Survival Rate Prediction Model of Cardio Vascular Disease Patients by Quantifying the Risk Profile using SVM

Conference Paper · December 2020

DOI: 10.1109/ICCIS49240.2020.9257638

CITATIONS
2

READS
70

3 authors:

Ali Javed
University of Engineering and Technology Taxila
186 PUBLICATIONS **4,595** CITATIONS
SEE PROFILE

Hayet Fatima Zohra Belhaouari
21 PUBLICATIONS **406** CITATIONS
SEE PROFILE

Hussain Dawood
Skyline University College
81 PUBLICATIONS **1,114** CITATIONS
SEE PROFILE

Survival Rate Prediction Model of Cardio Vascular Disease Patients by Quantifying the Risk Profile using SVM

Fatima Zohra
Dept. of Software Engg
UET Taxila
Pakistan
fatimasworld@gmail.com

Ali Javed
Oakland University
MI, USA
alijaved@oakland.edu

Hussain Dawood
College of CSE
University of Jeddah
Saudi Arabia
hdaoud@uj.edu.sa

Abstract – **This research paper provides a clinical analysis of heart patient data to predict the survival rate of patients suffering from cardio-vascular diseases (CVD). The proposed solution will facilitate medical specialists in terms of providing quality health services to patients including the intensive treatments. Our model determines the chances of survival of patients suffering from any cardiovascular disease by analyzing the risks associated with them and other factors of their life style, physical activity, smoking habit, etc. The proposed survival rate prediction model is efficient and economical solution to facilitate the medical specialists in terms of following the most appropriate medical procedures for given symptoms. This paper presents an improved Stochastic Gradient Descent (iSGD) approach along with Hinge Loss Function of Support Vector Machine (SVM). Experimental results illustrate the effectiveness of the proposed prediction model in terms of predicting the survival rate of CVD patients.**

Index Terms- **Artificial intelligence; cardio vascular disease; knowledge representation; Machine Learning; Prediction methods.**

I. INTRODUCTION

Cardiovascular disease (CVD) is a class of diseases identified with the heart or veins and is the main source of mortalities around the globe. Numerous hazard factors are observed to be related with the CVD i.e., family history, age, physical activity, obesity, lifestyle, smoking, etc. [1]-[4]. In the last few decades, we have witnessed an exponential growth in CVDs in both developed and under-developing countries for various reasons [5]-[6]. The statistics of heart diseases and consequently heart failures are very alarming. According to the heart association of America [7], an American citizen goes through a heart attack after every 40 seconds. Around 720,000 heart attacks are recorded annually only in the US and approximately 335,000 new heart attacks are registered [7]. The rate of heart diseases is continuously intensifying and affecting people of all age groups. Rate of CVDs is high in both developed and developing countries due to unhealthy diet and life style habits. An average of 2% adults undergo heart failure in developed states; furthermore, 6 to 10 % old age citizens over 65 years suffer from heart failure [8].

After the detection of CVD/HF, the identification of patients' survivability/mortality rate is important. The treatments are very expensive and require consultation of various staff such as intensivists, cardiologists, radiologists, pathologists, physiotherapists, anaesthetics, nursing staff, dieticians, pharmacists, therapists, etc. Manual treatments are unable to achieve the required level of accuracy in terms of survival rate prediction of CVDs. The prediction of heart disease survivability is a challenging research problem and demands a need to propose more effective solutions. The application of sophisticated machine learning algorithms in medicine have significantly increased the performance of the treatment of various diseases. The incorporation of advanced machine learning techniques in CVD treatment can also improve the survival rate prediction of CVD patients. The proposed prediction model ensures that the time, energy and cost is focusing on the most deserving patient first, thus increasing the mortality rate.

Existing literature have proposed various automated approaches based on machine learning to predict cardio-vascular diseases and survival rate of patients suffering from these diseases. Kokol [9] proposed an intelligence-based system that uses machine learning techniques to increase the survival rate of patients. This method [9] empowers the doctors to escalate the intensive care of patients. This research formulates multimethod approach to extract useful information from the data. Miao [10] employed an improved Random Survival Forest (iRSF) model to predict the chances of survival of patients. Lakshmi [11] presented a comparative analysis of different machine learning classifiers to predict the survival rate of patients suffering from cardio-vascular disease. Similarly, Tripoliti [15] provided a comprehensive comparison of machine learning classifiers to predict the heart failures. Ordonez et al. [12] proposed an automated technique based on association rules to predict the frequency of heart diseases. Priyanka [13] proposed a hybrid classification model based on Naive Bayes and Decision Trees to predict the heart disease. Similarly, Shahed [14] proposed a hybrid model based on SVM, Decision Trees and Naïve Bayes to predict the heart diseases.

In [16], iRSF approach highlights the classification of various risk factors and predicts the mortality of patients with cardiovascular diseases. MAGGIC study [17] used a universal database to generate an effective model for death rate prediction of patients suffering from the cardiovascular diseases. Vazquez et al. [18] used the multivariable cox model to create predictive models for Cardiac Heart Failure (CHF) patients. In [19], authors proposed a threat model based on

978-1-7281-5467-1/20/$31.00©2020 IEEE

SENIORS dataset comprising of older patients with ages of ≥70 years. In Seattle Heart Failure Model [20], a multivariate threat model using the multivariate cox model was proposed to predict 1, 2, and 3 year endurance in cardiovascular breakdown patients. In [21], random survival forests (RSF) was used to predict important risk factors for survival in patients with systolic heart failure.

In this paper, we proposed a novel prediction approach using SGD with hinge loss function of SVM to improve the accuracy of survival rate prediction of cardio-vascular patients. The proposed model effectively separates the CVD patients with high chances of survival with those of least chances. The fact that existing prediction models are still unable to achieve better accuracy, we aim to improve the accuracy, precision, and recall of existing approaches. For this purpose, we propose a novel model of improved stochastic gradient descent along with SVM to predict the survival rate of CVD patients. Additionally, our model provides the prediction using minimum number of risk factors.

The rest of the paper is organized as follows. Section II presents the proposed methodology. Section III provides the discussion on the experimental results along-with comparative analysis with existing methods. Finally, Section IV concludes the proposed work.

II. PROPOSED METHODOLOGY

This section provides the methodology of the proposed prediction model. We propose an improved stochastic gradient descent algorithm with the Hinge Loss function in order to identify the survival cases of CVD patients. Shown in Fig. 1 is the modelling architecture of the proposed iSGD based method.

A. Dataset Acquisition

We used Cleveland data of UCI Repository [21] which contains a total of 4000 CVD patients' records. Among the data of 4000 patients, 3599 cases include the data of those patients that survived from the heart problems, whereas, remaining 401 cases are of deceased ones. A patient whose survival is not confirmed is considered as deceased for not showing up for follow-up from the last five years.

B. Dataset Processing

1) Replace Missing Values Filter: Missing values is a common problem in large-scale real-world datasets. The Cleveland dataset [21] also contains some records where values of certain attributes are missing. The missing values must be handled in the pre-processing step to prepare the data; failing to properly address the missing values problem degrades the classification performance and results in lower accuracy. Therefore, to address the issue of missing values, we employed the mean substitution-based imputation technique to replace the missing values with statistical estimates of the neighbouring values. Mean substitution-based imputation technique computes the mean value of the attributes and use this mean value to fill the missing values.

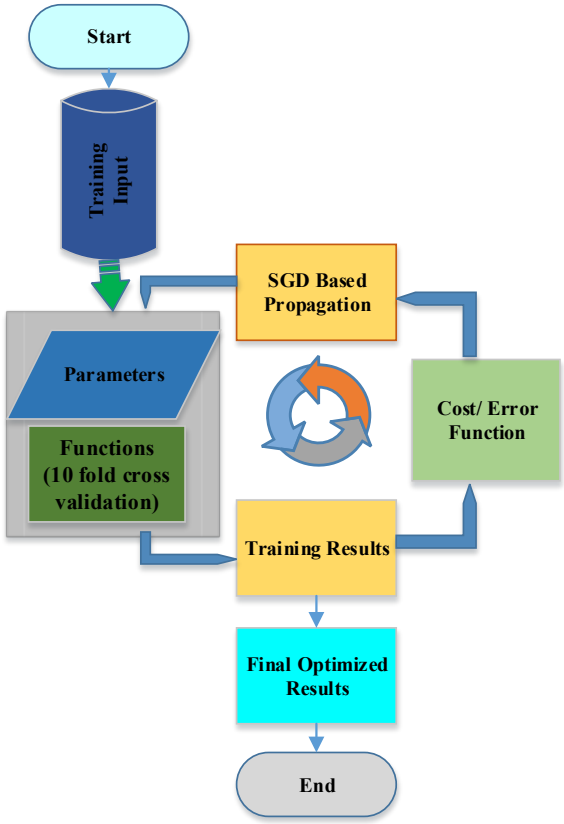

Fig. 1. MODELLING ARCHITECTURE OF ISGD.

2) Attribute Selection: Selection of relevant attributes is an indispensable requirement for effective classification. To select the most relevant attributes, we employed the correlation-based feature selection (CFS) attribute evaluator method in combination of Best First search method for calculating correlations among the best data attributes. Using this method, we attained the Trestbps, Restecg and Famhist attributes.

3) Dimensionality Reduction by Principal Components Analysis: In the next step, we applied the principle component analysis (PCA) to perform the dimensionality reduction on the selected twenty two attributes to remove the redundant attributes. PCA divided the dataset into eigenvectors that explains the variance in data in the best possible way. We selected the Eigen vectors on the basis of more positive values that are contributing positively in the final outcome. The resulting eigenvectors obtained after applying the PCA are mentioned in Table I.

TABLE I. EIGEN VECTORS

V1	V2	V3	V4	Description
0.0359	-0.0087	-0.7062	0.7071	Trestbps
0.0624	-0.8144	0.0187	0.0055	Restecg=0: normal
-0.7354	0.3517	-0.0422	-0.0005	Restecg=1: ST-T wave abnormality
0.673	0.4612	0.0235	-0.005	Restecg=2: left ventricular hypertrophy
-0.0321	0.0185	0.7061	0.7071	Famhist=yes

4*) Ten-Fold Cross Validation Resampling:* Along with the methods applied for dataset processing and classifiers, 10-fold cross validation resampling of dataset is used in conjunction. Ten-fold technique is used to split the input dataset into training and test data. Training data is used to train the dataset while test data is used to evaluate the trained model. In ten-fold cross validation, the input sample dataset is divided arbitrarily into 10 equivalent samples. More generically in any k-fold cross validation one sample is used as validation data/testing data while the remaining k-1 samples of data are used as training data. The whole process of cross validation is iterated 10-times and the average of all the results is used. The sampling technique is mathematically expressed as in (1).

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - a - \beta^T x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 \qquad (1)$$

Where $y_1 ... y_n$ are the response values and n represents the *p*-dimensional vector *covariates* $x_1 ... x_n$.

C. Classification

1) Stochastic Gradient Descent (SGD): We applied the stochastic gradient descent (SGD) algorithm on the dataset. Since our outcome i.e. the survivability is a binary value i.e. either 1 (true/ survived) or 0 (false/ not survived), so SGD is best suitable approach for optimizing the binary problems. In stochastic gradient descent, a number of iterations selects the data randomly known as a `batch`. The method then calculates the gradient for each iteration and finally combines the results to provide the optimized output. SGD act as more rapid and efficient on large datasets. SGD is applied on a small subset of entire dataset at a time followed by dataset shuffling to perform the next iteration. SGD approach is cost effective as compared to other optimization methods such as Batch Gradient Descent and Mini-Batch Gradient Descent. SGD algorithm can be expressed as as in (2).

$$\theta_j = \theta_j - \alpha(\hat{y}^i - y^i)x_j^i \qquad (2)$$

Where $(\hat{y}^i - y^i)$ represents the gradient of slope and x_j^i represents the cost function at each iteration, θ_j represents the initial point to start the iteration.

2) Support Vector Machine (SVM) Hinge Loss Function: In this work, SGD is combined with famous hinge loss function of linear SVM for better optimization and training. Linear SVM is also known as kernel less SVM. Both of them works best for classifying a binary classification problem. SGD in combination with Hinge loss function (SVM) works by globally replacing all the missing values and perform normalization after converting the nominal values into numeric values. The final output is based on these normalized values. A result of normalized values is shown in Table II.

III. PERFORMANCE EVALUATION

Performance of the proposed method is evaluated on Cleveland dataset [21] for survival rate prediction of cardio-vascular disease. This section provides the quantitative analysis of the proposed method. We also provided the performance comparison with existing state-of-the-arts. The proposed framework is implemented and evaluated in Weka 3.8.2.

TABLE II. NORMALIZED DATASET

S.N	Value	Variables
1	0.0372	(normalized) Age
2	0.12	(normalized) Sex=Female
3	-1.7492	(normalized) Cp=1: typical angina
4	-1.4293	(normalized) Cp=2: atypical angina
5	-1.4993	(normalized) Cp=3: non-anginal
6	-1.2994	(normalized) Cp=4: asymptomatic
7	10.7138	(normalized) Trestbps
8	0.1441	(normalized) Chol
9	-0.03	(normalized) Fbs=no
10	-4.248	(normalized) Restecg=0: normal
11	-3.9781	(normalized) Restecg=1: ST-T wave abnormality
12	2.249	(normalized) Restecg=2: left ventricular hypertrophy
13	0.1911	(normalized) Thalach
14	0.04	(normalized) Exang=no
15	0.0298	(normalized) Oldpeak
16	-0.09	(normalized) Smoke=yes
17	-0.07	(normalized) Cigs
18	0.1078	(normalized) Years
19	-0.1	(normalized) Famhist=yes
20	-0.9096	(normalized) Activity=Seldom
21	-0.6997	(normalized) Activity=Daily
22	-0.6297	(normalized) Activity=Yearly
23	-0.6396	(normalized) Activity=Once
24	-0.6797	(normalized) Activity=Monthly
25	-0.8696	(normalized) Activity=Often
26	-0.8296	(normalized) Activity=Weekly
27	-0.7197	(normalized) Activity=Never

In the first experiment, we used the objective evaluation metrics i.e. precision, recall, accuracy, mean absolute error (MAE), root-mean squared error (RMSE), and relative absolute error (RAE) for performance evaluation. The proposed method achieves the precision of 0.841, recall of 0.914, MAE of 0.158, RMSE of 0.398, and RAE of 59%. The results are presented in Table III. As it can be observed from Table III that the proposed method achieves more recall over precision as our method achieves high true positive rate.

TABLE III. PERFORMANCE EVALUATION OF THE PROPOSED METHOD

Overall Evaluation Summary	
Precision	0.841
Recall	0.914
Mean absolute error	0.1588
Root mean squared error	0.3984
Relative absolute error	59.4065 %

In the second experiment, we performed the confusion matrix analysis to present the classification accuracy of the proposed method as shown in Table IV. From the results, we can observe that the proposed method achieves remarkable true positives in comparison of false negatives. However, the proposed method achieves more false positives as compared to true negatives resulting in less precision as compared to recall. This is due to the fact that the dataset contains more true values as compares to false entries.

TABLE IV. CONFUSION MATRIX

	Survive	Not Survive
Survive	3065	101
Not Survive	534	300

In our third experiment, we provided a comparative analysis of the proposed and existing state-of-the-arts methods in terms of precision. The detailed comparison is presented in Table V. From the results, we can clearly observe that the proposed method outperforms all existing state-of-the-arts in terms of accurate survival rate prediction of cardio-vascular disease. The second-best method [16] achieves the precision of 0.82 and [20] is the worst that obtained the lowest precision of 0.70.

TABLE V. PERFORMANCE COMPARISON WITH EXISTING STATE-OF-THE-ART METHODS

Model	Statistical Method	Performance evaluation Method	Precision
iRSF [16]	iRSF	OOB C-Statistics	0.821
MUSIC risk Score [17]	Multivariable Cox Model	OOB C-Index	0.76
SENIORS study [18]	Multivariable Cox Model	C-Statistics	0.72
Seattle Heart Failure Model [19]	Multivariable Cox Model	AUC	0.729
RSF based Heart Failure Model [20]	RSF	OOB C-Statistics	0.705
Proposed Model	**iSGD with SVM (Hinge Loss)**	**Goodness-of-fit**	**0.84125**

IV. CONCLUSION

The proposed classification model successfully learned the trends and risk factors in the clinical record of heart patients and accurately predicts the survival rate. The proposed iSGD predict the survival rate with an enhanced precision of 0.84 and recall of 0.91. Performance comparison of the proposed model against the existing state-of-the-arts illustrate the significance of the proposed method in terms of accurate survival rate prediction of patients suffering from the CVD. The proposed model will assist the medical specialists in the field of medicine on many grounds. The proposed framework can be further enhanced to predict the exact medication and treatment that must be given to the CVD patients by encompassing all the historical context of treatment of each patient. This work can be further used for the implication of finding the exact root causes of CVD diseases.

REFERENCES

[1] Maton A, Hopkins J, McLaughlin CW, Johnson S, Warner MQ, LaHart D, Wright JD. (1993). Human Biology and Health. Englewood Cliffs, New Jersey: Prentice Hall.

[2] Bridget B. Kelly; Institute of Medicine; Fuster, Valentin (2010). Promoting Cardiovascular Health in the Developing World: A Critical Challenge to Achieve Global Health. Washington, D.C: National Academies Press.

[3] Dantas AP, Jimenez-Altayo F, Vila E (2012). "Vascular aging: facts and factors". Frontiers in Vascular Physiology 3 (325): 1–2.

[4] Meeting the Challenges in Developing; Fuster, Board on Global Health ; Valentin; Academies, Bridget B. Kelly, editors ; Institute of Medicine of the National (2010).

[5] Mendis, Shanthi, Pekka Puska, Bo Norrving, and World Health Organization. *Global atlas on cardiovascular disease prevention and control*. Geneva: World Health Organization, 2011.

[6] McGill HC, McMahan CA, Gidding SS (2008). "Preventing heart disease in the 21st century: implications of the Pathobiological Determinants of Atherosclerosis in Youth (PDAY) study". Circulation 117 (9): 1216–27.

[7] https:// www. healthline.com/ health/ heart - disease/ statistics#1.

[8] J.J. McMurray and M.A. Pfeffer, "Heart Failure" Lancet, vol. 365, no. 9474, pp, 1877-1889,2005.

[9] P Kokol, P Povalej and Z Pehnec., "Analysing Heart Attack Survival using Intelligent Systems", Computers in Cardiology Mo3;30:33S-338, 2003 lEEE.

[10] Miao, Fen, Yun-Peng Cai, Yu-Xiao Zhang, Xiao-Mao Fan, and Ye Li. "Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest." *IEEE Access* 6 (2018): 7244-7253.

[11] K.R. Lakshmi , M.Veera Krishna and S.Prem Kumar "Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability, International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013.

[12] Carlos Ordonez, Edward Omincenski and Levien de Braal "Mining Constraint Association Rules to Predict Heart Disease", Proceeding of 2001, IEEE International Conference of Data Mining, IEEE Computer Society, ISBN-0-7695-1119-8, 2001, pp: 433-440.

[13] Priyanka, N., and Pushpa RaviKumar. "Usage of data mining techniques in predicting the heart diseases—Naïve Bayes & decision tree." In *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pp. 1-7. IEEE, 2017.

[14] Sabab, Shahed Anzarus, Md Ahadur Rahman Munshi, and Ahmed Iqbal Pritom. "Cardiovascular disease prognosis using effective classification and feature selection technique." In *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)*, pp. 1-6. IEEE, 2016.

[15] Tripoliti, Evanthia E., Theofilos G. Papadopoulos, Georgia S. Karanasiou, Katerina K. Naka, and Dimitrios I. Fotiadis. "Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques." *Computational and structural biotechnology journal* 15 (2017): 26-47.

[16] Fen Miao, Yun Peng Cai, Yu-xiao Zhang, Xiao-Mao Fan, and Yeli "Predictive Modeling of Hospital Mortality for patients with heart failure by using an improved random survival Forest" IEEEAccess .2018. 2789898, vol 6, 2018.

[17] R. Vazquez et al., ''The MUSIC Risk score: A simple method for predicting mortality in ambulatory patients with chronic heart failure,'' Eur. Heart J., vol. 30, no. 9, pp. 1088 1096, 2009.

[18] L. Manzano et al., ''Predictors of clinical outcomes in elderly patients with heart failure,'' Eur. J. Heart Failure, vol. 13, no. 5, pp. 528-536, 2011.

[19] W. C. Levy et al., ''The Seattle heart failure model: Prediction of survival in heart failure,'' Circulation, vol. 113, no. 11, pp. 1424 1433, 2006.

[20] E. Hsich et al., ''Identifying important risk factors for survival in patient with systolic heart failure using random survival forests,'' Circulat., Cardiovascular Quality Outcomes, vol. 4, no. 1, pp. 39-45, 2011.

[21] Cleveland Dataset, https://archive.ics.uci.edu/ml/index.php, accessed on Oct, 2019.