardiac monitoring system was designed for preventive care [9]. The system was developed to provide preventive care services to cardiac patients. By calculating the information gain of features, highly related feature subsets were selected and SVM classifier was applied to them. The proposed prediction algorithm gives 87.5 % accuracy. F. Imran Kurt et al uses a real data set from VA Medical Center from Long Beach, California and compare performances of logistic regression, decision tree, and neural networks for predicting coronary artery disease [15]. Lift charts and error rates were used to compare the performances of these classification models. Prediction of coronary artery diseases by Neural Networks yields excellent results as it gives higher accuracy while classifying the data. Logistic Regression was found to be second most accurate classed whereas, decision trees show least accuracy and highest error rate.

A prototype of intelligent heart disease prediction system (IHDPS) [10] was developed by using Naïve Bayes, Decision Trees and Neural Network. this system is capable of mining unknown patterns and associations correlated to cardiac disease. IHDPS assists medical practitioners to make intelligent decisions as it can give response to simple as well as complex' what if' queries. Further, it delivers operative and inexpensive treatment and enhances visualization and develop better understanding. IHDPS has used the CRISP-DM methodology to make three models (Naïve Bayes, Decision Trees and Neural Network) and Data Mining Extension language is used to create, train, predict and access model content. Classification Matrix and Lift Chart methods are used to check which model gave a maximum percentage of right predictions. In this research study, five mining goals were set and assessed with respect to three trained models. These are:

- Predict those patients who have chances to get heart disease based on their medical profile.

- Find out the important influences and relationships between medical inputs and medical attributes related to the predictable state.
- Find out heart patient characteristics.
- Define attribute values that discriminate nodes favoring and disfavoring the predictable conditions.

Naïve Bayes appears most efficient by answering four out of the five goals and by identifying all important medical predictors; Decision Trees answered three and its results are easier to interpret; Neural Network two and in this the correlation between attributes is hard to interpret. Intelligent heart disease prediction system is constructed using 15 features, in categorical sample data of 909 patients. The authors of the paper suggested that more features and techniques like association rules, clustering and time series can be used by prolonging it.

TABLE I. COMPARATIVE ANALYSIS OF CLASSIFICATION TECHNIQUES USED IN LITERATURE

Ref #	Data Set (Real/ Artificial)	Classification Type (Categorical/ multinomial)	Tools	Techniques						Results
				ANN	LR	DT	NB	SVM	Other	
7	Artificial	Binomial categorical	WEKA	✓		✓		✓	✓	SVM 84% accuracy
8	Artificial	Binomial categorical	WEKA	✓	✓				✓	LR is considered to be best
6	Real	Binomial categorical		✓		✓		✓		SVM achieved highest accuracy
10	Artificial	Binomial categorical		✓		✓	✓			NB gives highest accuracy

Another widely used machine learning technique, Neural Network, is used to predict heart disease, BP and Diabetics with the help of. Their dataset consists of 78 records with 13 input variables on which various experiments are conducted and the system is trained. For heart disease diagnosis, the author has suggested supervised network, which is trained using a Back Propagation algorithm. When the system is input with new data, it will find it from the trained data and generate a list of probable diseases that may be occurred by the patient. The success rate of the system to give the desired output from the inaccurate input is 100%. The results of the study show that a neural network has the extreme capability to be used as an indexing function. It is used for modeling and prediction of experimental data, this is a fast substitute to classical statistical techniques. The system can avoid human error. Thus, the system is reliable and assists medical practitioners in making accurate decisions. Certainly, neural networks cannot replace the expert mind since the expert is more reliable, but it can assist human experts by cross checking their diagnosis. Table 1 shows the comparative analysis of the research carried out in the field of classification of cardiac data.

The literature discussed above reveals that artificial neural networks (ANN) and Decision trees most widely used classification algorithms for categorical data and Logistic regression is also being used widely for prediction. There are several intelligent heart disease prediction systems, which uses different approaches and propose various models implementing Naïve Bayes and ANN.

In countries like China, India and Malaysia and in some European countries much work has been done in medical data mining and specifically in cardiac data mining [11] [12] [13] [14] on the basis of real and artificial data sets.

All the research discussed above is based on either of these countries. Besides, most of the medical data mining work discussed above focuses on either clustering, classification or association rules mining while in some [11] [12] the details of the results are not discussed and visualized properly.

In 2004, a community-based survey was carried out in an ur-ban populated major city of Pakistan (Sunita Dodani et al.). This survey-based study was estimation of the occurrences and aware-ness of different risks associated with coronary heart disease (CHD). The notable risk factor of heart disease was hypertension, obesity and a sedentary lifestyle. Lack of awareness among people was a common reason of negligence about health. The statistical results of high prevalence factors of cardiac heart disease are well presented and it was suggested to develop some guidelines to manage the coronary heart disease. However, the authors of the study did not formulate any guidelines or model to manage and prevent the increasing risk factors of this serious death causing disease.

However, there is little research seen in Pakistan in building a framework specifically for cardiac data mining based on real data obtained from some renowned cardiac hospital. In addition, there is a need of framework that unifies the data mining tasks from data preparation to data visualization and the discovery of knowledge. In this work, analysis is based on the results of machine learning techniques like clustering, correlation and logistic regression to better and complete visualization of results.

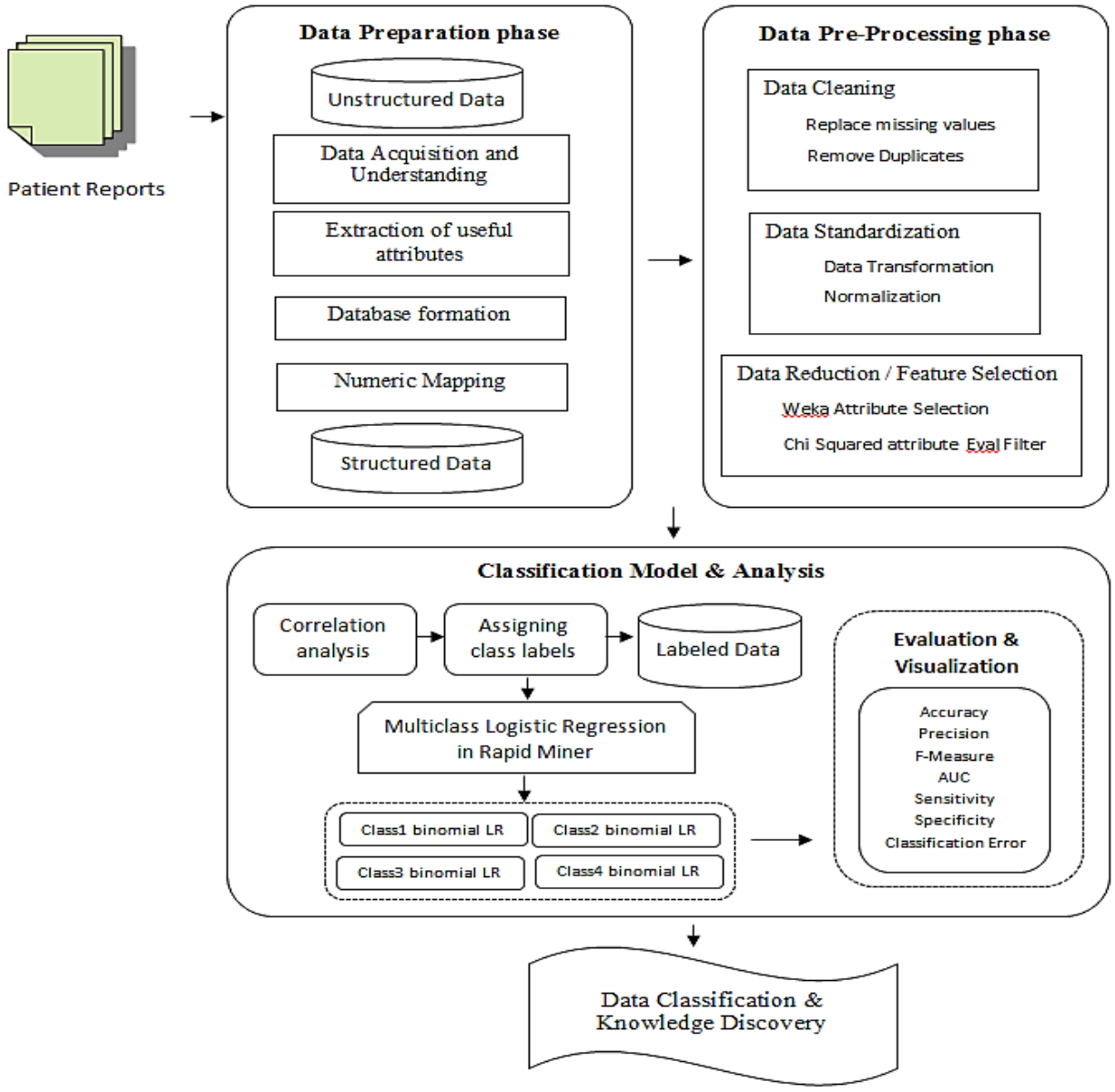


Fig. 1. Proposed Methodological Framework to Classify Unstructured Data

III. Proposed Framework

In this research study of heart disease data for predicting heart patients' condition, following a well-known CRISP-DM methodology we proposed a framework. It is used to classify the unstructured data of cardiac patients. The unstructured reports of cardiac patients were preprocessed manually and a structured database of patients' records was developed. An architectural model of our proposed classification approach is illustrated in figure 1.

It consists of four phases.

A. Data preparation

B. Data preprocessing

C. Feature selection

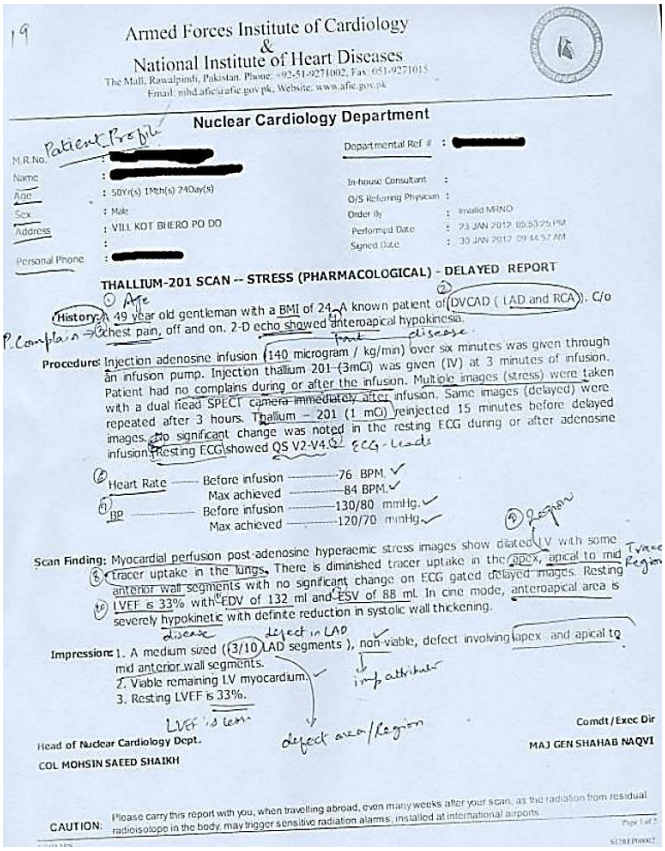
D. Classification model

A. Data preparation

The Data Preparation was carried out in five steps as discussed below.

1) Data Acquisition and Data Understanding

This research study was carried out in close collaboration with the Armed Forces Institute of Cardiology (AFIC), Pakistan. Previously, data used in this type of research study are mostly taken from an online data repository and different classification algorithms are applied on that. In this research, a real data set from AFIC was used. This collected data was unstructured historical records of 1500 patients. During manual preprocessing, data was transformed from unstructured reports to structured format and 50 plus attributes/features were identified. Figure 2 shows a sample patient's report from which attributes/features were extracted.



Armed Forces Institute of Cardiology
&
National Institute of Heart Diseases
The Mall, Rawalpindi, Pakistan. Phone: +92-51-9271002, Fax: 051-9271015
Email: nihd.afic@afic.gov.pk, Website: www.afic.gov.pk

Nuclear Cardiology Department

M.R.No. : [redacted]
Name : [redacted]
Age : 50Yr(s) 1Mth(s) 24Day(s)
Sex : Male
Address : VILL KOT BHERO PO DO
Personal Phone : [redacted]

Departmental Ref # : [redacted]
In-house Consultant :
O/S Referring Physician :
Order By : Invalid MRNO
Performed Date : 23 JAN 2012 05:53:25 PM
Signed Date : 30 JAN 2012 09:44:57 AM

THALLIUM-201 SCAN -- STRESS (PHARMACOLOGICAL) - DELAYED REPORT

History: A 49 year old gentleman with a BMI of 24. A known patient of (DVCAD (LAD and RCA). C/o chest pain, off and on. 2-D echo showed anteroapical hypokinesia.

Procedure: Injection adenosine infusion (140 microgram / kg/min) over six minutes was given through an infusion pump. Injection thallium 201 (3mCi) was given (IV) at 3 minutes of infusion. Patient had no complains during or after the infusion. Multiple images (stress) were taken with a dual head SPECT camera immediately after infusion. Same images (delayed) were repeated after 3 hours. Thallium – 201 (1 mCi) reinjected 15 minutes before delayed images. No significant change was noted in the resting ECG during or after adenosine infusion. Resting ECG showed QS V2-V4.

Heart Rate — Before infusion ——— 76 BPM.
Max achieved ——— 84 BPM.
BP — Before infusion ——— 130/80 mmHg.
Max achieved ——— 120/70 mmHg.

Scan Finding: Myocardial perfusion post-adenosine hyperaemic stress images show dilated LV with some tracer uptake in the lungs. There is diminished tracer uptake in the apex, apical to mid anterior wall segments with no significant change on ECG gated delayed images. Resting LVEF is 33% with EDV of 132 ml and ESV of 88 ml. In cine mode, anteroapical area is severely hypokinetic with definite reduction in systolic wall thickening.

Impression: 1. A medium sized (3/10 LAD segments), non-viable, defect involving apex and apical to mid anterior wall segments.
2. Viable remaining LV myocardium.
3. Resting LVEF is 33%.

Head of Nuclear Cardiology Dept.
COL MOHSIN SAEED SHAIKH

Comdt / Exec Dir
MAJ GEN SHAHAB NAQVI

CAUTION: Please carry this report with you, when travelling abroad, even many weeks after your scan, as the radiation from residual radioisotope in the body, may trigger sensitive radiation alarms, installed at international airports.

Fig. 2. Cardiac Patient's Report

2) Extraction of Useful Data and Collection of Attributes

We developed a thoughtful understanding of medical reports of patients with the help of medical practitioners and cardiac spe-cialists. The stop criterion was used to determine whether the extracted attributes are mature enough and comprises all the im-portant useful attributes to proceed further or still more reports needs to be filtered out to get valuable attributes. This manual extraction of attributes is a key advantage to get the insight on the problem domain and gives a deep understanding of cardiology.

3) Database Formation

Using MS SQL, database was created to store the extracted attributes after the information retrieval and the structuring of the collected data that was extracted from the past patient reports.

4) Mapping Data to Numeric Values

Once the patients' records are moved into the database, it was quiet easy to handle inconsistency in the data. Machine learning algorithms are applied to structured data. We have used mapping tables in MS SQL Server to transform textual data to numeric data compatible with machine learning algorithms. Once we have prepared the data into structured format and created a database, we move towards the next phase, data preprocessing.

B. Data Preprocessing

The data-preprocessing phase includes data cleansing, data transformation and data reduction.

C. Data Cleansing

In the process of data cleansing, missing, identical and inconsistent data are handled. When the data is in unstructured form, it is manually cleaned and identical and missing records are removed and replaced. All data preprocessing steps are recursive in nature, they are performed in cycles or iteratively. When the data was fed into the database, it was again cleansed with the help of Rapid Miner tool. Missing values were replaced by average value of the attributes with the help of 'Replace Missing Values' operator.

D. Data Transformation

With Rapid Miner 5, a powerful tool of data mining, data was normalized using its 'Normalize' operator. Data normalization standardized the data to a range of 0 to 1 to avoid the attributes with greater values from misbalancing the attributes with smaller values in the assessment procedure [17]. Once the data is gone through the preprocessing, featured selection technique was applied in WEKA to select the important features.

1) Data Reduction

In this step, we remove the irrelevant and redundant data.

E. Feature Selection

Initially the total number of features extracted from the patients' report was 53. Some of the features were irrelevant and redundant that needs to be removed, thus we have used feature subset selection technique. This is a common preprocessing step used in machine learning. Feature subset selection is a step in data preprocessing that helps in reducing the dimensionality and irrelevant data that further enhances the learning efficiency, increases analytical accuracy of classification models The resulted feature subset states those features that are useful in predicting class and thus produces higher classification accuracy [18] [19].

Weka 3.6.9, data mining tool was used to determine important and relevant feature subset using weka.attributeSelection.ChiSquaredAttributeEval technique. It evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. It uses weka.attributeSelection.Ranker search method with threshold 0.5. This search method ranks attributes by their individual evaluations. Figure 3 shows the top 36 selected attributes that are effective in classification accuracy.

Patient_ID	2DEchoResult_Part1	BP-MA_mmHg-uppLim
Age	2DEchoResult_Part2	BP-MA_mmHg-lowLim
Gender	2DEchoResult_Disease1	AffectedArea1
Protocol	2DEchoResult_Disease2	AffectedArea2
BMI	P_Complain1	AffectedArea3
Known_Disease1	P_Complain2	LV_Myocardium
Known_Disease2	P_Complain3	Defect_Size
Known_Disease3	RestingECGResult1	Defected_AreaSize
FstMI_Type	HeartRate-BI_BPM	Defect_Segment
Angiography_Result1	HeartRate-MA_BPM	Via/Non-via
2DEchoResult	BP-BI_mmHg-uppLim	IsDefected
2DEchoResult_Part	BP-BI_mmHg-lowLim	I_LVEF

Fig. 3. Useful Attributes Selected through Attribute Selection Technique

The modeling tools used to carry out this research work is Rapid Miner® 5.3, Microsoft SQL Server R2 2008, WEKA 3.6.9 and Microsoft Excel 2010. except at the beginning of a sentence: "Equation (1) is . . ."

IV. MODELING

Now when we have already gone through feature extraction, pre-processing and finally classification steps, this section will focus on modeling correlation, clustering and prediction model.

A. *Correlation -- Modeling*

The correlation coefficient measures the linear relationship between two attributes or columns of data. The value can range from -1 to +1 [20].

Correlation is the association between different attributes. It is a statistical measure, widely being used in data mining for finding the relationship between attributes. We have used The 'Correlation Matrix Operator' of Rapid Miner to calculate the correlation among all the attributes of our data set.

After running the process, the correlation matrix was generated as shown in Figure 4. The results of correlation show that there are no two attributes that can be statistically correlated to each other. Some attributes have a strong positive correlation whereas some have weak correlation. For example, the LVEF is an important attribute that has 0.733 weights. The attributes Via/Non-via and I_LVEF has -0.123 correlations that show the negative correlation with each other when the value of one increases, the second value decreases. In actuality, the low LVEF value indicates a critical or risky situation of cardiac patients and Via/Non-via has value 0 and 1. The 0 value indicated viable condition and 1 indicated non-viable. Similarly, the value of LVEF is negatively correlated with Age and IsDefected attributes.

To label the data for applying the logistic regression model along with intelligent guesses and these weights and relationship between attributes, the clustering results and trends shown in work [21] helped us to assign class labels to each record.

Attributes	Patient_ID	Age	Gender	label	Protocol	BMI	Known_Dis...	Known_Dis...
Patient_ID	1	0.037	-0.153	-0.153	0.219	-0.117	-0.264	-0.083
Age	0.037	1	-0.066	-0.066	-0.251	-0.108	0.162	0.133
Gender	-0.153	-0.066	1	1	-0.066	0.105	-0.037	0.086
label	-0.153	-0.066	1	1	-0.066	0.105	-0.037	0.086
Protocol	0.219	-0.251	-0.066	-0.066	1	-0.119	-0.129	-0.104
BMI	-0.117	-0.108	0.105	0.105	-0.119	1	0.059	0.038
Known_Disease1	-0.264	0.162	-0.037	-0.037	-0.129	0.059	1	0.215
Known_Disease2	-0.083	0.133	0.086	0.086	-0.104	0.038	0.215	1
Known_Disease3	-0.098	0.013	-0.033	-0.033	-0.121	0.059	0.095	0.435
Prev_Scan	0.268	-0.217	-0.044	-0.044	0.421	-0.020	-0.099	-0.046
PScan_Result	0.293	-0.281	-0.051	-0.051	0.399	-0.055	-0.127	-0.083
Prev_Procedure	0.139	-0.004	-0.049	-0.049	0.284	0.034	-0.092	-0.050
FstMI_Type	-0.207	0.135	-0.075	-0.075	-0.229	-0.110	-0.034	-0.008
2ndMI_Type2	-0.105	-0.086	0.056	0.056	-0.079	0.025	-0.080	-0.051
Angiography_Result1	0.126	0.032	-0.322	-0.322	-0.011	-0.067	-0.278	-0.202
Angiography_Result2	0.084	-0.088	-0.096	-0.096	0.123	0.039	-0.100	-0.064
2DEcho_EF-per	-0.028	0.191	-0.082	-0.082	-0.034	0.055	0.038	0.031
2DEchoResult	0.012	0.026	0.010	0.010	-0.056	0.111	0.053	0.081
2DEchoResult_Part	0.102	-0.011	-0.079	-0.079	-0.047	-0.116	0.201	-0.061
2DEchoResult_Part1	0.074	0.103	-0.115	-0.115	-0.189	-0.024	-0.007	-0.020
2DEchoResult_Part2	-0.087	0.078	-0.055	-0.055	-0.056	-0.048	-0.056	-0.036
2DEchoResult_Disease	-0.057	0.122	-0.133	-0.133	-0.274	-0.063	-0.047	-0.116
2DEchoResult_Part3	0.041	0.035	-0.077	-0.077	-0.078	-0.085	-0.022	0.077
2DEchoResult_Disease	0.035	0.037	-0.077	-0.077	-0.079	-0.087	-0.028	0.066
P_Complain1	0.012	-0.041	0.203	0.203	0.021	0.043	-0.076	-0.134
P_Complain2	0.133	0.010	0.016	0.016	-0.041	0.166	-0.107	-0.109

Fig. 4. Correlation matrix generated through RapidMiner®. Correlation coefficients are visible

We studied the correlation results, consulted the experts' views, sorted out some rules manually, and based on those rules we classified the data into four classes and allocated each row a target class label. The four classes are, Normal: patient's condition is fair, means he is normal, Moderate: patient is having a moderate type heart disease, Risk: Patients condition is risky, means having serious heart disease and Critical: patient condition is critical.

B. *Logistic Regression --- Modeling*

Similar to linear regression, logistic regression is a extrapolative analysis, but logistic regression implicates the likelihood of a dichotomous contingent variable, whereas the predictors can be continuous or dichotomous [22].

1) *Data preparation*

The major preprocessing and data preparation was done in the previous section. The problem under discussion was multi-class classification problem. However, Logistic Regression deals with binary classification problems. To perform multiclass classification, there are some simple methods for transforming multi-class problems into a set of binary problems. These techniques are known as class-binarization techniques [23].

Using this approach, the data sheet was arranged in four sets with four different class label attributes (i.e. Normal, Risk, Moderate and Critical). Let say we have dataset A, B, C and D where dataset A is Normal / not normal, dataset B is Moderate / not moderate, dataset C is Risk / not risk and dataset D is Critical / not critical. After classifying the data into four categories in previous section it has been noted that this data was not balanced. Each class has different number of patients.

2) Process Description

After being done with data preprocessing tasks, Logistic Re-gression was applied on the dataset in Rapid Miner. The 'Retrieve' operator, load the data set to be trained. In order to label the class attribute we have used 'Set Role' operator to set the role of 'Report-Category' attribute as class label. The attribute is to be predicted by the model. As the 'Logistic Regression' operator takes the binomial label, so we have used 'Numerical to Binomial' operator and selected the Report Category attribute.

The Logistic Regression operator is used with 'dot' kernel type and complexity constant 0.5. It receives the training dataset in the input port and runs the algorithm for logistic regression. The logistic regression model is delivered at the output port. In order to use the learned/trained model on unseen data 'Apply Model' operator is used to calculate the performance of training model using 'Performance' operator that provides some important performance measures. The results of trained model are saved in the excel sheet using 'Write Excel' operator.

TABLE II. CONFUSION MATRIX OF FOUR LOGISTIC REGRESSION MODELS

Classes		True Negative	True Positive
Normal Class	Pred. False	775	138
	Pred. True	144	443
Moderate Class	Pred. False	514	229
	Pred. True	397	360
Risk Class	Pred. False	870	80
	Pred. True	467	83
Critical Class	Pred. False	1119	104
	Pred. True	151	126

One by one, the data sheets with four class attributes were loaded into Rapid Miner and four models were generated respectively. Our concerned attribute was, "Report-Category". This was the categorical attribute of a patient, summarizing his complete impressions as concluded by the cardiologist. The categories involved are; Normal, Moderate, Risk and Critical. Along with training the data the validation of the models was done using 10-fold cross validation. To analyze the performance of a classifier, confusion matrix is obtained using 'Performance' operator. Con-fusion matrix obtained from our four logistic regression models as shown in table 2. The confusion matrix helps, we can obtain accu-racy of model with the help of some important accuracy measures. We evaluated our models based on some performance measuring criteria discussed in next section.

V. RESULTS AND ANALYSIS

A. Performance Measuring Criteria

Extensive effort was made in order to obtain optimal results of classification by playing-around with different values of attributes used in modeling to run the model. Large numbers of the dataset and parameter alterations were carried out to reach optimal results against each model. Models that generated graceful-predictions against our stipulated 'performance criterion' are covered below.

1) Accuracy Measures

To classify the heart disease using logistic regression models, the elementary phenomenon used in calculating the performance and accuracy of the classifier. Sensitivity and Specificity are used for computing the accuracy. These Sensitivity and Specificity are obtained from a confusion matrix resulted from classification model. The confusion matrix displays the number of true positive, true negative and false positive, false negative assessments. It shows the comparison of actual values in the test dataset with the predicted values in the trained model. To measure the performance of logistic regression models we have used the performance measurement criteria [8]. The accuracy measures are shown in table 3.

TABLE III. ACCURACY MEASURES [8]

Accuracy Measures	Description
Precision	Precision is the proportion of relevant documents in the results returned.
F-measure	F Measure is a way of combining recall and precision scores into a single measure of performance.
Area Under Curve (AUC)	The AUC is an estimate of the probability that a classifier will rank a randomly chosen positive instance, higher than a randomly chosen negative instance
Sensitivity	TP/(TP + FN) (Number of true positive assessment)/(Number of all positive assessment)
Specificity	TN/(TN + FP) (Number of true negative assessment)/(Number of all negative assessment)
Accuracy	(TN + TP)/(TN+TP+FN+FP) (Number of correct assessments)/Number of all assessments)

B. Experimental Results and Discussion

The models were evaluated on performance criterion discussed above. To evaluate the unbalanced dataset, the more pertinent measures are precision, F-measure, AUC, sensitivity and specificity.

TABLE IV. ACCURACY MEASURES FOR LOGISTIC REGRESSION MODELS

Logistic Regression Models	Precision %	F-measure %	AUC	Classification Error %	Sensitivity %	Specificity %	Overall Accuracy %
Normal Class	66.11	75.86	0.890	18.80	76.01	81.80	81.20
Moderate Class	48.50	51.65	0.576	41.73	63.47	56.72	58.27
Risk Class	13.82	28.34	0.67	35.07	65.82	64.46	64.93
Critical Class	45.49	49.70	0.810	17.00	37.14	87.77	83.00
Average	43.48	51.38	0.73	28.15	60.61	72.68	71.85

There exists a commonality in these metrics, as they are all class independent. These are calculated by confusion matrix The confusion matrix is obtained for all three models and detailed accuracy is shown in the table 4. The area under the curve (AUC) is an independent metric. It gives equal weight to both classes and the greater the value of AUC the better the classifier performance is.

Similarly, for f-measure, precision, sensitivity and specificity, larger values show better performance. Hence, the logistic regression classifier for Normal Class gives better performance. To easily understand and depict the results of the experiment graphical representation of the results are presented in the following figure 5.

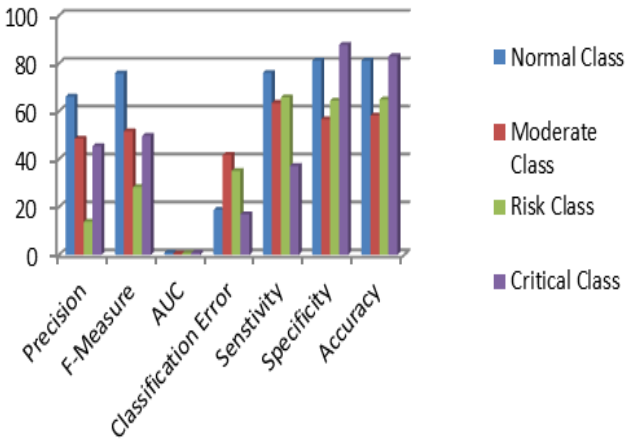


Fig. 5. Performance Measures of Four Logistic Regression Models

From table 4 it is noted that the critical class model performs best among all with 83% accuracy. Although it is showing least precision as compared to Normal and Moderate Class models, but it does not affect its accuracy. In the meantime, it has at least classification error. The second best model is Normal Class model with accuracy of 81.2 % and classification error of 8.8%. The risk Class model shows 64.93% accuracy with the lowest 35.07% classification error and least precision.

It is observed that models with high accuracy can have the least precision that shows precision is not dependent of accuracy. The model can be very precise, but inaccurate, as described above. Moreover, it can be accurate, but imprecise. Accuracy states the close relationship between measured value, standard or known value, and precision shows how close two or more measurements are to each other [24]. By computing the average accuracy of four models, we get to know the overall accuracy of logistic regression is 71.85%. However, here, we can see that for imbalanced dataset, accuracy measure is not an appropriate metric to evaluate the classifier. That is why we calculate other performance measures to better evaluate the performance of classifiers.

VI. CONCLUSION AND FUTURE WORK

The main motivation of this research was to propose a framework to extract the hidden valuable information from the unstructured medical reports of patients and to classify the data into important classes or patients' impressions that could assist the healthcare experts to make intelligent decisions and for predictive analysis.

The data set was taken from the Armed Forces Institute of Cardiology (AFIC), Pakistan. It was collected in unstructured form that was then preprocessed and maintained in a structured form in the database. We used Weka 3-6-4, a famous data-mining tool for classification. ChiSquaredAttributeEval technique was used to lessen the dimensionality and irrelevant data that increased the learning efficiency and analytical accuracy of classification models. Logistic Regression was used to predict four classes with the help of four models. To evaluate the performance of classifier models, accuracy measures were used. The aggregated results showed that when Logistic Regression was applied on four binomial models it gives a classification accuracy of 71.85 %.

The achievement of this research study is as follows:

- Useful data is extracted manually from unstructured records of patients.
- Database of heart patients is designed that could be further used for research and practical implementation of intelligent decision support systems.
- Logistic regression is used to classify the data into four classes of patients; Normal Class, Moderate Class, Fair Class and Critical Class.
- The proposed framework can be utilized for mining unstructured data in other health care centers

- The study on the whole gives new directions in the field of biomedical research in Pakistan.

The classification and predictive analysis of such data is of utmost importance nowadays. The future work of the research can be:

- These results can always be improved by improving the classification/prediction model. In the future, to improve the result by applying 'bootstrapping' technique that would balance the data and thus will give better results.
- Secondly, other important and better classification models like SVM and Artificial Neural Networks could be used to achieve high accuracy.
- The main future concern could be to design an inference engine for cardiac data that would assist the practitioner to make better decisions.
- The results of the study could be further improved by investigating other algorithms and by improving the data pre-processing techniques as well.
- The Text mining technique can be applied to mine the huge unstructured data in hospitals.
- Using the same data set to explore the reason, solution and precautionary measures of specific types of complaining diseases and problem occurring in specific type of patients.

ACKNOWLEDGMENT

We would like to thank and acknowledge the support and guidance of our supervisors. We are also thankful to Armed Forces Institute of Cardiology (AFIC), Pakistan for providing us the data set of heart patients and helping us to understand the research problem.

REFERENCES

[1] Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare, Journal of Healthcare Information Management—Vol, 19(2), 65

[2] Parpinelli, R. S., Lopes, H. S., & Freitas, A. A. (2001, July). An ant colony based system for data mining: applications to medical data. In Proceedings of the genetic and evolutionary computation conference (GECCO-2001) (pp. 791-797).

[3] Benko, A. & Wilson, B. (2003). Online decision support gives plans an edge. Managed Healthcare Executive, 13(5), 20.

[4] AbuKhousa, E., & Campbell, P. (2012, March). Predictive data mining to support clinical decisions: An overview of heart disease prediction systems. In Innovations in Information Technology (IIT), 2012 International Conference on (pp. 267-272). IEEE.

[5] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Applied statistics, 100-108..

[6] Xing, Y., Wang, J., Zhao, Z., & Gao, Y. (2007, November). Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In Convergence Information Technology, 2007. International Conference on (pp. 868-872). IEEE.

[7] Milan Kumari, Sunila Godara "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST Vol. 2, Issue 2, June

[8] Vijayarani, S., and S. Sudha. "Comparative Analysis of Classification Function Techniques for Heart Disease Prediction", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 3, May 2013

[9] Kwon, K., Hwang, H., Kang, H., Woo, K. G., & Shim, K. (2013, January). A remote cardiac monitoring system for preventive care. In Consumer Electronics (ICCE), 2013 IEEE International Conference on (pp. 197-200). IEEE.

[10] SH, M. I., & Sanap, S. A. (2013). Intelligent Heart Disease Prediction System Using Data Mining Techniques. International J. of Healthcare & Biomedical Research, 1(3), 94-101.

[11] Avci, E., & Turkoglu, I. (2009). An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases. Expert Systems with Applications, 36(2), 2873-2878.

[12] Rajeswari, K., Vaithiyanathan, V., & Amirtharaj, P. (2011). Prediction of Risk Score for Heart Disease in India Using Machine Intelligence. In 2011 International Conference on Information and Network Technology, IACSIT Press, Singapore IPCSIT (Vol. 4).

[13] Parthiban, L., & Subramanian, R. (2008). Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences, 3(3).

[14] Dangare, C. S., & Apte, D. S. S. (2012). A data mining approach for prediction of heart disease using neural networks. International journal of Computer Engineering & Technology (IJCET), 3(3), 30-40.

[15] Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, 34(1), 366-374

[16] Dodani, S., Mistry, R., Khwaja, A., Farooqi, M., Qureshi, R., & Kazmi, K. (2004). Prevalence and awareness of risk factors and behaviours of coronary heart disease in an urban population of Karachi, the largest city of Pakistan: a community survey. *Journal of public health*, *26*(3), 245-24

[17] Han, J., Kamber, M., & Pei, J., (2006), Data mining: concepts and techniques, Morgan kaufmann.

[18] Deekshatulu, B. L., & Chandra, P. (2013). Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection. Global Journal of Computer Science and Technology, 13(3).

[19] Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. arXiv preprint arXiv:0912.3924.

[20] http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.db2.udb.doc/admin/c0006909.htm.

[21] Fatima, M.; Basharat, I; Khan, S.A; Anjum, AR., "Biomedical (cardiac) data mining: Extraction of significant patterns for predicting heart condition," Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on , vol., no., pp.1,7, 21-24 May 2014.

[22] http://www.upa.pdx.edu/IOA/newsom/pa551/lectur21.htm] (Retrieved on 23 March 2014).

[23] Fürnkranz, J. (2002). Round robin classification. The Journal of Machine Learning Research, 2, 721-747.

[24] http://www.ncsu.edu/labwrite/Experimental%20Design/accuracyprecision.htm (Retrieved on 2nd April, 2014).