Detection of myocardial infarction on recent dataset using machine learning

Nusrat Parveen[1], Satish R. Devane[2], Shamim Akthar[3]
[1]Department of Computer Engineering, Datta Meghe College of Engineering, Maharashtra, India
[2]Department of Information Technology, Datta Meghe College of Engineering, Maharashtra, India
[3]Department of Pathology, NKP SIMS RC & Lata Mangeshkar Hospital, NKP Salve Institute of Medical Science, Maharashtra, India

Article Info

Article history:

Received April,2021
Revised Dec, 2021
Accepted Jan, 2022

Keywords:

Decision tree
Ensemble algorithm
Multi-layer perceptron
Myocardial infarction
Naïve Bayes
Neural network
Support vector machine

ABSTRACT

In developing countries such as India, with a large aging population and limited access to medical facilities, remote and timely diagnosis of myocardial infarction (MI) has the potential to save the life of many. An electrocardiogram is the primary clinical tool utilized in the onset or detection of a previous MI incident. Artificial intelligence has made a great impact on every area of research as well as in medical diagnosis. In medical diagnosis, the hypothesis might be doctors' experience which would be used as input to predict a disease that saves the life of mankind. It is been observed that a properly cleaned and pruned dataset provides far better accuracy than an unclean one with missing values. Selection of suitable techniques for data cleaning alongside proper classification algorithms will cause the event of prediction systems that give enhanced accuracy. In this proposal detection of myocardial infarction using new parameters is proposed with increased accuracy and efficiency of the existing model. Additional parameters are used to predict MI with more accuracy. The proposed model is used to predict an early diagnosis of MI with the help of expertise experiences and data gathered from hospitals.

Corresponding Author:

Nusrat Parveen
Department of Computer Engineering, Datta Meghe College of Engineering
Navi Mumbai-400708, Maharashtra, India
Email: np.cm.dmce@gmail.com

1. INTRODUCTION

The mortality rates of cancer and myocardial infarction (MI) are very high nowadays. MI is the clinical term describing a heart attack due to a lack of oxygenated blood to heart tissue due to a clogged artery. Patients who have survived an MI incident are at a greater risk of other heart-related health problems later in their lifetime. Amongst all harmful sicknesses, coronary heart attacks are taken into consideration as the most widely wide-spread. Medical practitioners' behavior so many surveys on heart sicknesses and accumulate records of coronary heart patients, their ailment development, and symptoms. Every year heart ailment reasons tens of millions of deaths globally. Many techniques and tools were developed for coronary heart disease prediction by using medical doctors. Researchers have made efforts to expand the automated diagnosis systems in order that accurate diagnosis ought to take place. Among these, the automated machine the usage of data mining and artificial intelligence (AI)-based totally approach is the recent one used in the automated prognosis. The motivation of the work is the lack of data available freely and really difficult to access patient's data from hospitals. Large datasets are required to find out the model accurately. It's also important to predict early MI to save lots of the lifetime of several.

In this research, the actual datasets are collected from the hospitals. This dataset is not sufficient to offer to the model. Providing limited information restricts the training of the model resulting in compromised results in terms of overfitting. To overcome this problem a new path is taken by creating a synthetic dataset to provide information in bulk to the model. For this, continuous discussions with expertise and rigorous study are done and a range of various parameters are calculated for early MI, MI, and non-MI. The datasets available on Kaggle are not recent and also it is not an Indian dataset. It is of utmost necessity to collect a recent dataset. Around 2149 patients' data is collected from three hospitals in pastoral areas of Nagpur. Machine learning models learn very well if datasets are in bulk. Therefore, the idea of the synthetic dataset is proposed and datasets are generated based upon the actual dataset. The accuracy of models is extremely high.

Figure 1 shows the myocardial infarction. An attack occurs when one among the heart's coronary arteries is blocked suddenly or has extremely slow blood drift. The foremost common MI is due to the bifurcation of the left arteria coronaria. The usual explanation for sudden blockage during an arteria coronaria is the formation of a thrombus. The grume typically forms inside an arteria coronaria that already has been narrowed by atherosclerosis, a condition during which fatty deposits (plaques) build up along the walls of blood vessels [1]. Risk factors that can be controlled are high cholesterol, high bp, diabetes, weight, family history, smoking, unhealthy diet, lack of physical activities, and metabolic syndrome.

Risk factors that cannot control are the age of men greater than 45 and in women, it is considered greater than 55. If father or brother diagnosed attack before 55 years aged or mother or sister diagnosed before 65 years aged [2]. This case history results in MI. Another factor is understood as Preeclampsia. This condition can develop during pregnancy. The 2 main signs of Preeclampsia are an increase in vital signs and excess protein within the urine [3]. The main purpose of this research is to find MI in an early stage by using the above risk factors which will save the life of mankind.

Figure 2 shows the diagrammatic representations of the research idea. Diagnosis relies upon many various sorts of (accurate) data, from patient history to physical examination to lab data to past medical records and radiographic findings. Each patients' lifestyle, body system, and history are different. It is vital to notice that if the first prediction is feasible then the death rate with MI will certainly lessen and the lifetime of mankind will upgrade. Most vital thing is to think about those parameters of MI that are not included in early research but are most vulnerable for MI in today's life.

There is always a scope to exit from the prevailing approach and explore beyond the limit of other findings. Therefore, there's a requirement for designing a model which can predict MI early supported the parameters fed to the model. To reinforce the accuracy of the prognosis of MI for clinicians and clinical scientists, in our system, the input is gathered from many doctors personally and therefore the patient's data through proper channel with history of MI and this data set is given to the predictive model which then verifies and validates the proposed model. Early detection of MI will save the lifetime of mankind. This technique is going to be helpful to the doctor's assistant, nurses to require timely action if the doctor is not available within the hospital [4].

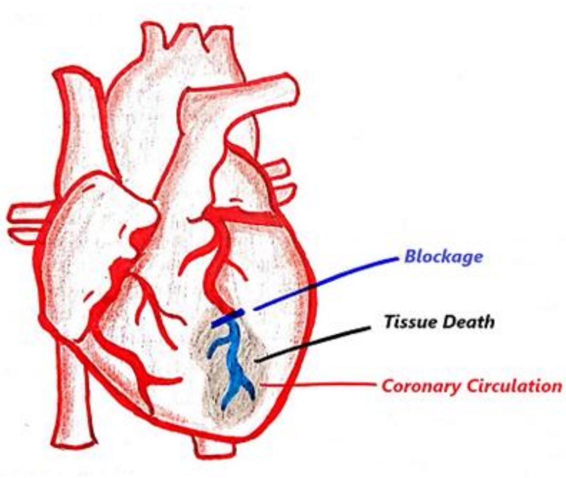

Figure 1. Myocardial infarction

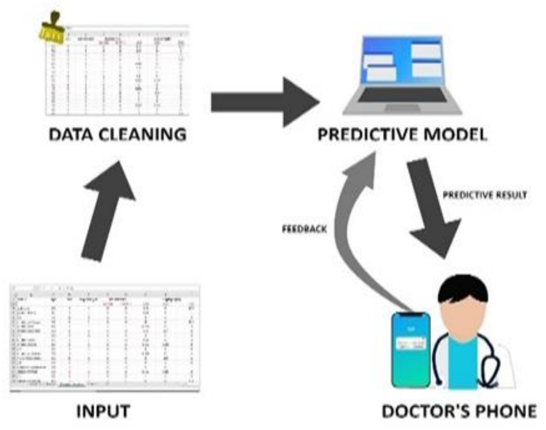

Figure 2. Proposed system [4]

2. RESEARCH METHOD

Timely hospital reporting and diagnosis are critical within the myocardial infarct. The prehospital delay could even be a significant explanation for increased morbidity and mortality within the myocardial infarct. This study finds a scarcity of realization and poor transportation facilities due to the main contributors to the delay within the management of myocardial infarction. Misjudgment of symptoms and transport delays still contribute foremost to pre-hospital delays. Systems of ST-segment–elevation myocardial infarction (STEMI) care will be got to concentrate on these variables to make an enormous impact on patient outcomes in ST-elevation myocardial infarction [5]. Atypical lipids, smoking, high blood pressure, diabetes, stomach obesity, psychosocial factors, eating fruits, vegetables, and alcohol, and regular physical activity account for several of the danger of myocardial infarct worldwide in both sexes and within the smallest amount ages altogether regions. This finding suggests that approaches to stop are often supported by similar principles worldwide and have the potential to prevent most premature cases of myocardial infarction [6]. Cardiologists Dr. Ashar Khan (DM) and Dr. Tamim Fazil (Medicine) and other experts have given tons of input during this research. All aspects of MI were discussed with the expertise. Many inputs are provided by them. There's a variable parameter that is liable for shown within Table 1. Firstly, MI features are excerpted from a rigorous study of literature review. Supported the literature review a survey is conducted and 20 expertise opinions are taken. This survey revealed the foremost important factors that ought to be considered during the research like diabetics, history of patients, diet, and stress. Still smoking, eating habits, and stress are not ready to include during this as they're vital features. The rationale is the unavailability of the info at the time of admission of the patients. And missing values affect the performance of the model. And filling missing values with mean and median is not suggested by expertise. Because the wrong values can cause misclassification of the model.

Table 1. Parameters list (literature review) [7]–[17]

Sr. No	Parameters
1	Age
2	High frequency of diabetes
3	Cigarette smoking
4	Overweight
5	Lethargy
6	Family history of early heart disease
7	A previous heart disease (PHF)
8	Depression
9	The ketone body oxidation increases MI
10	Non-pulsatile pulmonary blood flow in Fontan circulation
11	The HF with preserved ejection fraction (HFpEF)
12	Maternal mortality and morbidities
13	Thyroid dysfunction
14	In heart failure (HF), cardiac energy metabolism is deranged
15	Hormone replacement therapy
16	Illicit drug
17	A history of preeclampsia
18	An autoimmune condition
19	CKD (chronic kidney disease)
20	Stress.
21	Diabetes
22	Deficiency in Vit-D3
23	High blood pressure
24	ECG
25	High Cholesterol

2.1. Parameters excerpted from survey

Input features and their values are shown in Table 2 are extracted from the survey which is conducted during the research.

2.2. Statical analysis

This was an observational study conducted at two hospitals located in Nagpur (Kamptee). Data was collected prospectively of patients admitted within the hospital and treated for MI from March 2018 till Dec 2020. The information of patients is collected from the hospitals personally and analysis is completed. Employing a typical questionary, information was sought regarding the history of ischemic heart disease, coronary risk factors, time of onset of pain, pain type, patient's history, cholesterol, and blood pressure (BP). All parameters are considered and discussed the vulnerability of the parameters expertly and included during this research. As per the expertise, smoking and stress are the foremost important or responsible factors for MI.

Though they are not included within the research because the right information is not provided by the patients or not known by the relatives who are admitting the patients to the hospital.

Data is gathered from the hospitals from the patients' reports. Patients are evaluated with age, sex, ECG changes, biomarkers (CK-MB, TROP-I), angiography (LAD, LCA, RCA) cholesterol, BP (systolic, diastolic), chest paint type (acute, chronic), diabetics, chronic kidney disease (CKD), autoimmune condition (AC), family history (FH), hormone replacement therapy (HRT), thyroid dysfunction (TD), acute kidney injury (AKI). The evaluation is administered with the assistance of experts. Statistical analysis is completed using google form and therefore the graph generated during the survey for extracting the MI parameters. Patients' data are collected and transformed into the specified format.

Table 2. Parameters list (survey)

Sr. No	Parameters	Disintegrated_parameters and values
1	Age	Numeric
2	Sex	Male=1, Female=0
3	ECG	ECG Changes Yes=1, No=0
4	Biomarkers	CK-MB, TROP-I Changes Yes=1, No=0
5	Angiography	Left anterior descending (LAD), left coronary artery (LCA), right coronary artery (RCA) in percentage (Converted into 0.0 to 1.0)
6	Cholesterol	Numeric
7	Blood Pressure (Bp)	Systolic, Diastolic Numeric Values
8	Chest pain type	Acute, Chronic Acute=2, Chronic=1
9	Diabetic	Yes=1, No=0
10	History:	Chronic kidney disease (CKD), autoimmune condition (AC), previous heart failure (PHF), hormone replacement therapy (Hor_Rep), thyroid dysfunction (Thy_Dys), acute kidney injury (AKI). Yes=1, No=0
11	MI	Early MI=0, MI=1, non-MI=2

In this proposal, experiences and knowledge of experience are used. Victimization of data to answer queries alongside the study of various algorithms like SVM, NB, DT, LR, KNN, Ensemble, and NN and expert opinion is taken into account. Various data pre-processing techniques like data cleaning and pruning also the normalization of knowledge are important steps to use before feeding input to the model. Various steps are involved as:

- **Bucketization**
  It is used to make buckets for sub-features by disintegrating the main features into sub-features.
- **Normalization**
  Data are normalized converted into numeric with the help of experts.
- **Data cleaning and pruning**
  Data cleaning and pruning technique are performed on the chosen data in order that a correctly cleaned and pruned dataset provides far better precision than an unclean one with missing values. Data cleaning is the method of making data for the model by eradicating or altering data that is improper, imperfect, disparate, redundant, or inadequately formatted [18]–[20].

3. RESULTS AND DISCUSSION

In Figure 3 to Figure 21 graphs are created concerning each parameter vs the total number of patients count. A total of 565 patient data is collected from two hospitals. Of these, 65 patients' data have missing values. Therefore, it's not included in the research. Out of 500 data, there were 147 patients with angina, 150 were non-MI and 303 were of MI. To form data balanced each 150 approx. is taken into account for the research. Total 450 data is given to the model. Data analysis is carried out in Table 3.

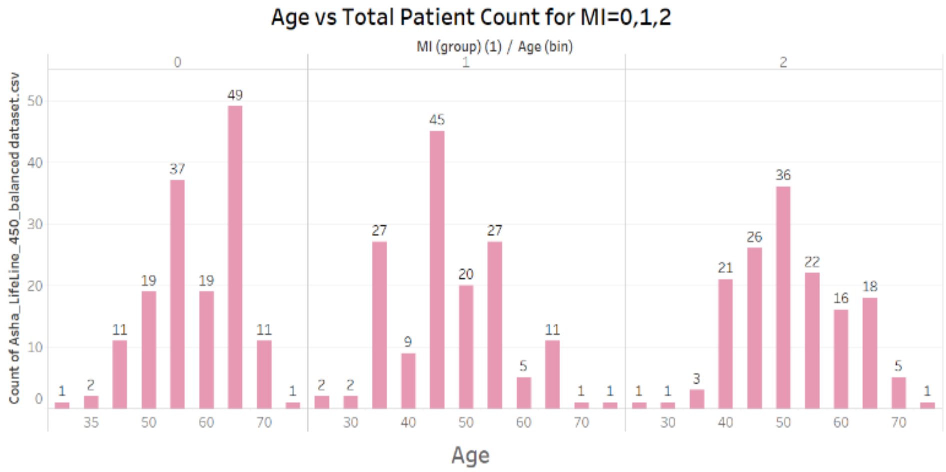

Figure 3. Graph between age vs total patient count

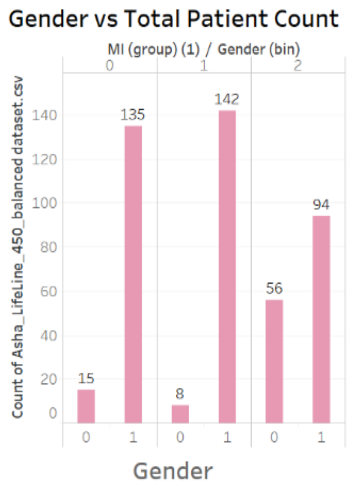

Figure 4. Graph between gender vs total patient count

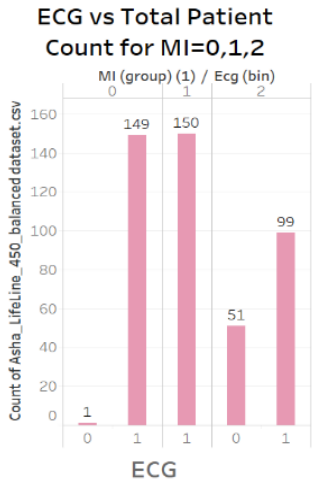

Figure 5. Graph between ECG vs total patient count

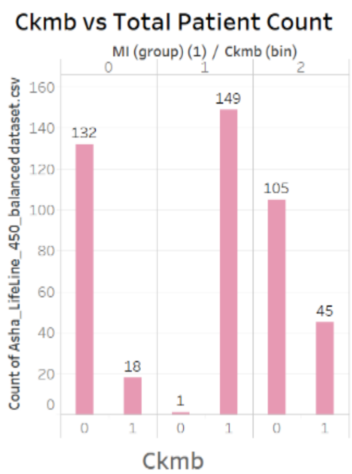

Figure 6. Graph between Ckmb vs total patient count

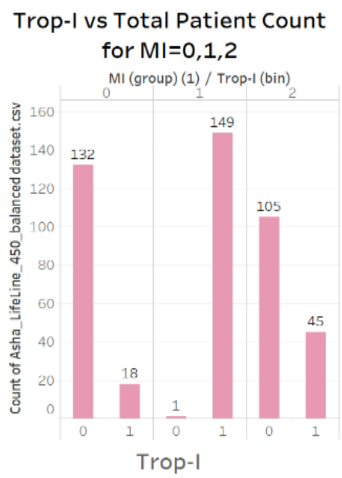

Figure 7. Graph between trop-i vs total patient count

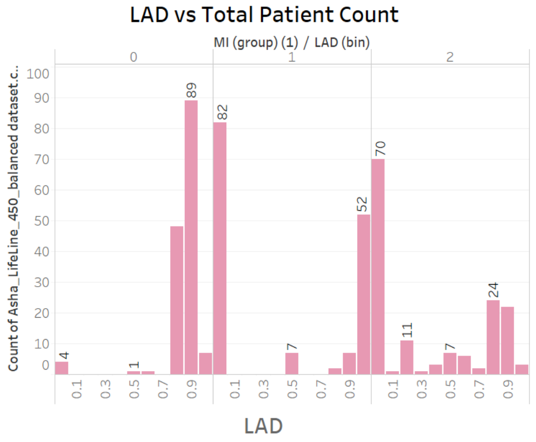

Figure 8. Graph between LAD vs total patient count

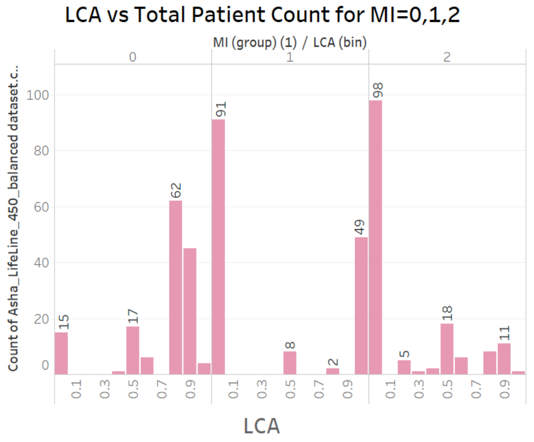

Figure 9. Graph between LCA vs total patient count

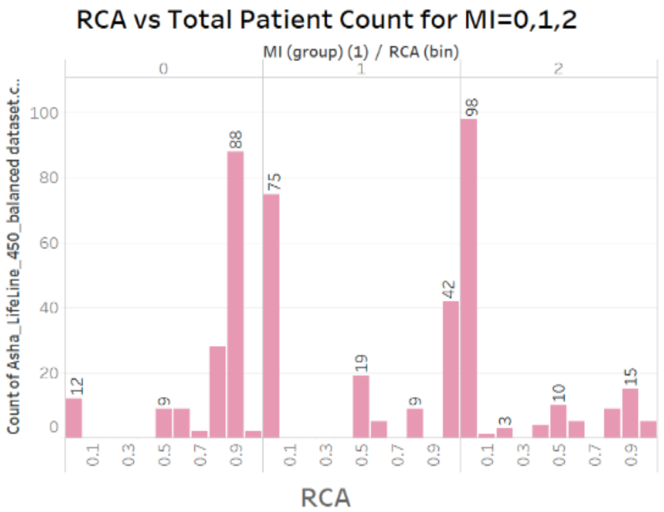

Figure 10. Graph between RCA vs total patient count

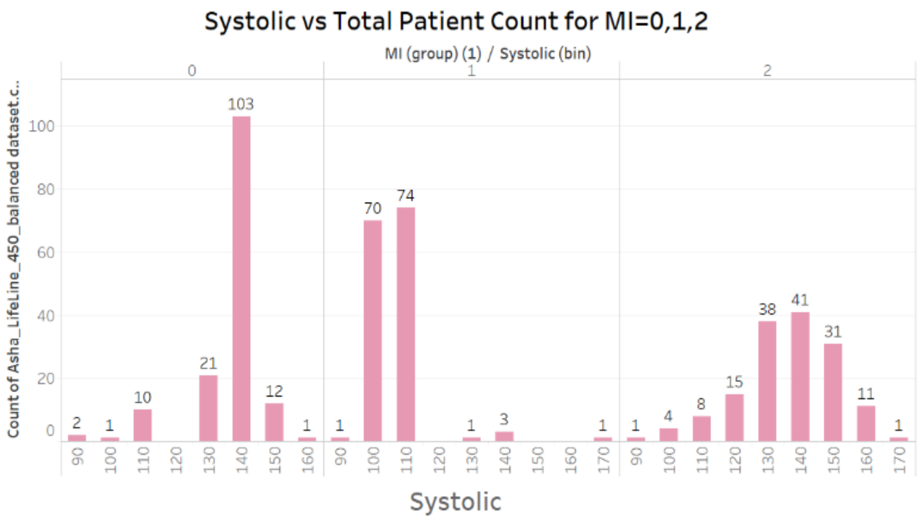

Figure 11. Graph between systolic vs total patient count

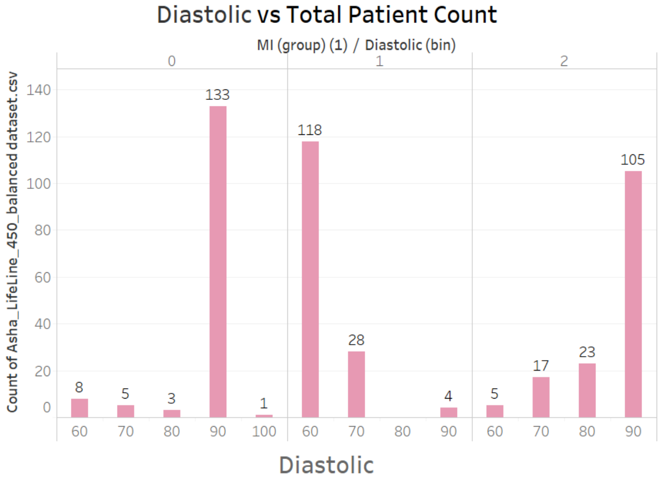

Figure 12. Graph between diastolic vs total patient count

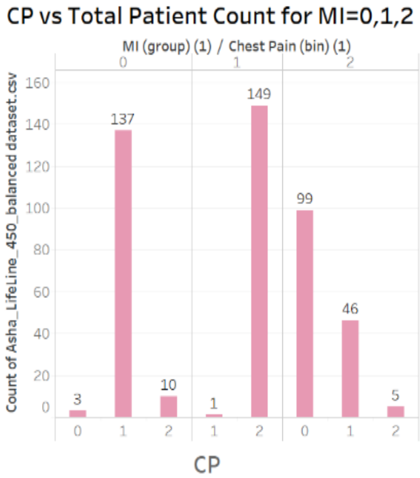

Figure 13. Graph between chest pain vs total patient count

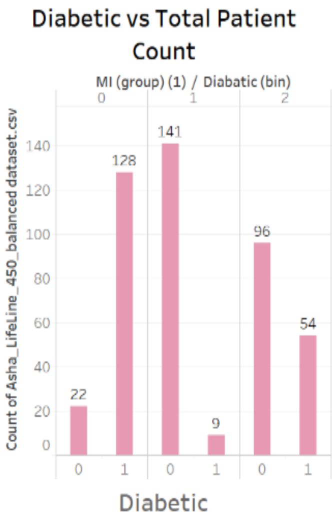

Figure 14. Graph between diabetic vs total patient count

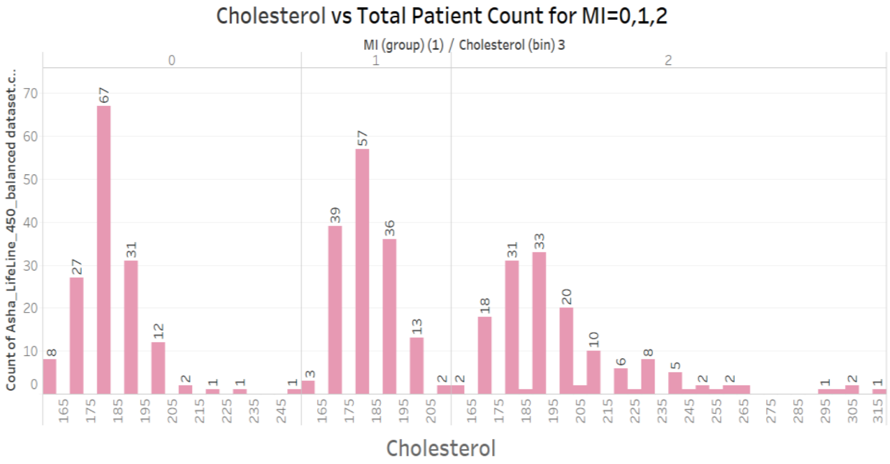

Figure 15. Graph between cholesterol vs total patient count

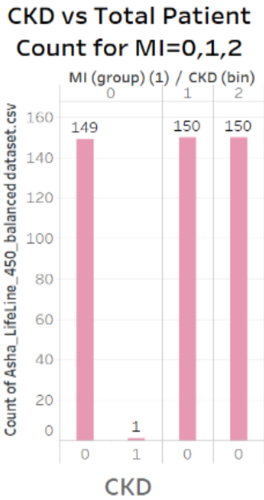

Figure 16. Graph between CKD vs total patient count

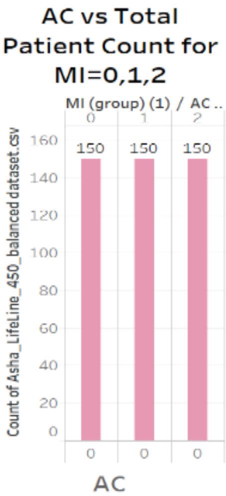

Figure 17. Graph between AC vs total patient count

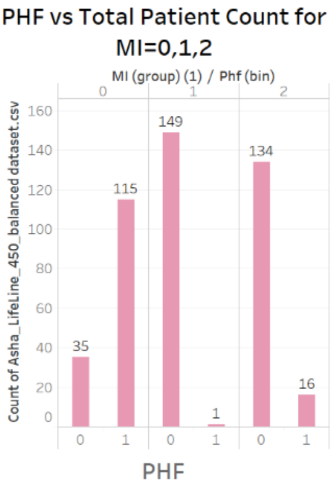

Figure 18. Graph between PHF vs total patient count

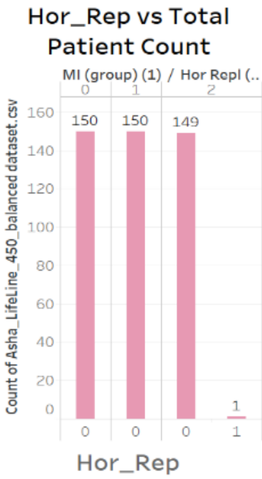

Figure 19. Graph between Hor_Rep vs total patient count

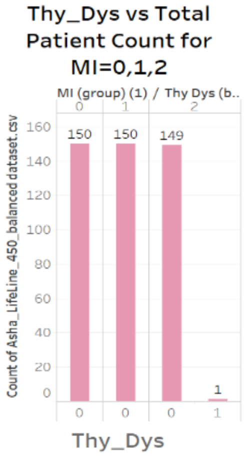

Figure 20. Graph between Thy_Dys vs total patient count

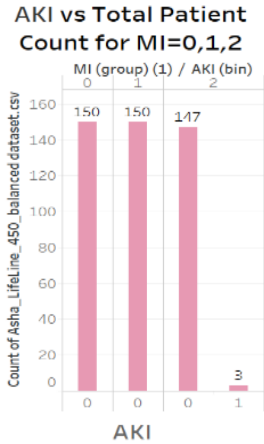

Figure 21. Graph between AKI vs total patient count

Table 3. Description of graph

Parameters	MI=0 Total patient count=150	MI=1 Total patient count=150	MI=2 Total patient count=150
Age	age>65 32%	age>45 30%	age>50 24%
Sex	Male=90% Female=10%	Male=95% Female=5%	Male=67% Female=33%
ECG	99% Yes 1% No	100% yes	66% yes 34% No
Biomarkers	Ckmb=88% yes Trop-I=88% yes	Ckmb=99% yes Trop-I=99% yes	Ckmb=70% yes Trop-I=70% yes
Angiography	LAD=60% patients having 90% blockage LCA=42% patients having 80% blockages RCA=59% patients having 90% blockages	LAD=35% patients having 100% blockages LCA=33% patients having 100% blockages RCA=28% patients having 100% blockages	LAD=16% patients having 80% blockages LCA=7% patients having 90% blockages RCA=1% patients having 90% blockages
Cholesterol	45% patients having 180	38% patients having 180	22% patients having 190
Bp	Systolic 69% patients having 140 Diastolic 89% patients having 90	Systolic 50% patients having 110 Diastolic 79% patients having 60	Systolic 28% patients having 140 Diastolic 70% patients having 90
Chest pain Type	Chronic 92% Acute 6% No pain 2%	Chronic 1% Acute 99% No pain 0%	Chronic 30% Acute 3% No pain 66%
Diabetic	85% diabetic	6% diabetic	36% diabetic
History	CKD=99% No AC=100% No PHF=76% yes Hor_Rep=100% No Thy_Dys=100% No AKI=100% No	CKD=100% No AC=100% No PHF=99% No Hor_Rep=100% No Thy_Dys=100% No AKI=100% No	CKD=100% No Ac=100% No PHF=90% No Hor_Rep=100% No Thy_Dys=100% No AKI=98% No

3.1. Experimental result

The dataset of two hospitals situated in Nagpur (Kamptee) is employed to classify three sorts of MI, i.e. Early MI (angina), Non-MI, and MI. Various algorithms are applied to the present dataset which has 450 patients' information. It is observed that the best results were achieved using MLP (alpha=0.7). Other's algorithms also are giving better accuracy within the training and testing phase. The output of algorithms can be seen in Table 4. Though the result's appreciable, it is suggested further to add more patient details to see the accuracy of the model. Because the data is especially from one region. It is going to vary from region to region because the lifestyle, eating habits and stress levels change. Though these parameters are not included within the research due to the unavailability of the knowledge. But expertise already emphasized this feature. Therefore, it is suggested to consider more datasets on this to predict accurately. For this a novel idea is proposed i.e., to generate synthetic datasets. The following steps are applied for the creation of a synthetic dataset.

Table 4. Output of algorithms

Algorithms	Training Set (%)	Testing Set (%)
Linear SVM	93%	91%
RBF SVM	98%	83%
Gaussian process	95%	91%
Naïve Bayes (NB)	80%	82%
Decision tree (DT)	96%	91%
Random forest (RF)	94%	91%
K-nearest neighbors (KNN)	94%	91%
Neural network (NN)	94%	91%
AdaBoost	88%	85%
Quadratic discriminant analysis	33%	33%
MLP classifier (alpha=0.1)	95%	91%
MLP classifier (alpha=0.2)	95%	92%
MLP classifier (alpha=0.7)	94%	92%

3.1.1. Function for generation of synthetic datasets

For a generation of synthetic datasets, firstly histogram of every feature is generated i.e., distribution of the information. Then normalized the histogram by scaling between zero and one. This distribution of data is then passed to the function that's used to prepare the synthetic datasets.

```
def genRand(l,u,n,d):
    print(l,u,(u-l)/n)
    return(np.random.choice(np.arange(l,u,(u-l)/n),n, p=d))
```

here:
l is lower limit of data
u is the upper limit of data
n is the number of samples to be generated
d is the distribution based on actual dataset

3.1.2. Graph for synthetic dataset

The distribution of actual datasets is passed to the function to get synthetic datasets. And 45000 patient report is generated from 2149 actual data gathered from patients' reports. The value of n is increased from 1k to 15k. 1k, 2k, 4k, 6k, 8k, 9k, 11k, 12k are giving NAN values. After 15k model accuracy is either constant or reducing. Therefore, the creation of synthetic data is stopped at 45000 samples.

3.1.3. The result on synthetic datasets

Table 5 listed the accuracy of the models for 15000 samples of synthetic datasets at the training and testing phase. In this KNN, RF is giving the highest accuracy.

Table 5. Algorithm accuracy at 15000 samples

Synthetic Datasets_15000		
Algorithms	Training	Testing
K-nearest neighbors	99.91	99.96
Linear SVM	99.99	100
RBF SVM	100	80
Decision tree	100	100
Random forest	98.42	98.1
Neural network	99.99	100
AdaBoost	100	100
Naïve Bayes	99.26	99.45
Quadratic discriminant analysis (QDA)	99.26	99.43

4. CONCLUSION

This study has attempted to research the dataset about the input features and customary reasons for early MI in patients presenting to the hospital within the urban area of Nagpur (Kamptee). There are previous studies shown only about MI not included Early MI. There's lagging in data also that was not recent data. It's also noticed that the Indian data is not available. This research has been done from scratch. Dataset is collected from the two hospitals and expert assistance is taken to incorporate some important features for early MI. After the gathering of knowledge from hospitals, the info is analyzed and it's discovered that in 450 patients there's almost no change in AC, Hor_Repl, Thy_Dys, AKI parameters. It'd be this pathological test is not referred to during this area due to expensive or could be not responsible most for MI during this region. As per expertise opinion, these parameters can be eliminated.

Feature selection is performed on 450 patients' data. More data is collected for the creation of synthetic datasets. 2149 patients' info is collected, Data cleaning and pruning technique is applied. A distribution graph is generated on this dataset and passed to the function to create synthetic datasets. This is done to create an authentic dataset. Expertise opinion is also taken on each step. Further work can be carried out by considering this opinion of experts. It is also suggested to collect more data from various regions of India to validate this work.

ACKNOWLEDGEMENTS

I am thankful and acknowledge the full support from Dr. Asher Khan (Cardiologist), Dr. Tamim Fazil (Medicine), Dr. Mehrosh Ghazal (Ped), Dr. Amera Ansari (Gyn), and Dr. Shamim Akhter (Path). I also thank them for allowing me to collect data from the hospitals. I am also thankful to all 20 doctors who had responded to my questionnaire through Google Form.

REFERENCES

[1] R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, *Braunwald's heart disease: a textbook of cardiovascular medicine*, 9th ed. Philadelphia: Elsevier Science, 2011.

[2] S. Tischler, "Does a family history of heart attacks increase your risk?," *UCI Health*. 2017, [Online]. Available: https://www.ucihealth.org/blog/2017/02/family-history-heart-attacks.

[3] J. Herndon, "Preeclampsia: causes, diagnosis, and treatments," *healthline*. 2021, [Online]. Available: https://www.healthline.com/health/preeclampsia.

[4] N. Parveen and S. R. Devane, "Efficient, accurate and early detection of myocardial infarction using machine learning," in *Disruptive Trends in Computer Aided Diagnosis*, 1st ed., R. Das, S. Nandy, and S. Bhattacharyya, Eds. New York: Taylor & Francis Group, 2021, p. 39.

[5] A. Khan, M. Phadke, Y. Y. Lokhandwala, and P. J. Nathani, "A study of prehospital delay patterns in acute myocardial infarction in an urban tertiary care institute in Mumbai," *Journal of Association of Physicians of India*, vol. 65, no. MAY, pp. 24–27, 2017.

[6] P. S. Yusuf *et al.*, "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study," *Lancet*, vol. 364, no. 9438, pp. 937–952, Sep. 2004, doi: 10.1016/S0140-6736(04)17018-9.

[7] J. Liu *et al.*, "Trends in outcomes of patients with ischemic stroke treated between 2002 and 2016: insights from a Chinese cohort," *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 12, Dec. 2019, doi: 10.1161/CIRCOUTCOMES.119.005610.

[8] E. Bertero, V. Sequeira, and C. Maack, "Hungry hearts," *Circulation. Heart failure*, vol. 11, no. 12, p. e005642, Dec. 2018, doi: 10.1161/CIRCHEARTFAILURE.118.005642.

[9] N. Parveen, S. R. Devane, and S. Akthar, "Synthetic datasets for myocardial infarction based on actual datasets," *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, vol. 10, no. 5, pp. 93–101, 2021.

[10] L. Kannan *et al.*, "Thyroid dysfunction in heart failure and cardiovascular outcomes," *Circulation: Heart Failure*, vol. 11, no. 12, p. e005266, Dec. 2018, doi: 10.1161/CIRCHEARTFAILURE.118.005266.

[11] M. F. Mogos, M. R. Piano, B. L. McFarlin, J. L. Salemi, K. L. Liese, and J. E. Briller, "Heart failure in pregnant women: a concern across the pregnancy continuum," *Circulation. Heart failure*, vol. 11, no. 1, p. e004005, Jan. 2018, doi: 10.1161/CIRCHEARTFAILURE.117.004005.

[12] T. Thorvaldsen *et al.*, "Predicting risk in patients hospitalized for acute decompensated heart failure and preserved ejection fraction: the atherosclerosis risk in communities study heart failure community Surveillance," *Circulation: Heart Failure*, vol. 10, no. 12, p. e003992, Dec. 2017, doi: 10.1161/CIRCHEARTFAILURE.117.003992.

[13] A. C. Egbe *et al.*, "Hemodynamics of Fontan failure: the role of pulmonary vascular disease," *Circulation: Heart Failure*, vol. 10, no. 12, p. e004515, Dec. 2017, doi: 10.1161/CIRCHEARTFAILURE.117.004515.

[14] M. Uchihashi *et al.*, "Cardiac-specific Bdh1 overexpression ameliorates oxidative stress and cardiac remodeling in pressure overload-induced heart failure," *Circulation: Heart Failure*, vol. 10, no. 12, p. e004417, Dec. 2017, doi: 10.1161/CIRCHEARTFAILURE.117.004417.

[15] M. Jessup and E. Antman, "Reducing the risk of heart attack and stroke: The American heart association/American college of cardiology prevention guidelines," *Circulation*, vol. 130, no. 6, Aug. 2014, doi: 10.1161/CIRCULATIONAHA.114.010574.

[16] K. T. Khaw and E. Barrett-Connor, "Family history of heart attack: a modifiable risk factor," *Circulation*, vol. 74, no. 2, pp. 239–244, Aug. 1986, doi: 10.1161/01.CIR.74.2.239.

[17] E. Barrett-Connor and K. T. Khaw, "Family history of heart attack as an independent predictor of death due to cardiovascular disease," *Circulation*, vol. 69, no. 6, pp. 1065–1069, Jun. 1984, doi: 10.1161/01.CIR.69.6.1065.

[18] "Data cleansing: what is it and why is it important?," *blue-pencil*. 2022, [Online]. Available: https://www.blue-pencil.ca/data-cleansing-what-is-it-and-why-is-it-important/.

[19] "Data Cleaning 101 – Towards Data Science." [Online]. Available: https://towardsdatascience.com/data-cleaning-101-948d22a92e4.

[20] N. P. M. Rafique, S. R. Devane, and S. Akhtar, "Early Detection of Myocardial Infarction Using Actual and Synthetic Datasets," in *2021 IEEE Bombay Section Signature Conference (IBSSC)*, Nov. 2021, pp. 1–6, doi: 10.1109/IBSSC53889.2021.9673210.

BIOGRAPHIES OF AUTHORS

Nusrat Parveen is pursuing Ph.D. She has 19 years of teaching experience. She is good in various subjects such as machine learning, web application and database. Nusrat's research is mainly focused on medical diagnosis using machine learning. She has published 8 papers in international conference, 3 international journals, 4 in national conferences and one chapter is publish in book under Tailor & Francis (CRC-press). She won cash prize in Indo-Korean festive competitions for outstanding innovator. She can be contacted at email: np.cm.dmce@gmail.com.

Satish R. Devane is an Academician of the IIT (Ph.D: Information Technology | M.E: Electronics | B.E.: Elcctronics |) and principal of KBTCOE, Nashik. Professor Devane is proficient in many technical areas such as networking, artificial intelligence and data minning. He has published 12 papers in international conferences. He can be contacted at email: srdevane@yahoo.com.

Shamim Akhtar is MBBS, MD (pathology), gold medalist and IOSR-JDM Global Editor. He has 30 years of experience. He published 17 papers in Int. journal. He has also published 3 books on "Solved question paper of pathology & genetics for B Sc nursing", "Essential to genetics and pathology", "Exam preparative manual for BDS students". Internationally invited as guest speaker and presenting recent research work at Montreal international translational medicine conference-2011. Invited as guest speaker and for presenting recent research work at Beijing international infectious diseases & antibiotics conference in Beijing (China)-2011. Best teaching and academics awards received. He can be contacted at email: akhtar_lmh@rediffmail.com