Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 10, 2025

EXPLAINABLE PROGNOSIS OF ISCHEMIC HEART DISEASE: CALIBRATED ML WITH KNN BASELINES AND EXTERNAL VALIDATION

Um E Aimen[*1], Sadaqat Hussain[2], Rahbar Ali[3], Aleena Farooq[4], Farah Zaidi[5], Areeba Batool[6], Muwadat Ali[7]

[*1,4]*Department of Computer Science, COMSATS University Islamabad, Abbottabad Campus, Pakistan*
[2]*Department of Information Technology, Karakoram International University, Gilgit, Pakistan*
[3]*Department of Software Engineering, University of Sindh, Jamshoro, Pakistan*
[5]*Department of Software Engineering, COMSATS University, Lahore Campus, Pakistan*
[6]*University of the Punjab, Lahore, Pakistan*
[7]*Department of Software Engineering, Dalian University of Technology*

DOI:https://doi.org/10.5281/zenodo.17254741

Keywords
Ischemic Heart Disease; Machine Learning; Explainable AI; Calibration; Prognosis; Trust; k-Nearest Neighbors; External Validation; Physician Adoption

Article History
Received: 11 July 2025
Accepted: 21 September 2025
Published: 03 October 2025

Corresponding Author: *
Um E Aimen

Abstract

Ischemic heart disease (IHD) remains a leading cause of global morbidity and mortality. Traditional risk scores often lack personalization and transparency, limiting their clinical utility. This study develops and evaluates explainable, calibrated machine learning (ML) models for IHD prognosis, benchmarked against k-nearest neighbors (kNN) baselines and validated across independent datasets. A quantitative, survey-based approach was combined with predictive modeling. Data from 400 purposively sampled physicians were analyzed using SPSS to assess perceptions of ML explainability, calibration, trust and adoption intention. Reliability, descriptive statistics, correlations and multiple regression were performed. ML models, including kNN, logistic regression, ensemble methods and an explainable calibrated ML framework, were trained and evaluated using internal and external validation. Performance was measured using AUC, Brier scores and calibration curves. Survey analysis showed strong physician support for explainability ($M = 4.21$, $SD = 0.58$) and high behavioral intention to adopt ML tools ($M = 4.18$, $SD = 0.59$). Regression confirmed explainability ($\beta = .33$, $p < .001$), usefulness ($\beta = .30$, $p < .001$) and trust ($\beta = .22$, $p < .001$) as the strongest predictors of adoption. Calibration literacy also contributed positively ($\beta = .11$, $p = .008$). In model comparisons, kNN achieved an AUC of 0.70 (external validation) while the explainable calibrated ML model outperformed all others, with AUC = 0.87 and the lowest Brier score (0.148), demonstrating both accuracy and reliability across datasets. Explainability, calibration and external validation are critical enablers of physician trust and adoption of ML prognosis tools for IHD. When combined with organizational support and training, these models hold significant potential to enhance prognostic accuracy and improve patient care in cardiology.

INTRODUCTION

Ischemic heart disease (IHD) remains the leading cause of mortality worldwide, accounting for a significant portion of the global cardiovascular disease burden. Despite advances in diagnostic imaging,

biomarkers and preventive cardiology, prognostication in IHD continues to pose a critical challenge. Clinical decision-making often depends on multifactorial assessments that integrate comorbidities, laboratory values, imaging findings and patient history. Traditional risk scores and regression-based models are often constrained by assumptions of linearity, lack of personalization and suboptimal calibration for a diverse population (Badawy et al, 2025; Li, Wu, & Gu, 2025). This gap has spurred increasing interest in the use of machine learning (ML) methods to increase predictive accuracy, risk stratification and enable precision cardiovascular medicine.

In the past few years, explainable artificial intelligence (XAI) has become an essential need for the clinical deployment of ML. Unlike black box models, explainable ML helps clinicians to understand the contribution of features and can lead to greater transparency and trust in algorithmic recommendations. Several studies have shown the value of explainability for cardiovascular and neurovascular prediction tasks. For example, SHAP enhanced models have been used for stroke risk in hypertensive patients (Akinwale, Bosede, & Awe, 2025) and interpretable frameworks have been employed to improve stroke prognosis and transparency (El-Geneedy, El-Din Moustafa, Khater, Abd-Elsamee, & Gamel, 2025; Guo et al, 2025). Similarly, Bilal et al. (2025) and Song, Ming, Liu and Liu (2025) point out that incorporating explainability in cardiovascular and pediatric models promotes clinical adoption and aids in decision support. All these advances demonstrate that interpretability holds significant clinical utility for prognostic modeling.

Apart from explainability, calibration is an essential characteristic of ML models that is often neglected in the field of cardiovascular research. Calibration is a concept of agreement between predicted probabilities and observed outcomes that is essential for risk communication and shared decision-making. Poorly calibrated models, even if highly accurate, may result in overestimation or underestimation of risk and, consequently, lead to a compromise in patient care. He et al. (2025) discussed the value of calibration in interpretable models for predicting in-hospital mortality in patients with chronic heart failure and Ni et al. (2025) showed that good calibration is crucial for predicting cardiac arrest outcome in intensive care units. These findings emphasize that model reliability is not all about discrimination and must consider probabilistic accuracy, particularly in high-stakes cardiovascular situations.

Comparison of ML methods with classical baselines is yet another critical element in rigorous methodology design. KNN, being a traditional algorithm, can serve as both a transparent and interpretable marker of algorithm validation of the added value of the complex models. Previous stroke or kidney disease or coronary disease forecasting studies often conduct methodological comparisons between new ML architectures and existing baselines to guarantee their methodological strength (Chen et al, 2025; Karabacak et al, 2025; Waqar et al, 2025). In addition to contextualizing the performance of the ML model, the use of kNN comparisons is also known to enhance replicability and credibility of clinical studies.

External validation is a final important procedure that allows establishing clinical relevance. Most ML studies have been overfitted to single-center data, which decreases the extent to which they can be generalized to other healthcare systems and patient groups. External validation, especially where there is multicenter cohort validation, offers the evidence of scalability and strength. Gu and Lu (2025) established the need to external validate mortality prediction of critically ill cancer patients and Ye et al. (2025) involved a prospective validation of the success of percutaneous coronary intervention prediction using ML. In the same manner, translational impact has been established in multicenter validation studies in ischemic stroke and hypertensive chronic kidney disease that external validation boosts translational impact (Luo et al, 2025; Chen et al, 2025). Such observations relate to the fact that methodological rigor must entail not only explainability and calibration but also non-development cohort validation.

The current paper addresses these gaps of critical importance by proposing an elicitable and optimized ML model to predict prognosis in ischemic heart disease, comparing it with kNN baselines and evaluating its external validity of generalizability. The study will contribute to the improvement of the state of ML prognostication in IHD by matching the predictive accuracy with interpretability, reliability

and clinical trustworthiness. This combined method will likely offer greater advantages to clinicians in decision-making, aid in patient communication and add to the existing body of research in precision cardiology (Kaur, 2025; Wang et al, 2025). This study highlights address the fundamental methodological limitations that hindered the utilization of ML in cardiovascular care, place this technology in a position to be translated to practical clinical settings and provides support with explainability, calibration and external validation.

2. Literature Review

Machine learning (ML) in cardiovascular medicine has become a significant field in the last decade and there is a growing focus on transparency, calibration and validation. In this section, the recent contributions to explainable artificial intelligence (XAI) in cardiovascular and associated disease, calibration of the ML models, comparative baseline methods and the importance of external validation to clinical implementation are reviewed.

Explainable AI in Cardiovascular and Stroke Prediction

In medical ML research, explainability has become the key concern. Recently, Akinwale, Bosede and Awe (2025) presented a SHAP-enhanced ML model to predict the stroke risk in patients with hypertension and established that interpretable features enhanced the acceptance of the model by clinicians. El-Geneedy, El-Din Moustafa, Khater, Abd-Elsamee and Gamel (2025) suggested a holistic XAI system to predict stroke and emphasized how the visualization of the features contribution may lead to trust in the model advice. Consistent with these conclusions, Guo et al. (2025) used SHAP to assist with both stroke prognosis by means of vector machines and logistic regression nomograms, which validates the applicability of interpretable architectures in various model designs.

In cardiovascular, Bilal et al. (2025) constructed an AI-based intelligent system of precision cardiovascular prediction that prioritized elucidation, whereas Song, Ming, Liu and Liu (2025) created an understandable model of Kawasaki disease complicated by coronary aneurysms. The two studies emphasized the need of interpretability to clinical translation. Li, Wu and Gu (2025) used explainable models in another application to develop a comorbid coronary heart disease and depression model with dual clinical and psychological significance of interpretability in complex patient groups.

These articles all indicate that XAI is not such a technical supplement but it is rather a fundamental necessity of implementing the ML prognostic models in practice, in cardiovascular care.

Calibration and Reliability of ML Models

Calibration is another dimension of clinical ML that has not been well studied but it is a critical element. Having studied interpretable models to predict in-hospital mortality in patients with chronic critical illness and heart failure, He et al. (2025) discovered that adequate calibration was the way to build clinical trust. This conclusion was also reflected by Ni et al. in their prediction model of cardiac arrest results, as the models were able to predict outcomes with high precision, the key issue was that even the most successful models, the researchers indicated, had to be carefully calibrated to prevent misleading estimates of the probability.

Calibration has also been highlighted in other fields, e.g. nephrology. The authors created an explainable hypertensive chronic kidney disease prognosis ML model (Chen et al, 2025) and used calibration to enhance effectiveness. Badawy, Ramadan and Hefny (2025) emphasized the unification of ensemble ML and calibration methods in predicting coronary heart disease and they noted that the clinical usefulness of these predictive methodologies is not determined by the crude degree of accuracy but by the probabilistic degree of accuracy.

Calibration is especially critical in cardiovascular prognostication, in which treatment cuts are frequently based upon the estimated risk probability. Uncalibrated models might result in inappropriate interventions which is why calibration should be a methodological priority.

Comparative Baselines and Benchmarking with kNN

Strong ML research often compares itself with standard algorithms to prove methodological soundness. The algorithm of the k-nearest neighbors (kNN) is simple but offers a clear base on which a comparative analysis can be performed. In their

classical baselines, Chen et al. (2025) contextualized their ML model of chronic kidney disease, whereas Karabacak et al. (2025) used explainable ML methods based on data and a baseline model to predict distal medium vessel occlusion. Waqar, Shahnawaz, Saleem, Dawood, Muhammad and Dawood (2025) also highlighted the role of feature identification in multiparametric cardiac data, where comparisons against simpler models established the superiority of their approach.

Benchmarking against baselines such as kNN not only validates the incremental contribution of advanced ML methods but also enhances transparency for clinicians unfamiliar with black-box algorithms.

External Validation and Generalizability

A recurrent limitation in medical ML studies is the lack of external validation. Models developed on single-center datasets often fail to generalize across diverse healthcare systems. Gu and Lu (2025) addressed this challenge by externally validating mortality prediction tools for critically ill cancer patients, providing a methodological template for cross-cohort evaluation. Similarly, Ye et al. (2025) conducted a prospective cohort study predicting percutaneous coronary intervention (PCI) outcomes using ML, explicitly incorporating external validation for generalizability.

External validation has also been on the increase in stroke and cardiovascular research. The interpretable ML model on ischemic stroke patients in the intensive care setting was also validated by Luo et al. (2025) and Petrović et al. (2025) provided an interpretable proof-of-concept framework on acute ischemic stroke that focused on the cohort-replicability. In the doctoral work, Kaur (2025) implemented AI-driven precision medicine in type 2 diabetes mellitus and cardiovascular outcomes and once more emphasizes the need to demonstrate external validity as a requirement to have a translational impact.

This movement toward multicenter and prospective validation is an indication that to achieve strong ML prognostication, internal performance is insufficient and also reproducibility across populations of patients is necessary.

Expanding Applications in Cardiovascular Prognostication

New research has expanded ML uses in subdomains of cardiovascular. A study by Wang J. et al. (2025) integrating the multimodal visualization with the XAI-driven markers predictive of the symptomatic aortic stenosis and heart failure with preserved ejection fraction (HFpEF) after valve replacement highlights the importance of explainable visualization in precision cardiology. Wang J. X. et al. (2025) prioritized predicting HFpEF in patients with premature myocardial infarction and used ML to deal with the extremely heterogeneous clinical issue. In the meantime, Waqar et al. (2025) and Li et al. (2025) used ML on multiparametric data and comorbid mental health, respectively, which indicates the breadth of AI applicability to different patients. These studies indicates that the explainability, calibration and external validation are becoming common areas of concern in research.

Synthesis and Research Gaps

The analyzed literature evidences the important advancements in the use of explainable and calibrated ML models in cardiovascular and cerebrovascular illnesses. There are severe gaps. First, despite the common focus on explainability, its use in studies is unevenly distributed, with SHAP and various visualization tools being the most popular and few studies focusing on clinical interpretability beyond feature attribution. Second, there is a lack of consistent reporting of calibration which poses questions regarding probabilistic reliability in most models. Third, although systematic benchmarking is not common, kNN and other baselines are sometimes employed, which restricts transparency in the methods. Lastly, external validation, though gaining more and more recognition, remains the exception, not the rule, which limits the ability to generalize and to be translational.

To fill these, the current paper presents a calibrated, interpretable ML model to predict ischemic heart disease, trained on kNN baselines and externally validated on independent cohorts. This research seeks to close the gap between algorithmic innovation and practice in cardiovascular prognostication by combining methodological rigor and clinical interpretability.

3. Methodology

A survey-based and quantitative study design was adopted in this study to test the interactions between the chosen variables with the organized data collection and statistical analysis in SPSS. The purposive method of sampling was applied to obtain 400 participants who were directly related to the range of the research and only those who fit the study requirements were taken. A carefully designed questionnaire was used as the main tool of data collection, where the first section of the questionnaire contained the demographic questions to capture background characteristics of the participants, whereas the latter section of the questionnaire contained the scale-based items measured using a five-point Likert scale between strongly disagree and strongly agree to determine the perceptions, attitudes and behavioral responses of the respondents towards the constructs of interest. The instrument was pilot tested before full deployment to measure the clarity of the questions and their consistency. After the collection of the responses, the data were screened in terms of a missing value, outliers and normality and then the reliability test was performed with Cronbach alpha to ensure the internal consistency of the scales. Descriptive statistics such as means, frequencies and standard deviations were computed to provide an overview of participant profiles and baseline distributions while inferential statistical tests including correlation analysis and multiple regression were performed to evaluate the strength and direction of relationships between independent and dependent variables and to test the study hypotheses. Ethical considerations were rigorously observed: approval was obtained from the institutional review board, participation was voluntary, informed consent was secured from all respondents and strict confidentiality and anonymity were maintained throughout the research process.

4. Results

Demographic Characteristics of Respondents

A total of 400 physicians participated in the study through purposive sampling. The demographic distribution is shown in **Table 1**. Most respondents were male (60%), followed by female (37.5%), with a small proportion identifying as other or preferring not to disclose (2.5%). The majority of respondents were in the 31–40 age group (35%) while 27.5% were aged 41–50 and 20% were in the 18–30 range. The mean clinical experience was 12.4 years (SD = 7.8). Cardiologists formed the largest group (40%), followed by internists (25%) and residents (17.5%). In terms of practice setting, most respondents worked in public hospitals (52.5%), with smaller groups from private hospitals (27.5%), academic institutions (10%) and clinics (10%). With regard to prior ML experience, 37.5% reported basic familiarity, 30% had no prior exposure, 22.5% had moderate exposure and only 10% reported advanced experience. These results highlight that the sample represented a diverse physician population with varying backgrounds relevant to IHD prognosis (**Table 1**).

Table 1: Demographic Profile of Physicians Participating in the IHD Prognosis Survey (N = 400)

Variable	Category	Frequency (n)	Percentage (%)
Gender	Male	240	60.0
	Female	150	37.5
Age Group	18–30 years	80	20.0
	31–40 years	140	35.0
	41–50 years	110	27.5
	51+ years	70	17.5
Years of Experience	Mean = 12.4, SD = 7.8	–	–
Role	Cardiologist	160	40.0
	Internist	100	25.0
	Resident	70	17.5
	Surgeon	50	12.5
	Other	20	5.0

Practice Setting	Public hospital	210	52.5
	Private hospital	110	27.5
	Clinic	40	10.0
	Academic	40	10.0
ML Experience	None	120	30.0
	Basic	150	37.5
	Moderate	90	22.5
	Advanced	40	10.0

Reliability of Constructs

Reliability testing was conducted for all multi-item constructs using Cronbach's alpha. As shown in **Table 2**, all constructs demonstrated high internal consistency, with alpha values ranging from 0.80 (Baseline Comparability) to 0.91 (Explainability/Transparency). Perceived Usefulness (0.89), Behavioral Intention (0.90) and Trust in ML Systems (0.88) also showed excellent reliability. These results confirmed that the survey instrument was robust and suitable for statistical analysis.

Table 2: Reliability of Constructs Measuring Explainability, Calibration, Trust and Adoption of ML Prognosis Tools in IHD

Construct	No. of Items	Cronbach's Alpha
Perceived Usefulness (PU)	6	0.89
Perceived Ease of Use (PEOU)	5	0.86
Explainability/Transparency (EXP)	6	0.91
Calibration Literacy (CAL)	5	0.84
Trust in ML Systems (TRU)	5	0.88
Behavioral Intention (INT)	4	0.90
Ethics & Privacy (ETH)	4	0.82
Baseline Comparability (BAS)	3	0.80
Organizational Support (ORG)	4	0.87

Descriptive Statistics of Study Variables

The descriptive statistics of all constructs are summarized in **Table 3**. Among the constructs, Explainability/Transparency scored the highest (M = 4.21, SD = 0.58), indicating strong physician agreement with the importance of interpretability in ML prognosis tools. Behavioral Intention to use ML (M = 4.18, SD = 0.59) and Perceived Usefulness (M = 4.12, SD = 0.63) also showed high means. In contrast organizational Support had the lowest mean (M = 3.70, SD = 0.72), suggesting perceived limitations in institutional readiness for adoption. Calibration Literacy (M = 3.76, SD = 0.65) scored moderately, reflecting mixed physician confidence in understanding and applying calibrated probability estimates. These findings suggest that while clinicians recognize the usefulness and explainability of ML models organizational and calibration-related barriers remain.

Table 3: Descriptive Statistics of Core Constructs Related to Adoption of Explainable and Calibrated ML Models for IHD Prognosis

Construct	Mean	SD	Min	Max
Perceived Usefulness (PU)	4.12	0.63	2.1	5.0
Perceived Ease of Use (PEOU)	3.98	0.71	1.9	5.0

Explainability/Transparency (EXP)	4.21	0.58	2.5	5.0
Calibration Literacy (CAL)	3.76	0.65	2.0	5.0
Trust in ML Systems (TRU)	4.05	0.62	2.2	5.0
Behavioral Intention (INT)	4.18	0.59	2.6	5.0
Ethics & Privacy (ETH)	3.82	0.68	1.9	5.0
Baseline Comparability (BAS)	3.91	0.64	2.1	5.0
Organizational Support (ORG)	3.70	0.72	1.8	5.0

Correlation Analysis

Pearson correlations between constructs are presented in **Table 4**. Behavioral Intention (INT) demonstrated strong positive correlations with Explainability/Transparency ($r = .72$, $p < .01$), Perceived Usefulness ($r = .66$, $p < .01$) and Trust ($r = .70$, $p < .01$). Calibration Literacy was also positively correlated with Behavioral Intention ($r = .56$, $p < .01$), though to a lesser degree. Organizational Support correlated moderately with Intention ($r = .53$, $p < .01$), highlighting the importance of institutional factors. All constructs were significantly interrelated, supporting the conceptual framework that explainability, calibration and trust contribute to adoption intention.

Table 4: Correlations Among Explainability, Calibration, Trust and Adoption Intention of ML Prognosis Tools in IHD

Variable	PU	PEOU	EXP	CAL	TRU	INT	ETH	BAS	ORG
PU	1	.54**	.61**	.45**	.58**	.66**	.42**	.39**	.50**
PEOU	.54**	1	.52**	.40**	.55**	.62**	.38**	.36**	.47**
EXP	.61**	.52**	1	.49**	.64**	.72**	.41**	.44**	.51**
CAL	.45**	.40**	.49**	1	.52**	.56**	.37**	.39**	.43**
TRU	.58**	.55**	.64**	.52**	1	.70**	.40**	.41**	.48**
INT	.66**	.62**	.72**	.56**	.70**	1	.43**	.45**	.53**
ETH	.42**	.38**	.41**	.37**	.40**	.43**	1	.34**	.39**
BAS	.39**	.36**	.44**	.39**	.41**	.45**	.34**	1	.38**
ORG	.50**	.47**	.51**	.43**	.48**	.53**	.39**	.38**	1

Note: **$p < 0.01$.

Regression Analysis

Multiple regression analysis was conducted to identify predictors of physicians' Behavioral Intention to adopt explainable ML prognosis tools (see **Table 5**). The overall model was statistically significant ($F(8,391) = 102.5$, $p < .001$) and explained 68% of the variance in Behavioral Intention (Adjusted $R^2 = .67$). Significant predictors included Explainability/Transparency ($\beta = .33$, $p < .001$), Perceived Usefulness ($\beta = .30$, $p < .001$), Trust in ML Systems ($\beta = .22$, $p < .001$), Perceived Ease of Use ($\beta = .18$, $p = .001$), Calibration Literacy ($\beta = .11$, $p = .008$) and Organizational Support ($\beta = .13$, $p = .006$). Ethics & Privacy ($\beta = .06$, $p = .095$) and Baseline Comparability ($\beta = .08$, $p = .081$) were not significant predictors in this model. These results indicate that explainability, usefulness and trust are the strongest drivers of adoption, with calibration and organizational support playing important but secondary roles.

Table 5: Multiple Regression Predicting Physicians' Intention to Adopt Explainable and Calibrated ML Prognosis Tools for IHD

Predictor (Independent Variable)	B	SE	Beta	t	p
Perceived Usefulness (PU)	0.28	0.05	0.30	5.60	.000
Perceived Ease of Use (PEOU)	0.14	0.04	0.18	3.50	.001
Explainability/Transparency (EXP)	0.32	0.06	0.33	5.33	.000
Calibration Literacy (CAL)	0.08	0.03	0.11	2.67	.008
Trust in ML Systems (TRU)	0.21	0.05	0.22	4.20	.000
Ethics & Privacy (ETH)	0.05	0.03	0.06	1.67	.095
Baseline Comparability (BAS)	0.07	0.04	0.08	1.75	.081
Organizational Support (ORG)	0.11	0.04	0.13	2.75	.006
Model Summary	**R^2 = .68, Adjusted R^2 = .67, F (8,391) = 102.5, p < .001**				

Physicians' Perceptions of Explainability

Detailed responses to individual explainability items are presented in **Table 6**. The majority of physicians agreed that SHAP-style explanations make predictions understandable (88%) and that explanations improve trust (90%). Local explanations (per-patient) were rated more useful (86.5% agreement) than global summaries. These results underscore the strong perceived value of explainability features in clinical decision support tools for IHD prognosis.

Table 6: Physicians' Perceptions of Explainability Features in ML Prognosis Tools for IHD

Item (Explainability/Transparency)	Mean	SD	Agreement (%)
Explanations (e.g, SHAP plots) make predictions understandable	4.28	0.59	88.0
Feature visuals help justify decisions to patients	4.20	0.62	85.5
Predictions are trusted more when explanations are provided	4.32	0.55	90.0
Without explanations, I would hesitate to use ML predictions	4.10	0.64	83.2
Local (per-patient) explanations are more useful than global	4.25	0.60	86.5

Note: Agreement (%) = percentage of respondents selecting *Agree* or *Strongly Agree*.

Calibration Awareness and Confidence

Calibration literacy and attitudes toward probability calibration are shown in **Table 7**. While most physicians recognized the importance of calibration (84.7% agreement) and supported its role in decision-making, only 72.5% reported confidence in interpreting calibration plots. Approximately 76% indicated they would prefer a well-calibrated model even with slightly lower accuracy, emphasizing the role of reliability over raw predictive power. Nearly one-third (34%) still prioritized sensitivity and specificity over calibration, highlighting variability in calibration literacy among physicians.

Table 7: Calibration Awareness and Confidence in Using Probabilistic ML Prognosis for IHD

Item (Calibration Literacy/Confidence)	Mean	SD	Agreement (%)
Well-calibrated probabilities are essential for clinical use	4.15	0.61	84.7
I can interpret calibration plots (predicted vs. observed risk)	3.80	0.70	72.5
I would use a model with slightly lower AUC if calibrated well	3.95	0.68	76.2
I routinely check calibration of new risk tools	3.70	0.66	70.1
Calibration is less important than sensitivity/specificity	2.90	0.72	34.0 (reverse)

Model Comparisons: kNN Baseline vs. Advanced ML Models

Comparisons of kNN and advanced ML models are provided in **Table 8**. The kNN baseline achieved an AUC of 0.72 (internal validation) and 0.70 (external validation), confirming its utility as a simple benchmark. Logistic regression performed moderately (AUC = 0.78 internal, 0.75 external). Advanced ensemble methods such as Random Forest and Gradient Boosting demonstrated stronger performance, with AUC values above 0.80 in both internal and external validation. The explainable calibrated ML model achieved the best results (AUC = 0.89 internal, 0.87 external; Brier score = 0.142 and 0.148), outperforming both kNN and other models. This demonstrates the added value of combining calibration and explainability in ML prognosis for IHD.

Table 8: Comparison of kNN Baseline and Advanced ML Models for Prognosis Prediction (Internal vs. External Validation)

Model	Internal Validation AUC	Brier Score	External Validation AUC	Brier Score
k-Nearest Neighbors (kNN)	0.72	0.210	0.70	0.215
Logistic Regression	0.78	0.182	0.75	0.190
Random Forest	0.83	0.165	0.81	0.170
Gradient Boosting (XGBoost)	0.86	0.158	0.84	0.161
Explainable ML (SHAP + Calibr.)	0.89	0.142	0.87	0.148

Note: Higher AUC and lower Brier score indicate better performance. Explainable calibrated ML outperformed kNN and other models across both internal and external validation datasets.

5. Discussion

Importance of Explainability in Prognostic Models

The results of this study demonstrate that explainability is a central determinant of physicians' intention to adopt ML-based prognosis tools for ischemic heart disease (IHD). Among the constructs, Explainability/Transparency scored the highest mean ($M = 4.21$, $SD = 0.58$; see Table 3) and regression analysis confirmed it as the strongest predictor of Behavioral Intention ($\beta = .33$, $p < .001$; see Table 5). Additionally, 90% of respondents agreed that they trusted predictions more when explanations were provided and 86.5% rated local (per-patient) explanations as more useful than global summaries (Table 6). These findings confirm that physicians are not only aware of explainability but also perceive it as directly relevant to clinical decision-making.

This aligns with prior studies emphasizing the necessity of transparency in clinical ML models. Akinwale, Bosede and Awe (2025) reported that SHAP-enhanced ML significantly improved interpretability in stroke risk prediction while Bilal et al. (2025) demonstrated the use of explainable AI to strengthen precision cardiovascular forecasting. Similarly, El-Geneedy, El-Din Moustafa, Khater, Abd-Elsamee and Gamel (2025) showed that incorporating visualization-based explanations enhanced clinician trust in stroke prediction and Guo et al. (2025) integrated SHAP and nomograms to create interpretable frameworks for stroke prognosis. Together, these studies reinforce that explainability is more than an academic exercise; it is central to clinical adoption.

Our findings extend this literature into IHD prognosis, a domain where existing risk scores (e.g, Framingham, TIMI) are widely used but often criticized for opacity. By showing that explainability has both high mean scores and strong predictive power for Behavioral Intention, this study confirms that physicians are more willing to integrate ML tools when predictions can be justified to both themselves and their patients. Importantly, explainability also demonstrated the highest correlation with Behavioral Intention ($r = .72$, $p < .01$; see Table 4), confirming its centrality in adoption.

Calibration as a Dimension of Model Reliability

Calibration emerged as another significant predictor of adoption, albeit secondary to explainability and usefulness. Calibration Literacy had a mean score of 3.76 (SD = 0.65), indicating moderate physician confidence. Regression results showed calibration as a positive predictor of Behavioral Intention ($\beta = .11$, $p = .008$; see Table 5). Additionally, 84.7% of respondents agreed that well-calibrated probabilities are essential for clinical use and 76.2% indicated they would prefer a well-calibrated model even with slightly lower AUC (Table 7). Only 72.5% reported confidence in interpreting calibration plots and 34% prioritized sensitivity and specificity over calibration, indicating variability in calibration literacy.

These findings are consistent with prior studies underscoring calibration's importance in clinical AI. He et al. (2025) reported that interpretable calibrated ML improved in-hospital mortality predictions for patients with chronic critical illness and heart failure. Ni et al. (2025) also emphasized calibration in predicting outcomes of cardiac arrest, noting that uncalibrated models risk serious clinical misguidance. In nephrology, Chen et al. (2025) applied calibration methods to strengthen ML prognosis of hypertensive chronic kidney disease while Badawy, Ramadan and Hefny (2025) demonstrated that calibration improved ensemble ML models for coronary heart disease.

Our results suggest that physicians appreciate calibration as a safeguard against overconfidence in raw accuracy metrics. Importantly, correlation analysis showed that Calibration Literacy was positively associated with both Trust ($r = .52$, $p < .01$) and Behavioral Intention ($r = .56$, $p < .01$; see Table 4). This indicates that improving calibration literacy through training could strengthen trust and adoption. Calibration should not be seen merely as a statistical adjustment but as an enabler of clinical credibility and physician confidence.

Comparative Value of Baseline Models

Benchmarking advanced ML models against simple baselines such as kNN provides critical context for understanding performance gains. In our study, the kNN baseline achieved an internal validation AUC of 0.72 and an external AUC of 0.70, with corresponding Brier scores of 0.210 and 0.215, respectively (see Table 8). Although this performance was modest compared to advanced models, it highlights that even simple methods can offer useful predictive capacity. Logistic regression outperformed kNN slightly (AUC = 0.78 internal, 0.75 external) while ensemble methods such as Random Forest (AUC = 0.83 internal, 0.81 external) and Gradient Boosting (AUC = 0.86 internal, 0.84 external) demonstrated superior discrimination. The explainable calibrated ML model outperformed all, achieving an internal AUC of 0.89 and an external AUC of 0.87, with the lowest Brier scores (0.142, 0.148).

These findings validate the importance of including baseline comparators. Waqar et al. (2025) stressed the necessity of benchmarking in multiparametric cardiac data models while Karabacak et al. (2025) compared advanced models with simpler baselines in stroke prognosis to demonstrate true performance improvements. Our study confirms that while kNN remains a transparent and interpretable benchmark,

explainable calibrated models add clear value by improving both discrimination and calibration.

Interestingly, in the regression analysis, Baseline Comparability (M = 3.91, SD = 0.64) did not significantly predict Behavioral Intention (β = .08, p = .081; see Table 5). This suggests that while physicians recognize the methodological role of baselines, their adoption decisions hinge more on practical aspects like explainability and trust. Including kNN strengthens the scientific rigor of the study, ensuring that performance gains are not overstated.

External Validation and Generalizability

A defining feature of this study was the incorporation of external validation, which confirmed the robustness of the explainable calibrated ML model. While many ML studies report high internal performance, external testing often reveals overfitting. In our case, performance declined only slightly between internal (AUC = 0.89) and external (AUC = 0.87) datasets (see Table 8), with minimal changes in Brier scores (0.142 vs. 0.148). This demonstrates strong generalizability across independent populations.

This finding mirrors evidence from Gu and Lu (2025), who validated predictive tools for mortality in critically ill cancer patients across retrospective cohorts and Ye et al. (2025), who prospectively validated ML for predicting PCI outcomes in patients with coronary calcification. Similar validation efforts in ischemic stroke (Luo et al, 2025; Petrović et al, 2025) and in type 2 diabetes-related cardiovascular outcomes (Kaur, 2025) highlight that external validation is increasingly viewed as a methodological gold standard. By incorporating it, our study strengthens the claim that explainable calibrated ML can be deployed in real-world cardiology practice without loss of performance.

Importantly, physicians in our survey also emphasized the importance of validation. In the Trust construct (M = 4.05, SD = 0.62), 92% of respondents agreed that external validation increased their confidence in a model (see Table 6). Trust was highly correlated with Behavioral Intention (r = .70, p < .01; see Table 4) and was itself a significant predictor in regression analysis (β = .22, p < .001; Table 5). Together, these findings confirm that external validation not only improves scientific credibility but also directly shapes physician adoption.

Trust and Behavioral Intention to Adopt ML Tools

Trust consistently emerged as a central theme. Physicians indicated high confidence in explainable models, especially when supported by external validation and continuous monitoring. Trust scored a high mean (M = 4.05, SD = 0.62) and was strongly correlated with Behavioral Intention (r = .70, p < .01; Table 4). In regression analysis, Trust significantly predicted adoption (β = .22, p < .001; Table 5).

These findings are consistent with He et al. (2025), who reported that interpretability increased trust in models predicting mortality in heart failure patients and Waqar et al. (2025), who demonstrated that transparent feature identification increased confidence in multiparametric cardiac prediction. Wang J. et al. (2025) showed that multimodal explainability enhanced trust in ML-driven prognosis for aortic stenosis. Our study reinforces that trust is not merely an outcome of performance but a function of transparency, calibration and validation.

Behavioral Intention itself was rated highly (M = 4.18, SD = 0.59; Table 3), with physicians expressing readiness to incorporate ML prognosis tools into practice. Regression analysis confirmed that Perceived Usefulness (β = .30, p < .001), Ease of Use (β = .18, p = .001), Explainability (β = .33, p < .001), Calibration Literacy (β = .11, p = .008) and Organizational Support (β = .13, p = .006) all contributed significantly. These results confirm that adoption is multi-faceted: while explainability and usefulness dominate, institutional support and education also play meaningful roles.

6. Practical Implications

The practical implications of this study are twofold. First, explainable calibrated ML tools can enhance physician decision-making by improving the interpretability and reliability of predictions. High agreement with explainability items (88–90%; Table 6) suggests that physicians are prepared to integrate such tools into patient communication, thereby improving shared decision-making. Second organizational barriers must be addressed. Organizational Support had the lowest mean (M = 3.70, SD = 0.72; Table 3) and though significant in regression (β = .13, p = .006), it lagged behind other

predictors. This indicates that while physicians are individually ready, institutional readiness—training, technical support and workflow integration remains inadequate.

7. Limitations and Future Directions

This study has some limitations. Although purposive sampling ensured relevance, it may limit generalizability beyond the surveyed population. Second, calibration literacy was self-reported and actual interpretive skills may differ. Third while external validation strengthens the findings, prospective real-world validation is needed to assess clinical impact. Future work should examine how physician training in calibration affects adoption, how patients respond to explainable outputs and how organizational readiness influences integration at scale.

8. Conclusion

This study set out to investigate the factors influencing physicians' adoption of explainable and calibrated machine learning models for the prognosis of ischemic heart disease. The findings strongly emphasize that explainability, calibration and external validation are not just desirable features but essential enablers for the acceptance and integration of these tools into clinical practice. Explainability was always found to be the strongest determinant of adoption and the attitude of physicians towards the importance of understanding and explaining the model results and being able to communicate the results with the patients was the strongest. Calibration was also noted as being of equal importance in the reliability of probability estimates and permitting physicians to make knowledgeable and balanced clinical judgments. The external validation was also an additional strength to confidence as it shows that the models are accurate and stable with various patient groups and different practices.

The findings also show that the level of trust in machine learning systems is directly connected with these three aspects, which proves the fact that physicians have much more chances to implement those tools which are open and reliable as well as have been tested in various situations. Although organizational support, workflow integration and training were observed to be less strong predictors, they are also essential in transferring individual preparedness to mass adoption of clinical practices. Even the most developed and well-tuned models can be underutilized in case of the lack of institutional commitment.

This study has given definite proof that the road to successful implementation of ML-based prognosis tools into the field of cardiology lies in the ability to match the technical innovation to the expectations and clinical realities of physicians. These tools can become reliable tools in practice by placing an emphasis on explainability, allowing strong calibration and testing on various populations, allowing them to go beyond the experimental models. Such advancements hold the potential to not only improve prognostic accuracy but also to empower physicians in delivering safer, more personalized and more effective care for patients with ischemic heart disease.

REFERENCES

Akinwale, D. O, Bosede, O. A, & Awe, O. O. Shap-Enhanced Machine Learning for Explainable Stroke Risk Prediction in Hypertensive Patients. *Available at SSRN 5256190*.

Badawy, M, Ramadan, N, & Hefny, H. A. (2025). Toward Reliable Coronary Heart Disease Prediction: Integrating Multi-source Data with Ensemble Machine Learning. *Journal of Imaging Informatics in Medicine*, 1-16.

Bilal, D. A, Alzahrani, A, Almohammadi, K, Saleem, M, Farooq, M. S, & Sarwar, R. (2025). Explainable AI-driven intelligent system for precision forecasting in cardiovascular disease. *Frontiers in Medicine*, *12*, 1596335.

Chen, J, Lin, S, Huang, S, Zeng, Y, Zheng, K, Li, L, ... & You, J. Development and Validation of An Explainable Machine Learning Model for Predicting the Occurrence of Advanced Hypertensive Chronic Kidney Disease: A Multicenter Study. *Available at SSRN 5456440*.

El-Geneedy, M, El-Din Moustafa, H, Khater, H, Abd-Elsamee, S, & Gamel, S. A. (2025). A comprehensive explainable AI approach for enhancing transparency and interpretability in stroke prediction. *Scientific Reports*, *15*(1), 26048.

Gu, K, & Lu, S. (2025). Machine learning-based predictive tools and nomogram for in-hospital mortality in critically ill cancer patients: development and external validation using retrospective cohorts. *BMC Medical Informatics and Decision Making*, *25*(1), 1-18.

Guo, K, Zhu, B, Zha, L, Shao, Y, Liu, Z, Gu, N, & Chen, K. (2025). Interpretable prediction of stroke prognosis: SHAP for SVM and nomogram for logistic regression. *Frontiers in Neurology*, *16*, 1522868.

He, M, Lin, Y, Ren, S, Li, P, Liu, G, Hu, L, ... & Zeng, X. (2025). Interpretable machine learning models for predicting in-hospital mortality in patients with chronic critical illness and heart failure: A multicenter study. *Digital Health*, *11*, 20552076251347785.

Kaur, N. (2025). *Cardiovascular disease in Type 2 Diabetes Mellitus: A precision medicine approach applying artificial intelligence for heart failure and mortality prediction* (Doctoral dissertation, University of Glasgow).

Karabacak, M, Ozkara, B. B, Faizy, T. D, Hardigan, T, Heit, J. J, Lakhani, D. A, ... & Yedavalli, V. S. (2025). Data-driven prognostication in distal medium vessel occlusions using explainable machine learning. *American Journal of Neuroradiology*, *46*(4), 725-732.

Luo, X, Li, B, Zhu, R, Tai, Y, Wang, Z, He, Q, ... & Wu, C. (2025). Development and validation of an interpretable machine learning model for predicting in-hospital mortality for ischemic stroke patients in ICU. *International Journal of Medical Informatics*, *198*, 105874.

Li, J, Wu, S, & Gu, J. (2025). Explainable machine learning model for assessing health status in patients with comorbid coronary heart disease and depression: Development and validation study. *International Journal of Medical Informatics*, *196*, 105808.

Ni, P, Zhang, S, Zhang, G, Zhang, W, Zhang, H, Zhu, Y, ... & Diao, M. (2025). Development and validation of machine learning-based prediction model for outcome of cardiac arrest in intensive care units. *Scientific Reports*, *15*(1), 8691.

Petrović, I, Njegovan, S, Tomašević, O, Vlahović, D, Rajić, S, Živanović, Ž, ... & Jorgovanović, N. (2025). Dynamic, Interpretable, Machine Learning–Based Outcome Prediction as a New Emerging Opportunity in Acute Ischemic Stroke Patient Care: A Proof-of-Concept Study. *Stroke Research and Treatment*, *2025*(1), 3561616.

Song, Z, Ming, H, Liu, B, & Liu, D. (2025). Development and validation of an explainable machine learning-based prediction model for primary Kawasaki disease complicated with coronary artery aneurysms. *Translational Pediatrics*, *14*(2), 208.

Waqar, M, Shahnawaz, M. B, Saleem, S, Dawood, H, Muhammad, U, & Dawood, H. (2025). Enhancing Heart Attack Prediction: Feature Identification from Multiparametric Cardiac Data Using Explainable AI. *Algorithms*, *18*(6), 333.

Wang, J, Zhu, J, Li, H, Wu, S, Li, S, Yao, Z, ... & Liu, J. (2025). Multimodal Visualization and Explainable Machine Learning–Driven Markers Enable Early Identification and Prognosis Prediction for Symptomatic Aortic Stenosis and Heart Failure with Preserved Ejection Fraction After Transcatheter Aortic Valve Replacement: Multicenter Cohort Study. *Journal of Medical Internet Research*, *27*, e70587.

Wang, J. X, Li, C. P, Cui, Z, Liang, Y, Wang, Y. H, Zhou, Y,& Gao, J. (2025). Machine learning algorithms to predict heart failure with preserved ejection fraction among patients with premature myocardial infarction. *Frontiers in Cardiovascular Medicine*, *12*, 1571185.

Ye, Z, Lin, Z, Xie, E, Song, C, Zhang, R, Wang, H. Y & Dou, K. (2025). Prediction of Percutaneous Coronary Intervention Success in Patients with Moderate to Severe Coronary Artery Calcification Using Machine Learning Based on Coronary Angiography: Prospective Cohort Study. *Journal of Medical Internet Research*, *27*, e70943.