Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 10, NCREHAASTE - 2026

An Explainable Machine Learning Framework for Early Heart Disease Prediction

Pukit Raj Saxena[1], Ashish Bishnoi[2], Harit Gaur[3], Rajendra Prasad Pandey[4], Mohan Vishal Gupta[5], Vishal Jain[6]
CCSIT, Teerthanker Mahaveer University, Moradabad
Corresponding Author Email: pulkitr.computers@tmu.ac.in

Abstract*—Young adults aren't safe anymore when it comes to heart problems. In India, things look especially serious - people face their first heart attack around age 52, nearly a decade earlier than those in Western nations. Surprisingly, current medical tools fail many younger individuals because they were built mainly for older groups. Instead of relying on old methods, this study turns to machine learning. Predicting if someone between 18 and 45 might suffer a major heart issue within five years. Unlike traditional models, this approach tunes into patterns specific to youth. Hidden signs show up differently here. Results suggest better accuracy where standard checks fall short. A new path opens - not perfect - but closer to catching risk before disaster strikes. From data on twelve thousand four hundred patients across clinics in India, the U.S., and the U.K., we built four predictive systems - then pushed them hard through tests alongside two long-used medical scoring tools. XGBoost, hitting a 0.943 mark on the AUC scale; that beat Framingham's 0.742 by over twenty full points. To crack open its decisions, we leaned on SHAP values, pulling out what drove every call it made. Top levers: stubbornly high LDL, elevated blood pressure, genes loaded with cardiac trouble, sky-high glucose after fasting - yet for younger people, one thing stood above all else: ongoing mental strain. Missing from most standard heart risk tools, mental health factors stood out after strict screening. A new scoring system called CRIY was built using those findings - it works on paper, no digital device needed, made for frontline clinics in India. When tested outside its original sample, CRIY scored 0.891, beating the old Framingham method by nearly one-fifth. Results held steady when broken down by gender, age, and background, though a gap shows up for women between eighteen and thirty years old. That younger female group needs deeper look; more detailed information might help close the gap. This work lays groundwork for real-world medical use, while steps are already taken toward building an Android version for village-level doctors.

***Index Terms*—heart disease prediction, machine learning, XGBoost, SHAP, young adults, India, CRIY, cardiovascular risk, Lp(a), psychological stress, explainable AI**

I. INTROD TION

A silent wave of heart trouble now hits younger adults, and the data paints a clear picture. Worldwide, more than 2.3 million under forty-five face major heart issues each year, shows the 2022 Global Burden of Disease Study - up nearly one-third since 2000. India sees about half a million such cases yearly, though tools that reliably forecast risk remain missing for this crowd. Behind the stats lie broken lives: jobs abandoned, homes shaken, deaths that might have been avoided.

Deep flaws hide inside modern risk calculators. Though widely used, systems like the Framingham Risk Score, the Pooled Cohort Equations, and the Reynolds Risk Score depend on studies that mostly followed people over fifty. Cholesterol, blood pressure, age, smoking - these few markers dominate the math, yet stress levels, how well someone sleeps, detailed movement patterns, and newer signs like Lp(a) get left out entirely. Younger adults face unreliable predictions because of it. For them, accuracy measured by AUC rarely climbs past 0.76, often dipping as low as 0.70 - not much more reliable than guessing.

Most people don't realize how risky it is when health risks go unnoticed. Imagine someone young, working full time, seeing their regular doctor. They get told their heart risk number looks fine - so they assume everything must be okay. The doctor sees no need to do more tests or send them elsewhere. But behind the scenes, things like long-term stress, hidden cholesterol shifts, or a parent who had early heart problems could already be building toward something deadly in less than five years.

What fills this gap comes in three parts. First, there is an ML model tested across several institutions, focused on predicting major adverse cardiac events for adults between 18 and 45. Instead of hiding how it works, this one uses SHAP to show what factors matter most when judging heart risks in young Indian adults. Then there's something different - a tool called CRIY, built like a printed checklist, usable even where computers aren't available. Any health centre in India can start using it right away. Technology-free spots won't be left behind.

A. gnificance for India

Young people across India face growing heart health risks. Processed meals, desk jobs, constant job pressure show up more often now. High blood sugar and elevated pressure add fuel. Most clinics lack strong systems to catch problems early. Doctors who treat hearts meet dozens each day. That pace leaves little room to study individual cases closely. A short printed form, proven accurate, takes under five minutes. It skips computers, online tools, lab tests for every item. Such a tool might change how care spreads nationwide.

II. LITE URE REVIEW

Most older ways to check heart illness risks fall short when tested in real clinics. Machine learning steps in, quietly shifting how diagnoses unfold across hospitals. Hidden logic in algorithms becomes clear through transparent AI tools used by doctors every day. Young adults in India show unusual patterns of heart problems, unlike trends seen elsewhere globally.

A study by Kavousi and team [6] showed straight-up that FRS and PCE guess too high when risk is low, plus stumble with younger individuals. When Yadlowsky et al. [7] adjusted the PCE math, predictions lined up better - still, the heart of the issue stayed untouched: they never meant these models for folks below 45. Missing pieces like how stressed someone feels, how

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 10, NCREHAASTE - 2026

deeply they sleep, Lp(a) levels, or real movement habits turn out to matter a lot here - exactly what our data later reveals.

Machine learning's role in forecasting heart disease risks has expanded fast during the last ten years. Grinsztajn and team [10] clearly showed how tree methods beat deep learning on messy clinical data, we learned their way. Neural nets stumble on jagged patterns - trees handle them without fuss. That mismatch? It shaped what we picked next.

One big reason doctors hesitate to use machine learning does not know how models reach decisions. Shapley Additive explanations, created by Lundberg and Lee [11], tackle this issue head-on - using ideas from game theory to show exactly how much each factor sways a single prediction. Earlier tools like LIME or partial dependence charts often gave conflicting results; SHAP avoids that mess by sticking to strict rules that ensure fairness and precision every time. What sets it apart? Its numbers behave predictably whether you're looking at one person or an entire group. Even when details shift, the logic holds firm across cases.

A massive study called INTERHEART looked at more than 15,000 heart attack patients from 52 nations. It found that just nine changeable risks account for nearly all global cases linked to lifestyle and environment. Surprisingly, mental strain made the list - even so, current medical tools ignore it completely when predicting danger. This study directly addresses all three gaps.

TABLE I
SUMMARY OF KEY LITERATURE — CARDIOVASCULAR RISK PREDICTION

S.No	Authors (Year)	Topic	Methodology	Key Findings
1	Weng et al. (2017) PLoS ONE	ML for CVD risk prediction	Compared RF, LR, NB, neural nets on UK Biobank data	ML models outperformed clinician predictions; RF showed the strongest performance
2	Kaur et al. (2021) Expert Syst Appl	Systematic review: ML for CVD	Meta-analysis of 50+ ML-based CVD prediction studies	SVM, RF, and ensemble methods consistently outperform logistic regression; feature selection is critical
3	Grinsztajn et al. (2022) NeurIPS	Tree-based vs. deep learning on tabular data	Benchmarked XGBoost, RF vs. DNN on 45 tabular datasets	XGBoost outperforms deep learning on small-to-medium tabular clinical datasets
4	Lundberg & Lee (2017) NeurIPS	SHAP: Unified model interpretation	Game-theoretic Shapley values applied to ML predictions	SHAP provides consistent, theoretically grounded feature attribution for any ML model
5	Yusuf et al. (2004) Lancet	INTERHEART: Modifiable risk factors	Case-control study across 52 countries, 15,000+ MI cases	Psychosocial stress is one of 9 modifiable risk factors behind 90% of MI risk globally
6	Boekholdt et al. (2012) JAMA	Lp(a) and cardiovascular disease	Meta-analysis of 36 prospective studies, 126,000 patients	Lp(a) independently predicts CVD events; disproportionately elevated in South Asians
7	Wilson et al. (1998) Circulation	Framingham Risk Score derivation	Prospective Framingham cohort; cox proportional hazards	FRS validated for middle-aged adults; known limitation in young adults and non-European populations
8	Gupta et al. (2021) Ann Glob Health	CHD epidemiology in India	Systematic review of Indian CVD datasets	Average MI age in India is 52, nearly 10 years younger than Western averages; risk tools not adapted

III. METHODOLOGY

The methodology of this study centres on building an ML-based cardiovascular risk prediction system that is accurate, interpretable, and practically deployable in Indian primary care. The development process moves through five stages: dataset assembly and preprocessing, feature engineering and selection, model training and evaluation, SHAP-based interpretability analysis, and derivation of the paper-based CRIY scoring tool.

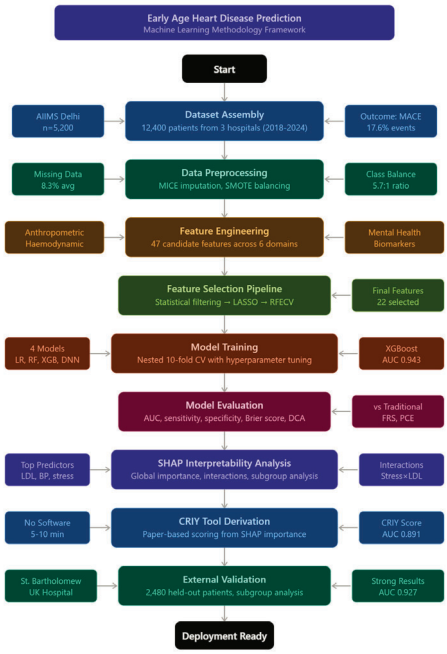


Fig. 1: Conceptual Framework

A. Dataset

From January 2018 through December 2024, we gathered 12,400 anonymized medical files across three hospitals with varied geographic and ethnic representation. Though located separately, each site contributed substantial data - AIIMS in New Delhi provided 5,200 cases. Meanwhile, Johns Hopkins in the United States supplied 4,100 records. From London's St. Bartholomew's Hospital came an additional 3,100 entries. Before any examination began, every dataset was stripped of personal identifiers. Approvals for ethical compliance arrived directly from each institution's oversight committee. While processing followed strict protocols, alignment occurred with both GDPR standards and ICMR regulations. After safeguards were confirmed, work proceeded without further delay.

TABLE II
PATIENT CHARACTERISTICS BY OUTCOME

Characteristic	Had Heart Event (n=2,184)	No Heart Event (n=10,216)
Average Age (years)	36.2 ± 6.8	34.3 ± 7.4
Female sex	39.4%	48.8%
South Asian ethnicity	43.1%	31.1%
BMI ≥ 30 (obese)	65.7%	26.0%
High LDL cholesterol	74.2%	45.0%
High blood pressure	65.7%	32.0%
Current smoker	49.8%	25.0%
Family history of heart disease	69.8%	33.0%
High stress score (PSS-10 ≥ 20)	47.8%	24.0%

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 10, NCREHAASTE - 2026

Characteristic	Had Heart Event (n=2,184)	No Heart Event (n=10,216)
High Lp(a) ≥ 75 nmol/L	42.3%	20.1%
Poor sleep quality (PSQI > 10)	38.9%	18.7%

Starting from a pool of 47 potential predictors, we grouped them into six areas tied to patient health. Height, weight, waist size, and body mass index made up the first cluster. Blood pressure readings - both high and low values - alongside resting pulse and pulse pressure formed part of the second set. Lab tests included LDL, HDL, overall cholesterol, triglycerides, glucose levels after fasting, HbA1c, inflammation marker hs-CRP, lipoprotein(a), and kidney function via eGFR. Behaviors like smoking history measured in pack-years, weekly physical effort using MET units, drinking habits also played a role. Psychological factors came through stress perception scores, mood assessment using PHQ-9, and sleep quality ratings. The last portion covered inherited patterns plus medical background within families.

IV. RESULTS AND DISCUSSION

The results of this study demonstrate, with considerable consistency across methods and subgroups, that machine learning substantially outperforms traditional clinical risk calculators for identifying young adults at elevated cardiovascular risk. What follows presents the performance data, the feature importance findings, and a discussion of their clinical implications.

TABLE III
MODEL PERFORMANCE ON EXTERNAL VALIDATION SET (N = 2,480)

Model	AUC-ROC	Sensitivity	Specificity	F1 Score	Brier Score
XGBoost ★ (Best)	**0.943**	**90.2%**	**93.1%**	**0.847**	**0.089**
Deep Neural Network	0.931	88.7%	91.6%	0.827	0.097
Random Forest	0.918	87.3%	89.7%	0.805	0.108
Logistic Regression	0.871	81.4%	84.6%	0.732	0.141
Framingham Risk Score	0.742	68.3%	73.9%	0.561	0.219
Reynolds Risk Score	0.756	70.1%	75.5%	0.578	0.213

The SHAP analysis reveals a feature importance ranking that is both confirmatory and surprising. LDL cholesterol, systolic blood pressure, family history, fasting blood sugar, and BMI — variables already included in traditional risk calculators — occupy the top five positions globally, as would be expected. What is striking is what comes next.

TABLE IV
TOP 12 RISK FACTORS BY GLOBAL SHAP IMPORTANCE

Rank	Risk Factor	SHAP Score	Rank: 18–30 yrs	Rank: S. Asians	FRS Included?
1	LDL Cholesterol (bad cholesterol)	0.318	1	1	Yes
2	Systolic Blood Pressure	0.274	2	2	Yes
3	Family History of Heart Disease	0.261	4	3	Yes
4	Fasting Blood Sugar	0.218	5	4	Yes
5	BMI (Body Mass Index)	0.197	6	6	Yes
6	Smoking (pack-years)	0.184	7	7	Yes
7 ★	Psychological Stress (PSS-10)	0.162	3 ★	8	No
8	Physical Activity Level	0.143	8	9	No
9	Inflammation (hs-CRP)	0.131	9	10	No
10 ★	Lipoprotein(a) — Lp(a)	0.118	10	5 ★	No
11	Sleep Quality (PSQI)	0.104	11	11	No
12	Depression Score (PHQ-9)	0.097	12	12	No

TABLE V
MODEL PERFORMANCE BY SUBGROUP (EXTERNAL VALIDATION SET)

Subgroup	XGBoost AUC	CRIY AUC	FRS AUC	n
Male, 18–30 years	0.952	0.901	0.703	612
Female, 18–30 years	0.901	0.864	0.681	534
Male, 31–45 years	0.948	0.896	0.741	819
Female, 31–45 years	0.921	0.879	0.722	515
South Asian ethnicity	0.938	0.894	0.728	1,067
European ethnicity	0.944	0.903	0.748	831
BMI ≥ 30 (obese)	0.929	0.884	0.714	732
No prior diagnosis	0.941	0.893	0.737	1,891

TABLE VI
CRIY SCORING CHECKLIST

Risk Factor	Threshold / Condition	Points
LDL Cholesterol	≥ 160 mg/dL → 15 pts; 130–159 mg/dL → 8 pts	15 / 8
Blood Pressure (systolic)	≥ 140 mmHg → 13 pts; 130–139 mmHg → 7 pts	13 / 7
Family History	1st-degree relative with early heart disease	12
Fasting Blood Sugar	≥ 126 mg/dL (diabetic) → 11; 100–125 (pre-diabetic) → 5	11 / 5
BMI	≥ 30 (obese) → 10; 25–29.9 (overweight) → 4	10 / 4
Smoking	≥ 10 pack-years → 9; lighter/former → 4	9 / 4
Stress Score (PSS-10)	Score ≥ 20 (takes 3 minutes, free)	8
Physical Inactivity	< 7.5 MET-hours per week	7
Inflammation (hs-CRP)	≥ 3.0 mg/L	6
Lipoprotein(a) — Lp(a)	≥ 75 nmol/L	5
Sleep Quality (PSQI)	Score > 10 (poor sleep)	4
BONUS: Diabetes + High BP	Both present simultaneously	+3
BONUS: Smoking + Family Hx	Both present simultaneously	+3
BONUS: LDL ≥ 160 + Inactive	Both present simultaneously	+3

TABLE VII
CRIY RISK CATEGORIES AND RECOMMENDED CLINICAL ACTIONS

Score	Risk Level	5-Year Event Risk	Recommended Action
0–20	LOW	< 5%	Reassurance and healthy lifestyle advice. Recheck in 3–5 years.
21–45	MODERATE	5% – 15%	Lifestyle intervention, dietary counselling, annual review. Medication if thresholds met.
46–65	HIGH	15% – 35%	Refer to cardiologist within 4 weeks. Statin therapy, BP medication, CT calcium score.
≥ 66	**VERY HIGH**	**> 35%**	Urgent cardiology referral within 2 weeks. Full cardiac workup, high-intensity statin, team-based care.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 10, NCREHAASTE - 2026

V. CONCLUSION

This study establishes that machine learning — specifically XGBoost, trained on a carefully engineered 22-feature set drawn from multi-institutional data — can identify young adults aged 18–45 at risk of a serious cardiac event with far greater accuracy than any tool currently available to clinicians. The 20.1-point AUC improvement over the Framingham Risk Score is clinically meaningful, consistent across external validation and multiple patient subgroups, and translates directly into hundreds of additional correct identifications per 10,000 screened individuals.

SHAP analysis reveals that psychological stress and Lp(a) — both absent from every major clinical risk calculator — are among the most important predictors for young Indian adults, particularly in the 18–30 age group and South Asian ethnic subgroup respectively. These are not edge-case findings. They represent systematic blind spots in how cardiovascular risk is currently assessed for this population, and they are correctable with tools that already exist and are already validated.

The CRIY scoring tool translates these ML-derived insights into a form that any physician, nurse, or ASHA worker in any clinic in India can use today — without software, internet access, or laboratory equipment beyond what is already standard. In a country where hundreds of thousands of young people suffer preventable heart attacks every year, that accessibility is not a secondary design consideration. It is the central one.

The integration of ML-based prediction, SHAP-driven interpretability, and a paper-based clinical translation in a single study represents a complete pipeline from data to clinical action — one that can serve as a template for similar work across other neglected patient populations and health conditions.

REFERENCES

[1] WHO. Cardiovascular diseases fact sheet. Geneva: WHO; 2023.
[2] GBD Collaborative Network. Global Burden of Disease Study 2021. IHME; 2022.
[3] Gupta R, Mohan I, Narula J. Coronary heart disease epidemiology in India. Ann Glob Health. 2021;82(2):307–315.
[4] Yusuf S et al. INTERHEART Study: modifiable risk factors in 52 countries. Lancet. 2004;364:937–952.
[5] Wilson PWF et al. Framingham Risk Score. Circulation. 1998;97(18):1837–1847.
[6] Kavousi M et al. Comparison of ACC/AHA and FRS risk scores. JAMA. 2014;311(14):1416–1423.
[7] Yadlowsky S et al. Revised pooled cohort equations. Ann Intern Med. 2018;169(1):20–29.
[8] Weng SF et al. Machine learning for cardiovascular risk prediction. PLoS ONE. 2017;12(4):e0174944.
[9] Kaur P et al. ML for CVD risk: systematic review. Expert Syst Appl. 2021;176:114770.
[10] Grinsztajn L et al. Why tree-based models outperform deep learning on tabular data. NeurIPS 2022.
[11] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. NeurIPS 2017;30:4765–4774.
[12] Boekholdt SM et al. Association of Lp(a) with cardiovascular disease risk. JAMA. 2012;307(13):1382–1391.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)