Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Missing Data Governance in Real-World Cardiovascular Risk Prediction: Effects on Machine Learning Performance and Interpretability

Gönül Kara
Pamukkale University
Leyla Özgür Polat
lozgur@pau.edu.tr

Pamukkale University

Research Article

Keywords: Heart disease prediction, Machine learning, Missing data handling, Real-world clinical data, Explainable artificial intelligence, Clinical decision support

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10755778/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Machine learning has become central to data driven cardiovascular risk prediction, yet the empirical foundation of the field rests heavily on a small set of curated, open benchmark datasets that are clean, complete, and drawn from populations unlike those in which models are deployed. Reported performance therefore often reflects idealized conditions rather than the irregular, incomplete data of routine clinical practice. This study reframes the heart disease prediction problem around the data themselves: the reliability of a clinical prediction model is governed not only by the choice of algorithm but also by how the incompleteness of real-world data is handled.

Methods

An original dataset was assembled directly from the cardiology records of a private hospital in Denizli, Türkiye, covering patients seen between 2019 and 2024. The records were retrieved without any prior check of whether individual fields were populated, and nineteen parameters were selected with cardiology specialists. Eleven supervised learning algorithms were evaluated under four contrasting strategies for treating missing data: imputation, retention of missing values, hybrid handling with an explicit unknown category, and complete case analysis. Models were tuned by grid search with five-fold cross validation, assessed on a held-out test partition using accuracy, precision, recall, and F1-score, and benchmarked against a majority class classifier. Explainable artificial intelligence analysis was used to examine whether the pattern of missingness itself carries information.

Results

Data quality governance proved decisive. Imputation and the hybrid strategy preserved predictive capability and stability, whereas deleting every incomplete record degraded performance markedly across all algorithms. Boosting ensembles were the most resilient to incomplete data, while instance based and simple probabilistic methods were the most adversely affected. Interpretability analysis showed that an indicator of unrecorded blood pressure ranked among the most influential features, so that the absence of a measurement itself acted as an informative signal.

Conclusions

Appropriate handling of incomplete records preserves model reliability, whereas uncritical deletion erodes it. The results argue for a shift from marginal accuracy gains on familiar benchmarks toward locally grounded, transparent models that can be trusted in the clinical environments they are meant to serve.

Introduction

Cardiovascular disease remains the foremost contributor to mortality worldwide, and the burden falls disproportionately on low- and middle-income countries, among which Türkiye is explicitly identified in international surveillance reporting [1]. A substantial share of premature cardiac events is preventable through timely detection and structured management, and contemporary clinical guidance emphasizes that the accurate and prompt evaluation of risk factors and diagnostic parameters is decisive for patient outcomes [2]. The clinical value of early identification, in turn, depends on the capacity to extract coherent risk signals from the multidimensional and frequently irregular information that accumulates around each patient demographic attributes, lifestyle factors, biochemical measurements, and the results of cardiac investigations.

Machine learning offers a principled means of meeting this challenge. By recognizing structure within large and complex datasets, learning algorithms can characterize disease risk in ways that complement, rather than replace, established clinical reasoning [3, 4]. In cardiology specifically, supervised models have demonstrated the ability to integrate heterogeneous clinical variables and to support earlier and more individualized risk assessment [5, 6]. The methodological repertoire is mature: across a large body of work, classical linear and probabilistic models, instance-based methods, decision trees, and most prominently in recent years ensemble and boosting frameworks have all been brought to bear on cardiovascular prediction.

Yet the empirical literature exhibits a structural limitation that is rarely confronted directly. The overwhelming majority of machine learning studies on heart disease prediction are conducted on a small number of curated public datasets most notably the Cleveland subset of the UCI repository, together with the Framingham, Statlog and Z-Alizadeh Sani collections several of which were assembled decades ago [7]. These datasets have been invaluable for methodological benchmarking, but they share three properties that limit their external relevance: they are comparatively clean and complete, they are demographically and geographically distant from many deployment settings, and their repeated reuse encourages a research culture oriented toward incremental gains on a fixed leaderboard rather than toward clinical generalizability. Performance figures obtained under these idealized conditions may therefore overstate the readiness of such models for the messier reality of routine care.

Real clinical data, by contrast, are pervasively incomplete. Measurements go unrecorded, investigations are deferred or performed elsewhere, and free text documentation is entered inconsistently across staff and time. How this incompleteness is handled is frequently relegated to a preprocessing footnote, despite evidence that the treatment of missing values can materially shape the behavior of the resulting model [8]. Two questions follow that the existing literature has only partially addressed. First, on genuinely real-world clinical data, how much does the choice of missing data strategy actually matter relative to the choice of algorithm? Second, when missingness is not random but systematically related to how care is delivered, can its very pattern carry predictive information that a model can exploit and that an analyst must interpret with caution?

This study addresses these questions through an original clinical case study from Türkiye. A dataset was constructed directly from the electronic records of the cardiology department of a private hospital in Denizli, covering patients seen between 2019 and 2024. Importantly, the dataset was not obtained as a clean export: every cardiology department encounter, together with all requested clinical variables, was retrieved from the hospital database without any prior check of whether individual fields were populated or empty. The study therefore deliberately confronts the raw incompleteness and inconsistency of operational clinical data the “dirty” data that accumulate in real information systems rather than the sanitized conditions of a curated benchmark. A broad family of supervised learning algorithms was evaluated, not once but under four deliberately contrasting strategies for handling missing data, so that the effect of data quality governance could be isolated as an experimental factor rather than fixed by convention. Model behavior was then examined through explainable artificial intelligence analysis, with particular attention to whether the absence of a measurement functions as an informative feature. On this basis, the study pursues the following research questions:

- RQ1. To what extent does the choice of missing data handling strategy influence the predictive performance of machine learning models for heart disease risk when those models are trained on real-world clinical data?
- RQ2. Which families of learning algorithms are most robust to incomplete clinical data, and how does that robustness vary across preprocessing strategies?
- RQ3. Does the pattern of missingness itself carry predictive information, and what does this imply for the clinical interpretation of model outputs?
- RQ4. How does building and modelling a locally sourced, real-world clinical dataset shape the conduct and clinical interpretation of machine learning based cardiovascular risk prediction, given the field’s prevailing reliance on curated public benchmarks?

The remainder of the paper is organized as follows. The rest of this Background section reviews the relevant literature on machine learning for cardiovascular prediction, the limitations of benchmark datasets, and the handling of missing data in clinical modelling, and from this synthesis derives the research gap. The Methods section presents the methodological framework, including the modelling pipeline, the construction of the Turkish clinical dataset, the four missing data scenarios, the algorithms, and the evaluation criteria. The Results section traces the analysis from the model comparison through to interpretable findings. The Discussion relates the findings to the research questions, considers their contribution, and sets out the limitations of the study, and the Conclusions close the paper.

Machine learning for cardiovascular risk prediction

The application of machine learning to heart disease prediction has produced an extensive and methodologically diverse literature. A recurrent design comparing several classifiers on a shared dataset and reporting the best performer has been instantiated many times. Thummala and Baskar [9] directly compared decision tree and k-nearest-neighbor classifiers and reported higher predictive accuracy for

the decision tree; Nikhar and Karandikar [10] reached a similar conclusion in favor of tree-based classification over Naive Bayes on the Cleveland data. Working on larger samples, Shorewala [11] evaluated an ensemble of classical learners on a coronary heart disease collection drawn from Kaggle, while Mahmoud et al. [12] reported random forest as the strongest model on the Framingham dataset and Dwivedi [13] identified logistic regression as the best performer on the Statlog data. Comparable studies by Anitha and Sridevi [14], Jindal et al. [15], Yadav et al. [16], Shah et al. [17], Ramesh et al. [18], Katarya and Meena [19], Erdoğan and Güney [20], Shukur and Mijwil [21], and Taşcı and Şamlı [22] reinforce both the breadth of the methodological toolkit and the consistency of the underlying design.

More recent work has shifted toward boosting and gradient based ensembles. Yang and Guan [23] combined sampling-based class balancing with an extreme gradient boosting model and reported superior performance over a panel of conventional learners; Nissa et al. [24] compared boosting techniques and found AdaBoost most effective on a large public sample; and El-Sofany [25] and Anjum et al. [26] likewise identified gradient boosting variants as leading performers, the latter on data collected manually from clinics in Bangladesh. Table 1 summarizes a representative cross section of this literature, highlighting the algorithms employed, the datasets and sample sizes, and the best performing model reported in each case.

Table 1
Representative machine learning studies on heart disease prediction

Study	Algorithms compared	Best model	Sample (n)	Variables	Dataset
Erdoğan & Güney [20]	SVM, K-NN, DT	SVM	303	13	UCI
Shukur & Mijwil [21]	LR, RF, ANN, SVM, K-NN	SVM	303	14	UCI
Jindal et al. [15]	K-NN, LR, RF	K-NN	303	14	UCI
Anitha & Sridevi [14]	K-NN, NB, SVM	NB	302	14	UCI
Yadav et al. [16]	LR, SVM, NB, K-NN, RF	K-NN	303	13	UCI
Vayadande et al. [27]	LR, NB, K-NN, SVM, ANN, DT, RF, XGBoost, CatBoost	LR / RF / XGBoost	303	13	Kaggle
Mahmoud et al. [12]	K-NN, SVM, DT, LR, RF	RF	4240	14	Framingham
Kumar et al. [28]	NB, MLP, RF, DT	NB	270	14	Statlog
Shorewala [11]	LR, SVM, K-NN, DT, NB, ANN	Stacked ensemble (KNN + RF + SVM)	70000	14	Kaggle
Dwivedi [13]	ANN, SVM, LR, K-NN, DT, NB	LR	270	12	Statlog
Ramesh et al. [18]	NB, SVM, LR, DT, RF, K-NN	RF	303	13	UCI
Shah et al. [17]	NB, DT, K-NN, RF	K-NN	303	14	UCI
Katarya & Meena [19]	LR, NB, SVM, K-NN, DT, RF, ANN, MLP	RF	303	14	UCI
Taşcı & Şamlı [22]	NB, J48, RF, MLP, K-NN, LR, SVM	K-NN	303	14	Kaggle
Nissa et al. [24]	Boosting ensemble techniques	AdaBoost	8763	13	Kaggle
El-Sofany [25]	NB, SVM, XGBoost, AdaBoost, DT, K-NN, RF, LR	XGBoost	503	13	UCI + private

Study	Algorithms compared	Best model	Sample (n)	Variables	Dataset
Anjum et al. [26]	LR, SVM, XGBoost, LightGBM, DT	XGBoost	600	13	Hospital data (Bangladesh)

ANN artificial neural network, CatBoost categorical boosting, DT decision tree, K-NN k-nearest neighbors, LightGBM light gradient boosting machine, LR logistic regression, MLP multilayer perceptron, NB Naive Bayes, RF random forest, SVM support vector machine, UCI University of California Irvine machine learning repository, XGBoost extreme gradient boosting.

Two observations emerge from this body of work. First, the field has effectively converged on a stable methodological vocabulary; the marginal contribution of yet another classifier comparison on the same data is correspondingly small. Second, and more consequentially, the great majority of these studies draw on a narrow set of public benchmarks, with sample sizes around three hundred and provenance concentrated in a handful of legacy repositories.

The limits of benchmark datasets and the case for real, local data

The reliance on curated public datasets is itself a recognized concern. In a systematic review of machine learning based cardiac diagnosis, Ahsan and Siddique [7] document the dominance of a small number of repositories and caution that data assembled under controlled conditions do not necessarily capture the full dynamics of real patient populations. The currency of the data compounds the problem: collections gathered decades ago may be poorly aligned with present day clinical practice, evolving diagnostic standards, and shifting demographic structure, all of which bear directly on a model's ability to generalize.

These limitations carry particular weight outside the populations from which the benchmarks were drawn. A model trained on a foreign, pre-standardized dataset inherits the demographic, genetic, nutritional and care delivery characteristics of that source, and its validity in a different national context cannot be assumed. Locally sourced clinical data, by contrast, reflect the demographic profile, comorbidity distribution, laboratory standards, and documentation practices of the setting in which a model is to be used, and thereby support both more representative training and a more honest assessment of clinical applicability. The argument is not that benchmarks are without value, but that conclusions drawn from them require validation on data that resemble the intended deployment environment a requirement that the present study takes as its point of departure. Recent regional evidence supports this emphasis on locally grounded cardiovascular modelling: Mehrabani-Zeinabad et al. [37] developed CVD incidence prediction models from a 16-year Eastern Mediterranean cohort, illustrating the value of population-specific data while explicitly addressing substantial missingness during model development.

Handling missing data in clinical machine learning

If benchmark datasets understate the heterogeneity of clinical populations, they understate their incompleteness still more. Routinely collected clinical data are missing for reasons that are rarely benign: measurements are not always performed, results are sometimes recorded in the wrong field or not at all, and patients may undergo investigations at other institutions and present only the outcome. The treatment of these gaps is a substantive modelling decision rather than a clerical one. Lin and Tsai [8], reviewing more than a hundred studies, show that the choice among imputation techniques, the assumed missingness mechanism, and the evaluation protocol all influence downstream classification performance.

The strategies adopted in the clinical literature are correspondingly varied. Nissa et al. [24] addressed missing cardiovascular measurements through imputation; El Hamdaoui et al. [29] used mode imputation for missing categorical variables; Omotehinwa et al. [30] employed multiple imputation by chained equations; Wu et al. [31] used a random-forest-based approach; and Zheng et al. [32] combined median imputation for numerical variables with mode imputation for categorical ones. A separate strand treats missingness as information rather than nuisance: representing an absent categorical value as an explicit “unknown” category allows a model to learn from the fact of absence itself, an approach that can improve sensitivity when the missingness is not random [8]. More recently, Sun et al. [38] showed in an EHR-based clinical prediction setting that missingness may be informative rather than incidental, and that the choice of missing-data strategy should reflect the underlying missingness mechanism when measurements are not missing at random.

Recent EHR-based studies have begun to compare alternative strategies for handling missingness and to demonstrate that patterns of data collection can themselves be informative [38]. Nevertheless, evidence remains limited in cardiovascular prediction using locally sourced routine hospital data, particularly in designs that vary missing-data strategy systematically while holding the dataset, algorithms, and evaluation protocol constant and then interrogate the predictive role of missingness through explainable analysis.

What this literature lacks are a controlled comparison of these strategies on a single, genuinely real-world clinical dataset, in which the strategy is varied while the data, the algorithms, and the evaluation protocol are held constant. Without such a comparison, the relative importance of missing data handling against the much-studied dimension of algorithm choice cannot be assessed.

Research gap and originality

The synthesis above therefore identifies a more specific threefold gap. First, although regionally grounded cardiovascular prediction studies demonstrate the value of local cohort data [37], the literature remains heavily concentrated on curated public benchmarks, leaving the behavior of machine-learning models on contemporary routine hospital records comparatively underexplored. Second, although EHR-

based work has begun to address informative missingness explicitly [38], missing-data handling is still seldom treated as a first-order experimental factor in cardiovascular prediction under a common modelling protocol. Third, the predictive contribution of missingness itself is rarely examined together with model interpretability, even though this distinction is essential for separating data-generating and documentation patterns from clinically causal signals.

This study addresses these gaps using an original dataset assembled from real Turkish cardiology records rather than a public benchmark. It compares four contrasting missing-data strategies under otherwise identical modelling and evaluation conditions, and it uses explainable artificial intelligence analysis to examine whether, and how, the pattern of missingness contributes to prediction. The contribution therefore lies not in proposing another classifier, but in providing controlled evidence on the interaction between real-world data quality governance, algorithmic robustness, and informative missingness in a locally grounded cardiovascular decision-support setting.

Methods

Overview of the modelling pipeline

The study follows an end to end pipeline designed to make the handling of real-world data quality explicit and reproducible. The pipeline proceeds from data acquisition and expert guided feature selection, through the standardization and clinical categorization of irregular records, to a structured audit of missingness, outlier treatment, and the systematic application of four missing data scenarios. The processed data then feed a panel of eleven supervised learning algorithms, each tuned and validated under a common protocol, after which performance is evaluated across all scenarios and the leading model is examined through explainable artificial intelligence analysis. The complete framework is summarized in Fig. 1.

Study setting, data source and study population

The study draws on local and contemporary clinical data obtained from patients who presented to the cardiology department of Egekent Hospital, a private hospital in Denizli, Türkiye, between 2019 and 2024. A dataset comprising 11706 patient records was assembled. Because patients underwent differing diagnostic investigations, the parameters relevant to cardiac risk prediction were selected in consultation with cardiology specialists and informed by a review of comparable machine learning studies; this process yielded nineteen clinically meaningful and interpretable parameters spanning demographic attributes, laboratory findings, and clinical measurements. The full set of variables, together with their data type and coding, is given in Table 2.

Table 2
Clinical variables in the dataset

No	Variable	Type	Values / coding
1	Sex	Categorical	Male = 1, Female = 0
2	Age	Numeric	14–97 years
3	Total cholesterol	Numeric	69–798 mg/dL
4	Triglycerides	Numeric	26–4198 mg/dL
5	Haemoglobin	Numeric	5.1–19.5 g/dL
6	Creatinine	Numeric	0.40–9.02 mg/dL
7	Diabetes	Categorical	Present = 1, Absent = 0
8	Family history	Categorical	Present = 1, Absent = 0, None reported = 2
9	Hypertension	Categorical	Present = 1, Absent = 0
10	Smoking	Categorical	Present = 1, Absent = 0
11	Chest pain	Categorical	Present = 1, Absent = 0
12	Palpitations	Categorical	Present = 1, Absent = 0
13	Ejection fraction (echocardiography)	Categorical	Unknown = 0, Normal = 1, Mild dysfunction = 2, Severe dysfunction = 3
14	Electrocardiography (ECG)	Categorical	Unknown = 0, Normal rhythm = 1, Rhythm disorder = 2, Conduction disorder = 3, Ischemic findings = 4, Device effect = 5, Prior myocardial infarction = 6
15	Blood pressure (systolic)	Categorical	Unknown = 0, Low = 1, Normal = 2, High = 3, Very high = 4
16	Glucose (fasting)	Categorical	Unknown = 0, Low = 1, Normal = 2, High = 3
17	HDL cholesterol	Numeric	15.9–131.9 mg/dL
18	LDL cholesterol	Numeric	25.2–473.2 mg/dL
19	Diagnosis (target)	Categorical	Heart patient = 1, Not a heart patient = 0

Numeric ranges denote the observed minimum and maximum; categorical levels are shown with their numerical codes. ECG electrocardiography, HDL high-density lipoprotein, LDL low-density lipoprotein.

The study was conducted on retrospectively collected clinical records and was approved by the Non-Interventional Clinical Research Ethics Committee of Pamukkale University (decision no. 10, dated 20 May 2025), in accordance with the principles of the Declaration of Helsinki. Patient records were de-identified prior to analysis.

Data construction and preprocessing

It must be emphasized at the outset that the data were not extracted in a pre-cleaned form. Rather than querying the database for complete records or for a curated subset of fields, the export retrieved the entire set of cardiology department encounters together with all requested clinical variables, with no prior check of whether individual fields were populated or empty. The analytic dataset therefore entered the study in a raw, unfiltered state, with the full burden of incompleteness and inconsistency intact. Working from such data entailed the well-documented difficulties of operating on the “dirty” data that accumulate in real information systems substantial manual and computational cleaning effort together with the additional challenge, relative to the curated datasets prevalent in the literature, of contending with extensive externally introduced missingness. These difficulties were deliberately borne rather than circumvented, since confronting them is precisely what distinguishes a real-world clinical modelling study from one conducted on a pre-sanitized benchmark.

Because the records had been entered at different times and by different clinical staff, the raw data exhibited substantial inconsistency, particularly in free text fields where semantically equivalent entries were expressed in many different forms. The smoking status field, for example, contained a wide range of non-standard entries (such as “sig+”, “sig-”, “smokes”, “does not smoke”) that were consolidated, according to their meaning, into two consistent categories. A further complication arose from the dispersion of information across fields: diagnostic information pertaining to cardiac disease appeared variously under “diagnosis,” “history,” or “background,” and physical findings such as electrocardiographic results were not always recorded in the expected field. Reconstructing a coherent record for each patient therefore required an intensive preprocessing effort, combining automated extraction with manual verification, in which semantically equivalent expressions were unified and each patient’s information was reassembled into a consistent structure. This stage was essential to data integrity and is itself representative of the practical realities that distinguish real clinical data from curated benchmarks.

Clinical categorization of variables

Several clinical variables electrocardiographic findings and echocardiographic ejection fraction among them presented numerous heterogeneous sub-values that would have introduced noise and instability into the modelling process. In collaboration with a cardiologist, and in line with accepted clinical classifications, these variables were grouped into clinically meaningful categories represented by numerical codes. Electrocardiographic findings were consolidated into six clinically grounded categories normal rhythm, rhythm disorder, conduction disorder, ischemic findings, device-related findings, and evidence of prior myocardial infarction alongside a separate code for absent records. Glucose, blood pressure, and ejection fraction were likewise categorized according to international reference ranges, with unknown values retained as a distinct category, and the target variable was defined as a binary distinction between patients with and without a cardiac diagnosis. Following grouping, categorical

variables were transformed into numerical form by one-hot encoding [36]. This categorization reduced model complexity while preserving and in clinical terms enhancing the interpretability of the variables.

Missingness profile and record cleaning

A structured audit revealed that incompleteness was concentrated in specific variables. The highest proportions of missing data were observed in blood pressure (35.13%) and electrocardiographic findings (33.78%), followed by echocardiography (20.47%) and high-density lipoprotein (11.78%); lower levels were recorded for glucose (2.88%), triglycerides (2.40%), creatinine (2.17%), haemoglobin (2.08%), cholesterol (1.89%), and low-density lipoprotein (1.81%). Demographic and core clinical variables age, sex, diabetes, family history, palpitations, chest pain, hypertension, smoking, and the target were complete. The audit suggested that this incompleteness arose from a combination of institutional factors, such as omissions in manual data entry, and patient level socioeconomic factors, such as investigations performed at other institutions to avoid cost an early indication that the missingness was not purely random.

Records exhibiting missingness in six or more variables were excluded to protect data quality, reducing the dataset from 11706 to 7680 observations, of which 4464 corresponded to patients with cardiac disease and 3216 to patients without. The effort test variable, whose missingness was extreme, was removed entirely to preserve data integrity and generalizability. Summary statistics for the numerical variables confirmed the expected clinical ranges age, for instance, spanned 14 to 97 years with a mean of approximately 64 while revealing pronounced right-skew and extreme values in several biochemical measurements, notably triglycerides, low-density lipoprotein, and creatinine.

Outlier treatment

Outliers in the numerical variables were identified using the interquartile-range method, chosen for its robustness to non-normal and skewed distributions. The analysis identified 1514 outlying observations, with the largest counts in creatinine (415), triglycerides (386), and high-density lipoprotein (253), and smaller counts in age, cholesterol, haemoglobin, and low-density lipoprotein. Rather than discarding these observations and incurring further data loss, the study constrained them through capping (winsorization) and subsequently transformed the numerical variables toward a normal distribution using a quantile transformer, thereby limiting the influence of extreme values while retaining the full sample.

The four missing data scenarios

The central methodological device of the study is the comparison of four contrasting strategies for handling incomplete records, applied to the same dataset and evaluated with the same models and metrics so that the effect of the strategy can be isolated:

- **Scenario 1- Imputation**: Variables with a high proportion of missing values are completed using a K-nearest neighbors imputer, while numerical variables with low missingness are completed by

median substitution. This strategy seeks to retain all observations while minimizing the information loss associated with deletion.

- **Scenario 2- Retention of missing values**: Missing entries are left unfilled and passed directly to the models, evaluating the native capacity of each algorithm to accommodate incomplete input. This scenario serves as a reference condition and is applicable only to those algorithms that handle missingness internally.
- **Scenario 3- Hybrid handling with an "unknown" category**: Missing categorical values are encoded as an explicit "unknown" class, allowing the fact of absence to be modelled directly, while numerical variables are completed using k-nearest-neighbors imputation for high missingness fields and median substitution otherwise.
- **Scenario 4- Complete case analysis**: All records containing any missing value are removed, so that analysis proceeds only on fully observed cases. This strategy avoids any imputation induced bias but at the cost of sample size and class balance, and functions here as a lower bound condition.

Learning algorithms

Eleven supervised classifiers were evaluated, chosen to span the principal families of the field and to permit a controlled comparison of their robustness to incomplete data. Linear and probabilistic baselines were represented by logistic regression and Gaussian Naive Bayes; instance-based learning by K-nearest neighbors; and single tree learning by a decision tree. The ensemble and boosting family of particular interest given its prominence in recent cardiovascular work was represented by random forest, gradient boosting, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), categorical boosting (CatBoost), and adaptive boosting (AdaBoost). The boosting methods are especially relevant to clinical data of mixed numerical and categorical type, and CatBoost in particular is designed to handle categorical features directly [33, 34].

Training protocol and evaluation criteria

For each scenario, the data were partitioned into training and test subsets in an 80:20 ratio, with the test partition reserved exclusively for assessing generalization. Hyperparameters were selected by grid search, and five-fold cross validation was applied during tuning to guard against overfitting and to provide a stable estimate of performance across data partitions. Model performance was assessed using four complementary criteria. Accuracy measures the overall proportion of correct classifications; precision measures the proportion of cases predicted positive that are truly positive; recall (sensitivity) measures the proportion of true positives that are correctly identified; and the F1-score, the harmonic mean of precision and recall, summarizes the balance between the two. Reporting all four is essential in a clinical setting, where the costs of false positives and false negatives differ and overall accuracy alone can be misleading. To establish that the models learn genuine structure rather than exploiting class frequencies, performance was additionally benchmarked against a baseline classifier that predicts the majority class.

Interpretability analysis

Because clinical decision support requires not only accurate predictions but also intelligible ones, the best performing configuration was analyzed using SHAP (SHapley Additive exPlanations), which attributes a model's output to its input features on a principled, additive basis [35]. Beyond ranking individual feature contributions, SHAP interaction analysis was used to assess whether the model's decisions depended on a few dominant variable pairs or drew on information distributed across many features an indicator of stability and generalizability. The interpretability analysis also provided the means to test, directly, whether indicators of missingness functioned as informative features in their own right. This emphasis is consistent with recent cardiovascular informatics research. Zhang et al. [39] developed interpretable models from irregularly spaced EHR data for advanced heart-failure therapy eligibility and evaluated the importance of clinical events at both population and individual levels. Jiang et al. [40] subsequently combined LightGBM with SHAP-based feature assessment for cardiovascular risk prediction, using global and local explanations to support clinically interpretable model outputs.

Results

Performance across the four scenarios

The eleven algorithms were evaluated under each of the four missing data scenarios using accuracy, precision, recall, and F1-score. The accuracy results are reported in Table 3; the precision, recall, and F1 results follow the same pattern and are reported in Tables 4, 5 and 6.

Table 3 Model accuracy across the four scenarios

Model	S1 (impute)	S2 (retain)	S3 (hybrid)	S4 (complete case)
Support Vector Machine	0.77	–	0.76	0.69
LightGBM	0.76	0.75	0.75	0.70
Random Forest	0.76	–	0.75	0.69
Gradient Boosting	0.76	–	0.76	0.71
XGBoost	0.76	0.75	0.76	0.71
CatBoost	0.76	0.75	0.76	0.72
Logistic Regression	0.75	–	0.76	0.68
AdaBoost	0.75	–	0.75	0.70
Decision Tree	0.72	–	0.72	0.68
K-Nearest Neighbors	0.72	–	0.72	0.66
Gaussian Naive Bayes	0.68	–	0.68	0.68

S1 imputation, S2 retention of missing values, S3 hybrid handling with an explicit unknown category, S4 complete case analysis. A dash indicates that a model is not applicable under Scenario 2, which retains raw missing values and is therefore restricted to algorithms that accommodate missingness internally.

Table 4 Model precision across the four scenarios

Model	S1 (impute)	S2 (retain)	S3 (hybrid)	S4 (complete case)
Support Vector Machine	0.77	–	0.76	0.69
LightGBM	0.76	0.75	0.75	0.70
Random Forest	0.76	–	0.75	0.69
Gradient Boosting	0.76	–	0.75	0.71
XGBoost	0.76	0.75	0.75	0.71
CatBoost	0.76	0.75	0.76	0.72
Logistic Regression	0.75	–	0.76	0.68
AdaBoost	0.75	–	0.75	0.70
Decision Tree	0.72	–	0.72	0.69
K-Nearest Neighbors	0.72	–	0.72	0.66
Gaussian Naive Bayes	0.68	–	0.71	0.69

S1 imputation, S2 retention of missing values, S3 hybrid handling with an explicit unknown category, S4 complete case analysis. A dash indicates that a model is not applicable under Scenario 2, which retains raw missing values and is therefore restricted to algorithms that accommodate missingness internally.

Table 5 Model recall across the four scenarios

Model	S1 (impute)	S2 (retain)	S3 (hybrid)	S4 (complete case)
Support Vector Machine	0.77	–	0.76	0.69
LightGBM	0.76	0.75	0.75	0.70
Random Forest	0.76	–	0.75	0.69
Gradient Boosting	0.76	–	0.76	0.71
XGBoost	0.76	0.75	0.76	0.71
CatBoost	0.76	0.75	0.76	0.72
Logistic Regression	0.75	–	0.76	0.68
AdaBoost	0.75	–	0.75	0.70
Decision Tree	0.72	–	0.72	0.68
K-Nearest Neighbors	0.72	–	0.72	0.66
Gaussian Naive Bayes	0.68	–	0.68	0.68

S1 imputation, S2 retention of missing values, S3 hybrid handling with an explicit unknown category, S4 complete case analysis. A dash indicates that a model is not applicable under Scenario 2, which retains raw missing values and is therefore restricted to algorithms that accommodate missingness internally.

Table 6 Model F1-score across the four scenarios

Model	S1 (impute)	S2 (retain)	S3 (hybrid)	S4 (complete case)
Support Vector Machine	0.77	–	0.76	0.69
LightGBM	0.76	0.75	0.75	0.70
Random Forest	0.76	–	0.75	0.69
Gradient Boosting	0.76	–	0.75	0.71
XGBoost	0.76	0.74	0.75	0.71
CatBoost	0.76	0.74	0.76	0.72
Logistic Regression	0.75	–	0.76	0.68
AdaBoost	0.74	–	0.75	0.70
Decision Tree	0.72	–	0.72	0.68
K-Nearest Neighbors	0.71	–	0.72	0.66
Gaussian Naive Bayes	0.66	–	0.63	0.68

S1 imputation, S2 retention of missing values, S3 hybrid handling with an explicit unknown category, S4 complete case analysis. A dash indicates that a model is not applicable under Scenario 2, which retains raw missing values and is therefore restricted to algorithms that accommodate missingness internally.

A consistent pattern is evident across all four metrics. When missing values are imputed (Scenario 1) or, equivalently, when the data are complete, the support vector machine attains the highest performance, with the boosting ensembles gradient boosting, XGBoost, and CatBoost close behind. Crucially, the hybrid strategy of Scenario 3, which imputes numerical variables and encodes categorical missingness as an explicit “unknown” class, preserves this performance almost entirely, indicating that information loss is minimized when absence is modelled rather than discarded. The complete case analysis of Scenario 4 tells the opposite story: deleting every incomplete record degrades performance markedly across the board the support vector machine falls from 0.77 to 0.69, and K-nearest neighbors and logistic regression decline to 0.66 and 0.68 respectively as the reduction in sample size and the disturbance of class balance erode the models’ capacity to learn. The boosting ensembles prove the most resilient to incomplete data, while K-nearest neighbors and Gaussian Naive Bayes are the most adversely affected, the latter consistent with the difficulty of satisfying its feature independence assumption on clinical data.

Interpretability and the informativeness of missingness

SHAP analysis of the leading model identified age, the absence of chest pain, an indicator of missing blood pressure, cholesterol, and the absence of hypertension among the most influential features, with several biochemical and echocardiographic variables also contributing (Fig. 2a). Interaction analysis showed that the model did not rely on a small number of dominant feature pairs but instead drew on information distributed across many variables; performance was optimized when a small set of strong interacting pairs was retained reaching an accuracy of approximately 0.76 whereas adding further interactions reduced it (Fig. 2b). This distributed pattern indicates a stable decision structure that is less prone to overfitting than one concentrated on a few dominant signals. To confirm that this structure reflects genuine learning, the model was compared against a majority-class baseline, which achieved an accuracy of approximately 0.58; the substantial margin over this baseline establishes that the models capture meaningful clinical structure rather than class frequencies.

The directionality underlying these importance rankings is shown in the SHAP summary (beeswarm) plot (Fig. 3), in which each point is a single patient, horizontal position gives the feature’s signed impact on the model output, and color encodes the feature value (red = high, blue = low). The indicator of missing blood pressure again ranks among the most influential features, consistent with the informative missingness interpretation developed below. Displaying the full distribution also reveals the direction of each effect higher age, for instance, shifts predictions toward higher risk while several one-hot-encoded “absence” indicators carry information whose direction reflects data patterns rather than direct clinical causation.

The most theoretically significant finding concerns the indicator of missing blood pressure, which emerged among the most important features. Because this indicator represents the absence of a measurement, its predictive value points to informative missingness: the pattern of what is not recorded is itself related to patient status, plausibly because it reflects systematic differences in how particular patient groups are investigated and documented rather than random omission. The same interpretation applies to the prominence of the “absence of chest pain” feature, which the one-hot encoding represents explicitly and which the model uses in combination with age, blood pressure, and cholesterol. These results must be read with care: the importance of a missingness indicator signals that absence carries information within this dataset, not that it bears a direct clinical-causal relationship to disease. The finding nonetheless demonstrates that, in real-world clinical data, how missingness is handled is not merely a question of data hygiene but one that can shape what a model learns and how its outputs should be interpreted.

Discussion

Contributions in relation to the research questions

The findings bear directly on the four research questions that motivated the study.

On RQ1 the influence of missing data strategy. The results establish that, on real-world clinical data, the choice of missing data strategy exerts an effect on predictive performance that is comparable in magnitude to the choice of algorithm. Holding the data, models, and protocol constant, the difference between a strategy that retains and model’s incompleteness (Scenarios 1 and 3) and one that deletes it (Scenario 4) reaches several percentage points of accuracy and F1-score for the strongest models a margin that exceeds the differences among the better performing algorithms themselves. Data-quality governance is therefore not a preparatory detail but a determinant of model reliability, and the common practice of fixing it by convention is shown to carry real cost.

On RQ2 robustness across algorithm families. The boosting ensembles gradient boosting, XGBoost, and CatBoost proved the most robust to incomplete clinical data, maintaining strong and stable performance across the imputation and hybrid scenarios and degrading least under complete case deletion. The support vector machine matched or led them under favorable preprocessing but, like most models, suffered under deletion. Instance-based and simple probabilistic methods (K-nearest neighbors and Gaussian Naive Bayes) were the least resilient. For heterogeneous clinical data combining numerical and categorical features, the boosting family thus offers the most dependable basis for deployment.

On RQ3 informative missingness. The interpretability analysis answers this question affirmatively and constitutes the study’s principal conceptual contribution. The pattern of missingness exemplified by the indicator of unrecorded blood pressure carries predictive information, reflecting systematic features of clinical practice rather than random noise. This finding reframes missing data from a defect to be

eliminated into a signal to be understood, while underscoring the interpretive caution required: informativeness is a property of the data generating process, not evidence of clinical causation.

On RQ4 the value of locally sourced data. By building and modelling an original Turkish clinical dataset rather than a public benchmark, the study demonstrates how real, local data alter both the conduct and the reading of cardiovascular prediction research. The pervasive missingness, the irregular documentation, and the more modest but more representative performance levels that characterize routine care are rarely surfaced by curated benchmarks. The resulting models are correspondingly more credible as a basis for clinical decision support in their intended setting, and the dataset itself constitutes a regionally representative resource of independent value.

Taken together, these contributions advance the literature less by adding another entry to the algorithm leaderboard than by relocating the locus of methodological attention. They show that, for clinical machine learning to be trustworthy, the governance of real-world data quality must be treated as a central scientific concern.

Limitations and future research

The study is subject to limitations that also define its future directions. It is based on retrospective data from a single center, and its performance levels, while honest, are modest by the standards of benchmark studies; external validation on data from other institutions and regions is required to establish generalizability. The target label was derived from documentation that was itself irregular, and the feature set, though clinically grounded, omits genetic, lifestyle, and environmental factors whose inclusion could sharpen individual risk estimates. Future work should pursue multi-site validation, incorporate richer feature sets, explore deep-learning approaches capable of modelling more complex variable interactions, and advance toward integration with hospital information systems so that the resulting models can function as practical clinical decision support tools.

Conclusions

This study examined heart disease prediction not as a contest among classifiers on familiar benchmarks but as a problem governed, in equal measure, by the quality of the data on which models are built. Using an original clinical dataset assembled from Turkish cardiology records, it compared a broad family of supervised learning algorithms under four contrasting strategies for handling missing data, and it interpreted the resulting models through explainable-artificial-intelligence analysis. The evidence is clear and consistent. Appropriate handling of incomplete records whether through principled imputation or through the explicit modelling of absence as an “unknown” category preserves predictive capability and model stability, whereas the uncritical deletion of imperfect data erodes both. Boosting ensembles emerged as the most robust models for heterogeneous clinical data, and the pattern of missingness itself proved to be informative, reflecting systematic features of clinical practice that a model can learn but that an analyst must interpret with care. By foregrounding the governance of real-world data quality, the study offers both a reproducible, regionally grounded framework for

cardiovascular risk prediction and a broader argument for treating data quality as a first-order scientific question in clinical machine learning.

Abbreviations

AdaBoost: Adaptive boosting

ANN: Artificial neural network

CatBoost: Categorical boosting

DT: Decision tree

ECG: Electrocardiography

HDL: High-density lipoprotein

K-NN: K-nearest neighbors

LDL: Low-density lipoprotein

LightGBM: Light gradient boosting machine

LR: Logistic regression

MLP: Multilayer perceptron

NB: Naive Bayes

RF: Random forest

SHAP: SHapley Additive exPlanations

SVM: Support vector machine

UCI: University of California Irvine (machine learning repository)

XGBoost: Extreme gradient boosting

Declarations

Ethics approval and consent to participate

All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the 1964 Declaration of Helsinki and its later amendments. The study was approved by the Non-Interventional Clinical Research Ethics

Committee of Pamukkale University (decision no. 10, dated 20 May 2025). The same committee waived the requirement for individual informed consent, as the study was retrospective and used de-identified clinical records. Patient records were de-identified prior to analysis.

Consent for publication

Not applicable. The manuscript contains no individual person's identifiable data.

Availability of data and materials

The clinical dataset analyzed in this study is not publicly available. The data contain sensitive patient information and were obtained from Egekent Hospital (Denizli, Türkiye) solely for the purposes of this study; authorization for their public release or redistribution was not granted. Access is therefore restricted and would require the explicit permission of Egekent Hospital and approval from the relevant ethics committee.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no specific funding for this work.

Authors' contributions

GK contributed to conceptualization, data curation, formal analysis, methodology, software, investigation and visualization, and wrote the original draft. LÖP contributed to conceptualization and methodology, supervised the study, validated the results, and reviewed and edited the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

The authors thank Özel Egekent Hospital (Denizli, Türkiye) for facilitating access to the clinical records used in this study. We are also grateful to Zehra Üstek Künarcı and to cardiology specialist Dr. Fatih Çam for their valuable support and contributions.

Clinical trial number

Not applicable. This study is not a clinical trial and was conducted as a retrospective, observational study using registered data from a database.

References

1. World Heart Federation. World heart report 2023. Geneva: World Heart Federation; 2023. https://world-heart-federation.org/world-heart-report-2023. Accessed 18 April 2026.
2. Visseren FLJ, Mach F, Smulders YM, Carballo D, Koskinas KC, Bäck M, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice: developed by the Task Force for cardiovascular disease prevention in clinical practice with representatives of the European Society of Cardiology and 12 medical societies with the special contribution of the European Association of Preventive Cardiology (EAPC). Eur Heart J. 2021;42(34):3227–337.
3. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.
4. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349(6245):255–60.
5. Mathur P, Srivastava S, Xu X, Mehta JL. Artificial intelligence, machine learning, and cardiovascular disease. Clin Med Insights Cardiol. 2020;14:1179546820927404.
6. Oyebode O, Fowles J, Steeves D, Orji R. Machine learning techniques in adaptive and personalized systems for health and wellness. Int J Hum Comput Interact. 2023;39(9):1938–62.
7. Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: a systematic literature review. Artif Intell Med. 2022;128:102289.
8. Lin WC, Tsai CF. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev. 2020;53(2):1487–509.
9. Thummala GSR, Baskar R. Prediction of heart disease using decision tree in comparison with KNN to improve accuracy. In: 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES). IEEE; 2022. p. 1–5.
10. Nikhar S, Karandikar AM. Prediction of heart disease using machine learning algorithms. Int J Adv Eng Manag Sci. 2016;2(6):617–21.
11. Shorewala V. Early detection of coronary heart disease using ensemble techniques. Inform Med Unlocked. 2021;26:100655.
12. Mahmoud WA, Aborizka M, Amer FAE. Heart disease prediction using machine learning and data mining techniques: application of Framingham dataset. Turk J Comput Math Educ. 2021;12(14):4864–70.
13. Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2018;29(10):685–93.

14. Anitha S, Sridevi N. Heart disease prediction using data mining techniques. J Anal Comput. 2019;8(2):48–55.
15. Jindal H, Agrawal S, Khera R, Jain R, Nagrath P. Heart disease prediction using machine learning algorithms. IOP Conf Ser Mater Sci Eng. 2021;1022(1):012072.
16. Yadav KK, Sharma A, Badholia A. Heart disease prediction using machine learning techniques. Inf Technol Ind. 2021;9(1):207–14.
17. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Comput Sci. 2020;1(6):345.
18. Ramesh TR, Lilhore UK, Poongodi M, Simaiya S, Kaur A, Hamdi M. Predictive analysis of heart diseases with machine learning approaches. Malays J Comput Sci. 2022:132–48.
19. Katarya R, Meena SK. Machine learning techniques for heart disease prediction: a comparative study and analysis. Health Technol. 2021;11(1):87–97.
20. Erdoğan A, Güney S. Heart disease prediction by using machine learning algorithms. In: 2020 28th Signal Processing and Communications Applications Conference (SIU). IEEE; 2020. p. 1–4.
21. Shukur BS, Mijwil MM. Involving machine learning techniques in heart disease diagnosis: a performance analysis. Int J Electr Comput Eng. 2023;13(2):2177–85.
22. Taşcı ME, Şamlı R. Veri madenciliği ile kalp hastalığı teşhisi. Avrupa Bilim ve Teknoloji Dergisi. 2020:88–95.
23. Yang J, Guan J. A heart disease prediction model based on feature optimization and SMOTE-XGBoost algorithm. Information. 2022;13(10):475.
24. Nissa N, Jamwal S, Neshat M. A technical comparative heart disease prediction framework using boosting ensemble techniques. Computation. 2024;12(1):15.
25. El-Sofany HF. Predicting heart diseases using machine learning and different data classification techniques. IEEE Access. 2024;12:106146–60.
26. Anjum N, Siddiqua CU, Haider M, Ferdus Z, Raju MAH, Imam T, et al. Improving cardiovascular disease prediction through comparative analysis of machine learning models. J Comput Sci Technol Stud. 2024;6(2):62–70.
27. Vayadande K, Golawar R, Khairnar S, Dhiwar A, Wakchoure S, Bhoite S, et al. Heart disease prediction using machine learning and deep learning algorithms. In: 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES). IEEE; 2022. p. 393–401.
28. Kumar M, Shambhu S, Sharma A. Classification of heart diseases patients using data mining techniques. Int J Res Electron Comput Eng. 2018;6(3):1495–9.
29. El Hamdaoui H, Boujraf S, Chaoui NEH, Alami B, Maaroufi M. Improving heart disease prediction using random forest and AdaBoost algorithms. Int J Online Biomed Eng. 2021;17(11):60–75.
30. Omotehinwa TO, Oyewola DO, Moung EG. Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease. Inform Health. 2024;1(2):70–81.

31. Wu J, Tao G, Xie S, Yang H, Qi F, Bao N, et al. Prediction of three-year all-cause mortality in patients with heart failure and atrial fibrillation using the CatBoost model. BMC Cardiovasc Disord. 2025;25(1):466.
32. Zheng C, Tian J, Wang K, Han L, Yang H, Ren J, et al. Time-to-event prediction analysis of patients with chronic heart failure comorbid with atrial fibrillation: a LightGBM model. BMC Cardiovasc Disord. 2021;21(1):379.
33. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Advances in neural information processing systems 2018;31.
34. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data. 2020;7(1):94.
35. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
36. Potdar K, Pardawala TS, Pai CD. A comparative study of categorical variable encoding techniques for neural network classifiers. Int J Comput Appl. 2017;175(4):7–9.
37. Mehrabani-Zeinabad K, Feizi A, Sadeghi M, Roohafza H, Talaei M, Sarrafzadegan N. Cardiovascular disease incidence prediction by machine learning and statistical techniques: a 16-year cohort study from eastern Mediterranean region. BMC Med Inf Decis Mak. 2023;23(1):72.
38. Sun M, Engelhard MM, Bedoya AD, Goldstein BA. Incorporating informatively collected laboratory data from EHR in clinical prediction models. BMC Med Inf Decis Mak. 2024;24(1):206.
39. Zhang Y, Golbus JR, Wittrup E, Aaronson KD, Najarian K. Enhancing heart failure treatment decisions: interpretable machine learning models for advanced therapy eligibility prediction using EHR data. BMC Med Inf Decis Mak. 2024;24(1):53.
40. Jiang ZZ, Jiang YF, Wang Y, Zhou Y, Peng RL, Ma CY, Liu XT. Development of an explainable machine learning model for 3-year cardiovascular risk prediction in new-onset type 2 diabetes using the TyG index and ultrasound features. BMC Med Inf Decis Mak. 2025;25(1):409.

Figures

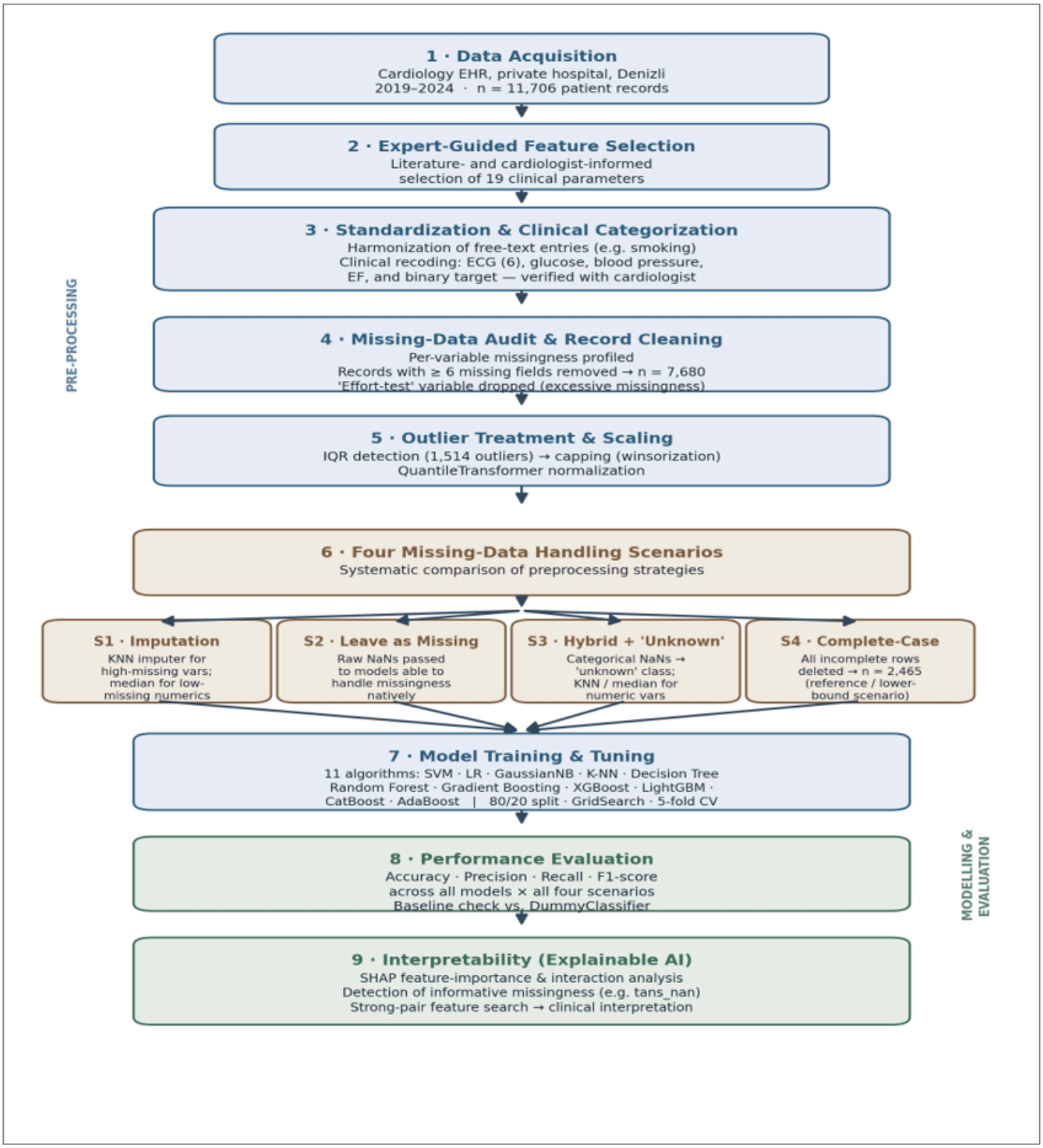


Figure 1

The methodological framework

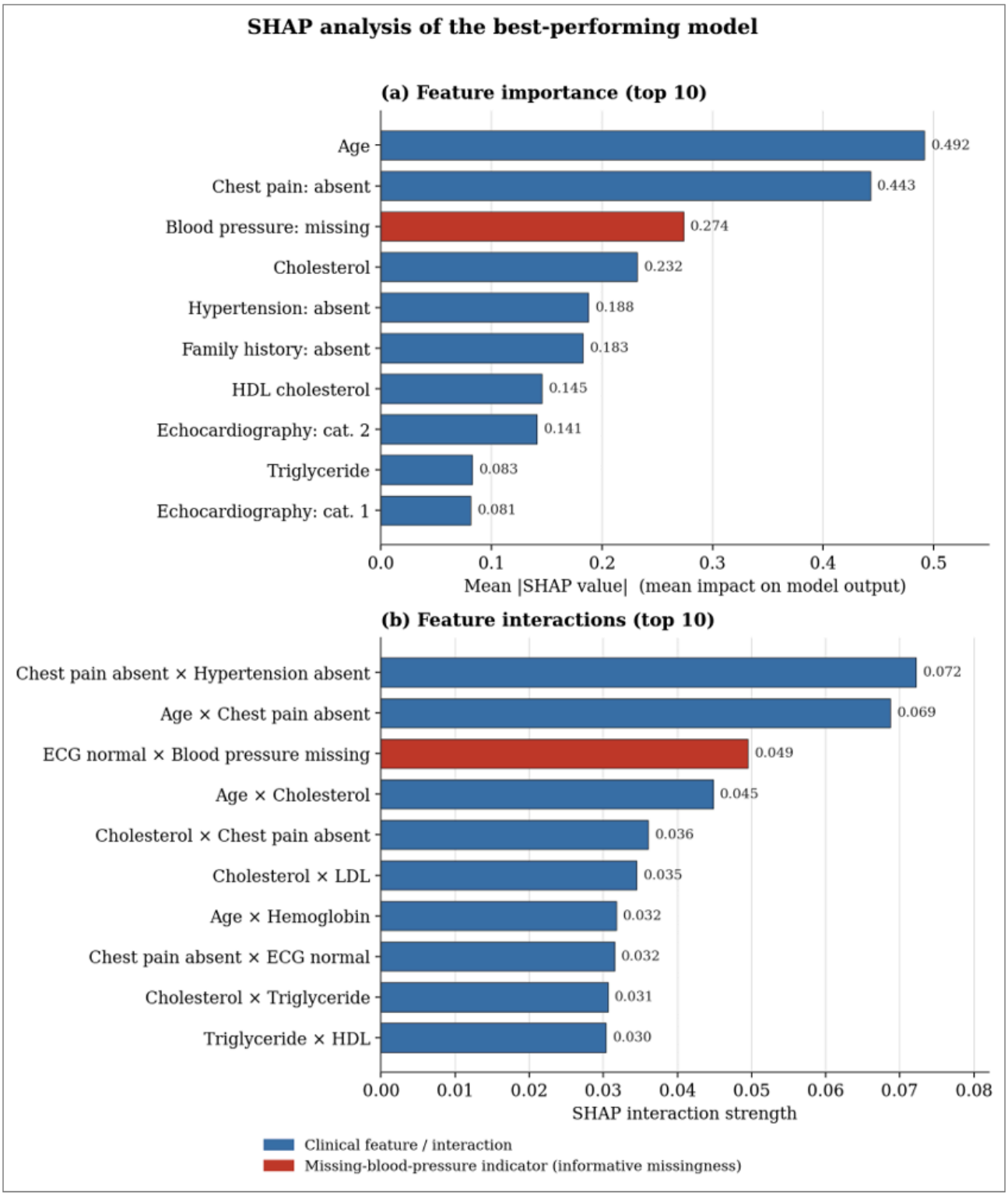


Figure 2

SHAP analysis of the best performing model

(a) Mean absolute SHAP values for the ten most influential features; (b) the ten strongest pairwise feature interactions. Features are derived from one-hot encoding, so the absence of a clinical finding (e.g., chest pain, hypertension) appears as a distinct feature. The indicator of missing blood pressure

(highlighted) ranks among the most important features and participates in a leading interaction with a normal electrocardiographic reading, illustrating that the pattern of missingness itself carries predictive information.

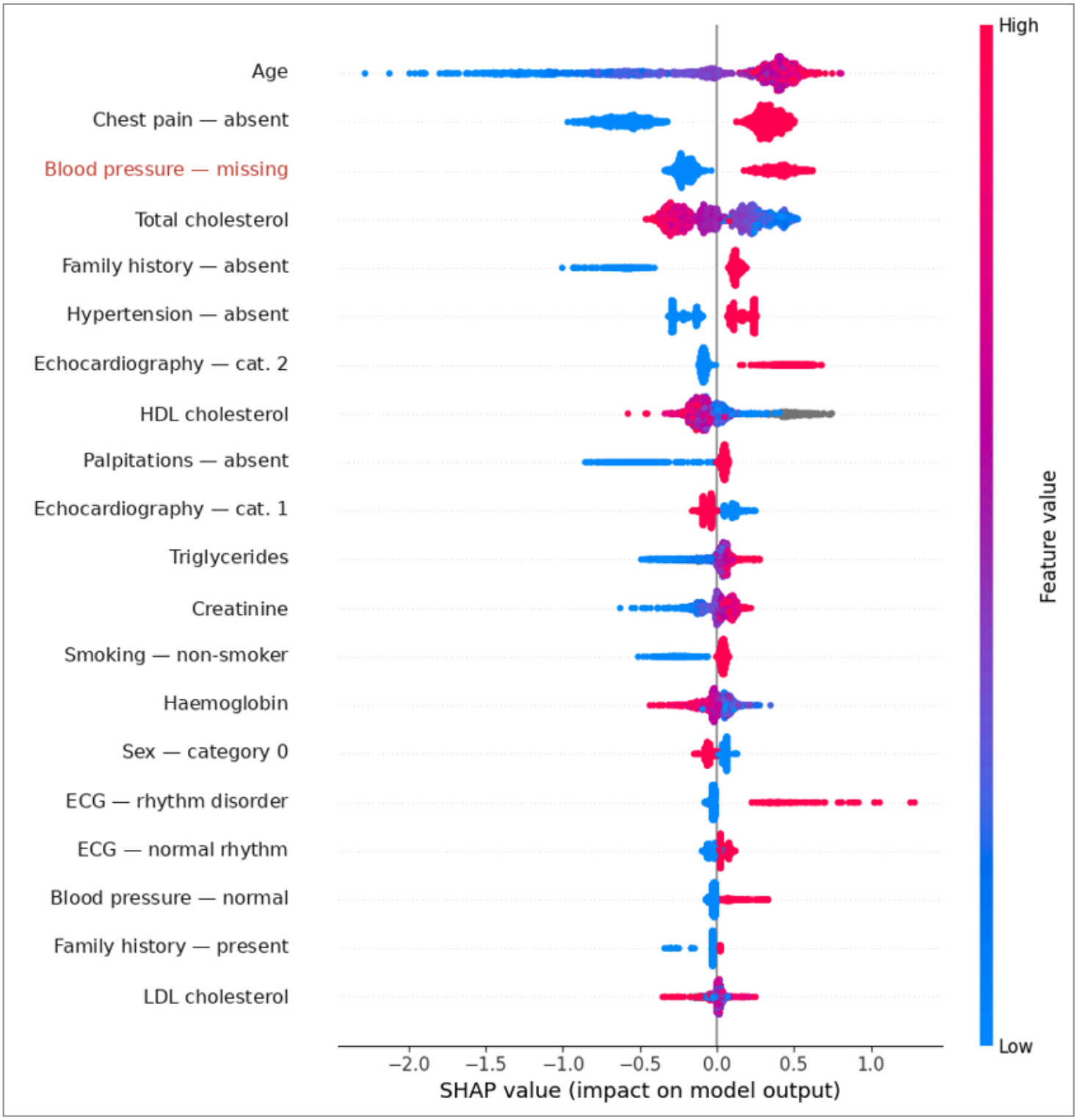


Figure 3

SHAP summary plot