JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Mortality Rate Prediction in ICU Using Logistic Regression Method

K V Sruthi [1*], K Manju [2], R K Rashmi [3], R Krishnamouli [4]

[1*] Computer Science Department, MVJ College of Engineering, VTU University,Bangalore-67, India
[2] Computer Science Department, MVJ College of Engineering, VTU University, Bangalore-67, India
[3] Computer Science Department, MVJ College of Engineering, VTU University, Bangalore-67, India
[4] Department of Big Data Analytics, St.Joesph's College, Autonomous, Bangalore, India

**Corresponding author: sruthi.krishna444@gmail.com, Tel: 0-8129647882*

Available online at: www.ijcseonline.org

Accepted: 16/May/2018, Published: 31/May/2018

Abstract— High risk of illness is observed for the patients admitted in hospital's cardiac intensive care units (ICU). Patient's dead/alive categorical outcome prediction would benefits for patients as well as medical professionals in creating awareness and making clinical decisions respectively. In this work, a model is proposed for predicting life outcomes of cardiac patients admitted in ICU. The model is prepared on the basis of data collected from the regular medication treatments and clinical laboratory test results. A logistic regression model is prepared and compared with two standard algorithms in machine learning such as artificial neural network (ANN) and random forest algorithms, which are the classifiers of decision tree. The performance parameters were compared for both Synthetic Minority Oversampling Technique and stratified sampling for all predictive learning models. It is concluded that logistic regression with stratified sampling techniques would be preferable as a predictive model for the inconsistent time series data set.

Keywords— Predictive learning, logistic regression, SMOTE sampling, stratified sampling, time series data

I. INTRODUCTION

Besides freedom from the illness, the good health helps to realize the true extend one's physical and mental abilities. Human services are regularly viewed as an immediate capacity of the wellbeing. The thought of heath being the unimportant nonattendance of isn't right. Other than opportunity from sickness, the great wellbeing understands the genuine expand one's physical and mental capacities. The wellbeing of the people is a vital issue openly arrangement talk in each enlightened society, deciding the sending significant assets the country over. Changes in people's life style and routines have a direct impact on their health. In the developing nations, the trend shows the uptick in the number of cardiac patients year by year. India's demographics have been undergoing a rapid change. The diabetic rate is found increasing year by year, becomes the highest rate ever reported number from anywhere in the world as per the data available to the International Diabetes Federation. The Cardio vascular cases in our country are witnessing a simultaneous uptick in incidence of heart failure, prevalent even among relatively younger men and women who were once regarded out of the risk bracket of this condition. However, general awareness about heart failure, its causes, and its diagnostic and treatment mechanisms remains low, implying that our healthcare system as well as society is unprepared to deal with this cardiovascular conundrum. Hence, a potential scope in information technology usage is found in this field for the public awareness and clinical decision preparation.

Patience admitted in the ICU observed a high risk of dying or critical illness. Therefore a continuous monitoring and treatment is necessary as a part of care. The treatments generally followed up in ICU are directly related to the recovery symptoms. Patients undergoing treatments will be directed to medication as well as laboratory tests, where professionals use laboratory test results to make clinical decisions. For heart related issues, there were many test procedures and standards to make judgments on conditions. Even though the treatments, tests, medications are same for the same kind of disease, but the recovery is completely independent to others and personal factors were dominated highly. The patient's outcomes or their status can be predicted by the technology of predictive analytics, which are the statistical and machine learning methods of producing the outcome form the information collected.

II. RELATED WORK

An increment in mortality due to cardiac diseases is found year by year. The various techniques and methodologies associated with health domain proposed by different researchers for the diseases like cancer, cardiac diseases, stroke and other health related issues are mentioned below. The literature trend shows the updates in international and national levels over the last ten years in the medical field. A comparison model for various machine learning techniques prepared by Rafiah Awang [1]. In this model, the author claimed that a model with precise data collection methods like computer based patient record and monitoring system would have enhance the patient safety and hence their outcome prediction. The issue of the proposed model is mainly associated with data set characteristics such as data set size, nature of data set. The analysis were limited for the smaller data set and needed to extract for any larger size data set. The inter patient similarity identification and cluster based prediction modelling for heart related problems are presented by Ebadollahi et al [2]. In this work, the data collected were time series data and outcomes were predicted from the list of logistic operations. A retrieved data from the particular patient group sets as standard for the future to give understandings for any uncertainty situation for the patient requisites. Very similar machining learning algorithms were by used by Huang et al. [3] for the prediction system. Even though the algorithm used were same, but the methodology taken were entirely different. A data mining method is adopted to identify the inconspicuous rules and the ill condition relationship from the actual system. The outcome of presence or absence occurrence is identified in this work, but not on the pathological indices information. Chignell et al. [4] proposed a predictive model with retained all privacy details of the patients. In this case study, they were able to identify the methods in grouping of patient and their importance in data mining techniques. From the summarized patient grouping and types, the clinical decisions were formulated. The grouping is here referred as clustering technology in data.

A clustering method using K means algorithms is used for the predictive decision making is proposed by Morissette et al [5.] In this approach the variance within the cluster found less and found more with other cluster segments. A centroid model represents a vector scheme with mean value would identify the nearest data for the segmentation. Rouzbahman et al [6] predicted the dichotomous life outcome for ICU stay hours calculations. In this approach an optimum cluster were identified for the row and clustered data for the accuracy prediction. Also another comparison is tested for the comparison of row data with linear and logistic regression approaches. The linear regression is failed in some occasion since it is of predicting dichotomous outcomes. Similar to the accuracy of death/live prediction, the time period in ICU ia also predicted using linear regression analysis. A heart stroke risk model is suggested by Jae-woo Lee et al. [7] for the Korean country cohort disease data set. This is hypothesis test method to predict the stroke risk within ten years of data collection. For the prediction of heart disease problems and their mortality survival probability, a novel method of Cox's proportional hazard regression model is used. This regression model is found good accuracy for type of time to event analysis. The goodness of estimated for 95% confidence level by using another technique of Hosmer–Lemeshow type $\chi 2$ test. A cancer diagnosis predictive modelling is suggested by Rouzbahman et al. [8] with health care data. They made a prediction on next visit date to ICU to diagnose the disease. The visit data is dependent on their admission date (ED visit) to the hospital with standardised symptoms and there after a symptom screened date. A symptom is score is calculated for each patient and their mortality is predicted based on the disgusted score.They also predicted mortality of patients with cancer diagnosis based on their last symptom scores.

For categorical outcome prediction, especially in patient life outcome, the life outcome would be more biased. This biasness in the data would not perform the modeling accurately. The sampling method hence preferred, could be able to reduce the biasness as well as time and cost of operation. There were many studies conducted on various sampling methods on machine learning problems. A study different sampling method called SMOTE (synthetic minority over sampling technique) sampling is conducted on the imbalanced data set by Nitesh et al. [9]. In this method, the data minority data set will be over sampled without changing much characteristics of the entire data set. This sampling technique made the data more balanced. Depending on the requirement the minority class is over-sampled at various multiplied percentage and of its original size. Each minority class and their neighbor in nearest is joined in segmented lines, and a synthetic sample will taken each minority class sample along all line segments. The selections were random for the nearest neighbors, and over sampling requirement has an effect on nearest neighbors. A study on the effect of stratified sampling on machine learning problem is conducted by Kevin Lang et al.[10]. In their approach, a future query is generalized from the past queries using different sampling and estimation methods.

A continuous monitoring and treatment is necessary as a part of ICU care. Patients undergoing treatments will be directed to medication as well as laboratory tests, where professionals use laboratory test results to make clinical decisions. For heart related issues, there were many test procedures and standards to make judgments on conditions. Even though the treatments, tests, medications are same for the same kind of disease, but the recovery is completely independent to others and personal factors were dominated highly. The purpose of the project is directed to propose a predictive model in predicting dead/alive outcomes from the real time inconsistent time series data set. The preliminary objective of the work is to prepare three prediction models such as logistic regression, as artificial neural network (ANN) and random forest algorithms. All the three models were formulated with two different sampling methods such as, SMOTE sampling and stratified sampling. The all six combinations mentioned above is compared for accuracy, which are measures of the models.

III. PREDICTIVE MODELING SYSTEM DESIGN

A model will be useful in reducing the manual task. Predictive model can be built for both prediction and classification. When using the data to make the decisions, the model that are empirically derived and statistically valid. As discussed before, the test result data is crucial for the patient admitted to hospital for their follow up. The Prognostic data for a query patient is conveyed to clinicians through a modified ICU monitoring system. A predictive model is proposed for predating the categorical outcome of death/ live conditions based on the ICU laboratory test data information.

A. Data collection

The inconsistent time series data collected from MIMIC database, a public database holds currently holds clinical data from over 40,000 stays in Beth Israel Deaconess Medical Center ICU's during the first 48 hours after admission. The database contrasted the severity of ICU patient conditions and outcomes, as well as treatment costs, across joining ICUs on the basis of relatively few, highly selected pieces of information. The patients were adults, were admitted for a wide variety of reasons to cardiac, medical, surgical, and trauma ICUs. The data has 42 variables were recorded at least once for each patient.

B. Data preprocessing

The quality of the machine in predictive accuracy is based completely on the collected data. A great deal of human effort is made on processing of the data The problem with in consistency in time series data was the biggest challenge in forwarding to the next level. Hence a multiple methods of data pre processing used include data cleaning, data normalization, data transformation, feature extraction and selection etc. The data transformation is made using principle component analysis. In this method consists of identifying data patterns and conveying the data in such a way as to show their similarities and differences. PCA found to be power tool for analyzing the data where data with high dimension failed in luxury graphical presentation. In PCA the data is compressed without losing much information's and redundancies. Fig.1 shows variance curve after principle component analysis. For the modeling, hence 30 components preferred with variance of 0.9.

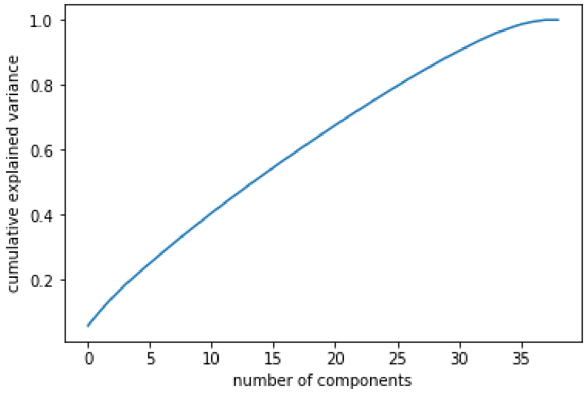

Fig 1 PCA variance plot

C. Modeling

Predictive modeling is a form of data mining technology that runs by historical data analysis and preparing a model to predict the future outcomes. Machine learning has developed thousands of algorithms to solve specific problems. The model based machine learning contrasts to create a best solution tailored to each new problem. A Logistic regression involves a better probabilistic view of classification. This technique is used to model dichotomous outcome variables. It takes 0 or 1 like dead/live condition for the ICU admitted patients. The ANN is a computational model. It works based on the functions and structures of biological neural networks. The interconnected neuron computes the values from the inputs.

It works very similar way like way of human brain works. Millions of neurons presents in the brain sends and process the signals on the form of chemical as well as electrical signal. The synapses are used to connect the neurons which pass information's from one to another. The ANN model has high variance and low bias. The random forest model can be model for comparison with logistic regression model with reduced variance. The low variance and low biased model produces a best comparator model versus logistic model. It is a fact that the dead outcome will be always lesser that the live outcome everywhere, this is reflecting in this data also.

The dead and alive nature outcome ratio of the data of shown in Fig.2. Out of 4000 patients, 3446 patients come under alive and the rest comes under dead outcomes. Because of higher variability in dead and alive outcome, the accuracy result would not come up correctly, so different methodologies of samplings are needed for the accuracy prediction.

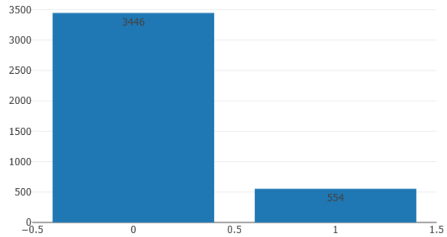

Fig 2 Patients dead/alive outcome graph

The approach of synthetic minority over sampling technique (SMOTE) is method of over sampling the minority and introducing synthetic examples rather than by over sampling by replacement. This sampling technique makes the data more balanced. The minority class can be over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the *k* minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the *k* nearest neighbors are randomly chosen. Since there may be a chance of losing characteristics of data set using SMOTE Sampling causes errors in predicting the outcome. An alternate stratified random sampling technique is adopted to capture the variation in modeling. The variability within the strata will be very less when compared to the entire data set, hence the accuracy can be improved. This sampling method helped in saving lot of time and effort in operation. The selection of this method would reduce the selection bias, that is the method ensuring the sample that reflects accurately about the population characteristics

IV. RESULTS

The model built by three different machine learning techniques such as logistic regression, artificial neural network and random forest algorithms. Each learning method is tested in two different sampling combinations of SMOTE sampling and stratified sampling. The results were tabulated and plotted for accuracy, ROC curve and confusion matrix. Accuracy reflects the correctness and misclassification represents the wrongness. The six combinations of model and their performance evaluations are listed below.

IV.a Case I: Logistic regression with SMOTE sampling

The ROC curve for the Logistic regression with over sampling using SMOTE technique is shown in Fig.3. The ROC curve is lying closer to the 45^o line, reflects less accurate and the values of accuracy estimates is 0.48. The accuracy result and misclassification for case I is 0.4815 and 0.518 respectively.

IV.b Case II : ANN model with SMOTE sampling

The ROC curve for the ANN (artificial neural network) model with SMOTE sampling is shown in Fig.4. The ROC curve is lying closer left side border and closure to top border, reflects better accuracy and the values of accuracy estimates is 0.68. The accuracy result and misclassification for case II is 0.679 and 0.324 respectively.

IV.c Case III: Random forest model with SMOTE sampling

The ROC curve for Random forest model with SMOTE sampling is shown in Fig.5. The area under the curve in ROC curve reflects better accuracy and the values of accuracy estimates is 0.81. The accuracy result and misclassification for case III is 0.876 and 0.192 respectively.

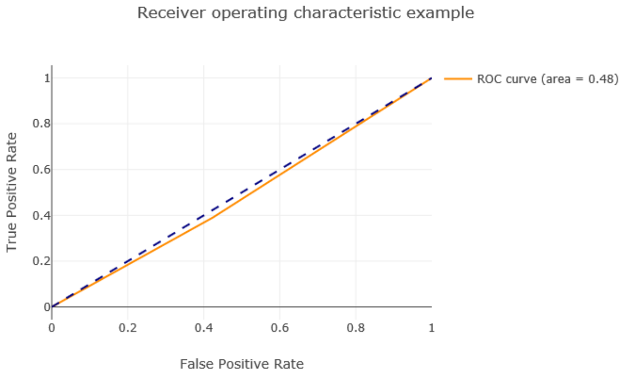

Fig.3 ROC curve for case-I

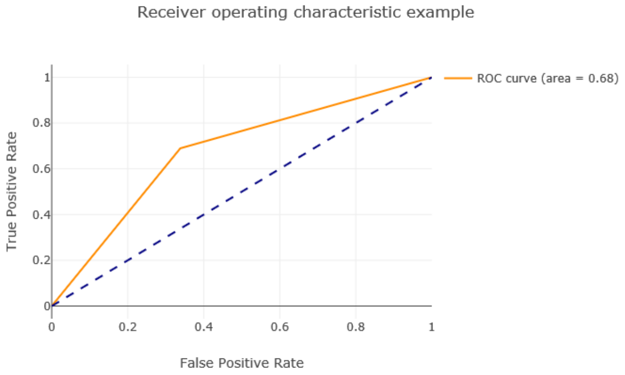

Fig.4 ROC curve for case II

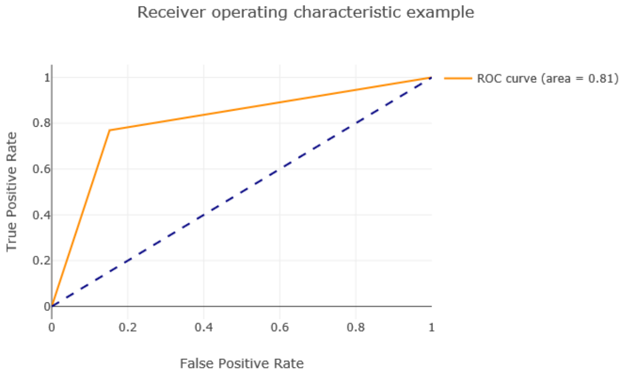

Fig 5 ROC curve for case III

IV.d Case IV: Logistic regression with Stratified sampling

The ROC curve for the Logistic regression with stratified sampling is shown in Fig.6. The AUC of ROC curve reflects better accuracy estimates of 0.81. From the confusion matrix the value of accuracy and misclassification rate is found 0.87 and 0.12 respectively.

IV.e Case V : ANN model with stratified sampling

The ROC curve for ANN model with stratified sampling shown in Fig.7. The AUC of ROC curve reflects better accuracy estimates of 0.80. The accuracy result and misclassification for case 5 is 0.886 and 0.1132 respectively.

IV.f Case VI : Random forest model with stratified sampling method

Random forest model with stratified sampling ROC curve is shown in Fig.8. The ROC curve reflected accuracy the values of 0.80. The model accuracy result and misclassification for case 6 is 0.876 and 0.123 respectively.

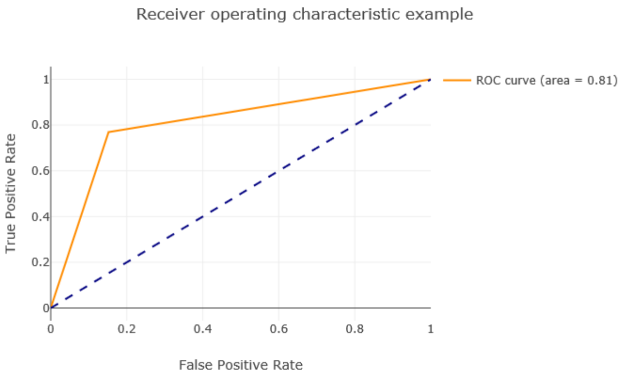

Fig.6 ROC curve for case IV

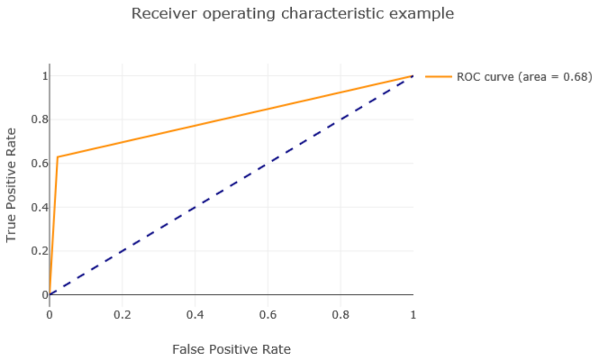

Fig.7 ROC curve for case V

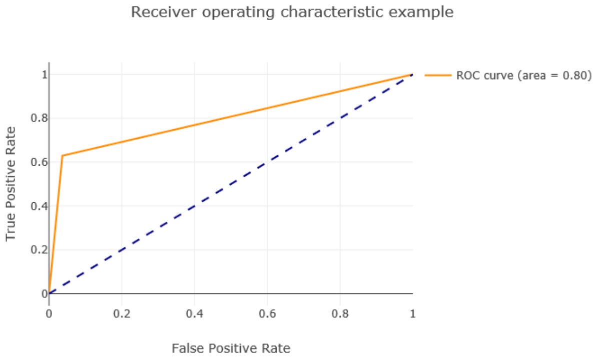

Fig.8 ROC curve for case VI

Table 1 Model comparison for SMOTE sampling

	SMOTE sampling		
	Accuracy	AUC under ROC curve	Misclassification rate
Logistic regression	0.4815	0.48	0.518
ANN	0.67858	0.68	0.3241
Random forest	0.80783	0.81	0.1921

Table 2 Model comparison for stratified sampling

	Stratified sampling		
	Accuracy	AUC under ROC curve	misclassification rate
Logistic regression	0.87	0.81	0.1293
ANN	0.8867	0.68	0.1193
Random forest	0.87601	0.8	0.123

The results were plotted in the chart for comparison in Fig.9 and Fig.10 for smote and stratified sampling respectively. Also a comparison table for confusion matrix for all the six cases is listed in table 3. A classification report is shown in Table 4and Table 5 for comparing the logistic regression model with stratified and smote samples.

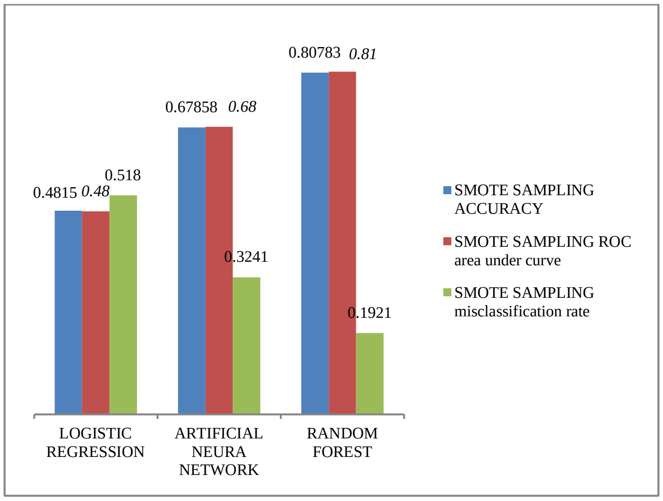

Fig.9 Model comparison for SMOTE sampling

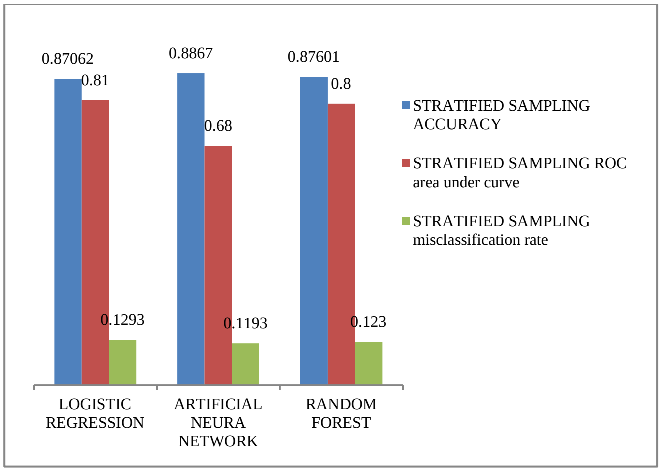

Fig. 10 Model comparison for stratified sampling

Table 3 Confusion matrices comparison for all cases

case-1			
Actual class		Predicted class	
		P	N
	P	390	287
	N	428	274

case-2			
Actual class		Predicted class	
		P	N
	P	448	229
	N	218	484

case-3			
Actual class		Predicted class	
		P	N
	P	574	103
	N	162	540

case-4			
Actual class		Predicted class	
		P	N
	P	266	8
	N	40	57

case-5			
Actual class		Predicted class	
		P	N
	P	268	6
	N	36	61

case-6			
Actual class		Predicted class	
		P	N
	P	264	10
	N	36	61

Table 4 classification report for logistic regression with stratified sampling

	Precision	Recall	f1 SCRE	support
0	0.87	0.97	0.92	294
1	0.88	0.5	0.70	97
Avg/total	0.87	0.87	0.86	371

Table 5 classification report for logistic regression with smote sampling

	Precision	Recall	f1 SCRE	support
0	0.53	0.12	0.19	677
1	0.51	0.90	0.65	702
Avg/total	0.52	0.51	0.43	1379

V. CONCLUSION

Medical diagnosis plays an important role in modeling but the task should be executed efficiently and accurately. This work is intended to predict the life outcome of patients admitted in ICU due to critical heart diseases. The prediction is completely based on the medical laboratory test results. A Predictive modeling is done with statistical logistic regression and machine learning models such as artificial neural network (ANN) and random forest. The performance of all the three mentioned models were compared for stratified sampling and SMOTE samplings. From the performance measures, it is indicated that the logistic regression with stratified sampling confirms the best fit model for the data collected. Even though the random forest model is identified with a near matching solution, but t the probabilistic value of outcome can be measured in logistic regression only

The model would be benefitted for the patients in creating awareness and medical research professional for critical thinking and decision making. Thus the accuracy of the model can improved further in future by using advanced technologies in data collection, which reduces the errors in data collection such as missing values, wrong entries etc.

REFERENCES

[1] R Awang, S Palaniappa, *"Intelligent heart disease prediction system using data mining techniques"*, IEEE/ACS International Conference on Computer Systems and Applications, ISSN: 2161-5322,2008.

[2] J.Sun, S. Ebadollahi, D. Gotz, J. Hu, D. Sow and C.Neti , *"Predicting Patient's Trajectory of Physiological Data using Temporal Trends in Similar Patients: A System for Near-Term Prognostics"*, AMIA Symposium Proceedings, pp-192-195, 2010.

[3] S.Wang , F.Huang, and C.Chan , *"Predicting Disease By Using Data Mining Based on Healthcare Information System",* IEEE International Conference on Granular Computing, 2012.

[4] M. Rouzbahman, R. Kealey, E. Yu, M. Chignell ,R. Samavi and T. Sieminowski, "*Development of Non-Confidential Patient Types for Use in Emergency Medicine Clinical Decision Support*," IEEE Securiy & Privacy, vol. 11, pp. 12-18, 2013.

[5] L.Morissette and S.Chartier, "*The k-means clustering technique: General considerations and implementation in Mathematica*", Tutorials in Quantitative Methods for Psychology, Vol. 9 (1), p. 15-24,2013.

[6] M. Rouzbahman, R. Kealey, E. Yu, M. Chignell ,R. Samavi and T. Sieminowski, *"Development of Non-Confidential Patient Types for Use in Emergency Medicine Clinical Decision Support*," IEEE Security & Privacy, vol. 11, pp. 12-18, 2013.

[7] J.Lee, H.Lim, D.Kim, S.Shin, J.Kim,B.Yoo, and K.Cho, *"The development and implementation of stroke risk prediction model in National Health Insurance Service's personal health record"*, Computer Methods and Programs in Biomedicine, Vol 153, pp. 253-257, 2018.

[8] M.Rouzbahman, A. Jovicic, and M.Chignell, "Can Cluster-Boosted Regression Improve Prediction of Death and Length of Stay in the ICU?", *IEEE Journal of Biomedical and Health Informatics,* Vol 21, pp. 851 – 858, 2016.

[9] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer , *"SMOTE: Synthetic Minority Over-sampling Technique",* Journal of Artificial Intelligence Research, vol.16, pp 321-357, 2002.

[10] Kevin Lang, Edo Liberty, Konstantin Shmakov, *"Stratified Sampling Meets Machine Learning",* International Conference on machine Learning, vol.48, pp 2320-2329, 2016.