ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(2), DOI: 10.17485/ijst/2019/v12i2/140252, January 2019*

Artificial Neural Network versus Binary Logistic Regression for Determination of Risk Factors of Myocardial Infarction

Muhammad Zubair Khan[1*], Salahuddin[2] and Muhammad Arif[3]

[1]BUITEMS, Quetta, Balochistan, Pakistan;
dr.zubair.statistics@gmail.com
[2]CECOS University of IT and Emerging Sciences, Peshawar, Khyber Pakhtunkhwa, Pakistan;
salahuddin_90@yahoo.com
[3]Iqra university, Islamabad; Muhammadarifimpcc@gmail.com

Abstract

Objectives: To identify the important (significant) risk factors of Myocardial Infarction (MI) and construction of statistical models using conventional technique of binary Logistic Regression (LR) and of artificial Neural Network (NN). Both the statistical models (LR vs NN) are compared in their predictive capabilities. A case-control study with the purpose of comparison of LR outcomes to NN outcomes. The research is covering the whole country. Therefore, the required data is collected from all parts of Pakistan (Peshawar, Quetta, Karachi, Lahore, Islamabad etc). The required data is collected in 13 months; starting from 01-Feb-2013 to 30-Mar-2014. **Materials and Methods:** The research is basically a case-control study. For this purpose a sufficient sample size of 2,000 is included containing 1,000 patients (cases) and 1,000 controls. The samples are collected from various places of the country. The sample involves male and female. AMOS and SPSS are used in the study to analyze the collected data. Two techniques are applied to the data to identify the significant risk factors of MI i.e. LR and Artificial NN. Results obtained from LR and NN are compared. **Findings:** Out of total 28 potential risk factors of MI, 16 variables are found significantly associated to the MI by LR analysis and 17 (16 are those selected by LR) are found significantly associated to MI by ANN model. Only one variable differs between the two outputs, i.e. fried food intake. The rest of 16 variables are exactly the same. These 16 risk factors are: hypertensive disorder, higher age, family history of CVDs, chest pain, atherosclerosis, psychosocial pressure, alcohol use, diabetes mellitus, income class, breathing problem, smoking, fish intake, obesity, male gender, physical activity, vegetable intake, and often intense anger. All the 16 risk factors are significant in development of the disease. In this study the most threatening etiology is found to be chest pain. **Applications:** The outcomes of the study has drawn the attention of the epidemiological investigators to consider other procedure (NN) alongside the orthodox method (LR) while examining risk factors of myocardial infarction for a better insight and comparison purpose. The results show that all the clinical and modifiable risk factors are important in context of Pakistan.

Keywords: Logistic Regression, Myocardial Infarction, Neural Network, Risk Factors, AMOS, SPSS

1. Introduction

Myocardial infarction occurs when the myocardial cell dies due to ischemia. Ischemia is a situation when an organ receives insufficient blood supply due to atherosclerosis and cell death is a process when blood forms clots and as a result blockage in blood supply towards heart occurs and finally the myocardial cell dies[1]. The most life threatening disease that is exposed by the research literature from around the globe is the CVD and the frequency of heart attack is the most among all CVDs. There are some definite risk factors those are explored in the world wide literature. These important risk factors are; hypoglycemic drug, vegetable intake, often intense

*Author for correspondence

anger, diabetes mellitus, atherosclerosis, tobacco use, sedentary life style, eating habit, gender, higher age, low income, disease history in the family, anti-hypertensive medication, hypertension, chest pain, alcohol consumption, psychosocial pressure and obesity etc. All of the mentioned variables are proved significant in one or the other study; moreover these are positively associated with the disease in most of the studies.

The maximum applied procedure used for discovering the prospective risk factors of myocardial infarction is multiple logistic regression models. The recent literature in the field of medical illustrates that ANN modeling procedureis also applied along side the LR model in examining the potential variables[2-4]. The most popular statistical method in use for years for identification of risk factors and a prediction tool is LR. This conventional technique is an asymptotic approach that's why it has some restrictions as well, like; problem of missing values or existence of outliers or managing of the interaction effects or problem of multicollinearity etc. In the face of such problems, another nonparametric techniqu has gain some popularity in recent past especially in the fields of medical research, i.e. Artificial Neural Network (ANN). This technique is working as a prediction tool instead of LR[2-9].

An ANN model is capable of capturing the concealed and complex associations in the data in a better way as compared to that of an LR model. In[10] proposed have discussed that an ANN is capable of tracing the hidden relationships among the variables and hence the overall performance of the model improves along with the enhanced predictive accuracy of the model. Another advantage of ANN over LR is that ANN models are capable of handling with variety of variables like; ordinal variable or nominal variable or continuous variable in same data analysis and NN models are more efficient in such cases because NN models don't follow strict statistical assumptions as an LR model does. Similarly another big advantage of NN over LR is that an NN model ranks the risk factors in their order of importance whereas an LR doesn't do this[11]. This quality of ranking of risk factors according to their significance is achievable with NN model only. On the other hand an LR model is not only capable of discovering the important risk factors from the list of available variables but also finds the strength of relationships between the disease and its etiologies/ risk factors, which is not possible by NN models. In[12] proposed has compared the outcomes obtained from both of these techniques in his medical research. In[13] proposed their MI case-control study presented this idea of relative importance of variables for the first time and recommended the use of NN as a prediction tool alongside conventional LR methods in medical research. Studies like have used both of these techniques (LR and NN) in their epidemiological studies (identification of risk factors of various diseases) and compared the results obtained by both the methods. In[14] proposed has used NN along with LR in his research of identifying the risk factors of uterine Myomas and discussed that NN's performance is better than LR in his research and declared NN as a powerful alternative to the conventional method of LR especially in studies involving analysis of risk factors of any disease. The studies carried[14-15] also reveals that NN models are atleast as good as an LR model if used for prediction purpose only and support the findings of the aforementioned studies.

Since ANN is non-parametric, therefore no statistical assumptions are involved. Being distributional free the use of ANN model becomes more suitable while analyzing complex large datasets containing some disease information (e.g. clinical data of heart disease). The application of the ANN models in the field of medical research has become widely prevailing to help physicians while analyzing complex medical information[16]. The relevant literature about the techniques carrying a list of a number of brands of NNs, but the most widely used NN is the one this study opted i.e. the ANN.

Artificial neural network is multilayer feed-forward neural networks (known as *multi-layer perceptrons*)[17]. The ANN transfers the data in only forward direction as an input and produces the required output (Figure 1).

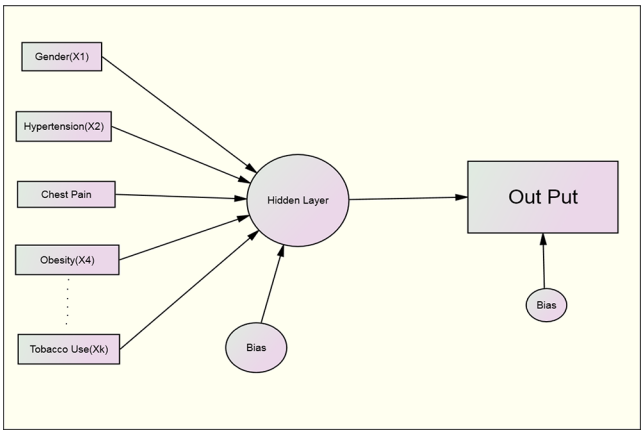

Figure 1. Architecture of neural network.

The aim of this study is to use both the methods on same data with same intensions to model and *scrutinize* the significant risk factors of heart attack in Pakistan, and not to demonstrate any *supremacy of ANN over the logistic regression modeling. The outcomes of the present study demonstrated that the application of neural networks is useful in selecting the important risk factors (ordered in sequence of importance) out of the lot and the findings are supported*[11].

2. Material and Methods

This cross-sectional study is carried at various hospitals to inspect the incidence of the myocardial infarction in Pakistan. For this study the samples are chosen from the patients (who had myocardial infarction) and the controls (free from myocardial infarction) from different parts of the country including the federal capital and all the four provincial capitals. For the purpose of collecting a more representative sample of the target population, hospitals from all four provincial capitals (Lahore, Peshawar, Karachi, and Quetta) are included in the sampled population. In this study sampled population is Lahore, Peshawar, Karachi, Rawalpindi, Quetta and Islamabad. The study included the biggest city from each of the four provinces as well as national capital i.e. Islamabad along with its twin city Rawalpindi. The hospitals from these cities are surveyed for collection of the required samples. Sample of size 2,000 (1,000 cases and 1,000 controls) is collected from the selected cardiac departments from the 4 provinces of the country. The case-control ratio is 1:1. For allocation of sample size to the four provinces sampling criterion of *Proportional Allocation* is used. Since the population of Punjab consists of more than 50% of the total population of the country, therefore 50% subjects (1,000) are chosen from Punjab province and remaining 1000 subjects are chosen from the three provinces of KPK (22.8%), Sindh (16.6%) and Baluchistan (10.6%). All the 2,000 samples are selected by the principal author in order to minimize the bias. The sampling Inclusion criteria for the subjects are as under:

1. Mandatory subject's/ individual's consent.
2. In case of married female subject, those were considered who didn't have pregnancy (because abnormal BMI, blood pressure, sleep or blood sugar etc readings could be due to pregnancy).
3. Age not less than 20 years (because the event of MI is likely to attack individuals after 20 years of age and the risk increases with increasing age).
4. Only complete and adequate questionnaires were included in the study.

The complete list of the variables in this study is; gender (GN), province (PR), age (AG), living environment (LIE), marital status (MS), headache (HD), income class (IC), ethnicity (EN), year of education (YOE), sleeping duration (SD), family history (FH), obesity (BMI), cholesterol level (HLDL), smoking/tobacco use (TOB), physical activity (PA), eating habit (EH), hypertension (HBP), diabetes mellitus (DM), breathing problem (BR), chest pain (CP), fried food (FF), fish eating frequency (FS), fruit eating (FT), vegetable eating (VG), soft drinks intake (SR), lipid lowering medication (LLM), high blood pressure medication (HBPM), daily aspirin intake (ASP), major sad event (ET) and easily angered (AR). Information about the risk factors like, marital status, age, eating habits, education, living environment, income class, physical activity etc were asked verbally (primary data) from the subjects while data for variables like, BMI, blood sugar, blood pressure, HDL, LDL, diabetes mellitus etc were taken from the patients (cases) files (secondary source) available at hospitals. Most of the controls of this study were the ones who were there in hospitals with their patients (cases). Data for obesity (measured from BMI), hypertension (measured from HBP), diabetes mellitus (measured from blood sugar) and cholesterol level (measured from HDL and LDL) is recorded in the line with *American Heart Association* criterion. All the analysis and statistical tests are performed using IBM *SPSS and AMOS* version 18. Logistic regression and neural network techniques are employed to catch the important / significant risk factors of myocardial infarction. Comparison of the results obtained by LR and NN is also made.

3. Results

In this study 2,000 subjects (1,000 cases + 1,000 controls) are analyzed of ages between 20 to 75 years. The sample of 2,000 includes both genders (1,116 females + 884 males). The study has chosen 30 risk factors/etiologies of myocardial infarction. These risk factors/ etiologies are based on Socio-demographic, Clinical, Eating behavior and life style practices, Medical history and medication and Psychosocial pressures and stress (Figure 2). All the data

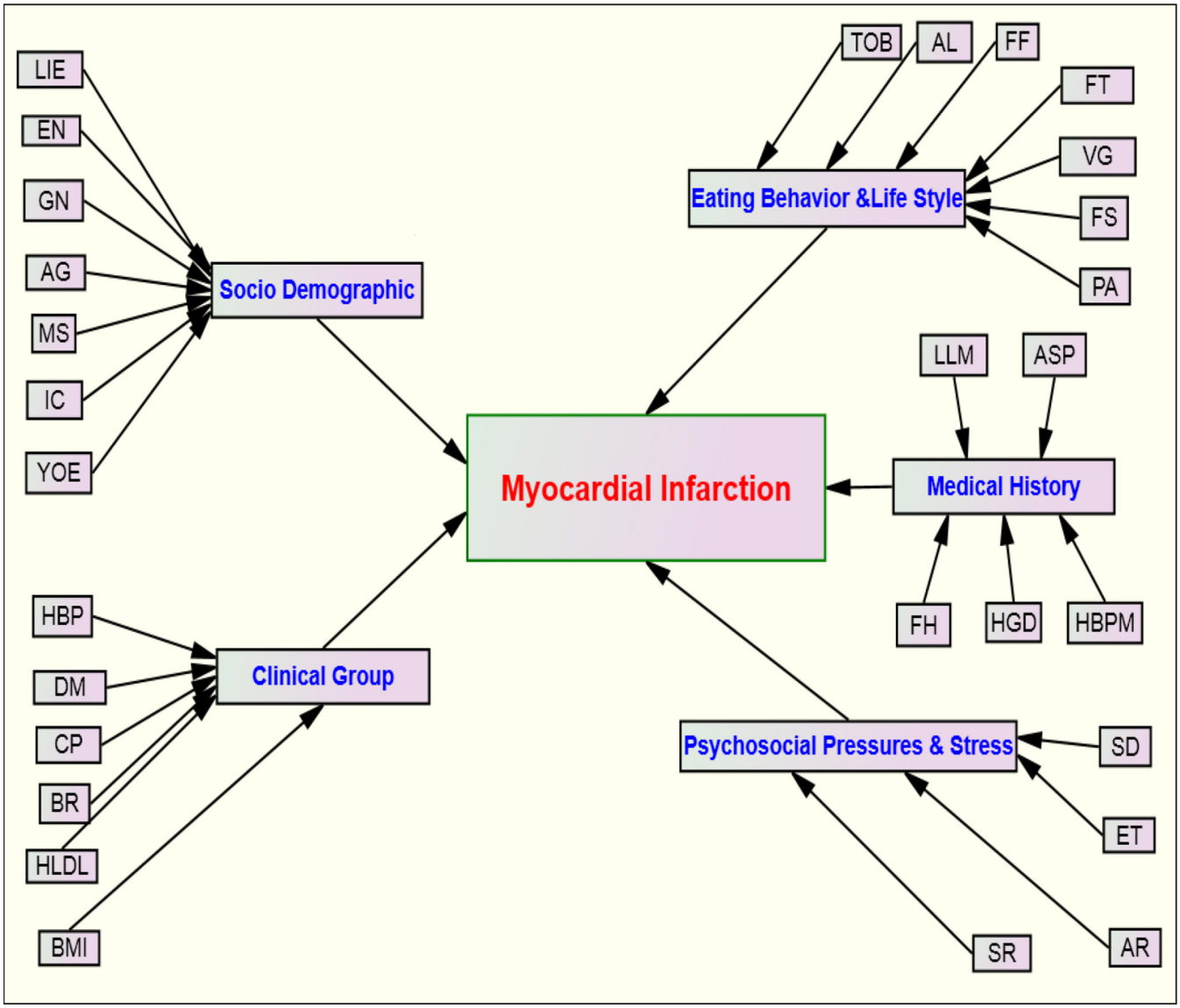

Figure 2. Risk factors/etiologies leading to MI.

for this study is collected on qualitative scale either nominal or ordinal.

Initially, total 36 etiologies/variables are included in the run of model construction on the complete data collected from socio-demographic group, clinical group, eating behavior and life style practices, medical history and medication, psychosocial pressures and stress characteristics. Later on, multivariate Logistic Regression model is run with 28 independent variables after the univariate analysis.

Table 1 presents the results (like; Odds ratios and Confidence Intervals) obtained from the LR model for the significant variables. The table shows that out of the total 36 variables (initially included) only 16 are finally selected in the multivariate analysis. The list of prospective risk factors is; anti-hypertensive medication (HBPM), older age (AG), chest pain (CP), vegetable intake (VG), obesity (BMI), atherosclerosis (HLDL), smoking (TOB), physical activity (PA), alcohol use (AL), hypertension (HBP), diabetes mellitus (DM), breathing problem (BR), family history (FH), fish intake (FS), gender (GN), psychosocial pressure (ET), often intense anger (AR). It is apparent from the positive sign of regression coefficients that all the variables are positively correlated to myocardial infarction except two variables i.e. fish intake and vegetable intake which are negatively associated with MI the odds ratio of OR=8.96 means that a person with obesity is at 8.96 times higher risk of an MI as compared to

Table 1. Output from multiple LR model for overall data carrying significant variables

Variable	B	SE	Wald	P-value	OR	95.0% C.I. for OR	
						Lower	Upper
GN	0.615	0.45	1.868	<0.001	1.850	1.079	7.845
FH	0.627	0.413	2.305	0.001	1.872	1.097	6.071
BMI	2.193	0.977	5.038	<0.001	8.968	4.973	28.859
HLDL	1.002	0.634	2.49	0.014	2.722	1.364	9.353
TOB	1.291	0.697	3.432	0.037	3.635	0.928	14.241
PA	0.230	0.988	0.053	0.019	1.256	0.478	5.187
AL	0.972	1.431	0.487	0.032	2.642	1.248	7.512
HBP	2.122	0.677	9.825	<0.001	8.345	2.334	30.648
DM	1.631	0.694	2.350	<0.001	5.118	2.981	21.327
BR	1.606	0.72	2.231	<0.001	4.985	1.652	25.488
CP	2.645	1.1	2.405	<0.001	14.089	4.285	41.458
FS	-2.452	0.687	12.722	<0.001	0.086	0.022	0.331
VG	-0.695	0.367	3.582	0.058	0.499	0.243	1.025
SR	1.012	0.53	1.901	0.013	2.745	1.326	8.581
ET	0.460	0.218	2.113	0.014	1.585	0.381	4.897
AR	1.981	0.646	9.411	0.002	7.247	2.045	25.685
Constant	-32.095	5.349	36	0	0		

an obesity free person, by keeping all the other risk factors constant. All the other OR of various risk factors are having the same interpretation for their respective OR values.

In the line with[18-19] the ANN is implied to the data, by taking the risk factors as inputs and the presence or absence of myocardial infarction as outputs. The output obtained from the ANN fitted model is shown below in Figure 3, which is run on all the factors mentioned in the Methodology section. The graph is attained (by ANN) after plotting the normalized importance of all the variables. The findings presented 17 important/significant risk factors of the disease. Figure 3 explores "Chest pain" as the most significant variable followed by Diabetes mellitus, Gender, Obesity, Physical inactivity, Less use of vegetables, disturbed cholesterol, hypertension, Fish intake, Psychosocial event, Breathing problem, Fried food, Family history, Alcohol intake, Soft drinks, Less intake of fruits, Easily angered, Tobacco use.

The normalized importance is the ratio of importance values (obtained from the ANN model) to the biggest importance value, and then presenting in percentages (Table 2). On comparison the findings of ANN and LR model are almost the same here because the important variables are the ones which are already found significant by LR model.

The LR presented 16 significant variables (risk factors/etiologies) whereas 17 (Table 2) are selected as important ones by the ANN (containing the same 16 which are chosen by the LR). This Table is showing normalize

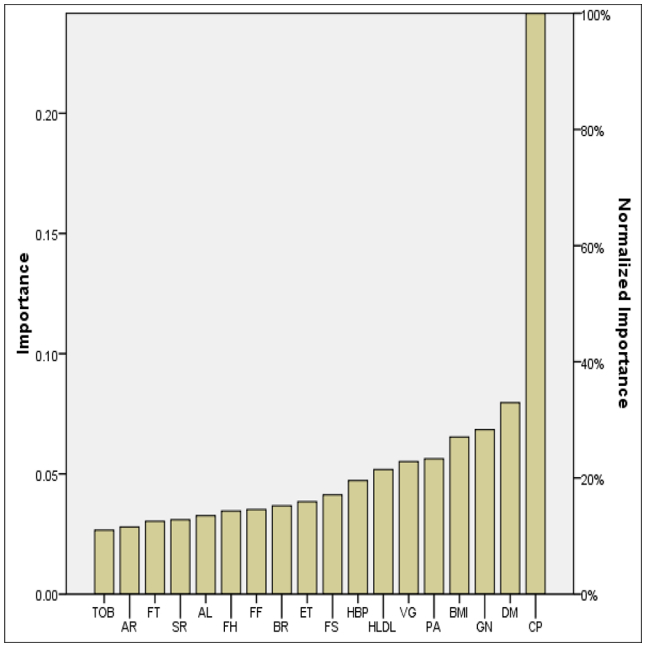

Figure 3. Independent variable importance chart from ANN model.

Table 2. Independent variable importance

Variables	Importance	Normalized Importance
Gender	0.068	28.310%
Family History	0.035	14.326%
Body mass index	0.065	27.049%
H/L Density Lipoprotein	0.052	21.465%
TOB	0.027	11.004%
Physical activity	0.056	23.302%
Alcohol use	0.033	13.502%
High Blood pressure	0.047	19.562%
Mellitus Diabetes	0.080	32.960%
Breathing problem	0.037	15.221%
Chest Pain	0.242	100.000%
Fried Food	0.035	14.566%
Fish eating habit	0.041	17.079%
Fruit eating habit	0.030	12.554%
Vegetable eating habit	0.055	22.831%
Psychosocial Factor	0.038	15.919%
Easily angered	0.028	11.567%

importance of each significant variable in percentages. All those variables are significant whose individual importance is atleast 10%.

4. Comparison of Logistic Regression with ANN Model

The comparison of both the outputs has become more interesting, as 17 important variables are chosen by the ANN containing all the 16 significant etiologies/risk factors which are selected by the LR model (Table 3). Only one variable differs between the two outputs, i.e. fried food intake. The ANN selected this variable as important one is development of the disease whereas LR dropped this one as insignificant. The rest of 16 variables are exactly the same. These 16 risk factors are: income class (IC), fish intake (FS), obesity (BMI), psychosocial pressure (ET) often intense anger (AR) smoking (TOB), physical activity (PA), alcohol use (AL), gender (GN), family history (FH), breathing problem (BR), chest pain (CP), vegetable intake (VG), diabetes mellitus (DM), hypertension (HBP), atherosclerosis (HLDL) and older age (AG) as shown in Table 2.

5. Discussion

The present study is exemplary in its application wherein the two different techniques (one parametric and other non-parametric) are used for analyzing the epidemiological data. On the other hand, it can also be said that this study presents how to apply an ANN versus LR in diagnosing the important risk factors/ etiologies in an epidemiological research like heart attack etc. The study does not aim to demonstrate superiority of the ANN over the LR except to compare and contrast the results obtained by the two methods. On comparison, results from LR and ANN models which are built on overall data, matching risk factors are identified by both models, apart from one variable i.e. fried food. On combining the results from both models, 16 risk factors are strongly associated with myocardial infarction. It can be concluded that a concordant set of risk factors, from both LR and ANN models is

Table 3. Summary of comparison of LR *vs* ANN

Model Parameter	Logistic Regression	Neural Network
Potential variable selected	1. Gender 2. Family History 3. Body mass index 4. Cholesterol level 5. Tobacco 6. Physical activity 7. Alcohol use 8. High Blood pressure 9. Mellitus Diabetes 10. Breathing problem 11. Chest Pain 12. Fish eating habit 13. Fruit eating habit 14. Vegetable eating habit 15. Sad event 16. Easily angered	1. Gender 2. Family History 3. Body mass index 4. Cholesterol level 5. Tobacco 6. Physical activity 7. Alcohol use 8. High Blood pressure 9. Mellitus Diabetes 10. Breathing problem 11. Chest Pain 12. Fried Food 13. Fish eating habit 14. Fruit eating habit 15. Vegetable eating habit 16. Sad event 17. Easily angered

identified which implies that ANN is a useful adjunctive method to identify risk factors for myocardial infarction in Pakistan. Overall both models equally performed. All the significant risk factors obtained using both the settings/models of patients are summarized/compared in Table 3 the same 16 risk factors are turned out as most consistently identified risk factors. Hence, these can be declared as the most common and general risk factors of a heart attack in the country.

6. Conclusion

The primary objective of the study is achieved by exploring the important risk factors of MI in context of Pakistan using both the techniques i.e. LR and NN. Moreover, the present research has also held the consideration of the investigators to look into the unconventional technique of ANN alongside the orthodox approach of logistic regression for analyzing the risk factors of any disease for better insight and comparison purpose.

7. References

1. Antman E, Bassand JP, Klein W, Ohman M, Sendon JLL, Rydén L, Tendera M. Myocardial infarction redefined-a consensus document of the Joint European Society of Cardiology/American College of Cardiology committee for the redefinition of myocardial infarction, Journal of the American College of Cardiology. 2000; 36(3):959-69. https://doi.org/10.1016/S0735-1097(00)00804-4.
2. Chowdhury DR, Chatterjee M, Samanta RK. An artificial neural network model for neonatal disease diagnosis, International Journal of Artificial Intelligence and Expert Systems (IJAE). 2011; 2(3):96-106.
3. Sgourakis G, Gockel I, Lyros O, Lanitis S, Dedemadi G, Polotzek U, Lang H. The use of neural networks in identifying risk factors for lymph node metastasis and recommending management of t1b esophageal cancer, The American Surgeon. 2012; 78(2):195-206. PMid: 22369829.
4. Jabbar MA, Deekshatulu BL, Chandra P. Classification of heart disease using artificial neural network and feature subset selection, Global Journal of Computer Science and Technology Neural and Artificial Intelligence. 2013; 13(3):1-11.
5. Jajoo R, Mital D, Haque S, Srinivasan S. Prediction of hepatitis c using artificial neural network 7th International Conference on Control, Automation, Robotics and Vision; 2002, 3. p. 1545-50. https://doi.org/10.1109/ICARCV.2002.1235004.
6. Pedersen SM, Jørgensen JS, Pedersen JB. Use of neural networks to diagnose acute myocardial infarction I Methodology, Clinical Chemistry. 1996; 42(4):604-12. PMid: 8605679.
7. Heydari ST, Ayatollahi SMT, Zare N. (2012). Comparison of artificial neural networks with logistic regression for detection of obesity, Journal of Medical Systems. 2012; 36(4):2449-54. https://doi.org/10.1007/s10916-011-9711-4.
8. Wan-dong H, Yi-feng J, Dang W, Tan-zhou C, Qi-huai Z. Use of artificial neural network to predict esophageal varices in patients with HBV related cirrhosis, Hepatitis Monthly. 2011; 11(7):544-47.
9. Xue H, Tatsumi N, Park K, Shimizu M, Kyojima T, SumiyaY, Sakano D. Searching for risk factors using multilayer neural network as a classifier, Medical Informatics. 1996; 21(3):229-32. https://doi.org/10.3109/14639239609025360.
10. Hakimpoor H, Arshad KAB, Tat HH, Khani N, Rahmandoust M. Artificial neural networks' applications in management, World Applied Sciences Journal. 2011; 14(7):1008-19.
11. How to measure importance of inputs? Date accessed: 23/06/2000. ftp://ftp.sas.com/pub/neural/importance.html.
12. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, Journal of Clinical Epidemiology. 1996; 49(11): 1225-31. https://doi.org/10.1016/S0895-4356(96)00002-9.
13. Vineis P, Rainoldi A. Neural networks and logistic regression: Analysis of a case-control study on myocardial infarction, Journal of Clinical Epidemiology. 1997; 50(11):1309-10. https://doi.org/10.1016/S0895-4356(97)00163-7.
14. Dussol B, Verdier JM, Le Goff JM, Berthezene P, Berland Y. Artificial neural networks for assessing the risk factors for urinary calcium stones according to gender and family history of stone, Scandinavian Journal of Urology and Nephrology. 2007; 41(5):414-18. https://doi.org/10.1080/00365590701365263. PMid: 17853052.
15. Flaherty CW, Patterson DA. Predicting child physical abuse recurrence: comparison of a neural network to logistic regression, Journal of Technology in Human Services. 2003; 21(4):93-111. https://doi.org/10.1300/J017v21n04_06.
16. Burke HB. Artificial neural networks for cancer research: outcome prediction, Seminars in Surgical Oncology. 1994; 10(1):73-79. https://doi.org/10.1002/ssu.2980100111. PMid: 8115788.
17. Maimon O, Rokach L. Data mining and knowledge discovery handbook, Springer. 2005; 1-1306. https://doi.org/10.1007/b107408, https://doi.org/10.1007/0-387-25465-X_1.
18. Mohamed N, Ahmad WMAW, Aleng NA, Ahmad MH. Assessing the efficiency of multilayer feed-forward neural network model: Application to body mass index data, World Applied Sciences Journal. 2011; 15(5):677-82.
19. Zurada J, Lonial S. Comparison of the performance of several data mining methods for bad debt recovery in the healthcare industry, Journal of Applied Business Research (JABR). 2011; 21(2):1-18. https://doi.org/10.19030/jabr.v21i2.1488.