Artificial Intelligence Evolution
http://ojs.wiserpub.com/index.php/AIE/

UNIVERSAL WISER PUBLISHER

Research Article

MLMI: A Machine Learning Model for Estimating Risk of Myocardial Infarction

Subhagata Chattopadhyay

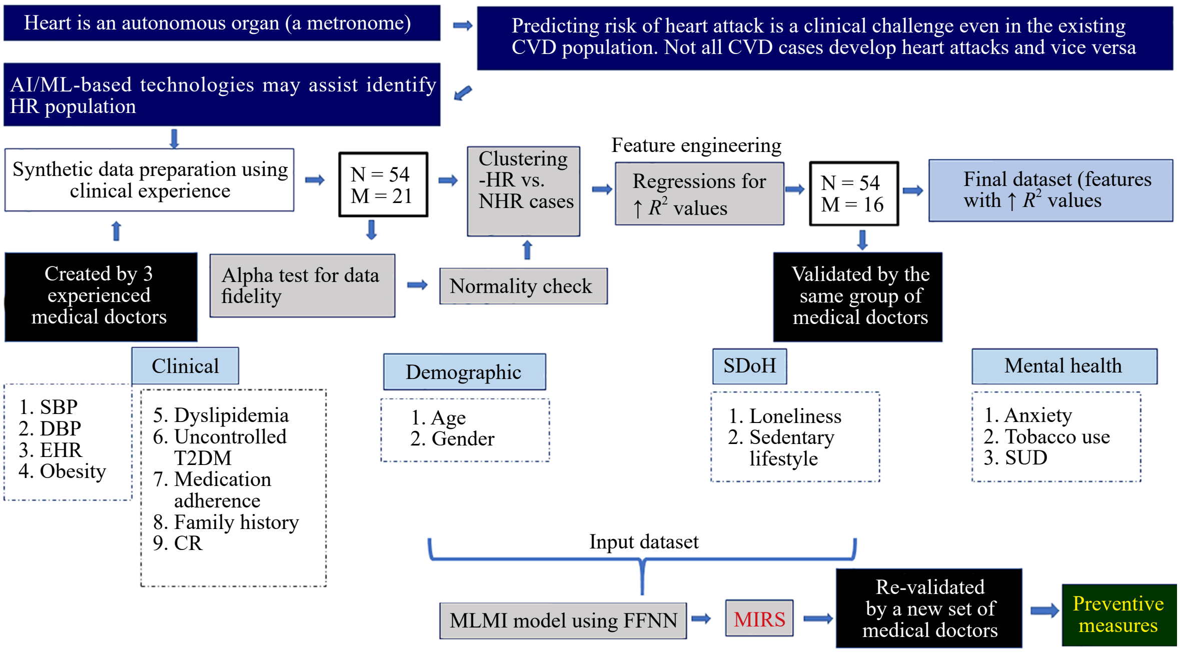
Formerly with Department of Computer Science and Engineering, Gandhi Institute of Technology and Management (Deemed to be) University, Bangalore 561203 KA India
Email: subhagata.chattopadhyay2017@gmail.com

Received: 26 September 2023; **Revised:** 9 January 2024; **Accepted:** 10 January 2024

Abstract: Cardiovascular diseases (CVD) are a global threat of high morbidity and mortality. Myocardial infarction (MI) due to coronary vessel malfunctions is one of the leading causes of mortality due to CVD. Interestingly, all CVD patients do not develop MI, and vice versa. Clinically, thus, it is a gray area. Therefore, an appropriate MI risk scoring (MIRS) tool could be useful to identify the high-risk (HR) population suffering from CVD. This research paper presents a hybrid machine learning (ML) model (MLMI) to identify MI risk where a) clustering of the CVD population with the help of the Gaussian mixture model (GMM) is used to identify the HR and not high-risk (NHR) groups, b) feature engineering of the members in both the HR and NHR populations using regression method that estimates the coefficient of determination (R^2) to explore significant features to create the model by c) leveraging the R^2 values > 0.7 as the key features of the input dataset to a d) Feed-forward neural network (FFNN) for scoring the risk on a set of synthetic patient data, created by three experienced medical doctors. The myocardial infarction risk scores (MIRS) would assist users in prioritizing the patients needing monitoring and treatment. Finally, the MIRS values are validated by another group of three medical doctors to curb the research bias. The sensitivity, specificity, precision, F_1 scores, and accuracy of the MLMI model are computed to measure its efficiency. With limited input data, the proposed model shows an average accuracy, and precision of 77.33% each, while sensitivity and F_1 score are 100% and 88%, respectively.

DOI: https://doi.org/10.37256/aie.5120243714

Graph abstract


Heart is an autonomous organ (a metronome)
Predicting risk of heart attack is a clinical challenge even in the existing CVD population. Not all CVD cases develop heart attacks and vice versa
AI/ML-based technologies may assist identify HR population
Synthetic data preparation using clinical experience
N = 54
M = 21
Clustering -HR vs. NHR cases
Feature engineering
Regressions for ↑ R^2 values
N = 54
M = 16
Final dataset (features with ↑ R^2 values
Created by 3 experienced medical doctors
Alpha test for data fidelity
Normality check
Validated by the same group of medical doctors
Clinical
1. SBP
2. DBP
3. EHR
4. Obesity
5. Dyslipidemia
6. Uncontrolled T2DM
7. Medication adherence
8. Family history
9. CR
Demographic
1. Age
2. Gender
SDoH
1. Loneliness
2. Sedentary lifestyle
Mental health
1. Anxiety
2. Tobacco use
3. SUD
Input dataset
MLMI model using FFNN
MIRS
Re-validated by a new set of medical doctors
Preventive measures

Keywords: artificial intelligence, machine learning, myocardial infarctions, cardiac risk scoring, myocardial infarction risk scoring, high-risk population

Abbreviations

AI	Artificial intelligence
ANN	Artificial Neural Network
ANS	Autonomic nervous system
CDSS	Clinical decision support system
CI	Confidence interval
CR	Cardiac risk
CRS	Cardiac risk scoring
CVD	Cardiovascular diseases
DBP	Diastolic blood pressure
EHR	End heart rate (average of 2-minutes hear rate)
FFNN	Feed-forward neural network
FP	False positive
FN	False negative
GMM	Gaussian mixture model
HR	High-risk
HRV	Heart rate variability
MI	Myocardial infarction
MIR	Myocardial infarction risk
MIRS	Myocardial infarction risk scoring
NHR	Not high risk
SBP	Systolic blood pressure

SDoH	Social determinant of health
SED_LIFE	Sedentary lifestyle
SUD	Substance use disorders
TP	True positive
TN	True negative

Notations

α	Cronbach's Alpha
i	Number of cases varying from 1 to N
j	Number of predictors varying from 1 to M
m	denotes the number of predictors under each construct
p	Probability value (statistical significance)
R^2	coefficient of determination
W	Weight

1. Introduction

Cardiovascular diseases (CVD) are the leading causes of mortality across the globe. About 17.9 million people die every year due to CVD [1]. Although the prevalence of CVD is high in those over 65 years of age [2], in the present day, the younger population (less than 40 years of age) shows a rising trend of CVD and cardiac death [3]. Myocardial infarctions (MI) are the principal cause of CVD-related deaths [4]. In MI, the heart muscles suffer from a lack of blood supply due to the obstruction of coronary vessels as a consequence of atherosclerotic plaques and/or narrowing of the lumen leading to muscle death eventually [5]. There are several contributing factors behind the development of CVD, which may be broadly divided into genetic predispositions such as a strong family history of CVD, environmental factors, such as tobacco smoking, substance abuse, sedentary lifestyle, comorbidities e.g., uncontrolled diabetes, hypertension, dyslipidemia, associated lung diseases, psychiatric illnesses, and so forth that increase the vulnerability of MI [6]. CVD management is a costly affair and showing an increasing trend with an estimated amount of 1.1 trillion USD by 2035 from 555 billion USD in 2015 [7].

Interestingly, all CVD cases do not develop MI and vice versa and vice versa. Therefore, who will develop MI and who not is a gray area and may require technology support. The normal living heart is a metronome and is run by the seamless commands of the autonomic nervous system (ANS) [8]. The ANS has multimodal inputs both from the environment and within the physiological activities, which are stochastic in nature and might be one of the explanations for this uncertainty. Here, predictive analytics using AI/ML may play a crucial role in unfolding both the diagnostic and prognostic uncertainties.

Artificial intelligence (AI) is to mimic human intelligence using algorithms that learn patterns iteratively [9]. Machine learning (ML) is a subset of AI where machines are trained by algorithms to learn patterns [10]. As the healthcare domain is highly nonlinear and the heart is a metronome, hybrid ML modelings are becoming popular methods for early MI risk estimations, called Myocardial Infarction Risk Scoring (MIRS) both at the individual and population levels.

Current knowledge of Cardiac Risk Scoring (CRS) tools can be found in the comprehensive work of Sofogianni et al. [11], hence it is not repeated in this paper. Interested readers can read the article. However, it is important to note the tools mentioned in that article predict 10-year cardiac risk (CR) based on the scores. However, they have some drawbacks, such as (i) these tools are rule-based and thus, rigid, and a particular rule may not fit into other cases, (ii) following up an individual for 10 years is a stupendous task and some information might be missed, and (iii) within that large window period, many may encounter several other influencing risk factors which were absent during the onset of the studies based on which patients were initially scored. Moreover, a normal heart is a metronome as explained by several works on HRV parameters [12-13] and heart health depends on various external parameters, which are discussed using several data mining techniques [14]. Therefore, instead of possessing predisposing CR factors, not everyone develops MI in real life and vice versa and the work of Sofogianni et al. [11] could have missed these important contexts.

The objective of this paper is to build and implement a hybrid ML model to predict the risk of MI in the population with CVD. The model is termed Machine learning in Myocardial infarction (MLMI). The aim is to develop an MLMI system to classify the CVD population and then mine who would require more intensive care.

2. Methodology

The research had originally been carried out at GITAM (Deemed to be) University Bangalore campus in 2021 during the author's tenure as an Associate Professor in the Department of Computer Science and Engineering. This section discusses the data and metadata for model building, its steps, and validation (refer to Figure 1).

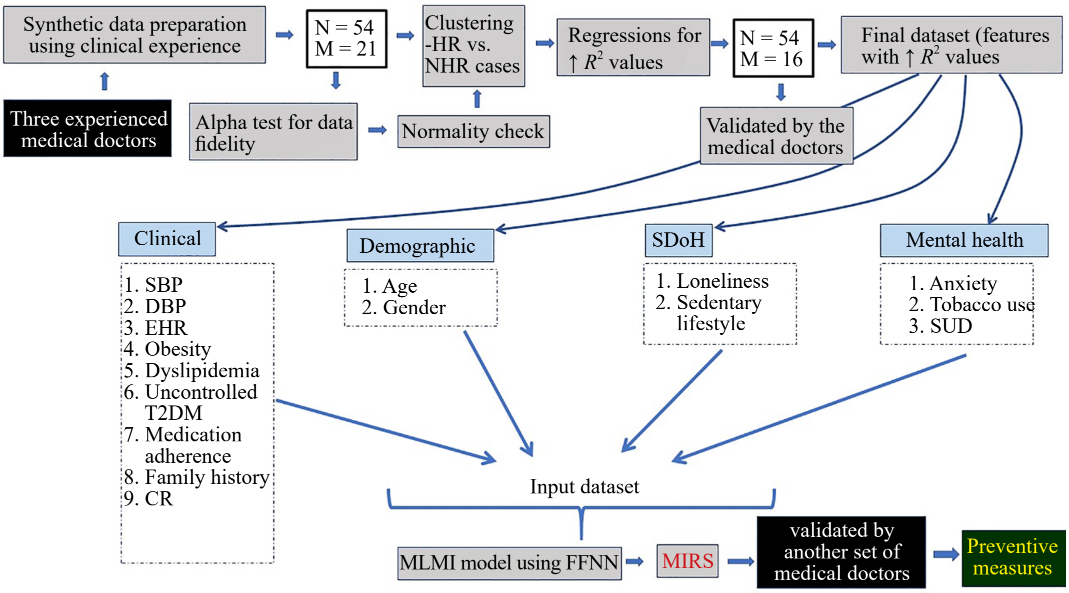


Figure 1. Methodology of MIRS modeling after feature reduction

The above figure depicts the MIRS modeling methodology step-by-step. It shows all the significant features/ predictors under individual constructs to create the final dataset, which has been explored and analyzed to create the FFNN model that calculates the MIRS for each patient. The patients, in turn, are then prioritized based on the calculated MIRS.

2.1 *Data preparation*

Availability of real-world clinical data is a major constraint due to various privacy and compliance-related matters. Thus, in this work, a set of synthetic data (N = 54) is prepared for the study. Three experienced medical doctors put their knowledge into creating it [15]. The aim is to mimic the way doctors practically diagnose and grade a disease using minimal but significant features or parameters obtained from patients' history, examinations, and clinical interpretations (together called the 'clinical eye') rather than including the whole set of features, which is mimicked in this hybrid MLMI model [16]. As desired by them, their identities are not disclosed in the paper.

Each case has 21 predictors (refer to all the variables shown in Figure 2). Later, the number of features is reduced by excluding primary care physician and nurse interactions, depression, financial strength, and city using regressions, as described below. The exclusion process is validated by the medical doctors and just not rely on the statistical technique as they bring vast experience in feature selection while diagnosing clinically. The final set of variables (M = 16) is used for MLMI model-building [MIRS and CRS are synonymous terms. Predictors, features, and variables are synonymous terms]. These variables are divided into four constructs as mentioned below.

(A) Clinical ($m = 8$): (i) SBP, (ii) DBP, (iii) EHR, (iv) Obesity, (v) Dyslipidemia, (vi) Uncontrolled T2DM (UNC_T2DM), (vii) Medication adherence (MEDS), (viii) FAM_HIST (Positive family history of CVD). Here, SBP, DBP, and EHR contain true values while binary values are assigned for the remaining variables, e.g., present is '1' and absent is '0'.

(B) Demographic ($m = 2$): (i) Age and (ii) Gender. Here, Age is the true value, while Males are assigned 1 and Females are assigned 0.

(C) SDoH ($m = 2$): (i) Sedentary lifestyle, (ii) Living alone/single. Here binary values are assigned as present is '1' and absent is '0' for both variables. Finally,

(D) Mental health ($m = 3$): (i) Anxiety, (ii) SUD, and (iii) Tobacco use. Here binary values are assigned as present is '1' and absent is '0' for all three variables.

The Dependent variable ($m = 1$) CR is also assigned binary values as mentioned above (present is '1' and absent is '0'). These values are probabilistic, and determined by the medical doctors.

2.2 *Data mining and model building*

Step 1: Data fidelity test (Cronbach's α test [17]) to note whether the data are internally consistent and suitable for model-building. It is the foremost important step. Equation 1 explains how Cronbach's α test works.

$$\alpha = \frac{\left(r - \overline{c}\right)}{\overline{\upsilon} + \left(r - 1\right) \times \overline{c}} \tag{1}$$

In the above equation 'r' refers to the number of data, $\overline{c}$ is the mean of all covariances between data points, and $\overline{\upsilon}$ is the average variance. The consistency score 'α' is expressed as a value between 0 and 1, where $\alpha \geq 0.8$ is considered ideal while $\alpha \leq 0.5$ is deemed "unacceptable" [18].

Step 2: Normality check is the second most important step to check the data distribution. Here, using the Shapiro-Wilk test (see Equation 2), it is checked [19] where Gaussian distribution refers to normal distribution.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \tag{2}$$

In this Equation, 'i' is the number of observations that vary from 1 to 'n', 'x' is the value of the ordered sample, and 'a' represents tabulated coefficients.

Step 3: Clustering of the population (using GMM) [20] to get two clusters, one for the 'HR' sub-population and the other for the 'NHR' sub-population, based on the patterns based on CR as 'no or 0' or 'yes or 1'. GMM is a probabilistic clustering [21] and CR metadata has a probabilistic distribution. Hence, GMM is chosen in this paper. The train-test split is 9:1. As unsupervised learning algorithms require more training data, 90% is allocated for training of the GMM clustering algorithm. It is important to note that clustering is performed on all 21 variables before feature engineering and feature reduction.

Step 4: Regression of each independent variable on the dependent variable shows the influence of each of them (antecedence) on the outcome (consequence) using Equation 3. High influencing predictors, i.e., predictors having high R^2 or coefficient of determination values are called 'significant predictors' (see equation 3).

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} \tag{3}$$

In this Equation, SS_{RES} and SS_{TOT} refer to the 'sum of squares of residues' and 'total sum of squares', respectively. The term 'residue' refers to the difference between actual and predicted value. Therefore, R^2 values indicate the 'goodness of fit' of the model. In this work, values > 0.5 have been considered 'significant'.

Step 5: Finally, the significant predictors (having the R^2 values > 0.7) are chosen for developing a Feedforward neural network (FFNN) [22] to compute CRS for both NHR and HR clusters. The architecture of the FFNN is described below. The FFNN has been widely used in clinical decision support system (CDSS) design [23] to curb clinical subjectivity in screening and grading complex medical conditions [24].

2.3 *Design of the FFNN (M:1:1)*

• It consists of one Input layer (I) having a set of Input nodes equal to the number of predictors (M, which are 'significant' according to the respective R^2 values) carrying values of either '0' or '1' or true values (e.g., SBP, DBP, EHR) to the input side of the Input nodes. The output side of the Input node is assigned a linear transfer function (refer to Equations 4 and 5, respectively)

$$y(I) = f(x) = ax + c \tag{4}$$

Here 'x' and 'y' are the input and output, 'a' is the coefficient, and 'c' is the constant, respectively. In the case of 'a' equals 1, and 'c' equals '0', the Equation becomes

$$y(I) = x \tag{5}$$

• The proposed FFNN architecture has one Hidden layer (H) with one node. Each of the outputs of the Input node feeds into the input side of the Hidden node and is summated (see Equation 6). The output side of the Hidden node is assigned a linear transfer function (refer to Equation 5) for computing its output.

$$y(H) = \sum_{i,j=1}^{N,M} x \times \beta \tag{6}$$

• Between the I and H, there is a set of connecting weights $W(I/H)$ having the significant predictors β values, obtained from the regressions study. The number of connectors is equal to the number of Input nodes. Here, the input values (x) are multiplied by the β values (see Equation 6).

• Finally, the H is connected to one Output layer (O) having one Output node. There is one connector $W(H/O)$ between the H and O layers. Its value is kept at 1 to process the output of H as it is. The input side of the Output node gets the summated values, one by one case-wise. The output side of the Output node possesses a log-sigmoidal transfer function (refer to Equation 7) calculating the MIRS for each case that passes through the FFNN architecture.

$$y(I) = f(x) = \frac{1}{1+e^{-x}} \tag{7}$$

Here 'e' is the exponential function, while 'x' and 'y' are the input and output variables.

2.5 *Classification using MIRS*

The MIRS values, thus obtained help classify the cases of HR and NHR clusters. Here, MIRS values < 0.35, 0.35-0.65, and > 0.65 refer to 'No or low', 'Moderate', and 'High' risk, respectively. The ML model-derived classes are finally validated by three medical doctors and average Sensitivity (SN), Specificity (SP), Precision (P), F_1-score, and Accuracy (A) of the models are calculated using the following equations 8 through 12.

$$SN = \frac{TP}{TP + FN} \tag{8}$$

$$SP = \frac{TN}{FP + TN} \tag{9}$$

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$F_1 = 2\frac{P \times SN}{P + SN} \tag{11}$$

$$A = \frac{TP}{TP + FP + TN + FN} \tag{12}$$

3. Results

In this section, the results of the experiments and the performance of the FFNN model are shown. It is noteworthy that the synthetic dataset on which the experiment has been performed is created using the 'clinical eyes' of three experienced senior medical doctors. The dataset is structured into four constructs, each having predictors/features/variables (see section IIA and Figure 1). Under the 'Clinical' construct, the 'risk of MI' (Yes/No) is the dependent variable, while all other features under all constructs are the independent variables.

Data fidelity test: The Cronbach's α value for the dataset is 0.82 meaning that the dataset is reliable and internally consistent and can be used for modeling [18, 25].

Data normality test: Shapiro Wilk test result shows the stat value is 0.69 with p-value < 0.00 (CI = 95%), which means the data distribution is not Gaussian/normal.

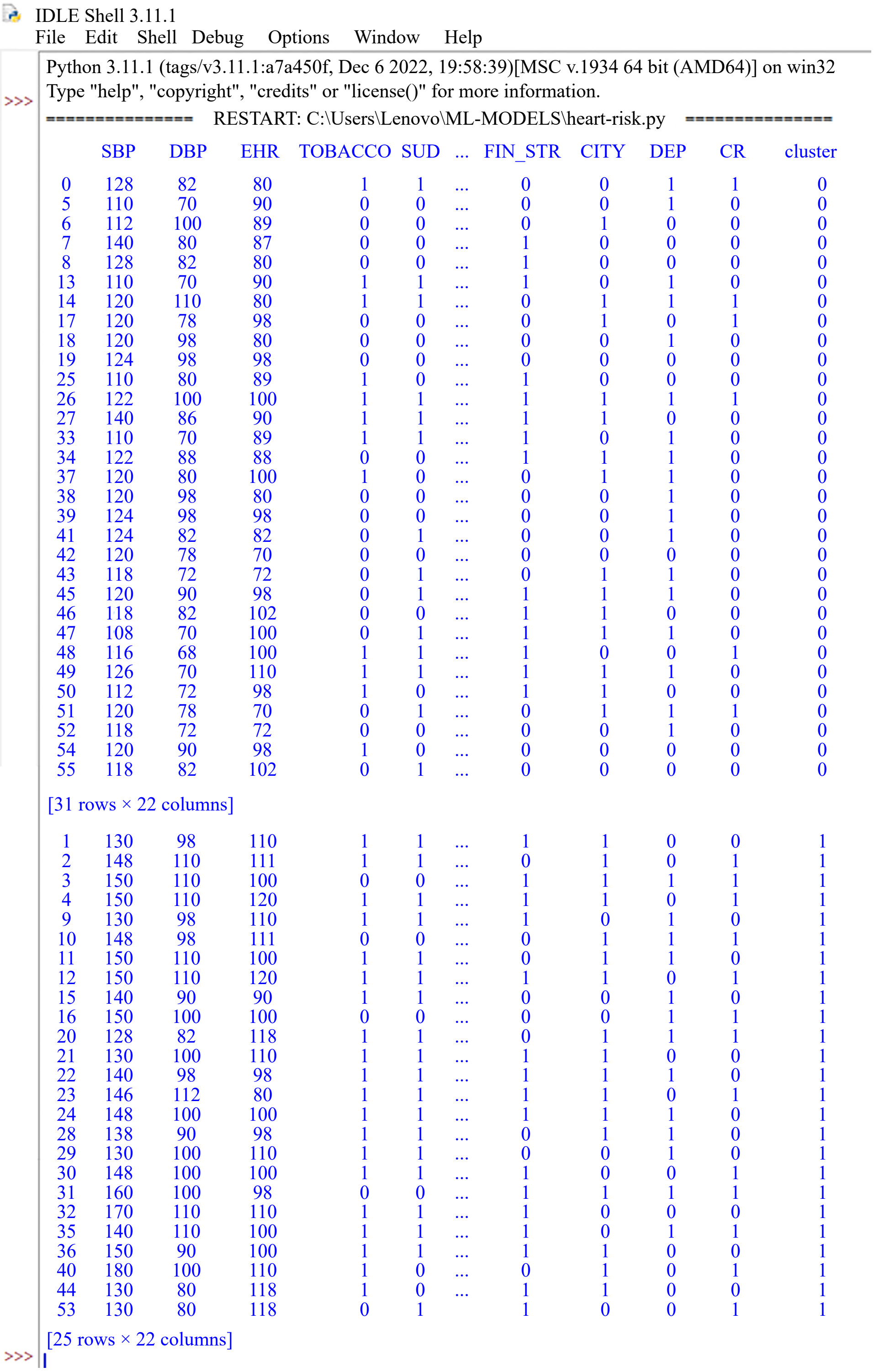
To realize the central tendency of the dataset, descriptive statistics can be seen in Table 1.

Table 1. Descriptive statistics of the final CR dataset

	Mean	Std	Min	25%	50%	75%	Max
SBP	132.28	16.35828	108	120	129	147.5	180
DBP	91.72	13.38906	68	80.5	98	100	112
EHR	97.24	12.39249	70	89	99	108	120
TOBACCO	0.6	0.494872	0	0	1	1	1
SUD	0.6	0.494872	0	0	1	1	1
ANXIETY	0.64	0.484873	0	0	1	1	1
AGE	55.16	10.22274	32	46.25	55	63.25	78
GENDER	0.48	0.504672	0	0	0	1	1
SINGLE	0.64	0.484873	0	0	1	1	1
MEDS	0.54	0.503457	0	0	1	1	1
UNC_T2DM	0.54	0.503457	0	0	1	1	1
DYSLIPIDEMIA	0.52	0.504672	0	0	1	1	1
FAM HIST	0.56	0.501427	0	0	1	1	1
OBESITY	0.52	0.504672	0	0	1	1	1
SED_LIFE	0.52	0.504672	0	0	1	1	1
CR	0.48	0.504672	0	0	0	1	1

Table 1 shows the central tendency of the features in the synthetic dataset. It can be highlighted that the average SBP, DBP, and EHR are higher than the normal values. On average, 60% of the population has a history of SUD and tobacco use, 64% are single and suffer from anxiety, while the remaining features are marginally high (little above 50%). The gender distribution and the CR are close to equal (48% each) in the study population.

Clustering: GMM on an original dataset (M = 21) yields two clusters-the NHR (cluster-0 or the first cluster) and HR (cluster-1 or the second cluster) that represent the second cluster. Figure 2 shows the members of each cluster.



```
    Python 3.11.1 (tags/v3.11.1:a7a450f, Dec 6 2022, 19:58:39)[MSC v.1934 64 bit (AMD64)] on win32
>>> Type "help", "copyright", "credits" or "license()" for more information.
    ================ RESTART: C:\Users\Lenovo\ML-MODELS\heart-risk.py ================
         SBP   DBP   EHR  TOBACCO  SUD  ...  FIN_STR  CITY  DEP  CR  cluster
    0    128    82    80        1    1  ...        0     0    1   1        0
    5    110    70    90        0    0  ...        0     0    1   0        0
    6    112   100    89        0    0  ...        0     1    0   0        0
    7    140    80    87        0    0  ...        1     0    0   0        0
    8    128    82    80        0    0  ...        1     0    0   0        0
    13   110    70    90        1    1  ...        1     0    1   0        0
    14   120   110    80        1    1  ...        0     1    1   1        0
    17   120    78    98        0    0  ...        0     1    0   1        0
    18   120    98    80        0    0  ...        0     0    1   0        0
    19   124    98    98        0    0  ...        0     0    0   0        0
    25   110    80    89        1    0  ...        1     0    0   0        0
    26   122   100   100        1    1  ...        1     1    1   1        0
    27   140    86    90        1    1  ...        1     1    0   0        0
    33   110    70    89        1    1  ...        1     0    1   0        0
    34   122    88    88        0    0  ...        1     1    1   0        0
    37   120    80   100        1    0  ...        0     1    1   0        0
    38   120    98    80        0    0  ...        0     0    1   0        0
    39   124    98    98        0    0  ...        0     0    1   0        0
    41   124    82    82        0    1  ...        0     0    1   0        0
    42   120    78    70        0    0  ...        0     0    0   0        0
    43   118    72    72        0    1  ...        0     1    1   0        0
    45   120    90    98        0    1  ...        1     1    1   0        0
    46   118    82   102        0    0  ...        1     1    0   0        0
    47   108    70   100        0    1  ...        1     1    1   0        0
    48   116    68   100        1    1  ...        1     0    0   1        0
    49   126    70   110        1    1  ...        1     1    1   0        0
    50   112    72    98        1    0  ...        1     1    0   0        0
    51   120    78    70        0    1  ...        0     1    1   1        0
    52   118    72    72        0    0  ...        0     0    1   0        0
    54   120    90    98        1    0  ...        0     0    0   0        0
    55   118    82   102        0    1  ...        0     0    0   0        0

    [31 rows × 22 columns]

    1    130    98   110        1    1  ...        1     1    0   0        1
    2    148   110   111        1    1  ...        0     1    0   1        1
    3    150   110   100        0    0  ...        1     1    1   1        1
    4    150   110   120        1    1  ...        1     1    0   1        1
    9    130    98   110        1    1  ...        1     0    1   0        1
    10   148    98   111        0    0  ...        0     1    1   1        1
    11   150   110   100        1    1  ...        0     1    1   0        1
    12   150   110   120        1    1  ...        1     1    0   1        1
    15   140    90    90        1    1  ...        0     0    1   0        1
    16   150   100   100        0    0  ...        0     0    1   1        1
    20   128    82   118        1    1  ...        0     1    1   1        1
    21   130   100   110        1    1  ...        1     1    0   0        1
    22   140    98    98        1    1  ...        1     1    1   0        1
    23   146   112    80        1    1  ...        1     1    0   1        1
    24   148   100   100        1    1  ...        1     1    1   0        1
    28   138    90    98        1    1  ...        0     1    1   0        1
    29   130   100   110        1    1  ...        0     0    1   0        1
    30   148   100   100        1    1  ...        1     0    0   1        1
    31   160   100    98        0    0  ...        1     1    1   1        1
    32   170   110   110        1    1  ...        1     0    0   0        1
    35   140   110   100        1    1  ...        1     0    1   1        1
    36   150    90   100        1    1  ...        1     1    0   0        1
    40   180   100   110        1    0  ...        0     1    0   1        1
    44   130    80   118        1    0  ...        1     1    0   0        1
    53   130    80   118        0    1             1     0    0   1        1

    [25 rows × 22 columns]
>>> |
```

Figure 2. Members within the NHR and HR clusters

The first cluster (HR cluster) has 30 members, while the second cluster (NHR cluster) possesses 24 members. On analysis, it can be seen that the NHR cluster has 6 outliers (accuracy of 81%) while the HR cluster has 12 outliers (accuracy of 52%). Therefore, out of 54 members, 38 members (i.e., 70%) have been correctly clustered by the GMM algorithm.

Significant CR predictors are engineered using R^2 values for both the NHR and HR clusters. For both clusters, 16 predictors can fit well with the model in explaining CR (R^2 values are > 0.7). Therefore, all these predictors are a good fit for the model and are leveraged to create the input dataset for the FFNN classifier (refer to Figure 1).

One FFNN-based CRS calculator model is created with all the predictors as the input to it. Figure 3 shows the architecture of the proposed FFNN and the hand calculation of the CRS values (also known as MIRS values) can be seen in the following part of the text.

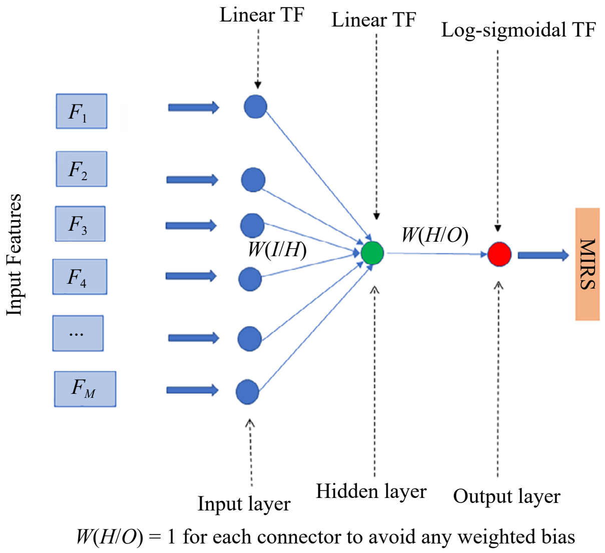


Figure 3. The FFNN architecture where features having higher R^2 values consist of the input dataset

In this figure, the input dataset to the FFNN model has the final sixteen features to compute MIRS for each case.

3.1 *FFNN in CRS/MIRS calculations resembling human cognition*

Step 1: Binarized or true values of all 16 predictors are passed through each Input node, case-wise in ‘I’ using the linear transfer function. This is the layer to capture the information from external sources.

Step 2: These values are multiplied by the respective R^2 values (which are the human perception) assigned to the $W(I/H)$ connectors. Here these values are > 0.7 as mentioned above. The information-perception product represents the thought (metacognition) that is a product of perception towards the environmental inputs.

Step 3: The multiplied value becomes the input to the Hidden node in the Hidden layer (H), where the thought is processed by summating all thoughts (representing human cognition).

Step 4: The output of the Hidden node is assigned a linear transfer function assigned to it.

Step 5: The outputs of the Hidden nodes are multiplied by the value of one $W(H/O)$ connector, which is again ‘1’ to avoid any ‘bias’ (the belief). However, bias may be assigned when the target is available for iterative learning, which is not happening here.

Step 6: The cognition is the input to the Output node (O), and finally.

Step 7: The output of the Output node passes through the log-sigmoidal conversion (decision/action) and this is the calculated CRS/MIRS value for each case.

Step 8: The above values thus obtained are validated by three medical doctors, and finally.

Step 9: The Sensitivity, Specificity, Precision, F_1-score, and Accuracy of the proposed model are calculated (see Tables 2, 3, 4).

3.2 *Observation*

Out of 56 cases, 6 cases (25%) show high MIRS and are the High-need, High-risk subpopulation. They are case numbers 18 and 20 (MIRS 0.7831), 19 (MIRS 0.7776), 4 and 7 (MIRS 0.7748), and 10 (MIRS 0.7692) represent the top most vulnerable cases with increased risk of MI and need a continuous monitoring and high-quality preventive care so that they do not develop MI attacks. Tables 2-4 shows the cases with predictors and the model's output validated by three different doctors' opinions (double-blind control) to curb the 'interpretation bias'. They are separate from those who built the dataset.

Tables 2, 3, 4 show doctor-wise confusion matrices. Using equations 9-12, SN, SP, P, and A are calculated. As there are no TN cases, SP has not been calculated. In the tables, the numbers indicate the respective population size diagnosed as CR (whether negative or positive) by the model and the doctors.

Table 2. Confusion matrix (Model Vs. D1)

	Positive risk (D1)	Negative risk (D1)
Positive risk (Model)	5	1
Negative risk (Model)	0	0

SN = 100%, P = 83%, A = 83%, F_1 = 90.71%

Table 3. Confusion matrices (Model Vs. D2)

	Positive risk (D2)	Negative risk (D2)
Positive risk (ML model)	4	2
Negative risk (Ml model)	0	0

SN = 100%, P = 66%, A = 66%, F_1 = 79.51%

Table 4. Confusion matrices (Model Vs. D3)

	Positive risk (D1)	Negative risk (D1)
Positive risk (ML model)	5	1
Negative risk (Ml model)	0	0

SN = 100%, P = 83%, A = 83%, F_1 = 90.71%

Therefore, the models' average SN, P, F_1, and A are 100%, 77.33%, 88%, and 77.33%, respectively.

4. Discussion

The prevalence of MI is 23.3% in the population and rising [26]. It is a life-threatening event due to coronary artery malfunctions [27] leading to fatal arrhythmias and sudden death [28]. The pathophysiology involves multimodal causative factors, ranging from positive family history, and various lifestyle issues to addictions and comorbidities, such as essential hypertension, type-2 diabetes mellitus, dyslipidemia, and so forth. The heart is an autonomous organ, a metronome. Thus, it functions abnormally, which is considered 'normal' during one's state of life unless the abnormality reaches a pathological state. Not everyone suffering from CVD develops MI and vice versa. This is the gray area and the challenge here is to identify the vulnerable population using technology. Medical doctors can not handle thousands of these populations in real-time and that is why ML models can be useful. This paper showcases a simple generic method of screening this population using a hybrid ML model consisting of clustering, finding high coefficients of determination (i.e., the R^2 values) for each feature, and finally leveraging the high coefficients to develop the input datasets of the FFNN (having the architecture of M:1:1), which eventually linearly scores the CR of each case/patient/member. This observation can help doctors, nurses, and caregivers to prioritize who should be given more aggressive measures to prevent MI in due course of time. This study has identified the factors raising the vulnerability of CR. Therefore, the authors postulate to give equal importance to all the features/predictors initially and then eventually reduce the feature numbers based on statistical measures (finding the coefficients of determinants using regressions) mentioned in this work. Appropriately institution of medications, such as antihypertensives, beta-blockers, lipid-lowering agents, antidiabetic drugs, judicious use of anxiolytics, measures to minimize use of tobacco and substance use, and regular measured workouts (contrary to sedentary lifestyles) are necessary to reduce such risk.

The contribution of this study is to leverage coefficients of determinants as the $W(I/H)$ to calculate the scores mimicking human perception-led decision-making. The model shows an average accuracy of 77% on a limited dataset, which demonstrates its efficacy. It is expected that with more cases, the model's accuracy would excel further. The scope of research bias has been curbed using a 'double-blind' validation methodology where a group of three doctors have created the clinical dataset and another group of three doctors validated the results, obtained by the model.

However, it is to reiterate that the accuracy could be higher with a larger dataset and more predictors, which is the limitation of this work, and the author is working on it as the extension of this research. Limited access to relevant published research papers on CR predictions is another limitation of this paper.

5. Conclusion

The proposed MLMI hybrid model can be used by both the provider and payer organizations to identify the high-need populations, e.g., in this study, case numbers 18 and 20 (MIRS 0.7831), 19 (MIRS 0.7776), case numbers 4 and 7 (MIRS 0.7748), and case number 10 (MIRS 0.7692), respectively. Using the scoring system, the doctors, nurses, social workers, and others can prioritize the patients/members having the highest risks, which helps develop optimal care plans, e.g., controlling hypertension (average SBP, DBP, and heart rate 132.28 mmHg, 91.72 mmHg, and 97.24, respectively), controlling T2DM and dyslipidemia, ensuring medication adherence, and taking care of anxiety, SUD and tobacco use, and other factors mentioned in this work. The objective would be to prevent emergency admissions and curb the morbidity, mortality risks, and cost.

The GMM clustering method can segregate NHR and HR sub-groups with 70% accuracy, which in turn, facilitates feature engineering and reduction with the help of regressions where high R^2 values (values > 0.7) signify the predominant variables/features/predictors. It is worth noting that about 24% feature reduction may proportionally reduce the algorithmic and computational complexities.

With a limited dataset, the MLMI model shows an average accuracy, and precision of 77.33% each, while sensitivity and F_1 score are 100% and 88%, respectively.

It is also noteworthy that the model is generic enough to score the risk of other chronic diseases, such as T2DM leading to nephropathy and retinopathy as two key complications, depression to suicide, hypertension to CVA and heart failure, etc.

It would be interesting to see how the model performs in real-world data with more cases and features in the future.

Funding

The study has not received any funding from any source.

System information

Python 3.11.1 is used for coding and algorithm execution on Windows 10 Pro 64-bit OS.

Conflict of interest

The author declares no competing financial interest.

References

[1] World Health Organization. *Cardiovascular Diseases*. Available from: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 [Accessed 26th June 2023].

[2] Yazdanyar A, Newman AB. The burden of cardiovascular disease in the elderly: Morbidity, mortality, and costs. *Clinics in Geriatric Medicine*. 2009; 25(4): 563-577.

[3] Napoli N. *Heart Attacks Increasingly Common in Young Adults*. American College of Cardiology; 2019. Available from: https://www.acc.org/about-acc/press-releases/2019/03/07/08/45/heart-attacks-increasingly-common-in-young-adults.

[4] Mechanic OJ, Gavin M, Grossman SA. *Acute Myocardial Infarction*. 2022 Available from: https://www.ncbi.nlm.nih.gov/books/NBK459269/ [Accessed 26th June 2023].

[5] Frangogiannis NG. Pathophysiology of myocardial infarction. *Comprehensive Physiology*. 2011; 5(4): 1841-1875.

[6] Sing CF, Stengård JH, Kardia SLR. Genes, environment, and cardiovascular disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2003; 23(7): 1190-1196.

[7] Cardiovascular Disease: A Costly Burden for America: Projections Through 2035. *American Heart Association*. 2017. [Accessed 13th August 2023].

[8] Acharya UR, Faust O, Ghista DN, Vinitha Sree S, Alvin APC, Chattopadhyay S, et al. A systems approach to cardiac health diagnosis. *Journal of Medical Imaging and Health Informatics*. 2013; 3(2): 261-267.

[9] Chattopadhyay AK, Chattopadhyay S. VIRDOCD: A VIRtual DOCtor to predict dengue fatality. *Expert Systems*. 2021; e12796: 1-17.

[10] Chattopadhyay S, Chattopadhyay AK, Aifantis AC. Predicting case fatality of dengue epidemic: Statistical machine learning towards a virtual doctor. *Journal of Nanetechnology in Diagnosis and Treatment*. 2021; 7: 10-24.

[11] Sofogianni A, Stalikas N, Antza C, Tziomalos K. Cardiovascular risk prediction models and scores in the era of personalized medicine. *Journal of Personalized Medicine*. 2022; 12(7): 1180.

[12] Chattopadhyay S, Das R. Statistical validation of cardiovascular digital biomarkers towards monitoring the cardiac risk in COPD: A lyfas case study. *Artificial Intelligence Evolution*. 2022; 3(1): 1-16.

[13] Faust O, Ramanan Prasad V, Swapna G, Chattopadhyay S, Lim TC. Comprehensive analysis of normal and diabetic heart rate signals: A review. *Journal of Mechanics in Medicine and Biology*. 2012; 12(5): 1240033.

[14] Satapathy S, Chattopadhyay S. Observation-prevention framework of cardiac risk factors: An indian study. *Journal of Medical Imaging and Health Informatics*. 2012; 2(2): 102-113.

[15] Chattopadhyay S. Mathematical modelling of doctors' perceptions in the diagnosis of depression: A novel approach. *International Journal of Biomedical Engineering and Technology*. 2013; 11(1): 1-17.

[16] Chattopadhyay S. A framework for studying perceptions of rural healthcare staff and basic ict support for e-health use: An indian experience. *Telemedicine and E-Health*. 2010; 16(1): 80-88.

[17] Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951; 16(3): 297-334.

[18] Goforth C. *Using and Interpreting Cronbach's Alpha*. 2015. Available from: https://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/ [Accessed 11th January 2021].

[19] Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. *Annals of Card Anaesthesia*. 2019; 22(1): 67-72.

[20] Reynolds D. Gaussian mixture models. In: Li SZ, Jain A. (eds.) *Encyclopedia of Biometrics*. Boston: Springer; 2009. p.659-663.

[21] Chattopadhyay S, Do NP, Flower DR, Chattopadhyay AK. Extracting prime protein targets as possible drug candidates: machine learning evaluation. *Medical & Biological Engineering & Computing*. 2023; 61(11): 3035-3048.

[22] Sildir H, Aydin E, Kavzoglu T. Design of feedforward neural networks in the classification of hyperspectral imagery using superstructural optimization. *Remote Sensing*. 2020; 12(6): 956.

[23] Chattopadhyay S, Acharya UR. A novel mathematical approach to diagnose premenstrual syndrome. *Journal of Medical Systems*. 2012; 36: 2177-2186.

[24] Ashish K, Dasari A, Chattopadhyay S, Hui NB. Genetic-neuro-fuzzy system for grading depression. *Applied*

Computing and Informatics. 2018; 14(1): 98-105.

[25] Forero CG. Cronbach's alpha. In: Michalos AC. (ed.) *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht: Springer; 2014. p.1357-1359.

[26] Dyrbuś K, Gąsior M, Desperak P, Osadnik T, Nowak J, Banach M. The prevalence and management of familial hypercholesterolemia in patients with acute coronary syndrome in the Polish tertiary centre: Results from the TERCET registry with 19,781 individuals. *Atherosclerosis*. 2019; 288: 33-41.

[27] Salari N, Morddarvanjoghi F, Abdolmaleki A, Rasoulpoor S, Khaleghi AA, Hezarkhani LA, et al. The global prevalence of myocardial infarction: A systematic review and meta-analysis. *BMC Cardiovascular Disorders*. 2023; 23(1): 206.

[28] Prati F, Gurguglione G, Biccire F, Cipolloni L, Ferrari M, Di Toro A, et al. Sudden cardiac death in ischaemic heart disease: Coronary thrombosis or myocardial fibrosis? *European Heart Journal Supplements*. 2023; 25(Supplement_B): B136-B139.