International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 11 | Nov - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

CardioMind AI: A Predictive Analytics Framework for Cardiovascular Disease Detection

Madhan K U[1], Karan S[2], Tejas V Maney[3], Vemula Sai Srikar[4], Komala D[5]

[1,2,3,4,5]Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India.

***Abstract*—Judging cardiovascular risk means reading blood pressure, cholesterol, glucose, chest-pain type and exercise response together, and a bare score from a model rarely helps the person who has to act on it. This paper describes CardioMind AI, a prototype decision-support web application that wraps a supervised risk model in a workflow for entering clinical attributes, inspecting the estimated probability and per-feature contributions, saving a report, finding cardiology centres, and asking a retrieval-grounded assistant for background information. Four classifiers, namely Logistic Regression, Random Forest, XGBoost and an SVM, were compared on the UCI Cleveland benchmark using a stratified 80:20 split and sigmoid (Platt) calibration. Logistic Regression gave the best accuracy (85.25%) and specificity (87.88%), while Random Forest gave the best ROC-AUC (0.9340), PR-AUC (0.9290) and Brier score (0.1199). The reported test partition appears to hold only 61 patients, so the differences between models are within sampling noise. The work is a prototype and a benchmark study. It is not clinically validated and must not be used for diagnosis.**

Keywords*—*cardiovascular disease, heart disease prediction, probability calibration, SHAP, explainable AI, clinical decision support, Cleveland dataset.

I. Introduction

Cardiovascular disease is a family of conditions that affect the heart and the vessels that serve it. In practice a clinician rarely relies on one number. Age, sex, resting blood pressure, serum cholesterol, fasting glucose, the character of chest pain and the heart rate reached during exercise each carry a part of the picture, and the parts have to be weighed against one another. Statistical models can do this weighing consistently across structured records, and several studies have used them to flag patients for closer review [1]. Their usefulness, however, depends on the population the data came from, the quality of the records, and, above all, on how the evaluation was carried out.

Most student and research prototypes in this area stop at the classifier. A model is trained, a table of accuracies is printed, and the work ends. That leaves a gap between the model and the person who would read its output: someone has to know what the probability means, which inputs pushed it up or down, where to go next, and where the model's authority ends. CardioMind AI was built to look at that gap. It places the prediction inside a small application with a data-entry form, an analysis view, a conversational assistant, a hospital locator and a report archive.

The contributions of this paper are limited on purpose: (1) a side-by-side comparison of four supervised classifiers on seven reported metrics, including two that measure probability quality; (2) a preprocessing design that handles numerical and categorical attributes along separate paths; (3) an interface concept that places the estimated probability next to feature-level explanations and a downloadable report; and (4) supporting modules for locating services and answering questions from retrieved material. The numbers reported here were transcribed from the project presentation and were not re-run for this paper, which is stated again where it matters.

II. Related Work

García-Ordás et al. [1] studied heart-disease risk prediction with deep networks and a feature-augmentation step, showing that a learned representation can help when the raw clinical variables are few. Their emphasis is on the representation. The present work looks instead at the other end of the pipeline: what happens after a probability exists. Detrano et al. [2] produced the clinical data behind the widely used Cleveland subset, and it remains a standard small benchmark [3], which is also why results on it should be read as preliminary.

The four model families used here are conventional baselines. Logistic Regression fits a linear decision function and is easy to audit. Random Forest [4] averages many decorrelated trees. XGBoost [6] adds trees sequentially to correct earlier errors, and a kernel SVM [5] learns a maximum-margin boundary. All four are available through scikit-learn [7], with XGBoost supplied as a separate library.

Probability quality matters once a number is shown to a person. Platt scaling [8] fits a sigmoid that maps raw classifier scores to probabilities, and Niculescu-Mizil and Caruana [9] compared how different learners behave before and after such a correction. The Brier score [10] summarizes the squared error of the stated probabilities. For explanation, SHAP [11] distributes a prediction among the input features. These attributions describe the behaviour of the model, not the biology of the patient, and should not be read as causes.

On the application side, CardioMind AI uses OpenStreetMap data [14] for facility search and retrieval-augmented generation [15] for its assistant. The distinction from earlier work is therefore one of scope. We do not claim better predictive accuracy than published systems. We combine a multi-model benchmark with explanation, reporting, service discovery and grounded question answering in one prototype, and we say plainly what has not been validated.

III. Proposed System

A. System Objective. CardioMind AI accepts a short set of clinical attributes, estimates the probability that heart disease is present, and shows how that estimate came about. It is meant

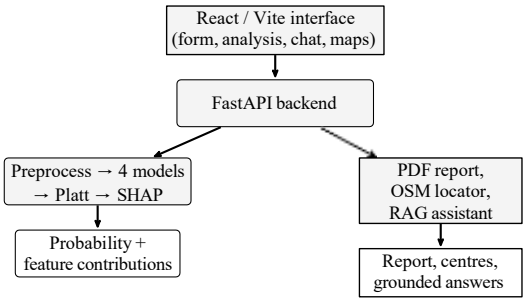


Figure 1: Overall architecture of CardioMind AI.

as a screening-support and teaching prototype. Its aim is not to diagnose, and it makes no claim of clinical readiness. The design goal is a transparent, inspectable chain from input to output, so that a reviewer can see the data, the probability, the contributing features and the limits in one place.

B. Architecture. The frontend is a React application built with Vite. It talks to a FastAPI backend that holds the trained pipelines and the supporting services, as shown in Fig. 1. Six views are exposed: Heart Risk Prediction, Live Analysis, CardioAI Chat, Hospital Locator, Reports and Public Reports. A request travels from the form to the backend, passes through preprocessing, is scored by the models, is calibrated, is explained, and is returned to the interface together with the material needed for a PDF report.

C. Data Preparation. The data are split 80:20 with stratification on the class label so that both partitions keep the positive-to-negative ratio. Numerical attributes are imputed with the median and then standardized,

$$z = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}, \tag{1}$$

where μ_{train} and σ_{train} are computed on the training partition only. Categorical attributes, such as chest-pain type, are imputed with the most frequent category and one-hot encoded. Fitting these transformations on the training data and applying them unchanged to the held-out records prevents test information from leaking into the model. The project material does not list the transformation attribute by attribute, so the assignment of each column to a numerical or categorical path is not reproduced here.

D. Models and Calibration. Logistic Regression, Random Forest, XGBoost and an SVM were trained on the same processed features. Tree ensembles and margin-based models often return scores that are poorly calibrated, so the raw scores are mapped to probabilities with Platt scaling [8]:

$$P(y = 1 \mid f) = \frac{1}{1 + \exp(Af + B)}, \tag{2}$$

where f is the uncalibrated score and the scalars A and B are fitted by maximum likelihood. The mapping has to be fitted on data that the classifier did not use for training, otherwise the probabilities look better than they are. The presentation

Algorithm 1 Handling one prediction request

1: **Input:** attribute vector **x** from the form
2: validate ranges; impute missing entries with training medians / modes
3: standardize numerical part by Eq. (1); one-hot encode categorical part
4: **for** each model $m \in$ {LR, RF, XGB, SVM} **do**
5: $p_m \leftarrow$ Platt-calibrated probability from Eq. (2)
6: **end for**
7: compute SHAP contributions for the primary model
8: return p, contributions and a plain-language note on the threshold

Table I
EXPERIMENTAL SETUP

Parameter	Value
Dataset	UCI Cleveland Heart Disease
Partition	Stratified 80:20 train / test
Numerical path	Median imputation, StandardScaler
Categorical path	Most-frequent imputation, one-hot
Models	LR, Random Forest, XGBoost, SVM
Calibration	Sigmoid (Platt)
Metrics	Acc., Sens., Spec., F1, ROC-AUC, PR-AUC, Brier

describes sigmoid calibration but not the number of folds, the random seeds, the hyperparameters or the rule used to choose the decision threshold. These items are therefore listed as undocumented rather than assumed.

E. Explanation, Report and Assistant. Feature contributions are computed with SHAP and shown as signed chips beside the probability. Positive values raise the estimate and negative values lower it. A report page collects the inputs, the probability and the contributions into a PDF. The locator queries OpenStreetMap for cardiac facilities near a typed place or a GPS fix. The assistant answers from a retrieved knowledge base through an online language-model path, with a local semantic-search fallback when the online service is not reachable. Algorithm 1 summarizes one prediction request.

IV. Dataset and Experimental Setup

A. Dataset. The benchmark is the UCI Cleveland Heart Disease collection [3, 2]. The processed version in common use has 303 patient records and 14 attributes, including a diagnosis label that is normally collapsed to presence versus absence of disease. Missing values occur in a very small number of rows, which the imputers described above handle. The record count is important for everything that follows: with this few patients, one misclassified case moves a percentage by more than a point.

B. Evaluation Metrics. With true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), the

threshold-dependent metrics are

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP}, \tag{3}$$

$$\text{F1} = \frac{2TP}{2TP + FP + FN}. \tag{4}$$

ROC-AUC and PR-AUC summarize ranking quality over all thresholds, so they do not depend on the cut-off chosen. The Brier score evaluates the stated probabilities directly over N test patients [10]:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2, \tag{5}$$

and lower is better. A model can rank patients well and still state poor probabilities, which is why the benchmark reports both groups of measures.

V. Results and Evaluation

A. Model Comparison. Table II gives the reported test results. Logistic Regression has the highest accuracy and specificity. Random Forest has the highest ROC-AUC and PR-AUC and the lowest Brier score, so it both ranks patients best and states the most reliable probabilities. XGBoost, which the application treats as the primary model, has the lowest accuracy and the highest Brier score in this particular comparison. Random Forest and the SVM share the same accuracy, sensitivity, specificity and F1, but differ in the probability-based measures.

B. How Much Weight the Differences Can Carry. The presentation does not state the size of the test partition, but the reported values allow it to be inferred. An 80:20 split of 303 records leaves 61 test patients, and every reported accuracy matches a whole number of correct cases out of 61: 52/61 = 85.25% for Logistic Regression, 51/61 = 83.61% for Random Forest and the SVM, and 49/61 = 80.33% for XGBoost. The sensitivity values fit 28 positive patients (23/28 = 82.14%, 22/28 = 78.57%) and the specificity values fit 33 negative patients (29/33 = 87.88%, 28/33 = 84.85%, 27/33 = 81.82%). Table III lists the counts this implies. They are our inference, not figures taken from the project.

The consequence is straightforward. Logistic Regression and XGBoost differ by three patients out of 61, and the confidence intervals of all four models overlap heavily. The accuracy ranking should therefore not be treated as established. Evaluating several models on one small split and then reporting the best one is also known to give optimistic estimates [13], so repeated stratified cross-validation, with the calibration fitted inside each training fold, is the appropriate next experiment.

C. Application Walkthrough. Fig. 3 shows four of the six views. The form collects age, sex, blood pressure, serum cholesterol, fasting blood sugar, chest-pain type, exercise-induced angina and peak heart rate. The analysis view returns a calibrated probability and signed contribution chips named after the Cleveland variables (for example `chol`, `trestbps`, `thalach` and the chest-pain indicator `cp_4. 0`). The assistant lists the clinical topics it can discuss, and the locator draws 18 cardiac facilities

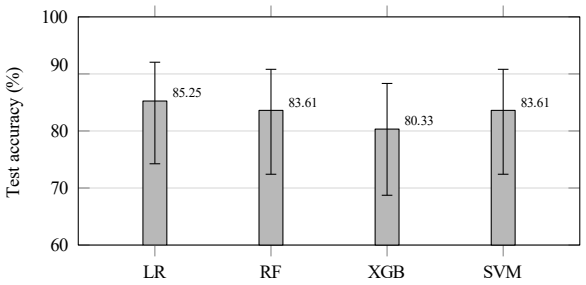


Figure 2: Reported accuracy with 95% Wilson intervals, assuming a 61-patient test set. The bars overlap almost completely.

around Bengaluru on a map, with filters for catheterization labs, emergency units and cardiac surgery.

We read the screenshots as illustrations of the interface and not as measured output. In the captured analysis screen, all four models display the same 93.9% and every pipeline stage still reads "pending" at 0% progress, which suggests that part of the view was captured before, or independently of, a live inference. The Public Reports page likewise shows a summary accuracy of 94.9% that corresponds to nothing in Table II. These values are not used as evidence anywhere in this paper.

VI. Discussion

Two observations stand out. First, the metric that leads depends on the question asked. If one cares about overall correctness at a fixed threshold, Logistic Regression comes first. If one cares about ranking and the honesty of the stated probabilities, Random Forest comes first. For a tool that shows a probability to a human reader, the second group of measures arguably matters more, and it favours Random Forest, which is not the model designated as primary in the application. The choice of the primary model should be revisited once cross-validated results exist.

Second, a simple linear model stays competitive with three more flexible learners on 14 attributes and a few hundred patients. That fits a pattern often reported for small tabular clinical data, where added model capacity tends to buy little, although one split cannot confirm it here. It also has a practical benefit: Logistic Regression coefficients can be read directly, which reduces the weight put on post-hoc explanations.

The explanation chips need careful framing. SHAP values show what the model used to arrive at its output. They do not say why a patient has a condition, and a large contribution from one variable is not a medical finding. The form also exposes eight of the thirteen standard predictors; how the remaining attributes, such as resting ECG, vessel count and thallium result, are supplied at prediction time is not documented, and it directly affects how far the displayed probability can be trusted. The assistant and the locator carry their own risks. An answer from a retrieval system is educational content and cannot replace individual medical advice, and a map marker does not guarantee that a service is open or equipped today. Chest pain with breathlessness, sweating or fainting is an emergency and calls for local emergency services, not a chat window.

Table II
MODEL METRICS REPORTED IN THE PROJECT PRESENTATION

Model	Accuracy	Sensitivity	Specificity	F1	ROC-AUC	PR-AUC	Brier
Logistic Regression	**85.25%**	82.14%	**87.88%**	**0.836**	0.9177	0.9055	0.1239
Random Forest	83.61%	82.14%	84.85%	0.821	**0.9340**	**0.9290**	**0.1199**
XGBoost (primary)	80.33%	78.57%	81.82%	0.786	0.9069	0.8796	0.1405
SVM	83.61%	82.14%	84.85%	0.821	0.9145	0.8976	0.1252

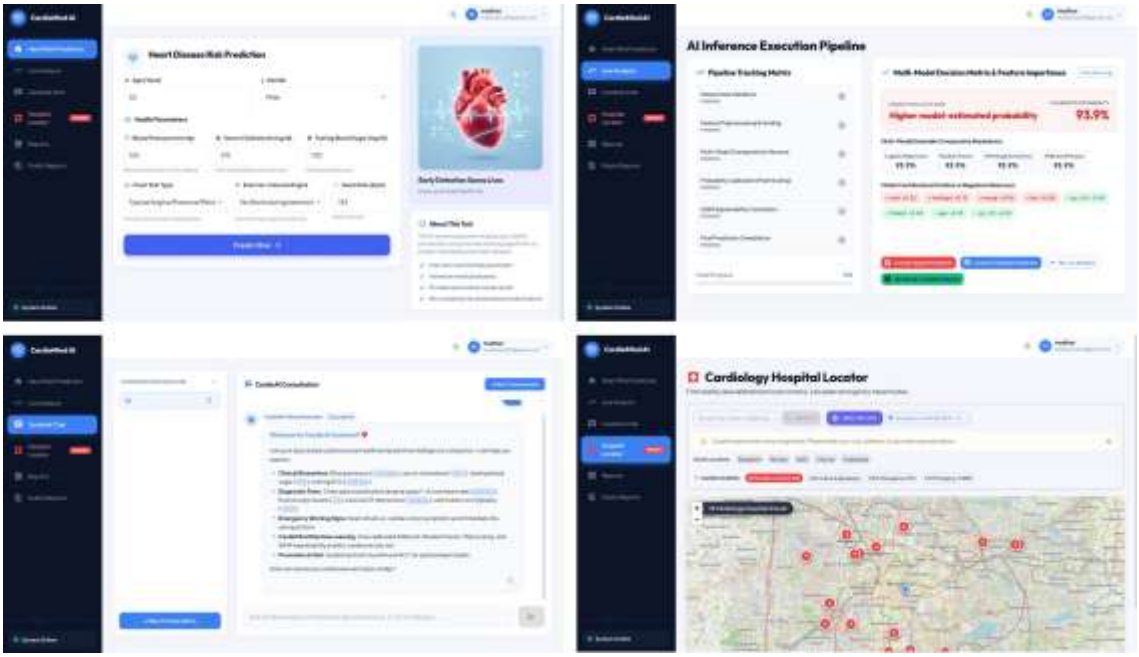


Figure 3: CardioMind AI interface: risk-entry form (top left), live analysis with the displayed probability and contribution chips (top right), conversational assistant (bottom left) and cardiology hospital locator (bottom right).

Table III
CORRECT CASES IMPLIED BY THE REPORTED VALUES (n = 61, INFERRED) WITH 95% WILSON INTERVALS FOR ACCURACY [12]

Model	TP/28	TN/33	Correct	95% CI
Log. Regression	23	29	52	74.3–92.0%
Random Forest	23	28	51	72.4–90.8%
XGBoost	22	27	49	68.7–88.4%
SVM	23	28	51	72.4–90.8%

VII. SAFETY AND ETHICAL CONSIDERATIONS

A prototype that prints a probability next to the word "risk" can be misread, so several safeguards belong in the design and not only in a disclaimer. The result screen should state what the number means (an estimate from a model trained on one historical dataset), the threshold at which a case is labelled higher risk, and that the estimate is not a diagnosis. A low estimate must never be presented as reassurance to a person who has symptoms. The assistant should refuse individual treatment advice, and it should answer descriptions of chest pain, breathlessness, sweating or fainting with a direction to contact local emergency services.

Privacy needs similar care. The form collects health attributes, and the prototype also stores archived reports under a user account. Before any real use, the project would need a clear statement of what is stored, for how long, who can read it, and how it is deleted, and the third-party calls made by the assistant and the locator would have to be reviewed for what personal data they receive. Finally, the Cleveland data were collected in a specific clinical setting several decades ago, so a model trained on them may behave differently for women, for younger adults, or for populations that are under-represented, and that has to be measured before the system is discussed in any wider context.

VIII. REPRODUCIBILITY NOTES

For another group to repeat this work, the following would have to be fixed and published: the exact data file and the rule that turns the diagnosis label into a binary target; the random seed of the split and of each model; all hyperparameters; the calibration scheme, including how many folds were used and on which partition the sigmoid was fitted; the decision threshold and the rule used to select it; and the library versions. The project



Figure 4: Clinical Reports Archive (left) and Public Clinical Audit & Hospital Latency Reports (right). Neither page is treated as experimental evidence.

Table IV
APPLICATION MODULES AND THE TESTING EACH STILL REQUIRES

Module	Function	Verification still needed
Risk prediction	Form, calibrated probability, model comparison	Cross-validation, calibration curves, threshold justification, independent data
Explanation	SHAP contribution chips per prediction	Stability of attributions across retrains and near-identical inputs
Report archive	PDF report with inputs, probability and contributions	Check that every printed value is generated from the current request
Hospital locator	OpenStreetMap search, GPS or typed location	Coverage, geocoding errors, data freshness, availability caveats
CardioAI chat	Retrieval-grounded answers, local semantic fallback	Retrieval quality, refusal of individual advice, emergency-symptom escalation
Public reports	Aggregate accuracy and latency summary page	Define the data source; separate illustrative values from measured ones

material supplies none of these in full. We therefore describe the results as a benchmark snapshot, and we consider the cross-validated rerun proposed in Section X the first thing that should be done with the code.

IX. LIMITATIONS

The evaluation rests on one benchmark with an inferred test set of 61 patients and no uncertainty reporting in the source material. There was no external validation, no prospective study and no subgroup analysis by age or sex, and the Cleveland data come from a single historical population, so the results cannot be assumed to hold in other hospitals. The decision threshold, calibration folds, seeds and hyperparameters are undocumented, and the calibration was not evaluated with a reliability diagram. The form does not collect all standard attributes. The assistant, the report generator and the locator have not been tested for retrieval quality, report correctness, data freshness, privacy or safe failure. Finally, the interface screenshots show inconsistent values and are not independent evidence.

X. FUTURE WORK

Several steps follow directly. The first is to rerun the benchmark with repeated stratified cross-validation and report intervals, fitting the preprocessing and the calibration inside each fold. The second is to add calibration diagnostics, namely reliability curves and the expected calibration error. The third is to test on independent datasets and to report performance by sex and age group. The fourth is to make every displayed number come from the current request, with each pipeline stage reflecting real progress. The fifth is to measure retrieval quality for the assistant, to add an explicit emergency-symptom rule, and to record the freshness date of locator data. Finally, any use that could affect patient care requires review by clinicians.

XI. CONCLUSION

This paper presented CardioMind AI, a prototype that places a calibrated heart-disease risk estimate inside a workflow of explanation, reporting, service discovery and grounded question answering. In the reported comparison, Logistic Regression led on accuracy (85.25%) and specificity (87.88%), and Random Forest led on ROC-AUC (0.9340), PR-AUC (0.9290) and Brier score (0.1199). Because the test partition most likely contains 61 patients, these gaps are small relative to their uncertainty, and the ranking is provisional. The system is a reasonable starting point for studying how model output can be presented responsibly. Cross-validated evaluation, independent data and clinical review are the main requirements before it can be considered for any real-world use.

ACKNOWLEDGMENT

The authors thank the faculty of the Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, for guidance and support during this project.

REFERENCES

[1] M. T. García-Ordás et al., "Heart disease risk prediction using deep learning techniques with feature augmentation," *Multimedia Tools and Applications*, vol. 82, pp. 31759–31773, 2023.

[2] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[3] A. Asuncion and D. J. Newman, "UCI Machine Learning Repository," Univ. of California, Irvine, 2007.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.

[7] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[8] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, MIT Press, 1999, pp. 61–74.

[9] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd ICML*, 2005, pp. 625–632.

[10] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *J. Am. Stat. Assoc.*, vol. 22, no. 158, pp. 209–212, 1927.

[13] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, art. 91, 2006.

[14] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[15] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.