ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

Machine Learning based Myocardial Infarction Risk Stratification as a Diagnostic Aid for Remote Areas with Limited Medical Resources.

Mr. Amol R. Patil[1], Dr. P. B. Bharate[2], Dr. Mohd. Junaid[3]

1. Research Scholar, Department of Statistics, Malwanchal University Indore(M.P.)
2. Professor, Department of Statistics, Malwanchal University Indore(M.P.)
3. Professor, Shri Shankaracharya Institute of Medical Sciences, Bhilai (C.G.)

Corresponding Author:
Mr. Amol R. Patil, Research Scholar, Department of Statistics, Malwanchal University Indore, Madhya Pradesh, India.
Email ID: arpatilstat@gmail.com

ABSTRACT

Background: This study highlights the vital role of Machine Learning in aiding myocardial infarction (MI) diagnosis, crucial in remote areas with limited medical resources. By leveraging ML algorithms and accessible patient data, it offers a valuable tool for early MI detection and risk assessment in underserved regions, potentially improving patient outcomes and healthcare delivery.

Methods: In this case-control study, data from 1,200 individuals (300 MI, 900 non-MI) were collected. Significant variables were identified using correlation. Eight ML models were built based on the patient's historical 24 variables and evaluated using the F1 score, Cohen's Kappa, and AUROC. We also conducted real-time clinical validation to assess the practical applicability of the model.

Results: In terms of training time, logistic regression (LR) with L2 regularization, AdaBoost, and XGBoost models showed significantly higher times (410ms, 520ms, and 220ms, respectively). LR had the lowest errors (1.67% training, 1.11% testing) and achieved a high accuracy of 96%, notable precision, recall, and an impressive AUC of 98.87%. In real-time clinical validation, LR and XGBoost performed exceptionally well, boasting F1 scores of 96.27% and 98.70%, respectively, solidifying their effectiveness for predictive accuracy in a clinical setting.

Conclusion: In real-time clinical validation, LR and XGBoost based on patient's historical data showcased impressive predictive power, highlighting their potential in clinical settings. These models can be helpful to improve the diagnosis of MI in Remote Areas with Limited Medical Resources.

Keywords: Myocardial Infarction, Machine Learning, Prediction, Diagnosis

ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

INTRODUCTION:

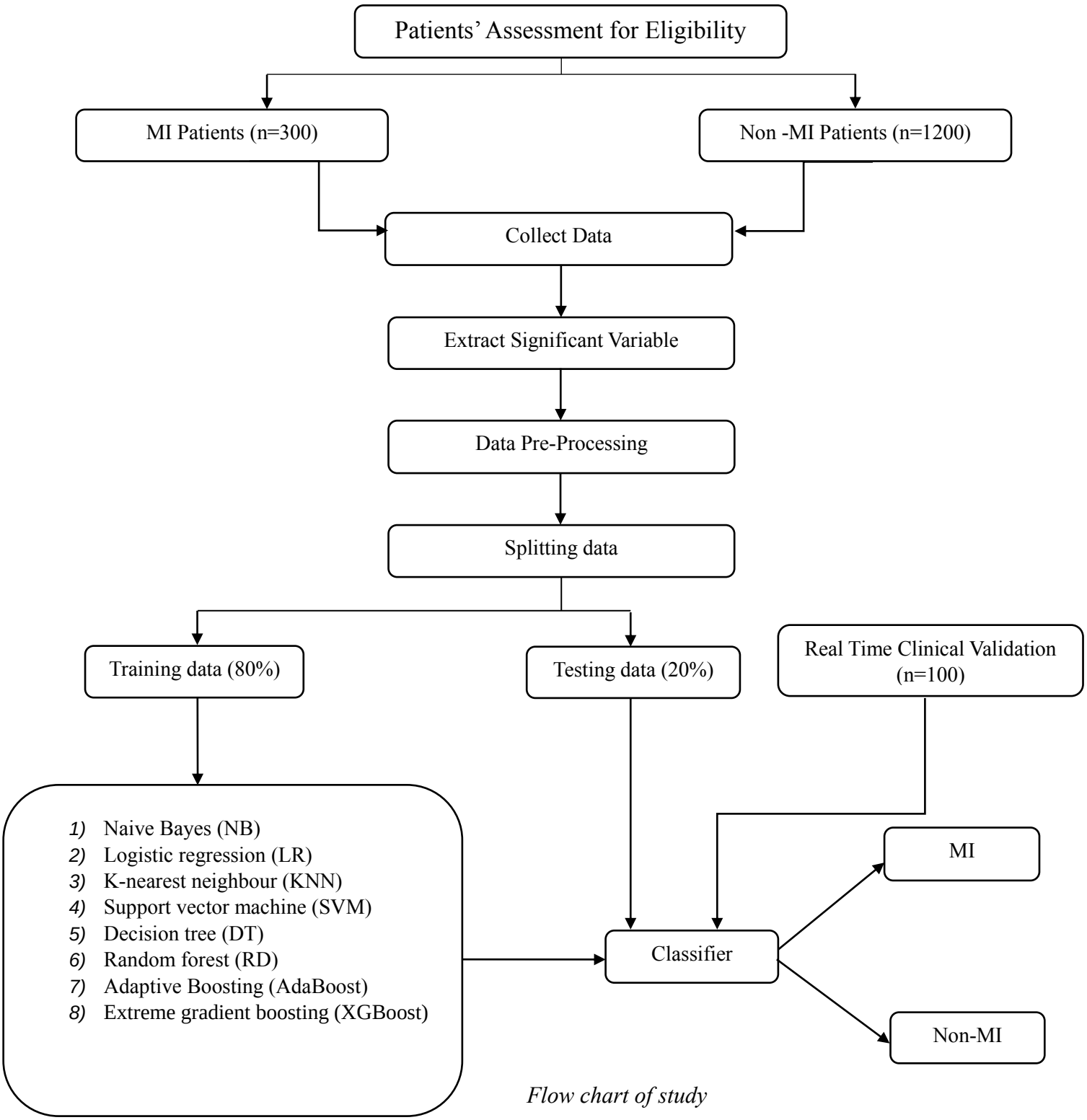
In the realm of healthcare, timely and accurate diagnosis of critical conditions such as myocardial infarction (MI) holds paramount importance in saving lives and improving patient outcomes [1]. However, in remote or resource-constrained areas where access to advanced medical facilities is limited, the task of efficient diagnosis becomes more challenging [2]. This predicament underscores the necessity for innovative machine learning solutions that can aid in early detection and risk stratification of MI, utilizing readily available data and computational methodologies [3,4]. According to data from the World Health Organization (WHO), India is responsible for approximately 20% of global fatalities, particularly among the younger demographic. The Global Burden of Disease study reveals that the age-standardized cardiovascular disease (CVD) death rate in India stands at 272 per 100,000 population, significantly surpassing the global average of 235[5]. This research delves into the realm of machine learning (ML), a burgeoning field that holds immense potential for transforming healthcare. In this study, we focus on the application of ML models for MI prediction, relying solely on historical patient data.

We explored a range of ML algorithms, evaluating their performance metrics, training time, and real-time clinical validation results. Logistic Regression, Support Vector Machine, Random Forest, AdaBoost, and XGBoost emerge as frontrunners, exhibiting significant promise in terms of predictive accuracy and reliability. Leveraging these models based on patient history could offer a pragmatic solution for early MI detection and risk assessment.

This study aims to shed light on the potential of history-based ML models as efficient diagnostic aids, presenting a stepping stone towards accessible and accurate MI prediction, ultimately contributing to improved patient care and healthcare delivery, regardless of the available resources.

ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

MATERIALS AND METHODS:



Flow chart of study

ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

This study, conducted between January 2021 and March 2023, employed a Case-Control design. The sample size of 1200 was determined using the formula provided by Riley et al [6] considering 20 predictors and an R^2 of 0.15. Maintaining a case-to-control ratio of 1:3, the research included 300 patients with myocardial infarction (MI) and 900 non-MI patients. Both cases and controls were selected from tertiary care hospitals in central India, and with valid written consent, comprehensive information encompassing detailed medical history, sociodemographic details, and lifestyle-related risk factors of MI were collected using a predefined structured questionnaire. Controls were matched with cases based on age (±5 years) and sex, with three controls selected for each case. Inclusion criteria stipulated that cases were above 18 years old and diagnosed with MI using established clinical criteria, while patients with severe illness were excluded. The primary data collected were complete without any missing values. Unique numerical labels were assigned to categories for nominal variables, and numerical labels were assigned based on predefined order for ordinal variables.

Statistical analysis:

The all-statistical analysis was done using R 4.3.1 software. The correlation of MI with continuous/ordinal, binary, and nominal (categories>2) risk factors was estimated using Point Biserial, Phi, and Cramer's V measures of correlation. The chi-square test was used to examine the relationship between two categorical variables.

Model building:

The dataset, consisting of 24 variables related history of patients significant with MI, underwent a random 80-20% split for training and testing, respectively. Within the training set, an additional 80-20% split was performed to create a training-validation dataset. Various machine learning models, including Naïve Bayes, Logistic Regression, Decision Tree, KNN, Support Vector Machine, Random Forest, XGBoost, and Adaptive Boosting, were constructed for each selected feature set. To control overfitting and stabilize coefficient estimates, Logistic Regression utilized Ridge (L2 regularization).

Evaluation Matrix:

Model performance was evaluated using both validation and testing datasets. Performance metrics encompassed Validation Accuracy, Testing Accuracy, Precision, Recall, Specificity, Negative Predictive Value, F1 Score, and Area under the ROC curve (AUC). Moreover, real-time clinical validation involved 100 patients from a tertiary care hospital, featuring 20 patients with myocardial infarction (MI).

Table 1: Description of variables

Variable	Description		Scale of Measurement
	Variable Name	Outcome	
Socio Demographical Factors			
X1	Myocardial Infarction	Yes=1, No=0	Nominal
X2	Age	Numbers	Ratio
X3	Gender	Male=1, No=0	Nominal
X4	Education	Primary=0, Secondary=1, High School=2, Graduate=3,	Ordinal

ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

		Post Graduate=4, Higher Education=5	
X5	Occupation	Related To Stress=5, Exposure to various Chemical=4, Exposure to Dust=2, Occupational Noise=1, Other=0	Ordinal
X6	Income	Numbers	Ratio
X7	Religion	Hindu=a, Muslim=b, Christian=c, Sikh=d, Buddhist=e, Other=f,	Nominal
X8	Marital Status	Married=1, Unmarried=2	Nominal
X9	Residential Status	Urban, Rural	Nominal
Symptoms of MI			
X10	Chest Pain	Yes=1 , No=0	Nominal
X11	Cold Sweat	Yes=1 , No=0	Nominal
X12	Dizziness Light headedness	Yes=1 , No=0	Nominal
X13	Fatigue	Yes=1 , No=0	Nominal
X14	shortness Breath	Yes=1 , No=0	Nominal
History of Disease			
X15	CKD	Yes=1 , No=0	Nominal
X16	COPD	Yes=1 , No=0	Nominal
X17	MI	Yes=1 , No=0	Nominal
X18	CVD	Yes=1 , No=0	Nominal
X19	DM	Yes=1 , No=0	Nominal
X20	RA	Yes=1 , No=0	Nominal
X21	HIV	Yes=1 , No=0	Nominal
X22	Thrombophilia	Yes=1 , No=0	Nominal
X23	HRT	Yes=1 , No=0	Nominal
X24	Preeclampsia	Yes=1 , No=0	Nominal
X25	PCOS	Yes=1 , No=0	Nominal
X26	Sedentary Lifestyle	Yes=1 , No=0	Nominal
Lifestyle related factors			
X27	Smoking	Non-Smoker=1, Former Smoker=2, Occasional Smoker=3, Light or moderate or high Smoker=4	Ordinal
X28	Alcohol	Non-Alcoholic=1, Former Alcoholic=2, Occasional Alcoholic=3, Light= 4, moderate=5, high	Ordinal

ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

		Alcoholic=6	
X29	Diet Score	Numbers	Ratio
X30	Stress	Never=1, Almost Never=2, Sometimes=3, Fairly Often=4, Very Often=5	Ordinal
X31	Sleep	Good=1, Moderate=2, Poor=3,	Ordinal
X32	Caffeine	Rarely or never=6 1-2 times per month=5 1-2 times per week=4 3-4 times per week=3 5-6 times per week=2 Daily=1	Ordinal
X33	NSAIDs	Yes=1 , No=0	Nominal
Family History of Disease			
X34	MI	Yes=1 ,No=0	Nominal
X35	DM	Yes=1 ,No=0	Nominal
X36	Hypertension	Yes=1 ,No=0	Nominal
X37	Hyperlipidaemia	Yes=1 ,No=0	Nominal
Physiological Traits			
X38	BMI	Numbers	Ratio

RESULTS:

Table 2: Correlation of Risk Factors with MI

Sr. No.	Variables	Correlation Coefficient
1	Age	0.113
2	Income	0.003
3	Smoking	0.002
4	Alcohol	0.243
5	Diet	0.388
6	Stress	0.294
7	Sleep	0.346
8	Caffeine	0.072
9	Body Mass Index (BMI)	0.636
10	Gender	0.05
11	Marital Status	-0.03
12	Residential Status	-0.04
13	Chest Pain	0.54
14	Cold Sweat	0.24
15	Dizziness Light Headedness	0.619

ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

16	Fatigue	0.281
17	Shortness Breath	0.595
18	Chronic Kidney Disease (CKD)	0.115
19	Chronic Obstructive Pulmonary Disease (COPD)	0.166
20	History Myocardial Infarction	0.231
21	History Cardio Vascular Disease	0.382
22	History Diabetes Militants	0.129
23	Rheumatoid Arthritis (RA)	0.125
24	Human Immunodeficiency Virus (HIV)	0.066
25	History Thrombophilia	0.174
26	Hormone Replacement Therapy (HRT)	0.007
27	Preeclampsia	0.026
28	Polycystic Overy Syndrome	0.102
29	Sedentary Lifestyle	0.373
30	Chromic Use of Nonsteroidal Anti-Inflammatory Drugs (NSAIDs)	0.18
31	Family History of Myocardial Infarction	0.346
32	Family History of Diabetes Militants	0.313
33	Family History of Hypertension	0.383
34	Family History of Hyperlipidaemia	0.397
35	Type A person	0.189
36	Education	0.052
37	Occupation	0.173
38	Religion	0.052

Table 2 illustrates correlation coefficients between various risk factors and myocardial infarction (MI). Strong positive correlations were found with BMI (0.636), shortness of breath (0.595), and chest pain (0.54). Moderately, diet (0.388), stress (0.294), sleep (0.346), and family history of hyperlipidaemia (0.397) showed positive correlations. Alcohol (0.243), cold sweat (0.24), and dizziness/light-headedness (0.619) displayed weak positive correlations. On the other hand, marital status (-0.03) and residential status (-0.04) exhibited weak negative correlations.

Table 3: Performance of ML Model for MI Prediction using Only History of Patients

Algorithm	Accuracy		Precision	Recall	Specificity	NPV	F1 Score	C Kappa	AUC
	Validation	Testing							
NB	82.05	90.39	90.67	92.49	84.53	89.56	90.57	84.55	91.3
LR	96	97.89	98.26	98.26	96.78	96.78	98.26	96.04	98.87
DT	92.79	86.50	87.70	90.20	78.47	83.89	89.93	85.51	88.26
KNN	88.99	87.78	97.41	87.67	88.33	58.89	92.28	63.33	78.15
SVM	92.81	92.78	92.89	92.16	90.59	88.44	92.52	88.03	90.67
RF	92.83	94.50	96.26	94.45	94.65	89.22	95.35	90.21	96.61
AdaBoost	97.81	97.33	97.52	98.25	94.65	96.78	97.88	94.59	98.86
XGBoost	96.43	97.22	99.63	96.76	98.78	90.00	98.18	92.37	94.81

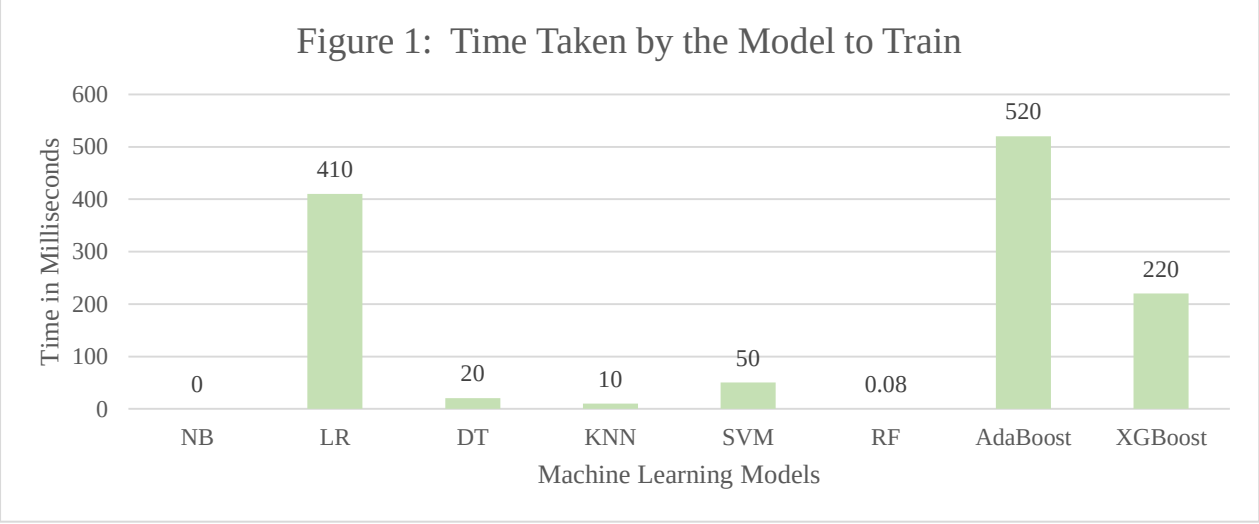
ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

Table 3 showcases the performance of various machine learning models in predicting myocardial infarction using patient history. Notable highlights include Logistic Regression's high accuracy of 96% and impressive precision and recall, along with a remarkable AUC of 98.87%. AdaBoost also excels with an accuracy of 97.81% and strong precision and AUC. XGBoost stands out with a recall of 99.63% and a noteworthy AUC of 94.81%. These models collectively demonstrate effective predictive capabilities, essential for accurate myocardial infarction prediction.

Table 4: Comparison between training and testing error for MI Prediction using Only History of Patients

ML Model	Training Error	Testing Error
NB	0.0417	0.0361
LR	0.0167	0.0111
DT	0.0694	0.0750
KNN	0.0456	0.1222
SVM	0.0348	0.0397
RF	0.0000	0.0250
AdaBoost	1.0000	1.0278
XGBoost	0.0000	0.0278

In Table 4, we compare training and testing errors for myocardial infarction (MI) prediction using patient history data across various machine learning (ML) models. Logistic Regression (LR) had the lowest errors (training: 1.67%, testing: 1.11%), indicating strong generalization. Support Vector Machine (SVM) showed low errors (training: 3.48%, testing: 3.97%), suggesting reliable predictive performance. Random Forest (RF) fit exceptionally well to training data (training: 0.00%), with a low testing error (2.50%), suggesting good generalization. However, AdaBoost exhibited significantly higher testing error (102.78%), indicating potential overfitting.



ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

Figure 1 presents the training time in milliseconds for different machine learning (ML) models using only patient history data. Naive Bayes and Random Forest had the fastest training times, both below 1 millisecond. Decision Tree, K-Nearest Neighbors, and Support Vector Machine showed moderate training times. Logistic Regression, AdaBoost, and XGBoost took relatively longer for training.

Table 6: Real-time Clinical Validation of Model-Based Only on History of Patients

Algorithm	Accuracy	Precision	Recall	Specificity	NPV	F1 Score
NB	91.11	95.56	92.81	85.37	77.78	94.16
LR	94.44	93.48	99.23	82.00	97.62	96.27
DT	91.11	95.19	93.12	84.52	78.89	94.14
KNN	86.67	95.93	87.50	82.81	58.89	91.52
SVM	96.67	97.78	97.78	93.33	93.33	97.78
RF	95.56	97.78	96.35	93.02	88.89	97.06
AdaBoost	95.83	98.52	96.03	95.18	87.78	97.26
XGBoost	98.06	98.52	98.88	95.60	96.67	98.70

In Table 6, the real-time clinical validation results for MI prediction using models based on patient history are presented. LR, SVM, RF, AdaBoost, and XGBoost exhibited high accuracy, precision, recall, and specificity, making them promising for real-world clinical use with accuracy ranging from 94.44% to 98.06%. NB and DT also performed well, although with slightly lower metrics, achieving an accuracy of 91.11%. KNN showcased lower specificity and AUC compared to other models.

DISCUSSION:

In this study, Strong positive correlations of occurrence of MI were found with BMI (0.636), shortness of breath (0.595), and chest pain (0.54). Moderately, diet (0.388), stress (0.294), sleep (0.346), and family history of hyperlipidaemia (0.397) showed positive correlations

Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), AdaBoost, and XGBoost showcased exceptional predictive performance, achieving accuracies ranging from 94.44% to 98.06%. Almost similar results were found in studies [7,8]. Logistic regression with L2 regularization stood out with high accuracy (97.89%) and recall (98.26%), making it vital for early MI detection. SVM demonstrated strong precision (92.78%) and specificity (90.33%), indicating its potential as a reliable diagnostic tool which was consistent with a study conducted by Ahmad et al [9]. Boosting models like Adaboost and XGboost also showed higher accuracy, precision, and AUC. RF showed impressive results, with a remarkable AUC of 96.61%, underlining its robust predictive ability. We achieved almost higher accuracy for RF, LR, and KNN than studies conducted for heart disease and MI prediction [10,11,12].

Analyzing training time, Naive Bayes (NB) and RF exhibited the quickest training durations, while AdaBoost and XGBoost required slightly longer training periods. Real-time clinical

ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

validation further validated the models' efficacy, affirming their practical use in real-world clinical settings.

This study emphasizes using patient history for myocardial infarction (MI) prediction with ML models. LR and XGBoost show potential for accurate diagnosis and real-time clinical support, especially in resource-constrained settings. The results highlight ML predictive models' effectiveness in enhancing healthcare accessibility and quality, suggesting further refinement for precise MI prediction and improved patient outcomes.

CONCLUSION:

In conclusion, machine learning models, notably Logistic Regression, XGBoost, and AdaBoost demonstrate great promise for myocardial infarction prediction using patient history alone. These models present high accuracy and reliability, making them valuable tools in clinical decision-making, particularly in resource-limited settings. Further optimization and integration of these models hold significant potential for enhancing healthcare outcomes and accessibility.

LIMITATION:

The data originated from a sole tertiary care hospital in central India, possibly restricting generalization. Also, the dataset displayed an imbalanced class distribution, potentially introducing bias in the outcomes.

CONFLICT OF INTEREST:

The authors confirm no conflicts of interest associated with this paper's publication.

REFERENCE:

1. Hilliard AL et al. Myocardial infarction classification and its implications on measures of cardiovascular outcomes, quality, and racial/ethnic disparities. Clin Cardiol. 2020 Oct;43(10):1076-1083.
2. Kumar A. The Transformation of The Indian Healthcare System. Cureus. 2023 May 16;15(5):e39079.
3. Lin A et al. Artificial Intelligence in Cardiovascular Imaging for Risk Stratification in Coronary Artery Disease. Radiol Cardiothorac Imaging. 2021 Feb 25;3(1):e200512.
4. Ihme, M., Chung, W. T., & Mishra, A. Combustion machine learning: Principles, progress and prospects. Progress in Energy and Combustion Science; 91 (2022, July 1). 101010
5. Sreeniwas Kumar A et al. Cardiovascular disease in India: A 360 degree overview. Med J Armed Forces India. 2020 Jan;76(1):1-3.
6. Riley RD et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38 :1276-96.10.1002/sim.7992 30357870
7. Sherazi SWA, Bae JW, Lee JY. A soft voting ensemble classifier for early prediction and diagnosis of occurrences of major adverse cardiovascular events for STEMI and

ISSN: 0975-3583, 0976-2833 VOL 14, ISSUE 10, 2023

NSTEMI during 2-year follow-up in patients with acute coronary syndrome. PLoS One. 2021 Jun 11;16(6): e0249338.

8. Huilin, Zheng & Sherazi, Syed Waseem Abbas & Lee, Jong. A Stacking Ensemble Prediction Model for the Occurrences of Major Adverse Cardiovascular Events in Patients with Acute Coronary Syndrome on Imbalanced Data. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3099795.
9. Ahmad, Ahmad Ayid, and Huseyin Polat. "Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm". Diagnostics ,2023,13, no. 14: 2392.
10. Jindal et al. Heart disease prediction using machine learning algorithms. IOP Conference Series, 2021, 1022(1), 012072.
11. Vaddi Niranjan Reddy et al, Myocardial Infarction Prediction Using Hybrid Machine Learning Techniques, Turkish Journal of Computer and Mathematics Education Vol.12 No.3(2021), 4251-4260
12. Izabela Rojek et al, AI-Based Prediction of Myocardial Infarction Risk as an Element of Preventive Medicine, Appl. Sci. 2022, 12, 9596.