ORIGINAL RESEARCH

Enhancing Myocardial Infarction Diagnosis with Efficient Machine Learning Techniques Through Combination of Correlation and Variance Threshold Feature Selection

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

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

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

Received: 26 Sep, 2023 Accepted: 15 Oct, 2023

ABSTRACT

Background: Delayed or misdiagnosis of myocardial infarction (MI) is a common occurrence in clinical settings. Timely detection of MI is crucial to prompt intervention which minimizes irreversible heart muscle damage, thereby reducing the risk of complications and heart failure. Thus, our objective was to develop a machine learning (ML) model to serve as a diagnostic aid, utilizing a minimal set of patient health parameters as features.

Methods: We collected data from 1,200 individuals (300 MI, 900 non-MI) using a case-control study with a 1:3 case-to-control ratio. Employing three feature selection methods, we identified significant variables. Six ML models (Naïve Bayes, Logistic Regression, Decision Tree, SVM, Random Forest, AdaBoost) were constructed for each technique, and their performance was evaluated using F1-Score, Cohen's Kappa, and AUROC. Additionally, clinical validation was conducted on real-time data for practical applicability.

Results: 17, 18, and 9 features were selected using variance threshold, correlation-based, and a combination of both techniques respectively. AdaBoost consistently showcased superior performance, followed by Random Forest. In real-time clinical validation, AdaBoost demonstrated remarkable performance with 94.12% accuracy, 98.86% precision, 98.58% recall, 93.11% F1 score, 96.49% Cohen's Kappa, and 94.12% Area under ROC.

Conclusion: The ML can serve in a timely, and precise diagnosis of MI, particularly AdaBoost. Furthermore, identified risk factors and their correlations emphasize the need for personalized preventive actions and lifestyle changes to mitigate myocardial infarction risks.

Keywords: Myocardial Infraction, Prediction, Feature Selection, Machine

INTRODUCTION

Myocardial infarction (MI), or heart attack, is a major global health issue, causing irreversible heart muscle damage [1]. Cardiovascular disease(CVD) accounts for a significant number of global deaths, with 85% attributed to heart attacks [2-3]. India faces a high burden of CVD, with a higher death rate than the global average, affecting younger populations [4]. Indian CVDs present unique challenges such as early onset, rapid progression, and elevated mortality, notably due to a high prevalence of coronary artery disease (CAD) [4]. Healthcare professionals struggle with early MI detection due to subtle symptoms. Machine learning offers promise for accurate disease diagnosis, with its autonomous learning and low error rates. Various advanced machine learning techniques, including logistic regression, KNN, decision trees, SVM, and algorithm ensembles, aid in early disease detection [5-6]. Despite numerous studies, precise MI prediction remains an ongoing challenge. Early MI detection is crucial for timely intervention, reducing muscle damage and complications [7-8]. Advanced machine learning (ML) can enhance accurate early detection and prediction compared to traditional diagnostic methods. This research aimed to assess ML algorithms for MI detection, using a combination of correlation-based and variance threshold feature selection and extraction methods to improve predictive accuracy. The goal was to create an integrated predictive model incorporating clinical investigations, medical history, lifestyle, and demographic data. The research aims to advance

cardiovascular knowledge and enhance personalized MI diagnosis and prognosis.

MATERIAL AND METHODS

This was a Case-Control study conducted during the period January 2021 to March 2023 after obtaining permission from Institutional Ethical Committee. The sample size of 1200 was determined using a formula given by Riley et al [9] predictors (20) and R^2(0.15). The case-to-control ratio was 1:3, so 300 MI patients and 900 non-MI patients were included in this research. Both cases and control were recruited from a tertiary care hospital in central India and with valid written consent the comprehensive information, including detailed medical history and laboratory test results, along with sociodemographic and lifestyle-related risk factors of MI were collected with a predesigned structured questionnaire. Three Controls were selected for each case after matching for age (±5 years) and sex. Only those cases were included who were above 18 years old and with MI diagnosed using standard clinical criteria. patients with severe illness were excluded.

Data pre-processing: The collected primary data had no missing values. For the nominal variables, we assigned a unique numerical label to each category, and for the ordinal variables, we assigned numerical labels according to the predefined order.

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 were estimated using Point biserial correlation, Phi correlation, and Cramer V respectively. The chi-square test was used to examine the relationship between two categorical variables. The normality of continuous variables was checked using the Shapiro-Wilk test. Continuous variables in MI vs. non-MI groups were compared using a t-test for normal, equal variance variables; else Wilcoxon Rank Sum test. Also, the correlation matrix checked multicollinearity among predictors.

Feature selection: Three feature selection methods were employed for prediction models. The first used the variance threshold, the second utilized correlation, and the third combined both. In the variance threshold, features with variance>2 were selected. For correlation-based, features with absolute correlation coefficient>0.40 with MI were chosen. The third method selected features with variance>2 and an absolute correlation coefficient> 0.4.

Model building: Among 45 variables, the variance threshold selected 17, the correlation-based picked 18, and the combination method chose 9 variables. The dataset was randomly split into training (70%) and testing (30%). Within the training set, a 80-20% split created a new training-validation dataset. Naïve Bayes, Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, and Adaptive Boosting models were built for each selected feature set. Logistic regression used Ridge (L2 regularization) to control overfitting and stabilize coefficient estimates. Six ML models predicted MI.

Evaluation Matrix: Model performance was assessed using validation and testing datasets. Metrics included Validation Accuracy, Testing Accuracy, Precision, Recall, Specificity, Negative Predictive Value, F1 Score, and Area under ROC (AUC). Additionally, real-time clinical validation involved 100 patients from a tertiary care hospital, including 20 MI patients.

RESULTS

Table 1: Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative(TN)	False Positive(FP)
Actual Positive	False Negative(FN)	True Positive(TP)

$$p_0 = \frac{TP+TN}{TP+TN+FP+FN} p_e = \frac{(TP+FN)\times(TP+FP)+(TN+FN)\times(TN+FP)}{(TP+TN+FP+FN)^2}$$

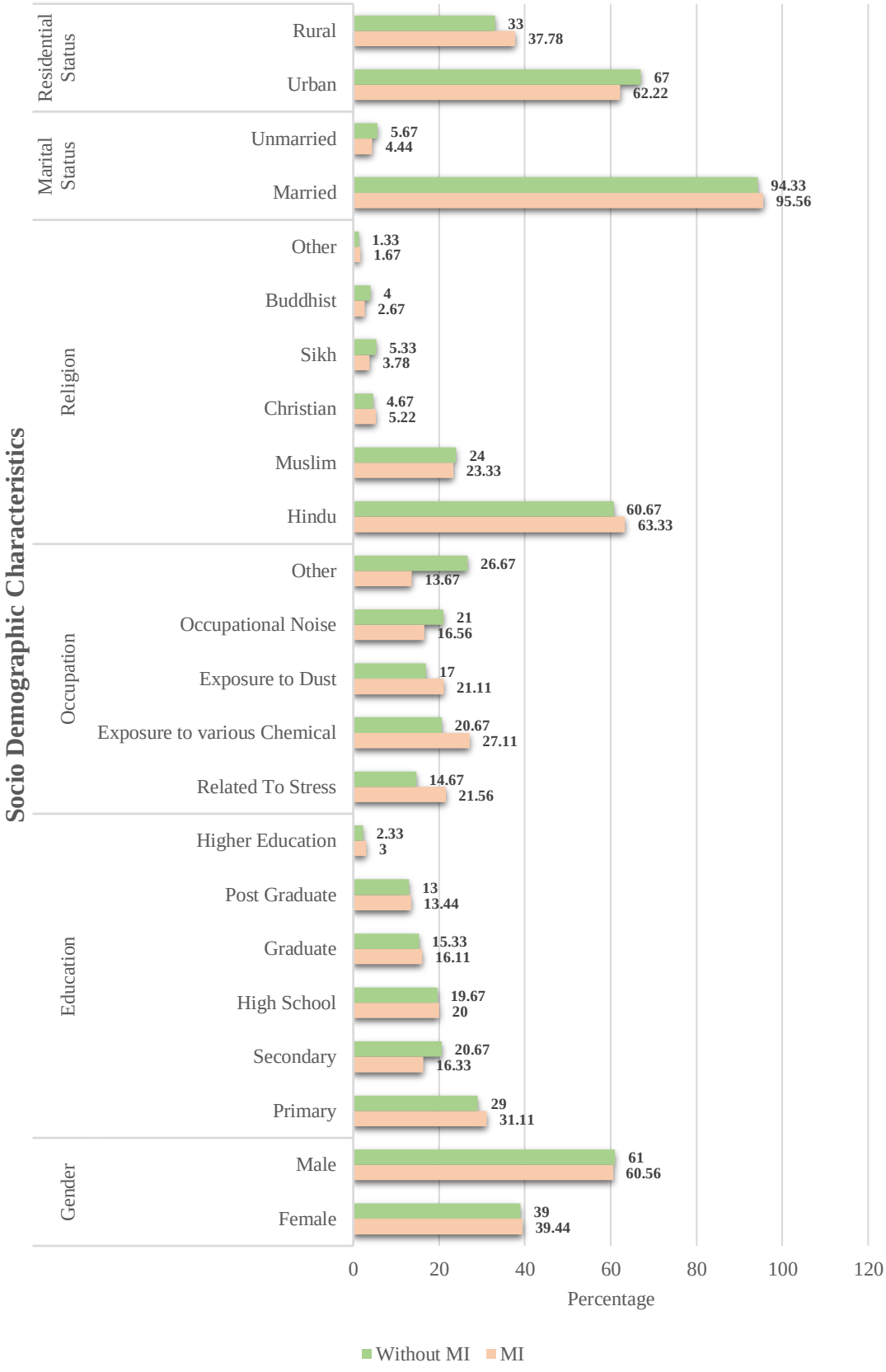
Cohen's Kappa (κ):$= \frac{p_0-p_e}{1-p_e}$ $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$

$$Positive\ Predictive\ Value(Precision) = \frac{TP}{TP+FP} \times 100\% Sensitivity(Recall) = \frac{TP}{TP+FN} \times 100\%$$

$$Specificity = \frac{TN}{TN+FP} \times 100\% Negative\ Predictive\ Value = \frac{TN}{TN+FN} \times 100\%$$

$$F1-Score = \frac{2(Precision \times Recall)}{Precision + Recall}$$

Figure 1 : Distribution of Socio Demographic Characteristics



:

Table 2: Comparison of Categorical Parameters between with & without MI Groups

Parameter			Non-MI (900) Frequency(%)	MI (300) Frequency(%)	P value
Symptoms					
1	Chest-Pain		180(20%)	174(58%)	0.000
2	Cold-Sweat		162(18%)	125(41.67%)	0.000
3	Dizziness & Light-Headedness		48(5.33%)	108(36%)	0.000
4	Fatigue		151(16.78%)	133(44.33%)	0.000
5	Shortness Breath		131(14.56%)	171(57%)	0.000
Medical History					
6	CKD		23(2.56%)	23(7.67%)	0.000
7	COPD		57(6.33%)	52(17.33%)	0.000
8	**MI**		**21(2.33%)**	**43(14.33%)**	**0.000**
9	CVD		119(13.22%)	150(50%)	0.000
10	DM		100(11.11%)	64(21.33%)	0.000
11	RA		59(6.56%)	44(14.67%)	0.000
12	HIV		2(0.22%)	1(0.33%)	0.111
13	Thrombophilia		35(3.89%)	41(13.67%)	0.000
14	HRT (for female only)		121(13.44%)	42(14%)	0.884
15	Preeclampsia (for female only)		14(1.56%)	7(2.33%)	0.525
16	PCOS (for female only)		25(2.78%)	22(7.33%)	0.001
17	NSAIDs		111(12.33%)	83(27.67%)	0.000
18	Type“A”person		96(10.67%)	78(26%)	0.000
Life Style					
19	Sedentary-Life-Style		173(19.22%)	175(58.33%)	0.000
20	Smoking	Never	638(70.89%)	178(59.33%)	0.0018
		Former	56(6.22%)	21(7%)	
		Occasional	55(6.11%)	25(8.33%)	
		Light /Moderate /Heavy	151(16.77%)	76 (25.33%)	
21	Alcohol	Never	726(80.67%)	158(52.67%)	0.000
		Former	45(5%)	46(15.33%)	
		Occasional	26(2.89%)	14(4.67%)	
		Light /Moderate /Heavy	103(11.44%)	82 (27.33%)	
22	Stress	Never	324(36%)	55(18.33%)	0.000
		Almost Never	225(25%)	25(8.33%)	
		Sometimes	118(13.11%)	50(16.67%)	
		Fairly Often	124(13.78%)	80(26.67%)	
		Very Often	109(12.11%)	90(30%)	
23	Sleep	Good	396(44%)	25(8.33%)	0.000
		Moderate	342(38%)	145(48.33%)	
		Poor	162(18%)	130(43.33%)	
24	Caffeine	Daily	756(84%)	215(71.67%)	0.000
		5-6 times per week	108(4.5%)	37(12.33%)	
		3-4 times per week	21(2.33%)	16(5.33%)	
		1-2 times per week	38(4.22%)	15(5%)	
		1-2 times per month	24(2.67%)	10(3.33%)	
		Rarely(never)	20(2.22%)	7(2.33%)	
Family History					
25	MI		40(4.44%)	29(29.00%)	0.000
26	DM		106(11.78%)	120(40.00%)	0.000
27	Hypertension		150(16.67%)	167(55.67%)	0.000
28	Hyperlipidaemia		133(14.78%)	163(54.33%)	0.000

Table: 2 & figure 1 displays MI associations with risk factors. Gender, residence, marital status, HIV, and hormone replacement therapy (females) show no significant link*(P>0.05)*. Symptoms (Chest Pain, Cold Sweat, Dizziness, Fatigue, Shortness of Breath) are significantly associated*(P<0.05).* Medical history (CKD, COPD, MI, CVD, DM, RA, thrombophilia) highly correlates with MI *(P<0.000).* PCOS history is significant for females*(P=0.001).* Type A personality is also significantly linked*(P<0.000).* Lifestyle habits (smoking, alcohol,

stress, sleep quality, caffeine intake) and family history (DM, Hypertension, and Hyperlipidaemia) are significantly associated*(P<0.000)*.

Table3: Comparison of Quantitative Parameters Between with and without MI Groups

Parameter	MI		Non-MI		P Value
	Mean	SD	Mean	SD	
Age	60.33	7.44	58.42	7.35	0.099
Income	14.13	15.17	14.04	13.65	0.9155
Diet Score	41.42	19.28	24.73	16.38	0.0001
Stress level	2.42	1.45	1.41	1.40	0.0000
BMI	27.97	2.27	24.06	1.99	0.0000
Iron level	151.37	28.16	105.92	9.10	0.0000
Homocysteine level	18.31	1.75	10.75	2.06	0.0000
CRP	4.16	2.10	1.27	0.28	0.0001
LDL	197.54	18.87	98.50	14.33	0.0000
HDL	34.82	5.38	50.62	6.70	0.0001
Triglyceride level	248.27	52.10	128.28	38.62	0.0000

In Table 3, we compare quantitative variables between MI and non-MI groups. Age and income were matched, showing no significant difference. Dietary practices, stress level, BMI, Iron level, Homocysteine, C-reactive protein, LDL, and TG levels were higher in the MI group*(P<0.000)*, positively correlating with MI occurrence. Conversely, HDL was significantly lower in the MI group, indicating a negative correlation with MI occurrence*(P=0.0001)*.

Table 4: Variable Selection Using Feature Selection Techniques

Feature Selection Techniques	Selected Variables	Number of Variables
Variance Threshold (Variance>2)	Age, Education, Occupation, Income, Religion, Smoking, Alcohol, Diet, Stress, Sleep, Caffeine, BMI, Iron level Levels of Homocysteine, CRP, LDL, HDL, TG level	17
Correlation Based (|Correlation>0.35|)	Chest Pain, Dizziness & Light-headedness , Shortness of Breath, History of CVD, SL, Diet, Sleep, Family History of MI, History of DM, Hypertension, Hyperlipidaemia, BMI, Iron level, Levels of Homocysteine, CRP, LDL, HDL, TG level	18
Combine correlation Based with Variance Threshold (Variance>2 & |Correlation| > 0.4)	Diet, Sleep, BMI, Iron level, Levels of Homocysteine, CRP, LDL, HDL, TG level	9

We selected 17 variables using variance threshold, 18 variables using correlation-based and 9 variables using a combination of variance threshold and correlation-based feature section techniques as shown in table 4.

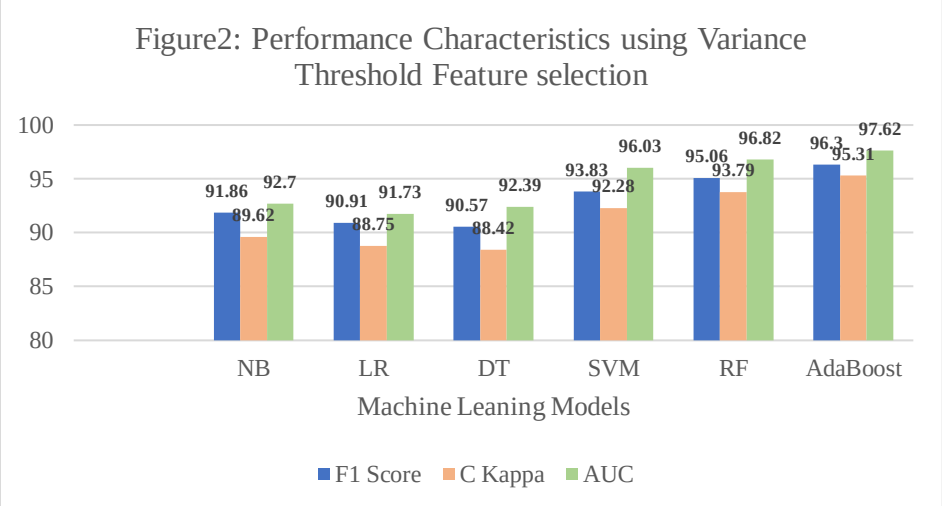
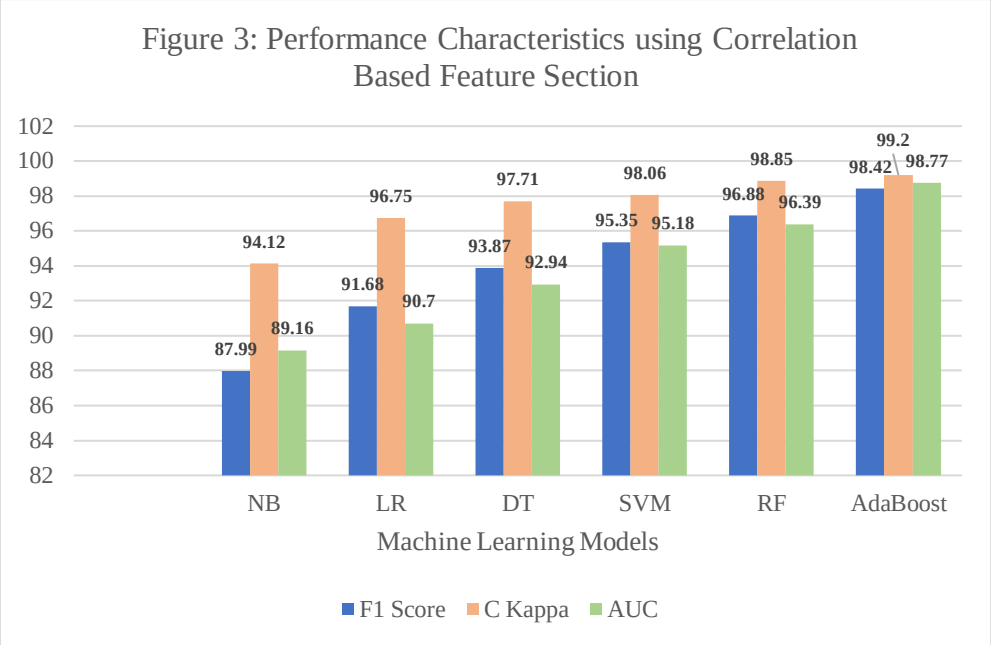


Figure: 2 shows the performance characteristics of models which were built using feature selected by variance threshold techniques. The SVM, RF, and AdaBoost models demonstrate excellent performance across various

metrics, F1-score(96.30% to 95.51%), C-Kappa(92.28% to 95.31%) and AUC (96.03% to 97.2%). Logistic Regression, while having a lower performance compared to these models, still shows reasonable performance. Naive Bayes, although having a high AUC(92.7%), has lower precision, recall, and C-kappa(89.62%) compared to other models. Figure 3 indicates the performance characteristics of models which were built using feature selected by correlation-based techniques. SVM, Random Forest, and AdaBoost performed exceptionally well across various metrics, achieving high F1 score(95.35% to 98.42%), Cohen's Kappa(98.06% to 99.20%), and AUC (95.18% to 98.77%) . Decision Tree and Logistic Regression, despite an unusually low accuracy value, show good performance in other metrics.



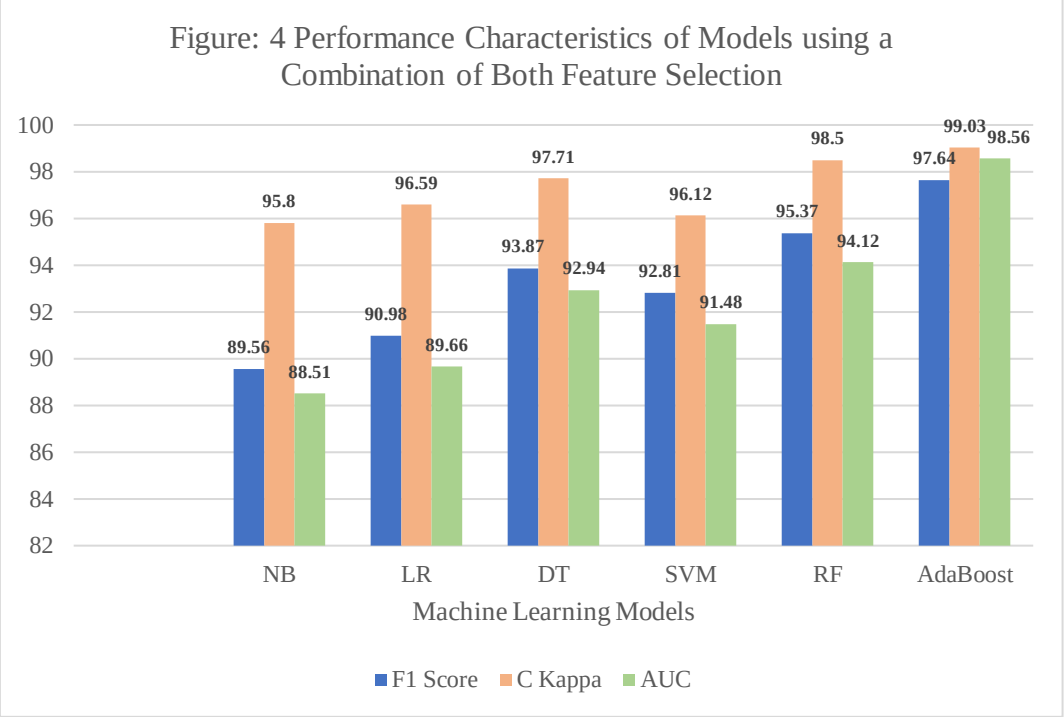
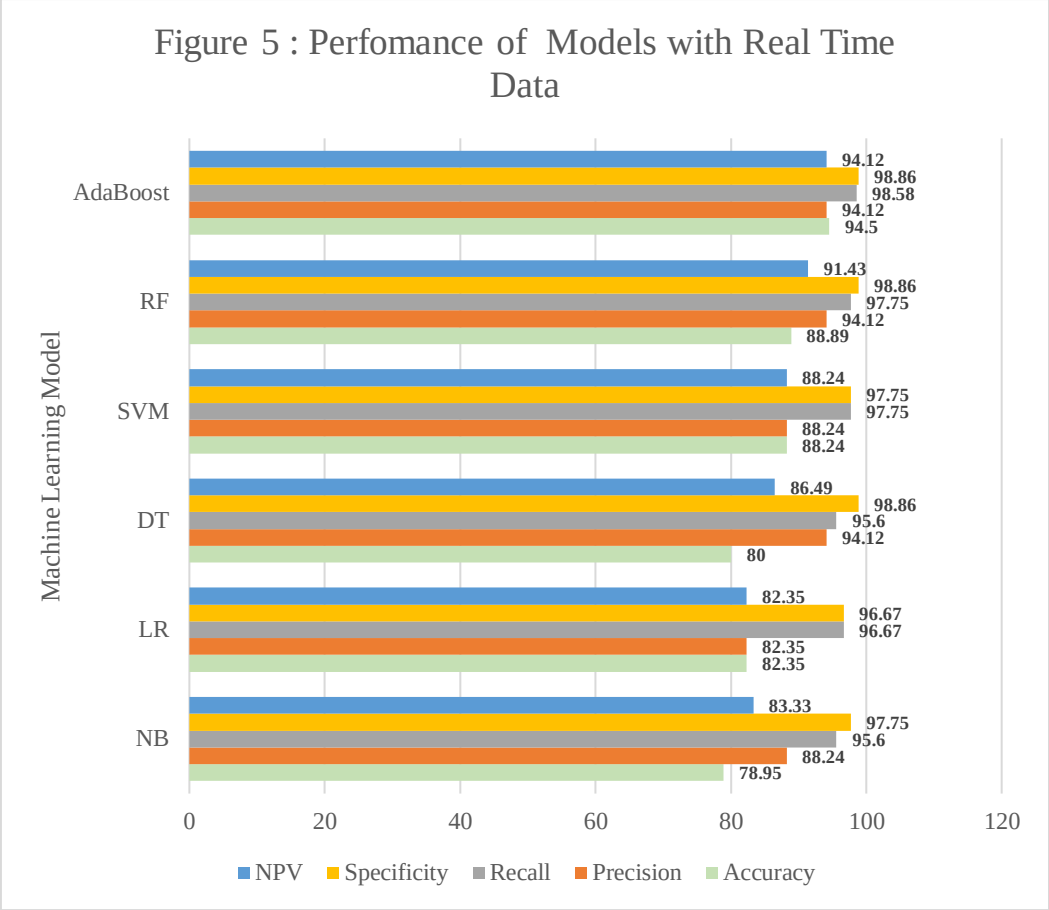
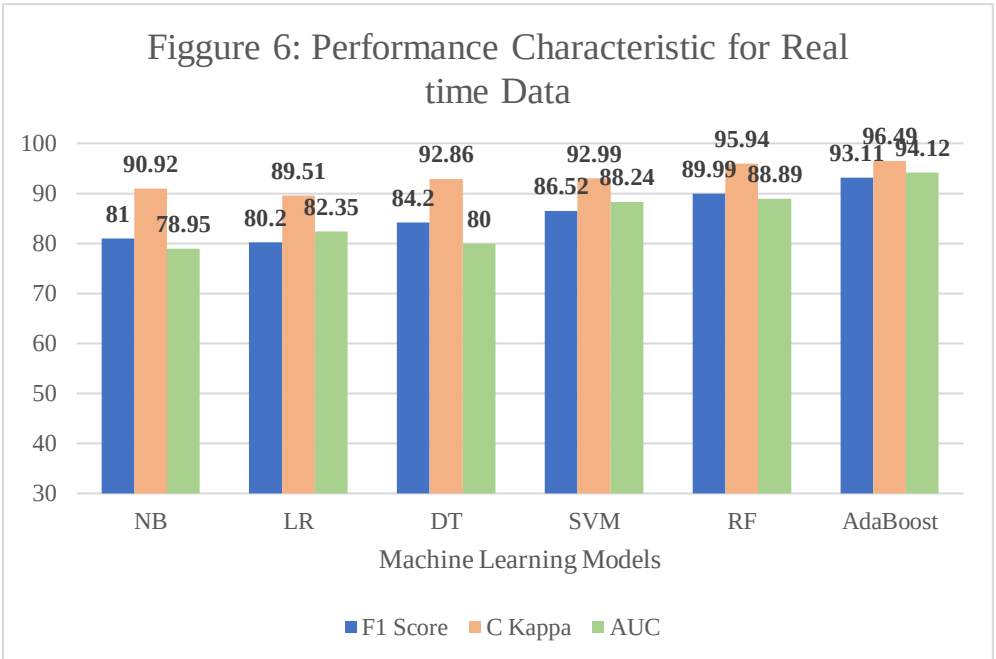


Figure: 4 shows the performance Characteristics of models based on a feature selection using combination of both. All models demonstrate high performance across various metrics. AdaBoost consistently shows the highest performance across most metrics(F1-score97.64%, C-Kappa=99.03%, AUC=98.06%) followed closely by Random Forest. Naive Bayes, Logistic Regression, Decision Tree, and SVM also demonstrate good performance, but slightly lower than AdaBoost and Random Forest.





Figures 5 & 6 shows the real-time clinical validation of models in which AdaBoost stands out as the top performer across most metrics, demonstrating high accuracy(94.5%), precision(94.12%), recall(98.58%), specificity(98.86%), NPV(94.12%), F1 score(93.11%), Cohen's Kappa(96.49%), and AUC(94.12%). SVM, Random Forest, and Decision Tree also perform well, showing strong capabilities in correctly classifying patients and achieving high values for various evaluation metrics. Naive Bayes and Logistic Regression, while having slightly lower performance compared to the top models, still demonstrate reasonable capabilities in classifying patients with a focus on sensitivity (recall) and specificity.

DISCUSSION

The average age of the study subjects was 58.88±7.39(years). In the MI group, males experienced MI at an earlier age compared to females, indicating early occurrence in males. This finding aligns with multiple studies [10]. For MI patients, prevalent symptoms were chest pain (58%), shortness of breath(57%), fatigue(44.33%),old sweats(41.67%), and Dizziness & Light Headedness(36%). Medical history of chronic kidney disease, chronic obstructive pulmonary disease, myocardial infarction, other cardiovascular diseases, diabetes militants, and thrombophilia significantly correlated with MI occurrence, consistent with previous studies [11-13]. Lifestyle-related factors like stress, sleep quality, alcohol consumption, and smoking were identified as major modifiable risk factors for MI, consistent with findings by Dugani [14]. we found that the elevated levels of BMI, Iron, Homocysteine, CRP, LDL, HDL and TG contribute to risk of MI since they are significantly correlated with MI.

In comparing ML models, those using correlation-based feature selection performed better compared to variance threshold techniques overall. The combined method, while using only half the variables, achieved similar accuracy and precision to correlation-based models. Notably, Random Forest and AdaBoost demonstrated superior performance, consistent with a study by Absar N et al. on Heart-Disease-Prediction [15].AdaBoost models excelled across metrics, displaying high F1 score, and AUC for all three feature selection methods. In real-time clinical validation, AdaBoost showcased superior accuracy(94.5%), F1 score(93.11%), Cohen's Kappa(96.49%), and AUC(94.12%) compared to other models. Utilizing AdaBoost, we achieved over 94% accuracy in predicting MI using just 9 health parameters associated with MI.

CONCLUSION

Our study uncovered essential MI insights, highlighting early occurrence in males and key symptoms such as dizziness, chest pain, and shortness of breath. Medical history and lifestyle were linked to MI. Biomarkers (BMI, iron, homocysteine, lipids) correlated with MI risk. AdaBoost showed strong performance, suggesting real-time MI prediction potential, underlining the need for comprehensive risk assessment and personalized interventions. Future research can enhance MI understanding and management by addressing overfitting and incorporating additional data for diverse testing is an important perspective noted by the authors.

LIMITATIONS AND FUTURE CONSIDERATIONS

Our data was sourced from a single tertiary care hospital in central India, that potentially limiting the generalizability. Notably, the dataset exhibited an imbalanced class distribution, that can potentially introduce bias in study's outcomes.

CONFLICT OF INTEREST

The authors assert that they have no conflicts of interest related to the publication of this paper.

REFERENCES

1. Shanthi Mendis et al, Writing group on behalf of the participating experts of the WHO consultation for revision of WHO definition of myocardial infarction: 2008–09 revision, *International Journal of Epidemiology*, Volume-40, Issue 1, February-2011, Pages 139–146
2. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019:a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;**396**(10258):1204–1222
3. World Health Organization:WHO.(2021). Cardiovascular diseases (CVDs). www.who.int. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
4. PrabhakaranD.,JeemonP.,RoyA. Cardiovascular diseases in India. Circulation. 2016;133:1605–1620.
5. Mohd Javaid, et al, Significance of machine learning in healthcare: Features, pillars and applications, International Journal of Intelligent Networks,Volume3,2022, Pages58-73,ISSN:2666-6030.
6. Ahsan MM, Luna SA, Siddique Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare (Basel).2022Mar15;10(3):541.
7. Mehta RH, Lopes RD, Ballotta A, Frigiola A, Sketch MH, Jr, Bossone E, et al. Percutaneous coronary intervention or coronary artery bypass surgery for cardiogenic shock and multivessel coronary artery disease? Am Heart J.(2010)159:141–7.
8. Bulluck H et al, Reducing myocardial infarct size: challenges and future opportunities.Heart.(2016)102:341–8.10.1136/heartjnl-2015-307855.
9. Riley RD, Snell KI, Ensor J, 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
10. Davis LL. A Qualitative Study of Symptom Experiences of Women with Acute Coronary Syndrome. J Cardiovasc Nurs. 2017 Sep/Oct;32(5):488-495.
11. Pierfranco Terrosu, Relapse of chronic obstructive pulmonary disease and myocardial infarction: what is the connection? *European Heart Journal Supplements*, Volume 22, Issue Supplement_L, November 2020, Pages L151–L154.
12. Draman MS et al. A silent myocardial infarction in the diabetes outpatient clinic: case report and review

of the literature. Endocrinol Diabetes Metab Case Rep. 2013; 2013:130058.
13. Bereczky Z, Balogh L, Bagoly Z. Inherited thrombophilia and the risk of myocardial infarction: current evidence and uncertainties. Kardiol Pol. 2019 Apr 18;77(4):419-429.
14. Dugani SB et al. Risk factors associated with premature myocardial infarction: a systematic review protocol. BMJ Open. 2019 Feb 11;9(2):e023647
15. Absar N et al. The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction. Healthcare (Basel). 2022 Jun 18;10(6):1137.