

Advancing cardiovascular disease diagnosis with an interpretable and responsible AI framework

Received: 1 July 2025

Accepted: 6 January 2026

Published online: 02 April 2026

Cite this article as: Hasan K.S. & Dhruvo I.S. Advancing cardiovascular disease diagnosis with an interpretable and responsible AI framework. *Sci Rep* (2026). <https://doi.org/10.1038/s41598-026-35451-3>

Kazi Sakib Hasan & Irfan Sadi Dhruvo

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

Advancing Cardiovascular Disease Diagnosis with an Interpretable and Responsible AI Framework

Kazi Sakib Hasan*

School of Data and Sciences

BRAC University

Kha 224, Bir Uttam Rafiqul Islam Ave, Dhaka 1212, Bangladesh

kazi.sakib.hasan@g.bracu.ac.bd

Irfan Sadi Dhrubo

School of Data and Sciences

BRAC University

Kha 224, Bir Uttam Rafiqul Islam Ave, Dhaka 1212, Bangladesh

irfan.sadi.dhrubo@g.bracu.ac.bd

15 June 2025

Abstract

Cardiovascular disease (CVD) remains a leading global health threat, responsible for one in five deaths worldwide. Early detection is critical to mitigate morbidity and mortality, yet traditional diagnostic methods often rely on reactive clinical assessments, missing opportunities for preventive intervention. In this study, a machine learning (ML) ecosystem is developed to enhance CVD diagnosis through two key approaches: (1) an early warning system using non-clinical, self-reported features for accessible risk stratification, and (2) specialized diagnostic models integrating clinical and non-clinical data. The framework leverages advanced ML techniques, including tabular neural networks (TabNet, TabPFN) and ensemble methods (XGBoost, Random Forest), validated on multi-regional datasets. Shapley Additive Explanations (SHAP) analysis identified ECG-related features as dominant predictors of CVD risk, with ST-segment slope (+0.93) and ST depression (+0.63) exhibiting the strongest effects. Counterfactual explanations from the non-clinical model further revealed actionable preventive measures: reducing exercise-induced angina and chest pain severity, alongside increasing exercise heart rate, could shift predictions from diseased to healthy, highlighting the model's utility for lifestyle interventions. To address ethical and clinical trustworthiness, interpretability tools (SHAP, counterfactuals), fairness mitigation (FairLearn), and uncertainty quantification (Bayesian Neural Networks) are incorporated. Causal inference identified key predictors and their Average Treatment Effects (ATEs) such as exercise-induced angina (ATE: 0.36) and ST slope (ATE: 0.33), informing a hybrid ensemble model that achieved 89% accuracy while reducing dimensionality. The system aligns with FDA

23 Good ML Practices and EU Trustworthy AI guidelines, offering a scalable solution for
24 early detection and equitable diagnosis.

25 1 Introduction

26 Cardiovascular disease (CVD) is among the most serious health conditions, claiming one
27 life every 33 seconds worldwide. It is also the leading cause of death for men, women, and
28 individuals across most racial and ethnic groups [1]. In 2022, approximately 702,880 deaths
29 were attributed to CVD, accounting for 1 in every 5 deaths [2]. Within the same year, nearly
30 1 in 5 deaths among adults under the age of 65 was caused by CVD [1]. It has been estimated
31 that about 5% of adults above the age of 20 are affected by coronary artery disease. Reports
32 from the CDC further indicate that approximately 1 in 5 heart attacks are silent, meaning
33 patients are unaware of the damage sustained [3]. Because CVD can present with both
34 symptomatic and asymptomatic characteristics, the disease often remains undiagnosed, which
35 increases the risk of fatality. CVD typically affects the heart and blood vessels, manifesting as
36 narrowed arteries, congenital structural problems, or malfunctioning heart valves. Mortality
37 rates vary across sex, race, and ethnicity, highlighting the influence of lifestyle factors on
38 disease prevalence [4]. Risk-enhancing lifestyle behaviors include smoking, poor diet, excessive
39 alcohol consumption, and physical inactivity. In addition, medical conditions such as hyper-
40 tension, diabetes, and hypercholesterolemia serve as major risk factors [3]. Environmental
41 factors also play a role; for example, air pollution has been identified by the World Health
42 Organization as a contributor to cardiovascular disease [5]. The complex interplay of these
43 factors, combined with the frequent asymptomatic nature of CVD, often results in delayed or
44 missed diagnoses. Such delays contribute to disease progression, increased healthcare costs,
45 and, in severe cases, premature death. Heart conditions with lower prevalence may remain
46 undiagnosed due to lack of familiarity or lower clinical suspicion, especially in pediatric prac-
47 tice. Misclassification of heart murmurs is also common in developing countries. Furthermore,
48 organ dysfunction resulting from cardiovascular instability has been strongly associated with
49 delayed or inaccurate diagnoses, leading to heightened morbidity and mortality [6]. For these
50 reasons, early detection and preventive strategies are essential in reducing the burden of
51 CVD.

52
53 Early identification of risk factors allows timely interventions through targeted medica-
54 tion and lifestyle modification, thereby preventing the progression of disease. It also reduces
55 the need for expensive and invasive treatments in later stages. Moreover, early detection
56 mitigates the risks associated with misdiagnosis, ensuring that conditions are identified before
57 they advance silently. In essence, proactive diagnosis provides an opportunity to manage risk
58 factors at the earliest phase, thereby limiting the severity of outcomes. Before examining
59 early diagnostic strategies, it is important to consider traditional diagnostic approaches.
60 Conventional cardiovascular workflows rely on established clinical protocols, where physicians
61 gather information through physical examination, patient history, and diagnostic tests. These
62 tests typically include electrocardiograms (ECGs), blood pressure measurements, cholesterol
63 profiling, and stress tests. In advanced cases, invasive procedures such as angiography or
64 fluoroscopy are employed. While these methods provide valuable insights, they are often

65 resource-intensive, requiring costly equipment, specialized personnel, and significant time.
66 Moreover, they are predominantly reactive, being deployed only after symptoms become
67 apparent [7]. This limits opportunities for early detection, particularly in asymptomatic
68 or atypical cases. Additionally, traditional methods rely on generalized risk models, which
69 may not adequately account for individual variations in lifestyle, genetics, or environmental
70 exposure, resulting in limited personalization in diagnosis and treatment planning. Machine
71 learning (ML) offers a transformative alternative by enabling data-driven, personalized, and
72 scalable diagnostic support. Trained on large datasets such as the UCI Heart Disease dataset,
73 ML algorithms can uncover complex nonlinear relationships across multiple patient variables,
74 patterns that might otherwise be overlooked in conventional assessments. These models
75 not only improve diagnostic accuracy but also predict disease risk before the onset of symp-
76 toms. Importantly, retraining ML systems on non-invasive or self-reported data enhances
77 accessibility and affordability. Furthermore, integration with wearables, lifestyle trackers,
78 and electronic health records enables continuous monitoring and timely risk alerts. This
79 shift supports low-cost, remote, and proactive care delivery, which is particularly valuable in
80 underserved regions. While ML is not intended to replace clinical judgment, it functions as a
81 powerful tool to augment decision-making, prioritize high-risk patients, reduce diagnostic
82 delays, and enable more precise interventions.

83
84 Despite significant progress, current AI-based diagnostic systems for CVD continue to
85 face several critical limitations. First, most existing models are diagnosis-oriented and seldom
86 provide early warnings based on non-clinical or self-reported features, leaving a crucial gap in
87 preventive care. Second, although AI adoption in healthcare is increasing, key issues such
88 as interpretability, fairness, and predictive uncertainty are often neglected, reducing clinical
89 trust and reliability. Third, the predominance of correlation-driven approaches has limited
90 the incorporation of causal inference, thereby constraining model generalizability. Finally,
91 advanced tabular neural networks (e.g., TabNet, TabPFN), which are particularly well-suited
92 for healthcare data, remain largely unexplored in the context of CVD prediction.

93
94 To address these challenges, we propose a comprehensive, ethically aligned AI framework for
95 CVD risk assessment with the following key contributions:

- 96 1. An early warning system that leverages non-clinical and self-reported features suitable for
97 deployment via mobile health applications or community screening kiosks for accessible,
98 cost-effective risk stratification.
- 99 2. Integration of interpretability tools (e.g., SHAP, counterfactual explanations), fairness
100 mitigation (e.g., FairLearn), and uncertainty estimation (e.g., Bayesian neural networks,
101 TabPFN).
- 102 3. Use of causal inference to identify mechanistic disease drivers beyond correlation.
- 103 4. Extensive benchmarking of advanced tabular neural networks against ensemble learning
104 methods.

105 This framework is designed in alignment with regulatory guidelines from both the FDA and
106 the European Union, with the goal of delivering a scalable, equitable, and clinically actionable

107 solution for cardiovascular care.

108 **2 Related Work**

109 Machine learning (ML) and deep learning (DL) methods have been widely applied CVD
110 prediction, with studies exploring feature selection, model optimization, and ensemble tech-
111 niques. Despite strong performance across many works, challenges persist in generalizability,
112 interpretability, and ethical deployment.

113 **2.1 Traditional Machine Learning and Ensemble Models**

114 Qadri et al. [8] proposed the Principal Component Heart Failure (PCHF) method, reducing
115 dimensionality to eight features and reporting 100% accuracy using a Decision Tree—an
116 outcome suggestive of overfitting. Kumar et al. [15] evaluated classical models on the UCI
117 Heart Disease dataset and identified logistic regression as the most reliable. These studies
118 underscore the utility of feature selection and simple classifiers but provide limited discussion
119 on external validity and ethical considerations. Subramani et al. [11] introduced a stacking
120 ensemble using IoT-based inputs, achieving 96% accuracy. Rohan et al. [14] compared 21
121 classifiers with 11 feature selection methods, identifying XGBoost as the top model (F1-
122 score: 98%). Although ensemble methods consistently yield high accuracy, most works focus
123 narrowly on performance metrics while overlooking transparency and fairness. Mohan et
124 al. [22] applied Decision Tree entropy for feature extraction and achieved 0.887 accuracy using
125 a Random Forest–Linear Regression hybrid. Ali et al. [23] reported perfect accuracy (1.0)
126 with DT and RFC, raising concerns about potential overfitting. Bhatt et al. [24] conducted a
127 more extensive evaluation involving DT, RF, XGBoost, and MLP with GridSearch tuning and
128 Huang clustering, reaching an average cross-validation accuracy of 0.8707. Although diverse
129 in methodology, these works tend to prioritize optimization over clinical interpretability.

130 **2.2 Deep Learning Models and Feature Engineering**

131 Deep neural networks (DNNs) have shown strong capacity for capturing nonlinear patterns
132 in CVD data. Almazroi et al. [10] developed a Keras-based dense neural network that
133 outperformed ensemble methods. Saeed et al. [12] combined DNNs with SelectKBest and
134 SMOTE, reporting accuracies up to 99%. However, these studies provide limited attention to
135 explainability. Al-Alshaikh et al. [9] proposed a hybrid system integrating genetic algorithms,
136 recursive feature elimination, and a convolutional neural network optimized using adaptive
137 elephant herd strategies. The approach delivered high recall (96.2%) but emphasized the
138 need for broader real-world validation to improve generalization.

139 **2.3 Innovative Neural and Bio-Inspired Techniques**

140 Novel architectures and bio-inspired optimizers have been explored for CVD prediction.
141 Nandy et al. [16] introduced the Swarm-Artificial Neural Network (Swarm-ANN), achieving
142 95.78% accuracy by using heuristic-driven weight adjustments. While effective, interpretability

143 remained limited. Eleyan et al. [13] developed Rhythmi, a CNN-based mobile diagnostic tool
144 for ECG analysis, reporting over 98% accuracy. Although promising for real-time diagnosis,
145 the work relied on a relatively small dataset and focused primarily on signal data rather than
146 tabular clinical datasets commonly found in EHRs.

147 2.4 Recent Advances in Limited Data Integration

148 Recent studies have examined ML and DL approaches for integrating heterogeneous or limited
149 clinical datasets. Mehdi et al. [29] distinguished normal and impaired cardiomyocytes using
150 sarcomere transients and calcium kinetics, achieving AUC values of 0.94–0.95 through an
151 ensemble classifier. Building on multimodal integration, Mehdi et al. [30] proposed a multi-
152 fidelity ML framework for myocardial infarction diagnosis that combined simulation-based
153 low-fidelity data with limited human CMR data, improving Dice scores from 0.39 to 0.72.
154 Their later work [31] estimated myocardial mechanical properties directly from anatomical
155 and hemodynamic features, reporting high predictive accuracy ($R_{af}^2 = 99.12\%$, $R_{bf}^2 = 93.31\%$).
156 Pressure–volume loading emerged as the most informative feature set. Collectively, these
157 studies highlight the promise of multimodal integration and highlight gaps in interpretability,
158 fairness, and uncertainty quantification—areas critical for developing clinically trustworthy
159 AI systems.

160 3 Methodology

161 Multiple models were developed for clinical and non-clinical applications. Non-clinical model-
162 ing used three ensemble methods, while clinical modeling employed tabular neural networks,
163 a causality-guided voting classifier (CGEVC), Bayesian neural networks, and the ensemble
164 baselines. The framework incorporated interpretability and uncertainty quantification.

165
166 A diverse set of machine learning and deep learning methods was chosen to align with
167 regulatory priorities (FDA, EU AI Act) emphasizing interpretability, uncertainty, causality,
168 and robustness. Ensembles (Random Forest, XGBoost, LightGBM) served as strong baselines,
169 tabular neural networks (TabNet, TabPFN, FT-Transformer) offered interpretable generaliza-
170 tion for small datasets, Bayesian networks quantified uncertainty, and CGEVC prioritized
171 causally relevant features. This multi-model framework enabled systematic evaluation of
172 robustness, interpretability, fairness, and uncertainty, while identifying models suitable for
173 real-world deployment. Reproducibility guidelines are provided in the GitHub repository [27].
174 The overall workflow is illustrated in Figure 1.

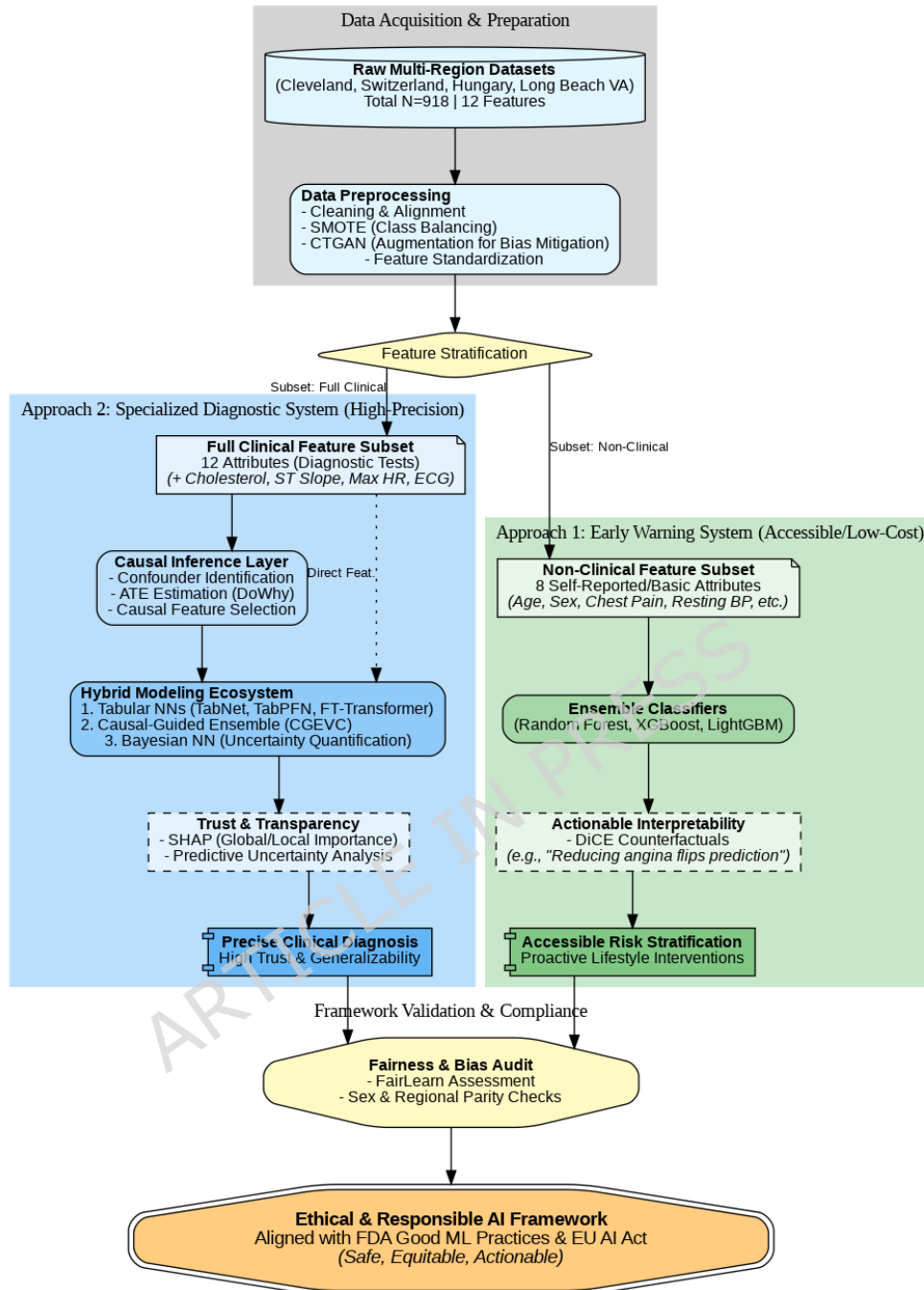


Figure 1: This diagram illustrates the methodology for the two key approaches described in the study. (1) Non-Clinical Stream: Designed for low-resource settings, this stream uses accessible patient history features to train interpretability-focused models (Random Forest, XGBoost) that suggest lifestyle changes via DiCE counterfactuals. (2) Clinical Stream: Designed for medical settings, this stream incorporates diagnostic test results (e.g., ST slope, Serum Cholesterol) into causality-guided and probabilistic neural architectures (TabPFN, BNN) to ensure clinical reliability. The shared evaluation layer ensures both streams meet ethical standards for fairness and transparency.

3.1 Dataset Information

Two datasets from Kaggle were utilized, both derived from the UCI Heart Disease dataset [28]. To avoid ambiguity, they are distinguished as follows:

- **Cleveland:** A Kaggle-provided dataset containing only the Cleveland Clinic data, with 303 instances and 14 features. This dataset has been widely adopted in heart disease prediction research.
- **Multi-Region-Combined:** A merged dataset comprising records from five sources: Cleveland, Switzerland, Long Beach VA, Hungary, and Statslog (Kaggle-provided). The initial version contained 1,190 instances.

The combined dataset exhibited several inconsistencies that required preprocessing:

1. The features `ca` (number of major vessels colored by fluoroscopy) and `thal` (thalassemia) were missing from the Switzerland, Long Beach VA, and Hungary subsets. To maintain feature consistency, these attributes were removed from the Cleveland and Statslog subsets as well.
2. Column names in the Statslog dataset were inconsistent with the Cleveland schema. These were standardized to ensure alignment.
3. Detailed inspection revealed that all 272 rows in Statslog were duplicates of rows from the Cleveland dataset, differing only in order. These duplicates were removed.

After preprocessing, the final **Multi-Region** dataset contained 918 unique instances aggregated from four sources: Cleveland, Switzerland, Long Beach VA, and Hungary. Table 1 summarizes the datasets, their sizes, and their differences, providing an overview of the data sources used in this study. Figure 2 shows how each dataset is derived and used in the research.

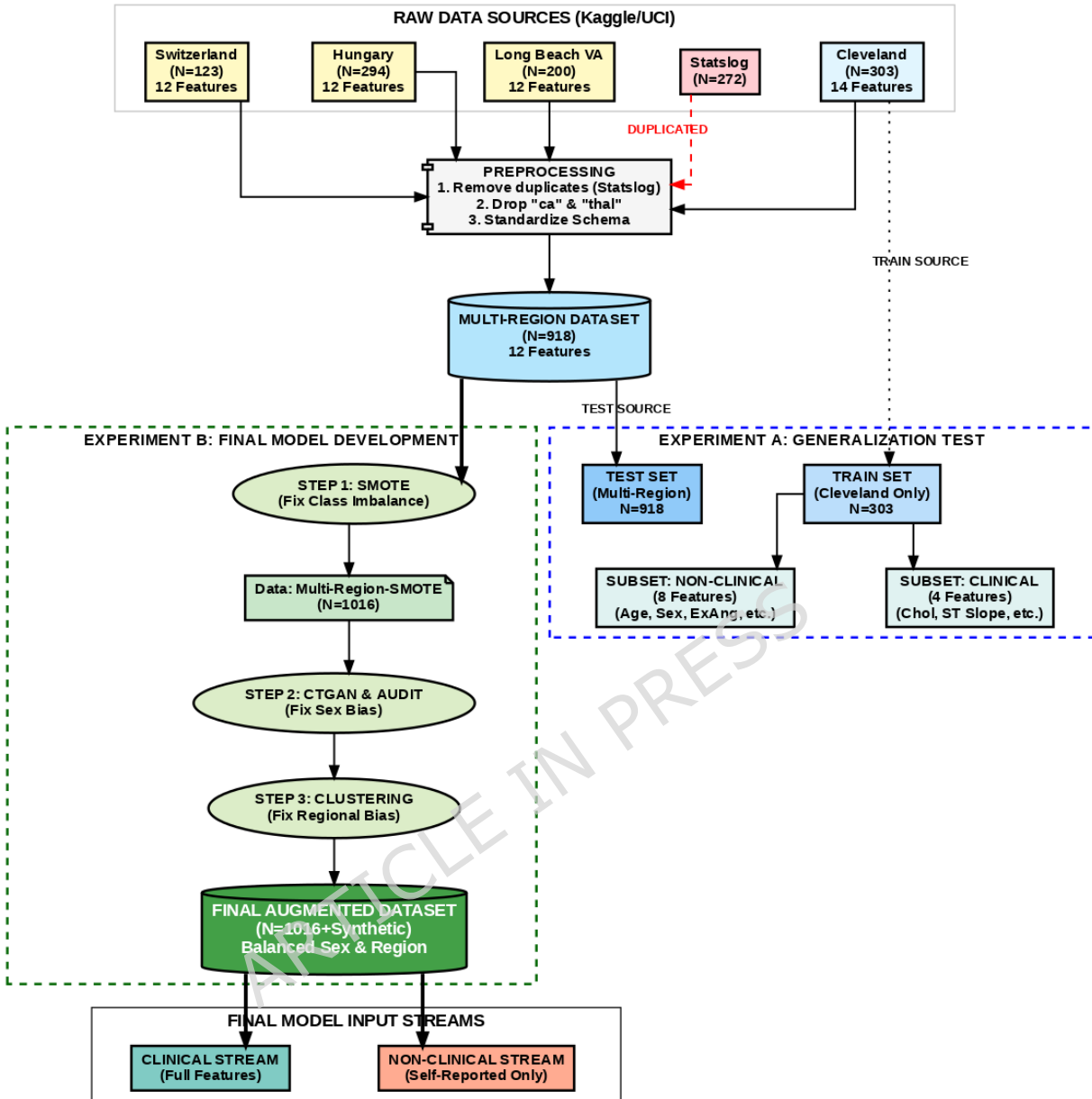


Figure 2: The diagram illustrates the consolidation of five raw data sources (Cleveland, Hungary, Switzerland, Long Beach VA, and Statslog) into a unified Multi-Region dataset ($N = 918$). Preprocessing involved removing the Statslog subset due to duplication and dropping inconsistent features (`ca`, `thal`) to ensure schema alignment. The workflow bifurcates into two experimental phases: **(A) Generalization Testing**, where models trained solely on Cleveland ($N = 303$) are evaluated on the combined dataset; and **(B) Final Model Development**, which employs a multi-stage bias mitigation pipeline. This pipeline integrates **SMOTE** for class balancing ($N = 1016$), **CTGAN** for demographic parity (Sex), and unsupervised **Clustering** to address regional bias. Finally, features are stratified into **Clinical** and **Non-Clinical** streams to support the framework's dual-model architecture.

198 3.1.1 Generalization Test

199 To evaluate generalization, models were first trained on the Cleveland dataset and then tested
200 on the Multi-Region dataset. For both datasets, features were divided into two categories:

- 201 • **Clinical features:** serum cholesterol, resting electrocardiographic result, ST depression
202 induced by exercise relative to rest, and slope of the peak exercise ST segment.
- 203 • **Non-clinical features:** all remaining attributes.

204 This division yielded four additional subsets: Cleveland-Clinical, Cleveland-Non-Clinical,
205 Multi-Region-Clinical, and Multi-Region-Non-Clinical. Models trained on non-clinical sub-
206 sets provide early-warning predictions, whereas clinical subsets incorporate diagnostic test
207 outcomes.

208
209 Three ensemble models (XGBoost, LightGBM, and Random Forest) were trained on the
210 Cleveland-Clinical and Cleveland-Non-Clinical datasets, and subsequently evaluated on the
211 corresponding Multi-Region subsets to assess generalization ability. Performance with accu-
212 racy above 0.80 and ROC-AUC above 0.85 was regarded as evidence of strong generalization.

213 3.2 Modeling on Multi-Regional Data

214 After confirming generalization from Cleveland to Multi-Region data, final model development
215 was conducted using the Multi-Region dataset. To mitigate slight class imbalance, the
216 Synthetic Minority Oversampling Technique (SMOTE) was applied.

Table 1: Summary of the datasets used in this study. The table categorizes data into raw sources, the unified multi-region set, and the specific subsets derived for generalization testing and final model development.

Dataset	Instances	Features	Description / Notes
<i>Raw Data Sources</i>			
Cleveland	303	14	Original Cleveland data (Kaggle).
Statslog	272	14	Removed (Duplicates of Cleveland).
Switzerland	123	12	Missing <code>ca</code> and <code>thal</code> features.
Long Beach VA	200	12	Missing <code>ca</code> and <code>thal</code> features.
Hungary	294	12	Missing <code>ca</code> and <code>thal</code> features.
<i>Merged Data</i>			
Multi-Region	918	12	Aggregation of Cleveland + 3 regional datasets.
<i>Experiment A: Generalization Test Subsets</i>			
Cleveland-Clinical	303	4	Train set for clinical modeling.
Cleveland-Non-clinical	303	8	Train set for non-clinical modeling.
Multi-Region-Clinical	918	4	Test set for clinical generalization.
Multi-Region-Non-clinical	918	8	Test set for non-clinical generalization.
<i>Experiment B: Final Modeling (Augmented)</i>			
Multi-Region-SMOTE	1016	12	Balanced via SMOTE. Basis for final models.
Clinical Dataset	1016	4	Final input for TabNet, TabPFN, etc.
Non-clinical Dataset	1016	8	Final input for Random Forest, Ensembles.

217 Figure 3 illustrates the class distribution before and after applying SMOTE. SMOTE gen-
 218 erates synthetic samples for the minority class to balance imbalanced datasets. Given that
 219 generalization has already been established, the use of SMOTE does not raise ethical con-
 220 cerns regarding data leakage. Unlike simple duplication, SMOTE creates new instances by
 221 interpolating between existing minority class samples, ensuring that the augmented data
 222 introduces meaningful variability while preserving the original data distribution.



Figure 3: Class distribution of the target variable (heart disease presence) in the multi-regional dataset before and after applying the SMOTE, demonstrating the effective balancing of the dataset.

223 Given a minority class sample \mathbf{x}_i , a synthetic sample $\tilde{\mathbf{x}}$ is generated as:

$$\tilde{\mathbf{x}} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_{z_i} - \mathbf{x}_i) \quad (1)$$

224 where \mathbf{x}_i is a minority class instance, \mathbf{x}_{z_i} is one of the k nearest neighbors of \mathbf{x}_i , and $\lambda \sim \mathcal{U}(0, 1)$
 225 is a random number drawn from a uniform distribution. This process generates synthetic
 226 data points along line segments connecting minority class neighbors, reducing overfitting and
 227 improving model generalization on imbalanced datasets. The oversampling procedure was
 228 applied once on the combined dataset, after which clinical features were removed to create
 229 separate clinical and non-clinical subsets. The three ensemble models were then trained and
 230 evaluated on both subsets.

231

232 While the initial performance was promising, the bias report from FairLearn indicated
 233 a tendency for the models to predict CVD more frequently for men, while predicting most
 234 women as healthy. This pattern may reflect the natural epidemiology of CVD, as men gener-
 235 ally exhibit higher incidence rates. However, the observed bias still raises ethical concerns.
 236 To address this, an additional experiment was conducted using data augmentation with a
 237 Conditional Tabular Generative Adversarial Network (CTGAN). Synthetic instances were
 238 generated to equalize the frequencies of diseased men and women, as well as healthy men
 239 and women. This experiment demonstrates that balancing the dataset can mitigate bias
 240 without requiring complex methods such as adversarial debiasing or algorithmic regularization.

241 Importantly, CTGAN was not applied to improve model performance in this study. The
 242 generative process does not introduce data leakage, as CTGAN is specifically designed for
 243 tabular data and effectively handles imbalanced, mixed-type datasets (both continuous and
 244 categorical features). By using a conditional generator, the model captures the distribution
 245 of minority classes and rare categories more accurately.

246

247 Integrating a conditional generator into a GAN architecture involves addressing three key
 248 challenges. First, an appropriate representation of the condition must be devised and pro-
 249 vided as input. Second, the generated rows must faithfully preserve the specified condition.
 250 Third, the conditional generator must learn the true conditional distribution of the real data,
 251 ensuring realistic and representative synthetic samples.

$$\mathbb{P}_G(\text{row} \mid D_{i_*} = k^*) = \mathbb{P}(\text{row} \mid D_{i_*} = k^*) \quad (2)$$

252 so that we can reconstruct the original distribution as

$$\mathbb{P}(\text{row}) = \sum_{k \in D_{i_*}} \mathbb{P}_G(\text{row} \mid D_{i_*} = k^*) \mathbb{P}(D_{i_*} = k) \quad (3)$$

253 To effectively model continuous columns, CTGAN employs mode-specific normalization using
 254 Variational Gaussian Mixture Models (VGM). For discrete columns, conditioning vectors \mathbf{c}
 255 are sampled to guide generation, ensuring better representation of underrepresented categories
 256 [25]. This conditional mechanism enables CTGAN to generate diverse and realistic synthetic
 257 tabular data, particularly in class-imbalanced scenarios. After applying CTGAN, the three
 258 ensemble models were retrained on the augmented dataset, resulting in balanced bias metrics
 259 across genders.

260

261 To fully address demographic bias, it was also necessary to evaluate potential regional
 262 bias. Although the final model does not include region as a feature, since real-world users are
 263 unlikely to correspond only to Switzerland, Long Beach VA, Cleveland, or Hungary, the goal
 264 was to generalize the model across all regions rather than rely solely on external validation. To
 265 achieve this, theoretical regions were extracted from the dataset using unsupervised learning.
 266 K-Means clustering, combined with Principal Component Analysis (PCA), was applied to
 267 identify four regions. The unsupervised approach produced a silhouette score of 0.21, and
 268 the resulting regional frequencies closely mirrored those of the original dataset, as illustrated
 269 in Figure 4.

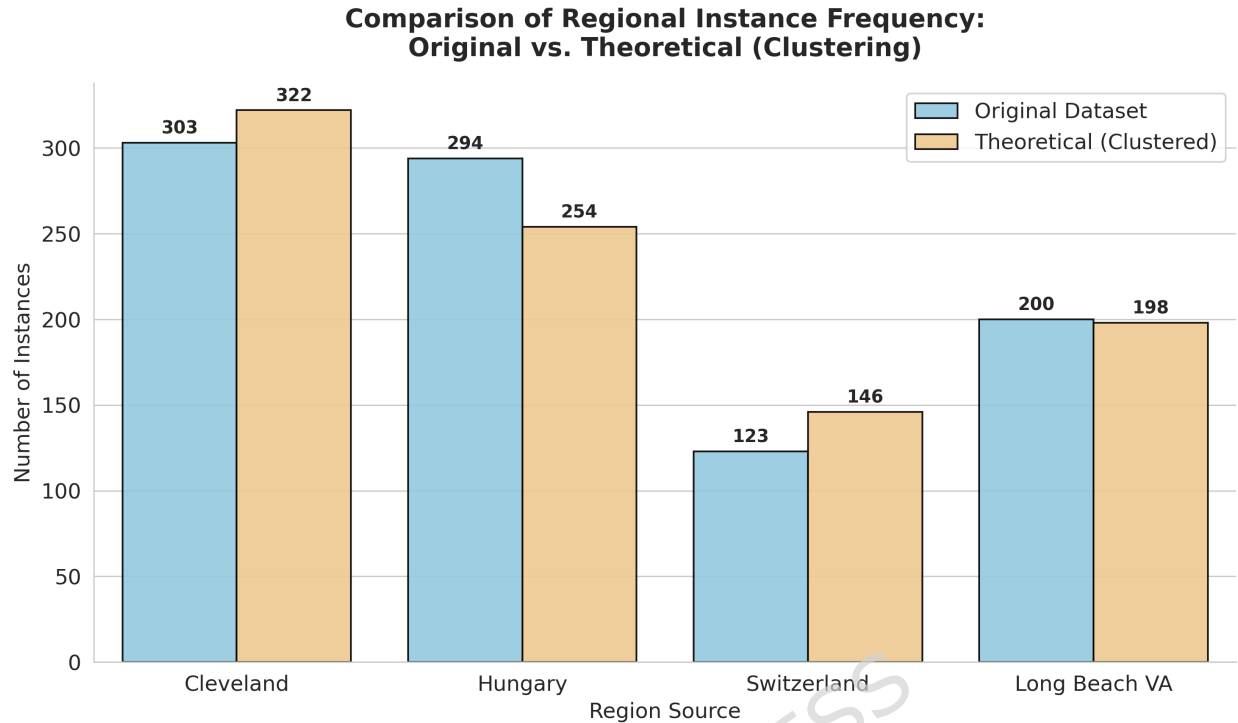


Figure 4: Distributional validation of the unsupervised regional clustering. The blue bars represent the actual instance counts from the four source datasets (Cleveland, Hungary, Switzerland, Long Beach VA), while the orange bars represent the clusters identified via PCA and K-Means. The close alignment in frequency counts (Silhouette Score: 0.21) confirms that the unsupervised theoretical regions effectively reconstruct the original geographical sources, validating their use for assessing regional bias.

270 The three ensemble models were subsequently trained on the dataset augmented with the
 271 additional theoretical region feature, and bias metrics were evaluated. The analysis revealed
 272 notable insights, which motivated the use of agglomerative clustering instead of K-Means
 273 with PCA, as detailed in the Results section. Following this, the combined dataset with
 274 SMOTE and CTGAN (excluding the region feature) was used to develop three tabular
 275 neural network models: TabNet, FT-Transformer, and TabPFN. A Bayesian Neural Network
 276 (BNN) employing Monte Carlo Dropout was also implemented for uncertainty quantification.
 277 Additionally, a causality-guided voting classifier, composed of XGBoost, LightGBM, Random
 278 Forest, K-Nearest Neighbors, and Logistic Regression, was developed to enhance prediction
 279 accuracy. Prior to training the neural network models, the dataset was standardized using
 280 scikit-learn's `StandardScaler()`, which performs z-score normalization on all features. The
 281 voting classifier was trained using only features with an Average Treatment Effect (ATE) of
 282 at least 0.1 on CVD diagnosis.

283 3.3 Model Selection

284 We employed three ensemble models, three tabular neural networks, and a Bayesian Neural
285 Network (BNN). Non-clinical models use only ensembles; clinical models use all architectures.

286 3.3.1 Ensemble Models

287 **XGBoost:** Boosted trees optimized via a regularized objective; well-suited for mixed tabular
288 data and tuned using Bayesian Optimization.

289 **Random Forest:** Averaged predictions from many bootstrapped trees; reduces variance
290 and overfitting in small clinical datasets.

291 **LightGBM:** Optimized tree learner using GOSS and EFB for efficient training; hyperpa-
292 rameters tuned via Optuna.

293 3.3.2 Tabular Neural Networks

294 **TabNet:** Sequential attention selects features sparsely, providing interpretability and strong
295 performance on mixed data.

296 **TabPFN:** A pretrained Bayesian few-shot learner that approximates posterior predictions
297 without tuning, ideal for small datasets.

298 **FT-Transformer:** Applies self-attention over feature tokens; default hyperparameters used
299 to prevent overfitting on small samples.

300 3.3.3 Bayesian Neural Network (BNN)

301 **BNN with MC Dropout:** Learns weight distributions and estimates predictive uncertainty
302 via multiple stochastic forward passes, beneficial for high-risk clinical prediction.

303 3.3.4 Causal-Guided Ensemble Voting Classifier (CGEVC)

304 A causal-guided ensemble classifier is designed to enhance the robustness and interpretability
305 of cardiovascular disease (CVD) prediction. The workflow integrates confounder discovery,
306 causal effect estimation, and ensemble learning. The pipeline is described below:

307
308 **1. Identifying Potential Confounders:** To estimate the causal effect of each feature
309 (treatment variable) on the target (CVD diagnosis), the potential confounders are firstly
310 identified using a data-driven approach. For each feature T_i :

- 311 1. Remove the target variable Y from the dataset.
- 312 2. Use a Random Forest Classifier (RFC) to predict T_i using the remaining features $\mathbf{X} \setminus T_i$.

313 3. Extract the feature importances from RFC and select the subset $\mathbf{Z}_i \subset \mathbf{X} \setminus T_i$ such that
 314 the cumulative importance satisfies:

$$\sum_{x \in \mathbf{Z}_i} \text{Importance}(x) \geq 0.80 \quad (4)$$

315 The selected features \mathbf{Z}_i are considered potential confounders for estimating the causal effect
 316 of T_i on Y .

317

318 **2. Estimating Average Treatment Effect (ATE):** The backdoor criterion is used
 319 from causal inference to estimate the Average Treatment Effect (ATE) of each treatment T_i
 320 on the outcome Y , conditioned on the identified confounders \mathbf{Z}_i . According to the backdoor
 321 adjustment formula:

$$\text{ATE}(T_i \rightarrow Y) = \mathbb{E}_{\mathbf{z} \sim \mathbf{Z}_i} [\mathbb{E}[Y \mid T_i = 1, \mathbf{Z}_i = \mathbf{z}] - \mathbb{E}[Y \mid T_i = 0, \mathbf{Z}_i = \mathbf{z}]] \quad (5)$$

322 This is implemented via the DoWhy framework with the `backdoor.linear_regression` esti-
 323 mator, which assumes the linear model:

$$Y = \beta_0 + \beta_1 T_i + \boldsymbol{\gamma}^\top \mathbf{Z}_i + \epsilon \quad (6)$$

324 The coefficient β_1 provides the estimated causal effect of T_i on Y , while controlling for
 325 confounders \mathbf{Z}_i .

326

327 **3. Causal-Guided Feature Selection and Ensemble Classification:** Features with
 328 statistically significant (0.1) ATE values are selected to train a soft voting ensemble classifier
 329 composed of:

- 330 • Random Forest Classifier (RFC)
- 331 • XGBoost
- 332 • LightGBM
- 333 • k -Nearest Neighbors (KNN)
- 334 • Logistic Regression (LR)

335 The final prediction \hat{y} is the weighted average of predicted class probabilities:

$$\hat{y} = \arg \max_{c \in \mathcal{C}} \sum_{m=1}^M w_m \cdot \hat{p}_m(c \mid \mathbf{x}) \quad (7)$$

336 where w_m is the weight of model m , and \hat{p}_m is the predicted probability for class c .

337

338 This pipeline integrates causal inference into model training, helping to focus on variables
 339 that have meaningful causal influence on the target. This improves interpretability, reduces
 340 bias from spurious correlations, and enhances performance.

341 4 Results and Discussion

342 Table 2 details the classification performance of the ensemble models across both experimental
 343 setups. It highlights that the application of SMOTE and multi-regional training data
 344 significantly enhanced model precision and recall compared to the baseline generalization
 tests.

Table 2: **Performance comparison across training phases.** The table aggregates results for the **Generalization Test** (models trained on Cleveland, tested on Multi-Region) and the **Final Evaluation** (models trained on the balanced Multi-Region dataset).

Experiment Phase	Feature Set	Model	Accuracy	AUC	F1 Score	Precision	Recall
Generalization Test (Train: Cleveland) (Test: Multi-Region)	Clinical	XGBoost	0.8028	0.8858	0.8297	0.79	0.87
		LightGBM	0.8159	0.8891	0.8425	0.80	0.89
		Random Forest	0.8292±0.002	0.9102±0.001	0.8499±0.002	0.84	0.85
	Non-Clinical	XGBoost	0.8137	0.8821	0.8452	0.78	0.92
		LightGBM	0.8082	0.8729	0.8365	0.77	0.91
		Random Forest	0.8038±0.001	0.8905±0.001	0.8381±0.001	0.76	0.92
Final Evaluation (Train: Multi-Region) (Balanced with SMOTE)	Clinical	XGBoost	0.9069	0.9578	0.9082	0.90	0.92
		LightGBM	0.9020	0.9537	0.9029	0.89	0.91
		Random Forest	0.8995±0.001	0.9509±0.001	0.9012±0.001	0.89	0.91
	Non-Clinical	XGBoost	0.8400	0.9055	0.8400	0.82	0.86
		LightGBM	0.8480	0.9028	0.9029	0.85	0.84
		Random Forest	0.8500±0.001	0.9085±0.001	0.8536±0.001	0.83	0.88

345

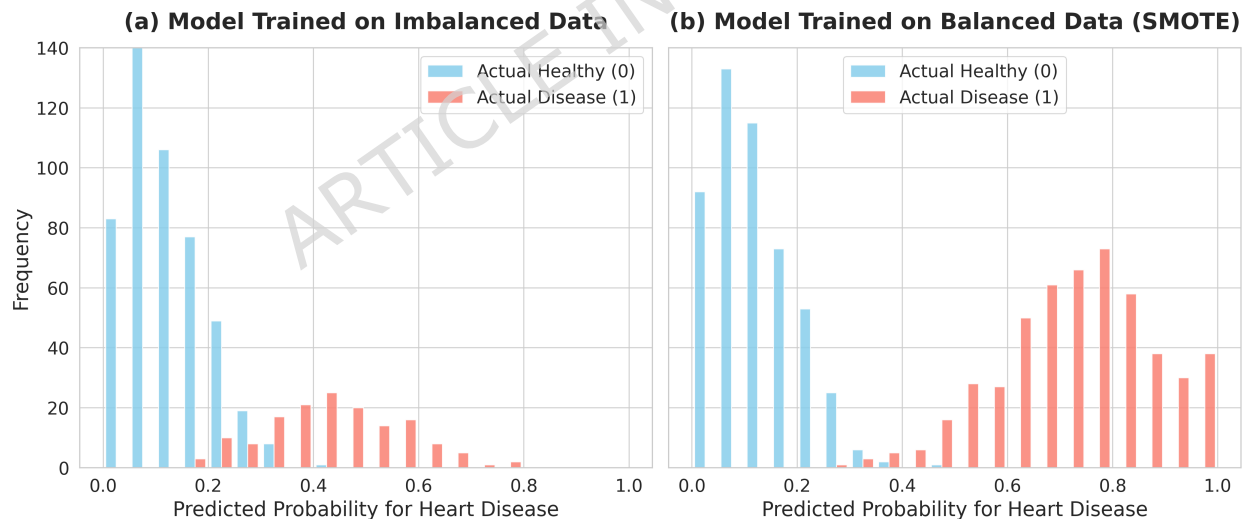


Figure 5: Histograms of predicted probabilities for the XGBoost clinical model. (a) visualizes the distribution on the original imbalanced dataset, and the SMOTE-balanced dataset is visualized in (b). The balanced dataset shows improved separation between healthy (blue) and diseased (red) classes.

346 Figure 5 visualizes that most blue bars (healthy) are clustered at low probabilities (0–0.3)
 347 for the imbalanced dataset. Many red bars (disease) are pushed toward lower probabilities

348 (0.2–0.4), instead of being clearly separated. This suggests the model (XGBoost trained
 349 on clinical dataset) struggles to predict disease correctly, indicating it’s biased toward the
 350 majority class (healthy) due to class imbalance. Conversely, the red bars (disease) are
 351 spread across higher probabilities (0.5–1.0), while blue bars (healthy) remain mostly at low
 352 probabilities (0–0.3) for the balanced dataset. There is better separation between healthy
 353 and disease, meaning the model can distinguish the classes more confidently. So, SMOTE
 354 helped the model learn patterns of the minority class, improving prediction for disease cases.

355 4.1 Fairness Metrics

Table 3: (A) Sex-wise performance metrics (Accuracy, Selection Rate, Precision, Recall) and (B) corresponding fairness metrics (Demographic Parity Difference/DPD, Demographic Parity Ratio/DPR, Equalized Odds Difference/EOD, Equalized Odds Ratio/EOR) for the primary clinical (XGBoost) and non-clinical (Random Forest) models, highlighting initial demographic disparities.

(A) Sex-wise Performance Metrics					
Model	Sex	Accuracy	Selection Rate	Precision	Recall
XGBoost (Clinical)	Female	0.9074	0.22	0.75	0.81
	Male	0.9066	0.62	0.95	0.91
RF (Non-Clinical)	Female	0.9242	0.2272	0.80	0.85
	Male	0.8617	0.6329	0.86	0.91
(B) Fairness Metrics					
Model	DPD	DPR	EOD	EOR	
XGBoost (Clinical)	0.40	0.35	0.11	0.51	
RF (Non-Clinical)	0.40	0.35	0.15	0.27	

356 From Table 3, Both clinical (XGBoost) and non-clinical (RF) models show noticeable dis-
 357 parities in sex-wise predictions. The selection rate for males is substantially higher (62% vs
 358 22% for XGBoost, and 63.3% vs 22.7% for RF), which directly contributes to a demographic
 359 parity difference (DPD) of 0.40 in both cases. This implies that males are much more likely
 360 to be predicted as positive compared to females. Likewise, the demographic parity ratio
 361 (DPR) is only 0.35, far below the commonly accepted fairness threshold of 0.8 (the “80% rule”
 362 [26]), suggesting potential adverse impact.

363
 364 Precision and recall values also reveal disparities: males tend to receive higher precision
 365 (0.95 for XGBoost, 0.86 for RF) while females exhibit lower precision but relatively balanced
 366 recall. Equalized odds differences (EOD) of 0.11 (XGBoost) and 0.15 (RF) further confirm
 367 discrepancies in true/false positive rates, with equalized odds ratios (EOR) well below 1
 368 (0.51 and 0.27, respectively). Ideally, both should be close to 0 (difference) and 1 (ratio),
 369 respectively. However, these disparities may not solely reflect algorithmic bias but could also
 370 be linked to the underlying prevalence of heart disease by sex, as illustrated in Figure 6.

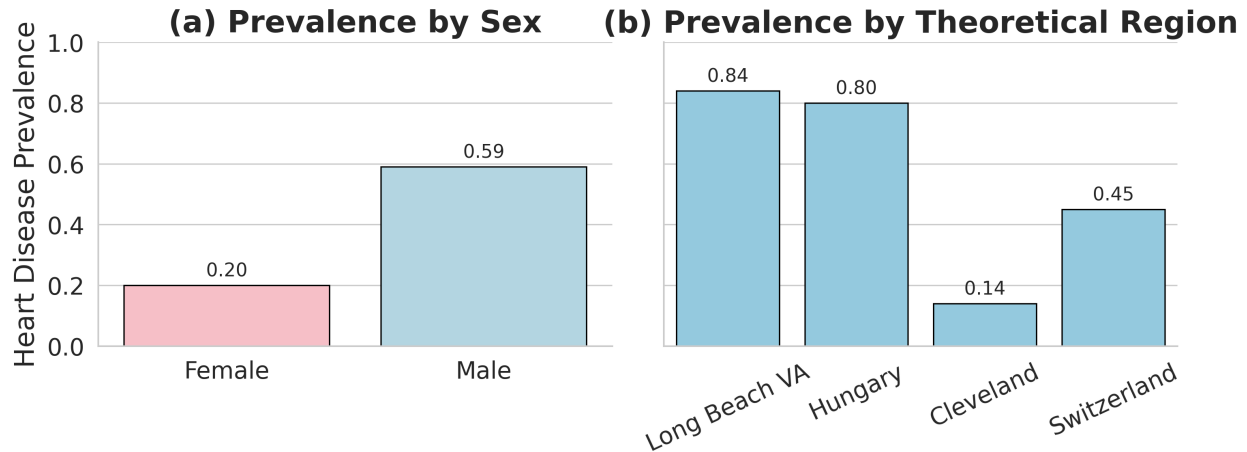


Figure 6: Bar chart illustrating the underlying prevalence of heart disease by sex and by the theoretically derived regions within the dataset, providing context for observed biases in model predictions.

371 Figure 6 also illustrates that, beyond sex-based fairness, regional bias could also exist because
 372 of Cleveland’s low prevalence rate. The selection rate and recall of Cleveland (theoretical)
 373 were significantly lower compared to other clusters, for both clinical (0.13) and non-clinical
 374 (0.082) models. Such imbalance could be due to differences in heart disease prevalence across
 375 regions, as shown in Figure 6(b), or the limitations of the PCA + K-Means clustering approach.

376
 377 To address this, agglomerative clustering was attempted, yielding four clusters with a silhou-
 378 ette score of 0.44. One cluster, however, contained only 8 instances and shared distributional
 379 similarity with another cluster. It indicates that one region (e.g., Cleveland) closely resembles
 380 to some other region (e.g., Longbeach VA), and therefore we can drop it to analyze disparities
 381 with more transparency. Hence, we retained three clusters instead of four and retrained the
 382 models, which improved balance in both performance and fairness metrics across regions. The
 383 final selection rates for the three regions were 0.52, 0.92, and 0.42 respectively, supporting
 384 the robustness and generalizability of the models under external validation.

385 4.1.1 Bias Mitigation Results

386 After applying CTGAN and re-training the models, fairness metrics improved significantly.

387

388 For the clinical model XGBoost:

- 389 • Demographic parity difference = 0.05
- 390 • Demographic parity ratio = 0.91
- 391 • Equalized odds difference = 0.09
- 392 • Equalized odds ratio = 0.55

393 For the non-clinical model Random Forest Classifier:

- 394 • Demographic parity difference = 0.07
- 395 • Demographic parity ratio = 0.892
- 396 • Equalized odds difference = 0.089
- 397 • Equalized odds ratio = 0.715

398 The fairness evaluation of both models shows acceptable performance in terms of demographic
399 parity than before. However, it reveals some concerns regarding equalized odds. The general
400 model exhibits a demographic parity difference of 0.05 and a ratio of 0.91, indicating a minor
401 disparity in selection rates between groups. Similarly, the Random Forest (non-clinical)
402 model has a slightly higher demographic parity difference of 0.07 and a ratio of 0.892, both
403 within commonly accepted fairness thresholds. However, equalized odds metrics are a bit
404 problematic: the general model shows a difference of 0.09 and a low ratio of 0.55, suggesting
405 an imbalance in error rates across demographic groups. The Random Forest model performs
406 slightly better with an equalized odds difference of 0.089 and a ratio of 0.715, though still
407 below the 0.8 fairness threshold. These results imply that while group-wise selection rates
408 are relatively fair, there is disparity in how accurately the models treat different groups,
409 warranting mitigation efforts focused on reducing outcome disparities. Future research may
410 look into this further.

411 4.2 Model Interpretability and Counterfactuals

412 SHAP interpretations and counterfactual explanations were incorporated into both clinical
413 and non-clinical models, enabling comprehensive global and local interpretability for each
414 prediction. In clinical models, SHAP identified variables such as chest pain type, exercise-
415 induced angina, and ST slope as critical contributors. In non-clinical models, behavioral and
416 symptom-related features, including chest pain and heart rate response, were emphasized.
417 Counterfactual explanations, generated using DiCE, complemented SHAP by illustrating
418 minimal actionable changes capable of altering a prediction from diseased to non-diseased.
419 These interpretability tools enhance transparency and model trustworthiness while providing
420 clinicians and users with actionable insights into potential lifestyle or symptom-based inter-
421 ventions. This aligns with ethical AI principles and increases the suitability of the models for
422 real-world healthcare deployment.

423
424 The SHAP summary plot and instance-level interpretation for the clinical model, shown in
425 Figure 7(b), depict the global feature importance and their impact on XGBoost predictions.
426 Features such as slope, oldpeak, and chest pain type exert the greatest influence, with wider
427 distributions indicating stronger effects. Positive SHAP values (right side) increase predicted
428 CVD risk, whereas negative values (left side) reduce risk estimates. The plot also indicates
429 that age and sex have relatively smaller, yet consistent, impacts compared to other clinical
430 features such as maximum heart rate and serum cholesterol.

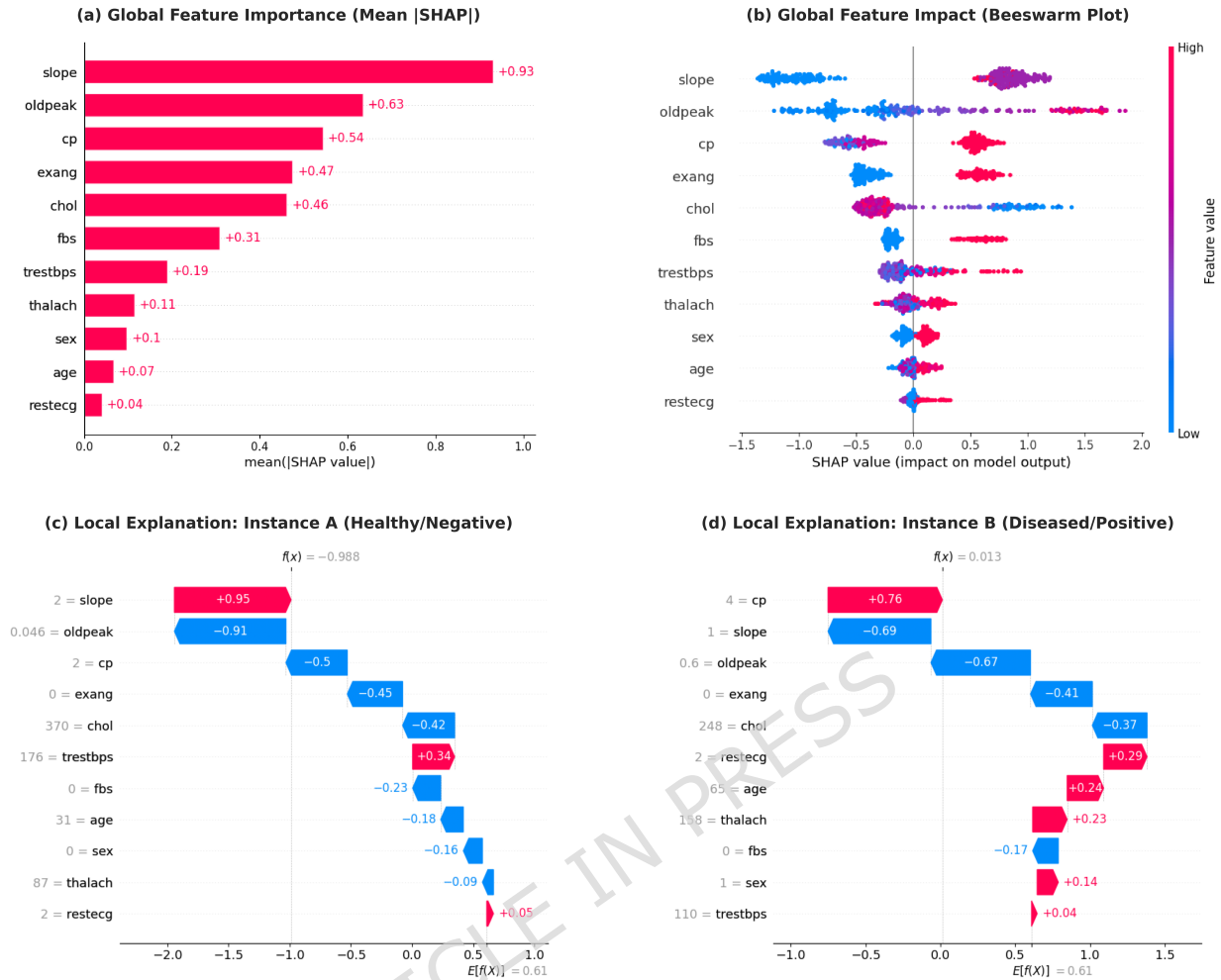


Figure 7: SHAP analysis for the XGBoost clinical model. (a) shows global feature importance, (b) illustrates the global feature impact (beeswarm plot) on model outcome, and the waterfall plot in (c) and (d) provides a local explanation for a specific instance predicted as healthy and diseased, respectively. These visualizations highlight the contribution of each feature to the prediction.

431 Figure 7(d) visualizes that the model's base value, or average prediction across the dataset,
 432 was 0.61, and the prediction for this specific instance rose to approximately 0.623, crossing
 433 the classification threshold. The strongest contributor to the positive prediction was the chest
 434 pain type ($cp = 4$), typically associated with asymptomatic or severe angina, contributing
 435 $+0.76$ to the final score. Additional risk factors included abnormal resting ECG results
 436 ($restecg = 2$), older age (65), male sex, and a high maximum heart rate ($thalach = 158$).
 437 Despite several features that reduced the risk, such as a flat ST slope ($slope = 1$), low ST
 438 depression ($oldpeak = 0.6$), absence of exercise-induced angina ($exang = 0$), and near-normal
 439 cholesterol levels- the cumulative weight of high-risk indicators led the model to classify the
 440 patient as having heart disease. This example highlights the model's ability to weigh clinical
 441 features in a nuanced manner, with significant emphasis on symptom severity and ECG

442 abnormalities, aligning well with medical understanding. Similarly, Figure 7(c) shows similar
 443 interpretation for a healthy individual. Figure 7(a) reveals the global feature importance for
 444 the XGBoost CVD prediction model:

445 • **Top Influencers:**

- 446 – `slope` (ST-segment slope, +0.93) and `oldpeak` (ST depression, +0.63) are the
 447 strongest predictors, indicating ECG-related features dominate CVD risk.
- 448 – `cp` (chest pain type, +0.54) and `exang` (exercise-induced angina, +0.47) follow
 449 closely, highlighting the importance of clinical symptoms.

450 • **Moderate Contributors:**

- 451 – `chol` (cholesterol, +0.46) and `fbs` (fasting blood sugar, +0.31) show measurable
 452 but smaller impacts.

453 • **Minimal Influence:**

- 454 – Demographic factors (`age` +0.07, `sex` +0.10) and `restecg` (resting ECG, +0.04)
 455 have relatively low importance.

456 The model prioritizes direct cardiac indicators over traditional risk factors, with ECG-related
 457 features (`slope`, `oldpeak`) showing nearly twice the impact of cholesterol levels. This aligns
 458 with clinical intuition that direct physiological markers are more predictive for CVD diagnosis
 459 than demographic variables.

460
 461 The local interpretation and mean SHAP values of **non-clinical model (RF)** is shown in
 462 Figure 8.

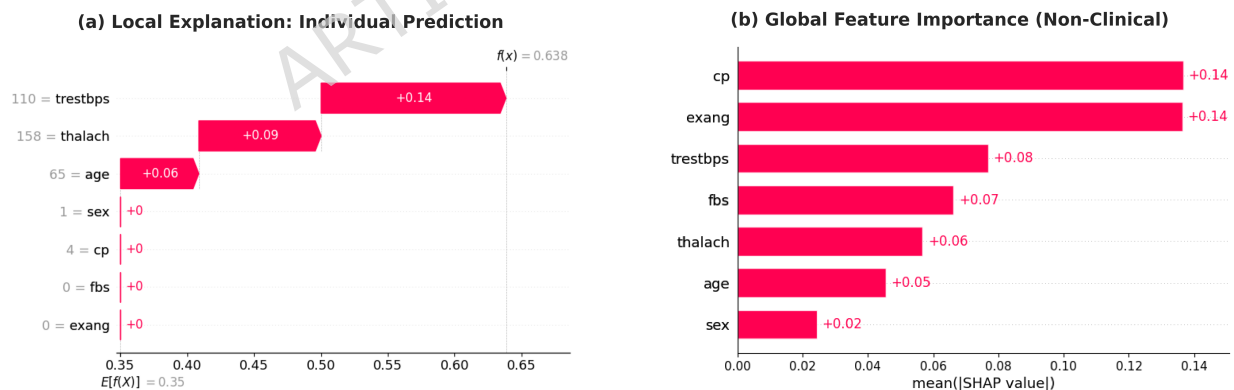


Figure 8: Interpretability analysis for the Non-Clinical Random Forest model. (a) Local waterfall plot explaining an individual prediction; basic health metrics like age and max heart rate (`thalach`) push the prediction toward disease, while the absence of exercise-induced angina (`exang`) lowers the risk. (b) Global feature importance (Mean |SHAP|) identifying chest pain type (`cp`) and exercise-induced angina (`exang`) as the most critical self-reported predictors for the early warning system.

463 The interpretation method for these plots are similar as XGBoost. Figure 8 shows that the
 464 individual's age, max heart rate, and resting blood pressure are the major factors behind
 465 getting diagnosed as CVD.

466

467 Figure 9(a) presents the global feature importance derived from TabNet, highlighting the
 468 overall contribution of each variable to the model's predictions. Figure 9(d) corresponds to a
 469 correctly predicted healthy individual, while Figure 9(c) represents a diseased case. These
 470 interpretations are made possible through TabNet's intrinsic sparse attention mechanism,
 471 which enables the model to selectively focus on the most relevant features for each individual
 472 prediction. TabNet also supports SHAP interpretation, as shown in Figure 9(b).

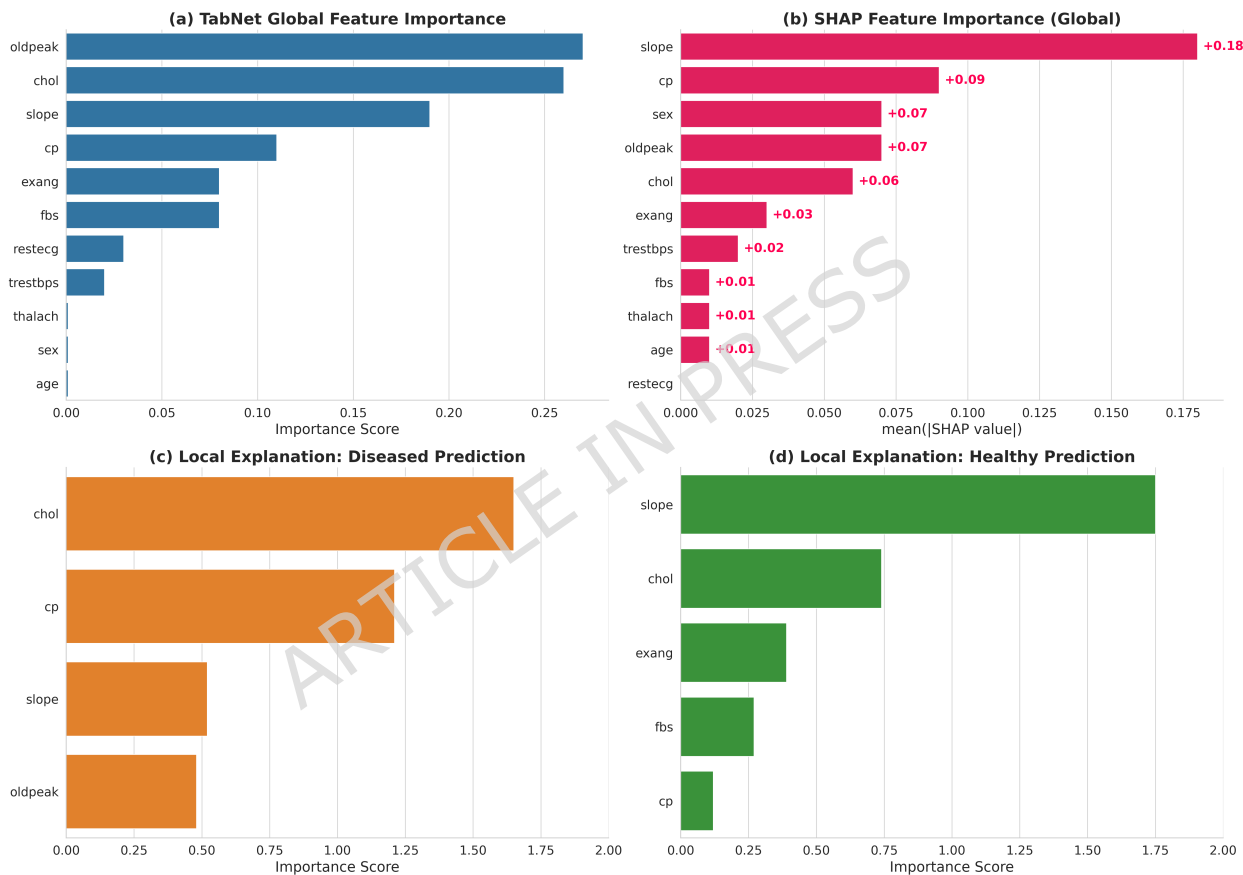


Figure 9: Interpretability of the TabNet model. (a) shows global feature importance from TabNet's intrinsic sparse attention mechanism. (c) and (d) provide local explanations for a diseased and a healthy prediction, respectively. (b) shows a SHAP summary plot for TabNet.

473 4.3 Counterfactual Analysis

474 Counterfactual analysis was used to explore how slight changes in input features could
 475 alter the model's prediction from **diseased** to **healthy**. By identifying minimal, plausible
 476 modifications such as reducing chest pain type or increasing ST slope, actionable insights are

477 revealed that may inform personalized interventions. This approach enhances interpretability
 478 by showing not only why a prediction was made, but also how it could change.

Table 4: **Counterfactual explanations generated using DiCE.** The table compares the original high-risk patient record (Age 38, Male, Resting BP 110) against generated counterfactuals. **Bold values** indicate the specific changes required to flip the prediction from Disease to Healthy. The values in parentheses in the Prediction column represent the model’s predicted probability of cardiovascular disease. Note that Age and Sex were held constant to ensure physiological plausibility. (*N/A indicates feature not used in that model*).

Model	Scenario	Key Features					Prediction
		CP Type	Ex. Angina	ST Slope	ST Depr.	Max HR	
Clinical	Original Patient	4	Yes (1)	3	1.5	105	Disease (0.91)
	Counterfactual 1	1	No (0)	3	1.5	105	Healthy (0.38)
	Counterfactual 2	4	Yes (1)	1	0.0	105	Healthy (0.42)
Non-Clinical	Original Patient	4	Yes (1)	N/A	N/A	105	Disease (0.86)
	Counterfactual 1	2	No (0)	N/A	N/A	105	Healthy (0.45)
	Counterfactual 2	4	No (0)	N/A	N/A	183	Healthy (0.41)

Abbreviations: CP = Chest Pain, ST Depr. = ST Depression (oldpeak), Max HR = Maximum Heart Rate (thalach).

479 Table 4 shows that, for the clinical model, two distinct pathways to a healthy prediction were
 480 identified. The values specified in parentheses correspond to the predicted probability of
 481 disease; a probability below 0.5 shifts the classification from ‘Disease’ to ‘Healthy’. The first
 482 pathway suggests a reduction in symptom severity—specifically, shifting chest pain present-
 483 ation from asymptomatic to typical angina (cp: 4 → 1) and eliminating exercise-induced
 484 angina (**exang**: 1 → 0). The second pathway highlights the model’s reliance on physiological
 485 markers; normalizing the ST slope (slope: 3 → 1) and ST depression (oldpeak: 1.5 → 0.0)
 486 alone was sufficient to flip the prediction, even if other risk factors remained constant.

487
 488 For the non-clinical Model, which relies on self-reported and basic health data, the counterfac-
 489 tuals provide actionable lifestyle targets. As shown in the bottom section of Table 4, increasing
 490 cardiovascular fitness—simulated by raising the maximum heart rate (**thalach**) from 105 to
 491 183, or strictly eliminating exercise-induced angina resulted in a healthy classification.

492 4.4 Neural Networks Performance, Interpretation, and Uncertainty

493 This section focuses on the performance of neural network architectures.

Table 5: Performance comparison of advanced neural models and the Causal-Guided Ensemble Voting Classifier (CGEVC) on the cardiovascular disease prediction task. TabPFN achieved the highest accuracy and F1-scores, while CGEVC maintained strong performance with reduced feature dimensionality.

Model	Accuracy	Macro F1-Score	Weighted F1-Score
TabNet	0.89	0.88	0.89
FT-Transformer	0.89	0.88	0.89
TabPFN	0.90	0.90	0.90
BNN	0.87	0.86	0.87
CGEVC	0.89	0.88	0.89

494 Table 5 shows that among all the evaluated neural models, TabPFN outperforms others
 495 with the highest accuracy (90%), macro F1-score (0.90), and weighted F1-score (0.90). This
 496 consistency across all key metrics indicates its superior balance in handling both positive and
 497 negative classes effectively. The model’s performance reflects not only strong predictive ability
 498 but also generalizability across patient groups. CGEVC’s balanced performance similar to
 499 other models indicate that the causal factors were truly important features for CVD diagnosis.
 500 Also, due to filtering out features that have an ATE of less than 0.1, CGEVC successfully
 501 reduced the dimensionality while preserving diagnosis performance.

502 4.4.1 Uncertainty Quantification

503 The outer histogram on Figure 10(a) illustrates the uncertainty distribution (measured as
 504 the standard deviation of predictive probabilities) for both CVD and non-CVD predictions
 505 using the Bayesian Neural Network.

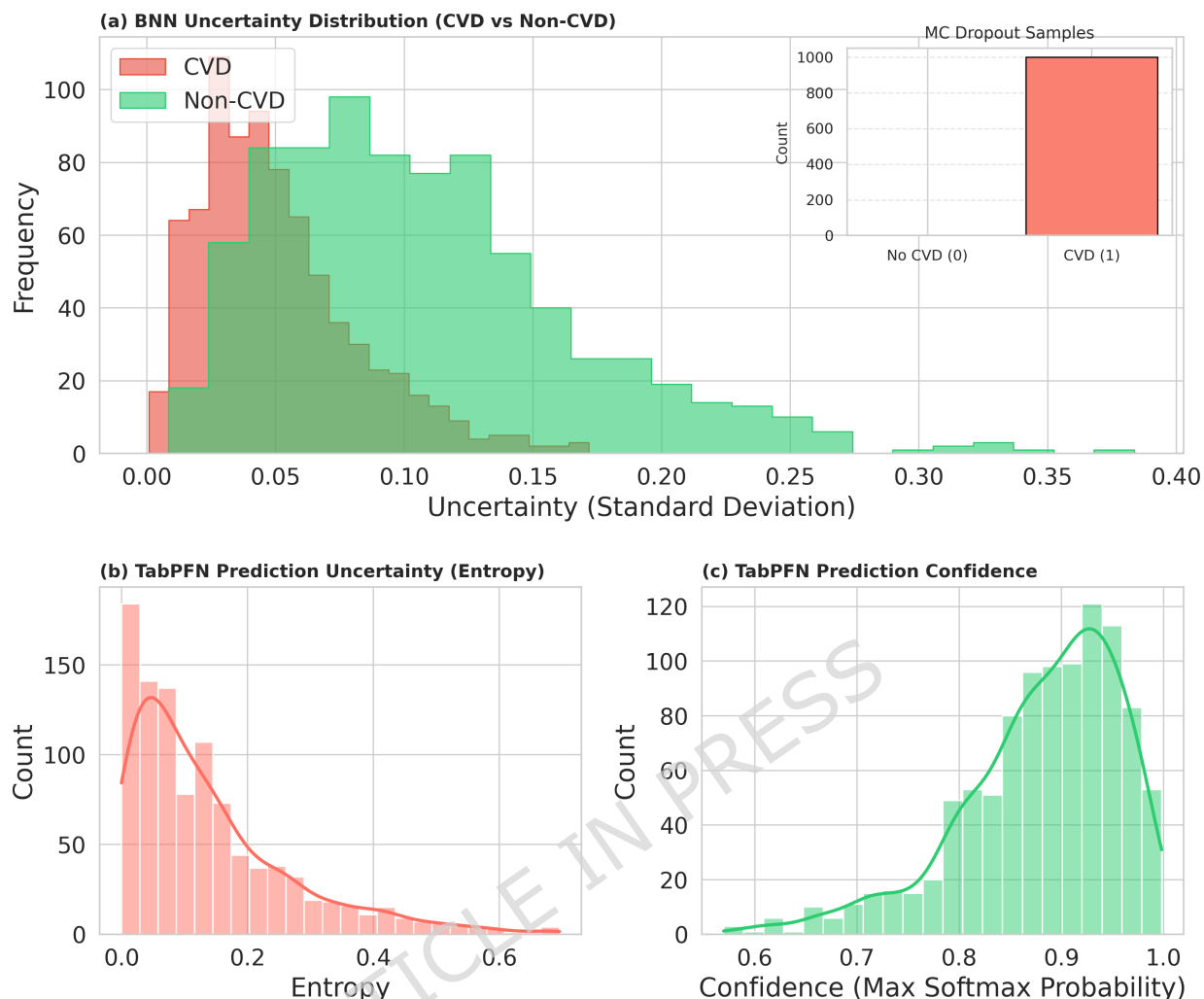


Figure 10: Uncertainty quantification for the Bayesian Neural Network and TabPFN. (a) shows the BNN’s uncertainty distribution (outer histogram) and prediction characteristics for a random instance (inner bar chart). (b) presents TabPFN’s predictive uncertainty via entropy and (c) shows confidence distributions.

506 Figure 10(a) visualizes in its outer histogram that, the majority of predictions, for both
 507 classes, exhibit low uncertainty, concentrated between 0 and 0.1, indicating the model was
 508 generally confident in its predictions. However, CVD predictions (in red) are slightly more
 509 concentrated in the very low uncertainty range (around 0.03–0.08), whereas non-CVD predic-
 510 tions (in green) display a slightly wider spread, with some extending up to 0.35. This may
 511 suggest the model is more certain when predicting diseased individuals than when ruling out
 512 disease. Nonetheless, both distributions taper off beyond 0.15, showing that high-uncertainty
 513 predictions are infrequent, which is a desirable trait for clinical reliability.

514

515 The inner bar chart of Figure 10(a) illustrates the prediction characteristics of the Bayesian
 516 Neural Network (BNN). Owing to its probabilistic nature, the BNN samples multiple weight

517 configurations during inference, resulting in a distribution of predictions for each instance.
 518 A sample is classified as class 1 (diseased) if the mean of these predictions exceeds 0.5. A
 519 prediction count concentrated near the total number of Monte Carlo (MC) samples indicates
 520 high confidence, while broader distributions reflect greater uncertainty. This approach offers a
 521 principled and robust quantification of uncertainty, a distinctive capability of Bayesian models.

522
 523 The uncertainty analysis for TabPFN is presented through two key metrics: entropy and
 524 prediction confidence. In the entropy histogram shown in Figure 10(b), we observe that
 525 most predictions are concentrated at very low entropy values (close to 0), indicating that
 526 the model often outputs highly confident and sharp class probabilities. However, there's a
 527 notable tail extending up to 0.7, suggesting that a minority of predictions remain highly
 528 uncertain, potentially due to ambiguous or borderline cases. The confidence histogram in
 529 Figure 10(c) complements this by showing that the majority of predictions have a softmax
 530 maximum probability above 0.90, further confirming the model's decisiveness on most samples.
 531 The sharp peak near 1.0 in the confidence plot supports the low-entropy pattern observed
 532 earlier. Together, these visualizations imply that TabPFN tends to be highly confident in its
 533 predictions, but retains the ability to flag uncertainty where necessary, which is essential for
 534 trustworthy AI in clinical settings.

535 4.5 Causal Inference

536 A clean summary and interpretation of causal inference results is provided in this section.

537 4.5.1 Estimated Causal Effects (Average Treatment Effect - ATE)

538 In the context of causal inference frameworks (such as potential outcomes), the term 'treatment'
 539 refers to the intervention variable or exposure whose effect is being estimated, rather than a
 540 clinical medical therapy. In this analysis, each feature—including immutable characteristics
 541 such as Age and Sex, is mathematically modeled as a 'treatment' variable. This allows us
 542 to estimate the ATE of that specific feature on the probability of CVD diagnosis, while
 543 controlling for relevant confounders.

544 1. Age

545
 546 - ATE: 0.0045

547 - Confounders: oldpeak, thalach, trestbps, chol

548 Interpretation: Age has a small but positive causal effect on heart disease. As age increases,
 549 the likelihood of disease slightly rises. However, the effect is relatively modest, likely due to
 550 its indirect interaction through other clinical variables. Statistically, on average, one unit
 551 change in age results in 0.45% point increase in the probability of heart disease, assuming
 552 other confounders are held constant.

554 2. Sex

555
 556 - ATE: 0.1274

557 - Confounders: oldpeak, thalach, trestbps, age

558 Interpretation: Being male is associated with a higher risk of heart disease, which aligns with
559 epidemiological findings. The effect is moderate, and appears independent of core vitals and
560 exercise-induced markers. ATE indicates that one unit change in Sex (practically, male to
561 female, or female to male) results in 12.74% point increase in the probability of heart disease,
562 assuming the confounders are held constant.

563

564 **3. Chest Pain Type (cp)**

565

566 - 0.1651

567 - Confounders: oldpeak, thalach, trestbps, chol, age

568 Interpretation: Chest pain type exerts a strong causal influence. This confirms its clinical
569 importance as a diagnostic feature, especially when chest pain is atypical or exertional.
570 Statistically, on average, a one-unit change in the cp (chest pain type) results in a 16.51%
571 point increase in the probability of heart disease, assuming other confounders are held constant.

572

573 **4. Resting Blood Pressure (trestbps)**

574

575 - ATE: 0.0011

576 - Confounders: oldpeak, thalach, chol, age

577 Interpretation: Despite its clinical relevance, resting blood pressure demonstrates a minimal
578 causal effect on heart disease in this analysis. The ATE value suggests that, on average, a
579 one-unit increase in resting blood pressure leads to a 0.11% point increase in the probability
580 of heart disease, assuming confounders are held constant. Its effect may be mediated or
581 overshadowed by stronger cardiovascular markers.

582

583 **5. Serum Cholesterol (chol)**

584

585 - ATE: -0.00089

586 - Confounders: oldpeak, thalach, trestbps, age

587 Interpretation: Surprisingly, cholesterol exhibits a very weak negative causal effect on heart
588 disease in this dataset. This could indicate that its influence is either non-linear, confounded,
589 or not directly contributing in the presence of other variables. A one-unit increase in
590 cholesterol is associated with a 0.089% point decrease in heart disease probability, assuming
591 confounders remain fixed.

592

593 **6. Fasting Blood Sugar (fbs)**

594

595 - ATE: 0.2566

596 - Confounders: thalach, trestbps, chol, age

597 Interpretation: Fasting blood sugar shows a strong positive causal relationship with heart
598 disease, highlighting the impact of glucose metabolism and diabetes on cardiovascular risk. On
599 average, a one-unit increase in fbs (indicating elevated blood sugar) increases the probability
600 of heart disease by 25.66% points, holding other factors constant.

601

602 **7. Resting ECG (restecg)**

603

604 - ATE: 0.0375

605 - Confounders: thalach, trestbps, chol, age

606 Interpretation: Resting ECG abnormalities contribute moderately to heart disease probability.

607 While not the most significant factor, it remains relevant in clinical contexts. A one-unit

608 change in ECG findings leads to a 3.75% point increase in the likelihood of disease, assuming

609 confounders are constant.

610

611 **8. Max Heart Rate Achieved (thalach)**

612

613 - ATE: -0.0023

614 - Confounders: oldpeak, trestbps, chol, age

615 Interpretation: Maximum heart rate achieved is negatively correlated with heart disease

616 risk, suggesting that better cardiovascular fitness may be protective. Statistically, a one-unit

617 increase in maximum heart rate results in a 0.23% point decrease in heart disease probability,

618 given fixed confounding variables.

619

620 **9. Exercise-Induced Angina (exang)**

621

622 - ATE: 0.3598

623 - Confounders: oldpeak, thalach, trestbps, chol, age

624 Interpretation: This is one of the most influential features. Patients experiencing angina

625 during exercise are significantly more likely to develop heart disease. The ATE suggests

626 that a one-unit increase in exercise induced angina leads to a 35.98% point increase in heart

627 disease probability, when other confounders are controlled.

628

629 **10. ST Depression Induced by Exercise (oldpeak)**

630

631 - ATE: 0.0925

632 - Confounders: thalach, slope, trestbps, chol, age

633 Interpretation: ST depression during stress testing moderately increases the risk of heart

634 disease. On average, each unit increase in oldpeak corresponds to a 9.25% point rise in heart

635 disease probability, assuming other variables are fixed.

636

637 **11. ST Slope (slope)**

638

639 - ATE: 0.3260

640 - Confounders: oldpeak, thalach, trestbps, chol, age

641 Interpretation: The slope of the ST segment is a strong causal indicator of heart disease,

642 particularly in exercise ECG tests. A one-unit increase in the slope of the ST segment leads

643 to a 32.60% point increase in disease probability, controlling for related confounders.

644

645 Overall, exercise-induced angina (exang), ST slope, fasting blood sugar, and chest pain

646 type demonstrate the strongest causal effects. Variables traditionally considered important,

647 like cholesterol or resting BP may not have strong causal impact in this dataset, potentially

648 due to mediation or correlation with stronger signals. This causal analysis not only supports
649 known clinical relationships but also justifies the feature selection used in the voting classifier.

650 **5 Conclusion**

651 This research presents a comprehensive and ethically aligned AI ecosystem for cardiovascular
652 disease diagnosis to address critical gaps in early detection, clinical applicability, and Re-
653 sponsible AI practices. By integrating advanced tabular neural networks, ensemble models,
654 Bayesian Neural Network, and causal model, robust diagnostic performance is achieved
655 with 0.90 accuracy while ensuring interpretability, uncertainty quantification, and bias mit-
656 igation. The work stands out by developing non-clinical models for accessible, early risk
657 assessment, enabling proactive interventions before symptoms manifest. The non-clinical
658 model is designed to be integrated into patient-facing mobile applications or web portals,
659 allowing individuals to assess their risk using self-reported data before seeking specialized
660 medical care. This serves as a filter to prioritize high-risk individuals in resource-constrained
661 healthcare systems. The aim was to enhance transparency, fairness, and trustworthiness
662 in clinical decision-making, by incorporating SHAP, DiCE counterfactuals, FairLearn, and
663 Bayesian uncertainty. Demographic and regional biases are addressed through unsupervised
664 clustering to ensure equitable performance across diverse populations. A novel ML-system
665 (CGEVC) guided by causal inference is also proposed for diagnosing purposes, obtaining an
666 accuracy and precision of 0.89, while reducing dataset dimensionality. Compared to prior
667 works that focus narrowly on accuracy, this approach aligns with FDA Good ML Practices
668 and EU AI ethics, offering a holistic solution for CVD diagnosis. Future directions include
669 expanding datasets for rare CVD subtypes, integrating multimodal data (e.g., ECG, imaging),
670 validating data augmentation procedures in clinical sectors, and refining causal inference for
671 personalized treatment insights. Future research should also focus on reducing disparities
672 in equalized odds while preserving overall model performance. By combining technological
673 innovation with ethical rigor, this research advances AI-driven healthcare toward safer, more
674 equitable, and actionable outcomes.

675 **Declarations**

676 **Ethical Approval**

677 Not applicable.

678 **Funding**

679 The study did not receive any funding.

680 **Data availability statement**

681 Data used in this research, materials, and reproducibility guidelines are provided in the GitHub
682 repository: <https://github.com/SakibHasanSimanto/CVD-AI-Research>. Original source of

683 data is UCI Machine Learning Repository: <https://archive.ics.uci.edu/dataset/45/heart+disease>

684 References

- 685 [1] Centers for Disease Control and Prevention. (2025). *National Center for Health Statistics*
686 *mortality data on CDC WONDER*. CDC WONDER. <https://wonder.cdc.gov/mcd.html>
- 687 [2] Martin, S. S., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L.,
688 Baker-Smith, C. M., Barone Gibbs, B., Beaton, A. Z., Boehme, A. K., et al. (2024). 2024
689 heart disease and stroke statistics: A report of US and global data from the American
690 Heart Association. *Circulation*, *149*(8). <https://doi.org/10.1161/CIR.0000000000001209>
- 691 [3] Centers for Disease Control and Prevention. (2024). *Heart disease facts*.
692 <https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html>
- 693 [4] Cleveland Clinic. (2022). *Cardiovascular disease*. [https://my.clevelandclinic.org/health/diseases/21493-](https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease)
694 [cardiovascular-disease](https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease)
- 695 [5] World Health Organization. (n.d.). *Cardiovascular diseases*. [https://www.who.int/health-](https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_2)
696 [topics/cardiovascular-diseases#tab=tab_2](https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_2)
- 697 [6] Karatza, A. A., Fouzas, S., Gkentzi, D., Kostopoulou, E., Loukopoulou, C., Dimitriou,
698 G., & Sinopidis, X. (2025). Missed or delayed diagnosis of heart disease by the general
699 pediatrician. *Children*, *12*(3), 366. <https://doi.org/10.3390/children12030366>
- 700 [7] Mandal, A. (2023). *Cardiovascular disease diagnosis*. News-Medical.net.
701 <https://www.news-medical.net/health/Cardiovascular-Disease-Diagnosis.aspx>
- 702 [8] Qadri, A. M., Raza, A., Munir, K., & Almutairi, M. S. (2023). Effective feature engi-
703 neering technique for heart disease prediction with machine learning. *IEEE Access*, *11*,
704 56214–56223. <https://doi.org/10.1109/ACCESS.2023.3281484>
- 705 [9] Al-Alshaikh, H. A., Prabu, P., Poonia, R. C., Saudagar, A. K. J., Yadav, M., Al-
706 Sagri, H. S., & AlSanad, A. A. (2024). Comprehensive evaluation and performance
707 analysis of machine learning in heart disease prediction. *Scientific Reports*, *14*, 7819.
708 <https://doi.org/10.1038/s41598-024-58489-7>
- 709 [10] Almazroi, A. A., Aldhahri, E. A., Bashir, S., & Ashfaq, S. (2023). A clinical decision
710 support system for heart disease prediction using deep learning. *IEEE Access*, *11*,
711 61646–61659. <https://doi.org/10.1109/ACCESS.2023.3285247>
- 712 [11] Subramani, S., Varshney, N., Vijay Anand, M., Soudagar, M. E. M. S., Al-keridis, L. A.,
713 Upadhyay, T. K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K., & Rohini,
714 K. (2023). Cardiovascular diseases prediction by machine learning incorporation with deep
715 learning. *Frontiers in Medicine*, *10*, 1150933. <https://doi.org/10.3389/fmed.2023.1150933>

- 716 [12] Saeed, M. H., & Hama, J. I. (2023). Cardiac disease prediction using AI algorithms
717 with SelectKBest. *Medical & Biological Engineering & Computing*, *61*, 3397–3408.
718 <https://doi.org/10.1007/s11517-023-02918-8>
- 719 [13] Eleyan, A., AlBoghbaish, E., AlShatti, A., AlSultan, A., & AlDarbi, D. (2024).
720 RHYTHMI: A deep learning-based mobile ECG device for heart disease prediction.
721 *Applied System Innovation*, *7*(5), 77. <https://doi.org/10.3390/asi7050077>
- 722 [14] Rohan, D., Pradeep Reddy, G., Pavan Kumar, Y. V., Purna Prakash, K., & Pradeep
723 Reddy, Ch. (2025). RHYTHMI: An extensive experimental analysis for heart dis-
724 ease prediction using artificial intelligence techniques. *Scientific Reports*, *15*, 6132.
725 <https://doi.org/10.1038/s41598-025-90530-1>
- 726 [15] Kumar, A., Singh, K. U., & Kumar, M. (2023). A clinical data analysis based diagnostic
727 system for heart disease prediction using ensemble method. *Big Data Mining and*
728 *Analytics*, *6*(4), 513–525. <https://doi.org/10.26599/BDMA.2022.9020052>
- 729 [16] Nandy, S., Adhikari, M., Balasubramanian, V., Menon, V. G., Li, X., & Za-
730 karyya, M. (2023). An intelligent heart disease prediction system based on swarm-
731 artificial neural network. *Neural Computing and Applications*, *35*, 14723–14737.
732 <https://doi.org/10.1007/s00521-021-06124-1>
- 733 [17] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *arXiv*
734 *preprint arXiv:1603.02754*. <https://arxiv.org/abs/1603.02754>
- 735 [18] Arık, S. Ö., & Pfister, T. (2020). TabNet: Attentive interpretable tabular learning. *arXiv*
736 *preprint arXiv:1908.07442*. <https://arxiv.org/abs/1908.07442>
- 737 [19] Hollmann, N., Müller, S., Eggenberger, K., & Hutter, F. (2023). TABPFN: A trans-
738 former that solves small tabular classification problems in a second. *arXiv preprint*
739 *arXiv:2207.01848*. <https://arxiv.org/abs/2207.01848>
- 740 [20] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revis-
741 iting deep learning models for tabular data. *arXiv preprint arXiv:2106.11959*.
742 <https://arxiv.org/abs/2106.11959>
- 743 [21] Goan, E., & Fookes, C. (2020). Bayesian neural networks: An introduction and survey.
744 *arXiv preprint arXiv:2006.12024*. <https://arxiv.org/abs/2006.12024>
- 745 [22] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease pre-
746 diction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542–81554.
747 <https://doi.org/10.1109/ACCESS.2019.2923707>
- 748 [23] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A.
749 (2021). Heart disease prediction using supervised machine learning algorithms: Per-
750 formance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672.
751 <https://doi.org/10.1016/j.combiomed.2021.104672>

- 752 [24] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart
753 disease prediction using machine learning techniques. *Algorithms*, 16(2), 88.
754 <https://doi.org/10.3390/a16020088>
- 755 [25] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019).
756 Modeling tabular data using conditional GAN. *arXiv preprint arXiv:1907.00503*.
757 <https://arxiv.org/pdf/1907.00503>
- 758 [26] U.S. Equal Employment Opportunity Commission, U.S. Department of Labor, U.S.
759 Department of Justice, & U.S. Civil Service Commission. (1978). *Uniform guidelines*
760 *on employee selection procedures*. [https://www.eeoc.gov/laws/guidance/questions-and-](https://www.eeoc.gov/laws/guidance/questions-and-answers-about-uniform-guidelines-employee-selection-procedures)
761 [answers-about-uniform-guidelines-employee-selection-procedures](https://www.eeoc.gov/laws/guidance/questions-and-answers-about-uniform-guidelines-employee-selection-procedures)
- 762 [27] Hasan, K. S. (2025). *CVD-AI-Research: A diverse ML ecosys-*
763 *tem for cardiovascular disease diagnosis* [Computer software]. GitHub.
764 <https://github.com/SakibHasanSimanto/CVD-AI-Research>
- 765 [28] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). *Heart Disease* [Dataset].
766 UCI Machine Learning Repository. <https://doi.org/10.24432/C52P4X>
- 767 [29] Mehdi, R. R., Kumar, M., Mendiola, E. A., Sadayappan, S., & Avazmohammadi, R.
768 (2023). Machine learning-based classification of cardiac relaxation impairment using
769 sarcomere length and intracellular calcium transients. *Computers in biology and medicine*,
770 163, 107134. <https://doi.org/10.1016/j.combiomed.2023.107134>
- 771 [30] Mehdi, R. R., Kadivar, N., Mukherjee, T., Mendiola, E. A., Bersali, A., Shah, D.
772 J., ... & Avazmohammadi, R. (2025). Non-Invasive Diagnosis of Chronic Myocardial
773 Infarction via Composite In-Silico-Human Data Learning. *Advanced Science*, e06933.
774 <https://doi.org/10.1002/advs.202406933>
- 775 [31] Mehdi, R. R., Mendiola, E. A., Sears, A., Choudhary, G., Ohayon, J., Pettigrew, R., &
776 Avazmohammadi, R. (2023). Comparison of three machine learning methods to estimate
777 myocardial stiffness. In *Reduced Order Models for the Biomechanics of Living Organs*
778 (pp. 363-382). Academic Press. <https://doi.org/10.1016/B978-0-32-389967-3.00025-1>