Advancing Cardiovascular Disease Diagnosis: A Robust ML Ecosystem Integrating Early Detection, Responsible AI Framework, and Causal Inference

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, we developed a machine learning (ML) ecosystem 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. Our framework leverages advanced ML techniques, including tabular neural networks (TabNet, TabPFN) and ensemble methods (XGBoost, Random Forest), validated on multi-regional datasets. 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, we incorporated interpretability tools (SHAP, counterfactuals), fairness mitigation (FairLearn), and uncertainty quantification (Bayesian Neural Networks). Causal inference identified key predictors 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. Notably, our analysis revealed limitations of synthetic data augmentation in clinical contexts. The system aligns with FDA Good ML Practices and EU Trustworthy AI guidelines, offering a scalable solution for early detection and equitable diagnosis.

1 Introduction

Cardiovascular Disease (CVD) is one of the most terrifying diseases that takes the lives of one person in 33 seconds globally. It is also the main cause of death for men, women, and people of most racial

and ethnic groups [1]. Furthermore, approximately 702880 people died due to CVD in 2022, which is equivalent to 1 in every 5 deaths [2]. In the same year, among the deaths of adults aging less than 65 years, about 1 out of 5 deaths was caused by CVD [1]. It is found that about 5% of adults above age 20 have Coronary artery disease. CDC reports also insight that approximately 1 in 5 heart attacks are silent, where the patient is unaware of the damage [3]. Since CVD can be both symptomatic and asymptomatic, often the disease is gone undiagnosed, increasing the chances of fatality. These diseases usually affect heart and blood vessels that include narrowing of the blood vessels in the heart or other organs throughout the body, heart and blood vessel problems since birth, and malfunction of heart valves. Deaths caused by CVD often vary by sex, race, and ethnicity, pointing to the potential lifestyle factors that might fuel the disease [4]. For example, smoking, unhealthy diet, alcoholism, and physical inactivity are lifestyle factors that contribute to CVD. However, medical conditions like high blood pressure, diabetes, and cholesterol also significantly contribute as the key risk factors behind CVD [3]. Surprisingly, WHO reports that environmental risk factors like air pollution can also lead to CVD, contributing to another hidden factor behind the disease [5]. Now, due to the diverse factors and often asymptomatic characteristics of CVD, it leads to delayed or misdiagnosis, which results in gradual disease progression which causes increased illness, higher healthcare costs, and death at worst. Heart diseases with lower prevalence often remain undiagnosed due to the unfamiliarity and lower level of suspicion of general pediatricians. It is also found that the misclassification of heart murmurs is also very common in developing countries. Organ dysfunction with cardiovascular disintegrity is associated with such delayed or misdiagnosis that contributes to augmented mortality and morbidity [6]. For these reasons, early detection and preventive medicine is essential to lessen the harms of CVD as it identifies the risk factors. If heart disease is detected in its very early stage, then precise medication and change in lifestyle factors leading to risk factors would help to mitigate it permanently. It also prevents the need for more expensive and invasive treatments later in the future. Furthermore, early diagnosis assists in delayed or misdiagnosis cases, as it does not let the disease grow internally. In essence, early diagnosis enables proactive measures to identify and manage risk factors, in order to mitigate the risks of CVD at the beginning phase.

Before focusing on early diagnosis, it is essential to understand how traditional diagnosis deals with the disease. Traditional diagnostic workflows in cardiovascular care are largely built around well-established clinical protocols and tools where physicians gather information through a combination of physical exams, patient history, and clinical testing including electrocardiograms (ECGs), blood pressure readings, cholesterol levels, stress tests, and in advanced cases, invasive imaging techniques like angiography or fluoroscopy. These procedures provide vital insights, but they often involve costly equipment, trained personnel, and time-consuming processes. Furthermore, they are typically reactive in nature, meaning they are deployed once symptoms become evident [7]. This leaves a gap in early detection, especially for patients with atypical or asymptomatic presentations. Traditional approaches also follow generalized risk models that may not fully reflect individual differences in lifestyle, genetics, or environmental exposure, leading to a lack of personalization in diagnosis and treatment planning. To tackle these, machine learning (ML) offers a transformative upgrade to this model by enabling data-driven, personalized, and scalable diagnostic support. Trained on large datasets such as the UCI Heart Disease dataset, ML algorithms can analyze complex, nonlinear relationships across numerous patient variables- many of which might be overlooked in standard clinical assessments. ML models can not only enhance the accuracy of diagnosis but also assist in predicting disease risk before traditional symptoms manifest, particularly when retrained using non-invasive or self-reported data, which is a core focus in our research as well. Moreover, ML systems can integrate inputs from wearables, lifestyle trackers, and electronic health records, enabling continuous monitoring and timely risk alerts. This shift opens the door to low-cost, remote, and

proactive care delivery, especially valuable in underserved regions. While ML is not a replacement for clinical judgment, it serves as a powerful augmentation, helping clinicians prioritize high-risk patients, reduce diagnostic delays, and tailor interventions more precisely than ever before.

The utilization of such learning algorithms for diagnosing CVD increases the rise of AI in healthcare diagnosis, both for early and emergency diagnosis. A brief chronological research trend on this domain insights a lot about the advancements. Mohan et al. (2019) [22] shows a method of extracting important features behind CVD using Decision Tree (DT) entropy and achieved a prediction accuracy of 0.887 using hybrid Random Forest with Linear Regression. Similar research conducted by Ali et al. (2021) [23] utilized Decision Tree (DT) and Random Forest Classifier (RFC) alongside their feature importance to predict heart disease, and achieved an accuracy of 1. While it potentially refers to overfitting, the performance cannot still be ignored. Another paper by Bhatt et al. (2023) [24] presents a relatively complex set of models alongside DT and RF in diagnosing CVD including XGBoost, MLP with GridSearch hyperparameter tuning. Huang clustering improves the prediction accuracy, and on average the models reached a cross-validation accuracy of 0.8707. A decent work is shown by Al-Alshaikh et al. (2024) [9]. The researchers proposed a robust CVD prediction system utilizing Genetic Algorithm and Recursive Feature Elimination to select best features, USCOM to address data imbalance, and multi-layered deep convolutional neural network trained with adaptive elephant herd optimization. Undoubtedly, all these works published in top journals- from simple classifiers to deep neural networks, show immense technological improvement throughout the years, yet they overlooked some critical issues. To begin with, early warning systems are rarely found that are required for preventive measures, with minimal work on early diagnosis. Little attention has been paid to models that can make accurate predictions based solely on non-clinical features, which are self-reported or easily measurable at home, thus missing a crucial opportunity to democratize preventive diagnostics. Secondly, with the rise of AI, ethical concerns regarding AI utilization in healthcare are raised by regulatory agencies and political unions like FDA and European Union. Almost every research overlooks key pillars of ethical AI and Good Machine Learning Practice, which is crucial to ensure patients' better health outcome. These key pillars are associated with model interpretability, uncertainty, algorithmic bias mitigation, and external validation. Thirdly, almost all research in the literature lack causal inference. Causality is becoming increasingly important in recent machine learning (ML) research as it allows models to move beyond mere correlation and prediction to understand the underlying mechanisms of data to enable better generalization and explainability. Lastly, research in this domain is usually conducted with tabular data from UCI repository, yet there is no work with the newborn advanced tabular neural networks. Such models have high potentiality of balanced performance in tabular dataset, and often with inherent interpretability. These models are heavily underutilized in the healthcare domain, not only in CVD diagnosis. So, if these models are not applied, there is a huge chance we will lack technological advancements.

To fill in the gaps presented above, we attempted to develop a multi-faceted robust AI ecosystem for CVD diagnosis, combining technological advancements and ethical concerns. Our key contribution of the research is outlined below:

- Designing an early warning model based entirely on non-clinical features using interpretable and accessible methods.
- Developing specialized models using clinical and non-clinical data for diagnostic purposes, focusing on generalization across combined datasets.

- Integrating advanced tabular deep learning models (TabNet, TabPFN, FT-Transformer) and ensemble learners (XGBoost, LightGBM, Random Forest) with state-of-the-art tuning and augmentation techniques.
- Ensuring fairness and transparency through bias mitigation (FairLearn), local and global interpretability (SHAP, TabNet attention), and counterfactual analysis (DiCE).
- Quantifying model uncertainty using Bayesian Neural Networks and TabPFN entropy distribution, alongside interpreting performance under ambiguous inputs.
- Exploring causal inference for identifying truly impactful features and guiding the design of a causality-informed hybrid ensemble.

2 Related Work

Significant advancements have been made in applying machine learning (ML) and deep learning (DL) techniques to cardiovascular disease (CVD) prediction, with diverse approaches focusing on accuracy, feature selection, and model optimization. However, challenges remain in achieving clinically robust, interpretable, and ethically aligned diagnostic systems.

2.1 Traditional Machine Learning and Ensemble Models

Qadri et al. [8] introduced the Principal Component Heart Failure (PCHF) method for feature engineering, enhancing early-stage heart failure detection by reducing dimensionality to eight critical attributes. Their Decision Tree classifier achieved 100% accuracy, though concerns of overfitting arise due to the unusually high performance. Similarly, Kumar et al. [15] evaluated classical models such as Support Vector Machines (SVM) and logistic regression on the UCI Heart Disease dataset, concluding logistic regression as the most accurate. These works underscore the importance of tailored feature selection and classical models but lack discussion on broader ethical and generalization considerations.

Subramani et al. [11] proposed a stacking ensemble strategy for improved predictive resilience, integrating multiple classifiers with IoT-based data sources. This method demonstrated a prediction accuracy of 96%, highlighting the benefits of hybrid models. Likewise, Rohan et al. [14] conducted an exhaustive comparison of 21 classifiers and 11 feature selection methods, with XGBoost emerging as the top performer (F1-score: 98%). These ensemble methods excel in performance but often focus narrowly on accuracy, without addressing interpretability, fairness, or uncertainty.

2.2 Deep Learning Models and Feature Engineering

Deep neural networks (DNNs) have gained prominence for their ability to model complex patterns in healthcare data. Almazroi et al. [10] developed a deep diagnostic model using Keras-based dense neural networks, outperforming ensemble models in various metrics. Saeed et al. [12] applied DNNs with SelectKBest feature selection and SMOTE for balancing, achieving up to 99% accuracy across datasets. These studies affirm deep learning's potential but offer limited insights into transparency or clinical explainability.

Al-Alshaikh et al. [9] proposed a hybrid system combining genetic algorithms, feature elimination, and a multilayer deep CNN trained via adaptive elephant herd optimization. Their model achieved high performance (e.g., recall: 96.2%), illustrating the power of DL when paired with advanced

optimization and data preprocessing strategies. Yet, the study acknowledged the need for real-world validation and improved generalizability.

2.3 Innovative Neural and Bio-Inspired Techniques

Several works introduced novel neural architectures and biologically inspired models. Nandy et al. [16] proposed a Swarm-Artificial Neural Network (Swarm-ANN), which adapts neuron weights using heuristic strategies. It achieved 95.78% accuracy and outperformed standard learners. This method reflects growing interest in hybrid optimization strategies but lacks attention to interpretability and clinical integration.
Eleyan et al. [13] designed "Rhythmi," a mobile diagnostic tool using convolutional neural networks (CNNs) on ECG signals, achieving over 98% in various performance metrics. While the real-time mobile application adds practical value, its limited dataset calls for broader clinical testing. This effort shows the trend toward embedded diagnostic systems but remains detached from structured tabular data common in electronic health records (EHRs).

2.4 Current Gaps and Motivation for Our Work

Despite strong diagnostic performance in existing works, several key limitations persist:

- **Data Scope:** Most studies rely solely on clinical features, overlooking non-clinical data (e.g., lifestyle, behavioral) that could facilitate early warnings and wider accessibility.
- **Trustworthiness:** Ethical AI pillars such as fairness, interpretability, uncertainty quantification, and causal inference are largely neglected, limiting clinical trust and adoption.
- **Advanced Tabular Deep Models:** While tabular data is common in CVD prediction, state-of-the-art tabular deep learning models like TabNet, TabPFN, and FT-Transformer remain underexplored.
- **Validation and Reproducibility:** External validation and open reproducibility resources are often missing, raising concerns about real-world applicability.

To address these gaps, our study proposes a comprehensive AI-based framework that integrates both clinical and non-clinical features, enabling early and accurate CVD detection. We employ advanced tabular neural networks (TabNet, TabPFN, FT-Transformer) alongside classical ensemble models (XGBoost, RFC, LightGBM), with augmentation through CTGAN and Gaussian Copula and class balancing via SMOTE. Unsupervised clustering techniques such as PCA, K-Means, and agglomerative clustering are applied for regional generalization analysis, while causal inference with DoWhy identifies statistically significant drivers of heart disease. Our system integrates interpretability (via SHAP and sparse attention), counterfactual reasoning (DiCE), fairness evaluation (FairLearn), and uncertainty estimation (Bayesian neural networks and TabPFN entropy distribution), aligning with the European Union's Trustworthy AI guidelines and FDA's Good Machine Learning Practice (GMLP). This multifaceted approach aims to shift the paradigm from purely performance-driven research to clinically reliable, ethically grounded, and interpretable AI for cardiovascular healthcare.

3 Methodology

We developed several models for clinical and non-clinical usage. Three ensemble models are used for non-clinical modeling, and tabular neural networks, causality-guided voting classifier, Bayesian

Neural Network alongside the three ensemble models are developed for clinical modeling. We incorporated interpretability and uncertainty in our models as well. Actually, each model has its specialized strength, instead of being an all-rounder. Bayesian Neural Network and TabPFN provides uncertainty estimation, while ensemble models and TabNet provide interpretability, and the voting classifier itself is built on relatively significant (ATE ≥ 0.1) causal factors. Reproducibility guidelines are provided in the GitHub repository [27].

3.1 Dataset Information

We downloaded two datasets from Kaggle, however, the original source of the dataset is UCI [28]. One dataset is from Cleveland clinic (renewed), and the other one is combined with separate datasets from statslog, Switzerland, Long Beach VA, Hungary, and Cleveland. The dataset information is presented in Table 1.

Table 1: Description of UCI Heart Disease Dataset Columns

Column Name	**Definition**
`age`	Age of the patient (in years)
`sex`	Sex (1 = male; 0 = female)
`cp`	Chest pain type (0–3: typical angina to asymptomatic)
`trestbps`	Resting blood pressure (in mm Hg)
`chol`	Serum cholesterol (in mg/dl)
`fbs`	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
`restecg`	Resting electrocardiographic results (0–2)
`thalach`	Maximum heart rate achieved
`exang`	Exercise-induced angina (1 = yes; 0 = no)
`oldpeak`	ST depression induced by exercise relative to rest
`slope`	Slope of the peak exercise ST segment (0–2)
`ca`	Number of major vessels (0–3) colored by fluoroscopy
`thal`	Thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect)
`target`	Diagnosis of heart disease (0 = no disease; 1 = disease present)

The combined dataset consists of 1190 instances, and has 2 features less than Cleveland and Statslog. These features are `ca`, and `thal`. So, we deleted these 2 features from the Cleveland and Statslog dataset. Then, we renamed every column of Statslog and combined the dataset following the column names of Cleveland for easier understanding. It is found that 270 instances of Statslog get matched with the instances of Cleveland. Now, the combined dataset has 272 duplicated rows, whereas the total number of Statslog instances is also 272. It indicates a high probability that Statslog is a shuffled dataset consisting of instances from the combined dataset. Hence, we drop the duplicates from the combined data. Thus, Statslog data is now removed and our combined dataset now consists of 4 regional data on CVD. Now, the combined dataset has 918 instances.

3.2 Modeling on Cleveland Data

The combined dataset and Cleveland dataset is divided into four portions by dropping the clinical features. The features include serum cholesterol, resting electrocardiographic result, ST depression induced by exercise relative to rest, and the slope of peak exercise ST segment. By doing so, we obtained four datasets: Cleveland non-clinical data, Cleveland clinical data, combined non-clinical

data, and combined clinical data. Now, it is possible to train models on both non-clinical and clinical features. Models trained on non-clinical features can be used for early warning.

We then trained three ensemble models on the Cleveland dataset (non-clinical and clinical)- XGBoost, LGBoost, and Random Forest on the entire set, and then tested on the combined data. A decent performance, typically indicated by an accuracy over 0.80 and a ROC-AUC of 0.85 depicts good generalization to the global data. Because the Cleveland dataset itself has only 303 instances, where the combined dataset has 918 instances. We also did not tune the hyperparameters of the model, as it increases predictive power too. Hence, if the model trained on Cleveland renewed dataset with the baseline intelligence performs well on the large dataset, it becomes explicit that a model trained on the larger dataset with a 80:20 train test split and hyperparameter tuning would generalize very well to the global data, and this is what we did later on.

3.3 Modeling on Combined Data

Once the generalization is proved, we move further with the combined dataset to build our semi-final models. In this stage, we fixed the slight class imbalance of the dataset with Synthetic Minority Oversampling Technique (SMOTE). Figure 1 and Figure 2 shows the comparison of before and after applying SMOTE.

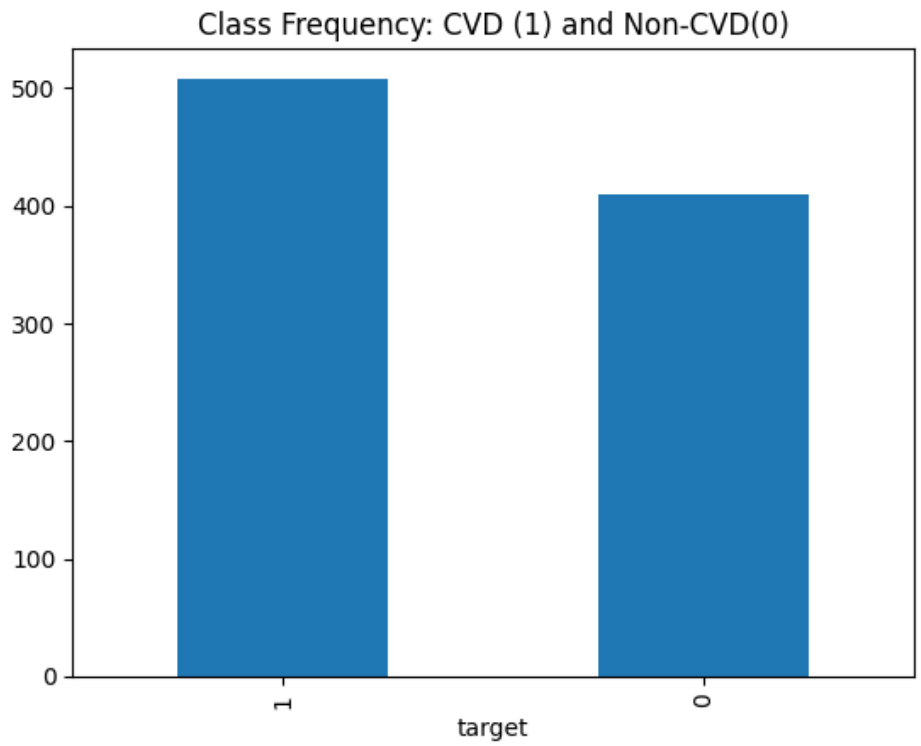

Figure 1: Class Frequency before applying SMOTE

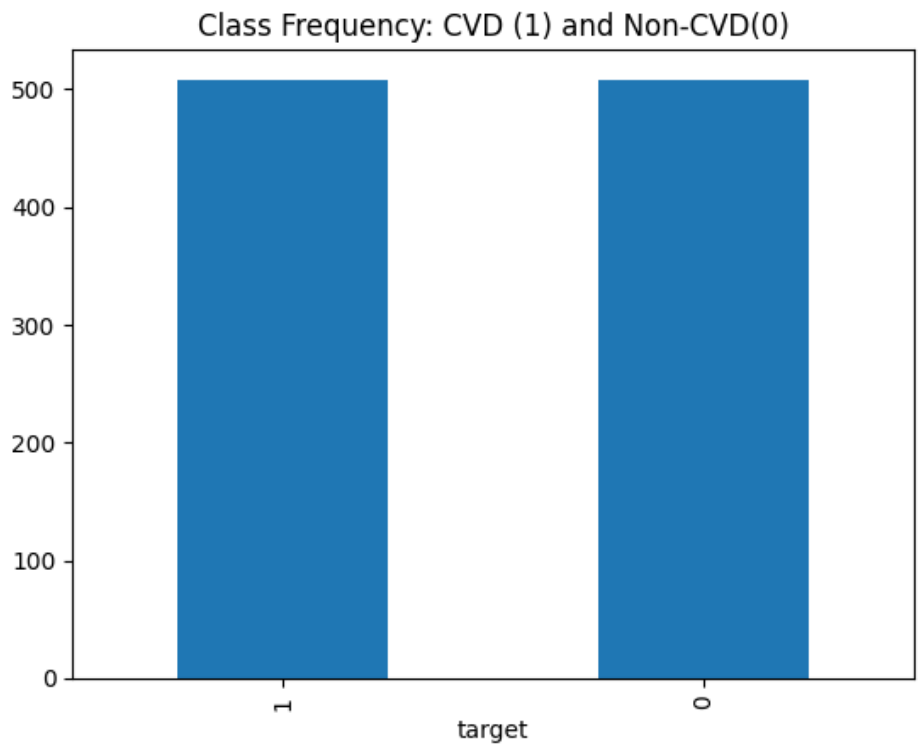

Figure 2: Class Frequency after applying SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic samples for the minority class to balance imbalanced datasets. Since generalization is proved already, it should not be an ethical concern that SMOTE may cause data leakage. Because instead of simple duplication, SMOTE creates new instances by interpolating between existing minority class samples. Given a minority class sample $\mathbf{x}_i$, a synthetic sample $\tilde{\mathbf{x}}$ is generated as:

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

where:

- $\mathbf{x}_i$ is a minority class instance,
- $\mathbf{x}_{zi}$ is one of the k nearest neighbors of $\mathbf{x}_i$,
- $\lambda \sim \mathcal{U}(0, 1)$ is a random number drawn from a uniform distribution.

This process effectively creates synthetic data points along the line segments joining minority class neighbors, thereby reducing overfitting and improving model generalization on imbalanced data. Note that, the oversampling process is repeated once in the combined dataset, and then the clinical features were dropped to separate the dataset into a clinical version, and into a non-clinical version. Then, again we run the three ensembles on the clinical dataset, and on the non-clinical dataset. While the performance was promising, from FairLearn's biasness report we noticed that the model is usually biased towards men, which means, most men are getting CVD as prediction, while most women are healthy. It might happen due to the natural pattern of CVD in both genders, that indicates the known fact that usually men suffer more with heart diseases than women. However, this still raises an ethical concern. To mitigate this, we performed another experiment with data

augmentation. We created several synthetic instances from the combined dataset using Conditional Tabular Generative Adversarial Network (CTGAN), so that the frequency of diseased men and diseased women is equalized, and so is the frequency of healthy men and healthy women. Again, CTGAN would not cause data leakage as well. It is a generative adversarial network tailored for tabular data, especially effective for imbalanced and mixed-type datasets (both continuous and categorical features). It uses a conditional generator to better capture the distribution of minority classes and rare categories.

Integrating a conditional generator into the architecture of a GAN requires to deal with three issues. Firstly, it is necessary to devise a representation for the condition as well as to prepare an input for it. Secondly, it is necessary for the generated rows to preserve the condition as it is given, and thirdly, it is necessary for the conditional generator to learn the real data conditional distribution, i.e.

$$\mathbb{P}_{\mathcal{G}}(\text{row} \mid D_{i*} = k^*) = \mathbb{P}(\text{row} \mid D_{i*} = k^*) \tag{2}$$

so that we can reconstruct the original distribution as

$$\mathbb{P}(\text{row}) = \sum_{k \in D_{i*}} \mathbb{P}_{\mathcal{G}}(\text{row} \mid D_{i*} = k^*)\, \mathbb{P}(D_{i*} = k) \tag{3}$$

To effectively model continuous columns, CTGAN uses mode-specific normalization via Variational Gaussian Mixture Models (VGM), and for discrete columns, it samples conditioning vectors $\mathbf{c}$ to guide generation, ensuring better representation of underrepresented categories. [25]

This conditional mechanism allows CTGAN to generate diverse and realistic synthetic tabular data, particularly for class-imbalanced scenarios. After applying CTGAN, we retrained the three ensemble models on the augmented dataset and found out that the bias metrics are balanced between the genders.

To entirely mitigate demographic bias, it was essential to quantify regional bias as well. Even though, in the final model, region will not be used as a feature because the users will obviously not represent the four regions only (Switzerland, Long Beach VA, Cleveland, and Hungary), rather it may be diverse. Still, we were trying to generalize the model over all the regions going beyond external validation only, that led us to extract theoretical regions from the dataset using unsupervised learning. We applied K-Means with Principal Component Analysis (PCA) to extract exactly four regions. With a silhouette score of 0.21, the unsupervised model extracted four theoretical regions, and surprisingly, the frequencies of regions in the processed dataset almost resembled the frequencies from the original dataset as shown in Figure 3.

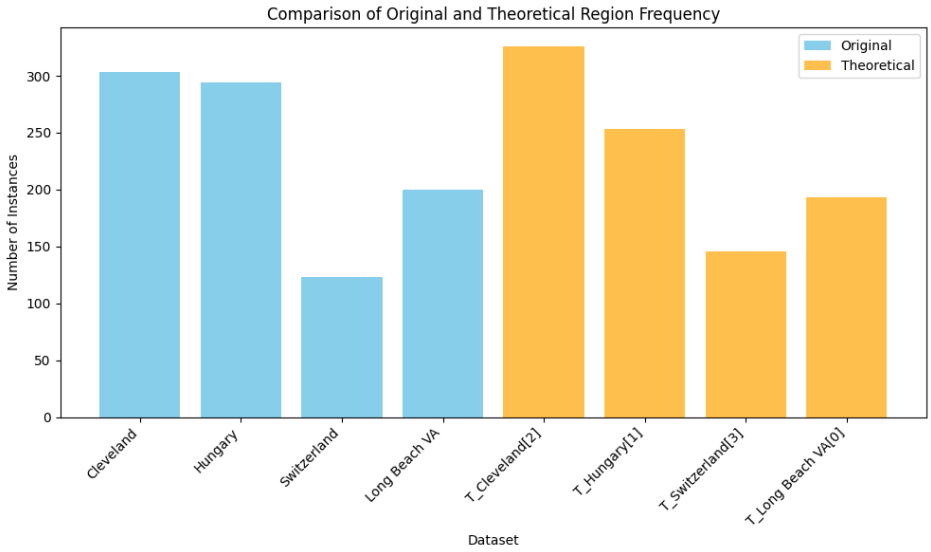

Figure 3: Comparison of Original and Theoretical Region Frequency

Then, we trained the three ensembles on this dataset with additional features (theoretical region), and checked the bias metrics. Interesting insights were found after that and it led to us for agglomerative clustering instead of K-Means PCA, which is presented in the results section.

After that, using the combined dataset with SMOTE and CTGAN (region excluded), we developed the three tabular neural network models- TabNet, FT-Transformer, and TabPFN. Bayesian Neural Network (BNN) with MC Dropout for uncertainty, and a causality guided voting classifier consisting of XGBoost, LGBoost, Random Forest, K-Nearest Neighbor, and Logistic Regression is also developed to push prediction accuracy. The dataset was scaled with scikit-learn's StandardScaler() before training the neural models, which basically performs z-score standardization on the features. The voting classifier is trained with the features (target) that have at least 0.1 Average Treatment Effect (ATE) on CVD diagnosis.

3.4 Model Selection

As mentioned before, we used three ensemble models, three tabular neural networks, and a Bayesian Neural Network. The non-clinical diagnosis system is only available with ensemble models. For clinical models, we incorporated the other models as well. The UCI Heart Disease dataset includes both categorical and numerical features, as well as non-linear interactions. Tree-based models like XGBoost, LightGBM, and Random Forest are well-suited for such structured data:

- **Mixed data types:** Handle categorical and continuous features.
- **Non-linearity:** Naturally capture non-linear patterns.
- **Feature selection:** Focus on informative splits.

XGBoost :
XGBoost minimizes the objective:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{4}$$

It builds trees sequentially, correcting prior errors, and uses L1/L2 regularization to reduce overfitting [17]. We tuned hyperparameters using Bayesian Optimization with 10-fold F1-score as the objective:

$$\max_{\theta \in \Theta} \frac{1}{K} \sum_{k=1}^{K} \mathrm{F1}(f_\theta^{(k)}(X_{\text{val}}^{(k)}), y_{\text{val}}^{(k)}) \tag{5}$$

Random Forest :
RF builds multiple trees on bootstrapped data and averages outputs:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{x}) \tag{6}$$

It reduces variance and overfitting, which is important in small/noisy medical datasets. Hyperparameters were tuned using Bayesian Optimization with F1-score.

LightGBM :
LightGBM introduces *GOSS* and *EFB* for fast training. GOSS focuses on high-gradient samples to accelerate learning. EFB bundles sparse features to reduce dimensionality:

$$\tilde{V}_j(d) = \frac{1}{n} \left(\frac{(\sum g_i)^2}{n_j^l} + \frac{(\sum g_i)^2}{n_j^r} \right) \tag{7}$$

This improves speed and efficiency on clinical data. Hyperparameters were optimized via Optuna with cross-validated F1-score.

These ensemble models are robust to feature scale, capture complex interactions, and resist overfitting. Hyperparameters were finely tuned using Bayesian Optimization for optimal performance on the UCI dataset.

3.4.1 Tabular Neural Networks

Neural models like TabNet, TabPFN, and FT-Transformer are designed for structured data with mixed feature types and non-linear interactions, as seen in the UCI Heart Disease dataset.

TabNet (Attention-based Selection) :
TabNet learns sparse, step-wise feature importance via sequential attention:

$$\mathbf{a}_t = \mathrm{Sparsemax}(P_t \cdot \mathbf{a}_{t-1}) \tag{8}$$

This makes it both interpretable and efficient [18]. Aggregate feature importance is computed by weighting masks at each step:

$$\mathbf{M}_{\text{agg-b},j} = \frac{\sum_{i=1}^{N_{\text{steps}}} \eta_{\mathbf{b}}[i]\, \mathbf{M}_{\mathbf{b},j}[i]}{\sum_{j=1}^{D} \sum_{i=1}^{N_{\text{steps}}} \eta_{\mathbf{b}}[i]\, \mathbf{M}_{\mathbf{b},j}[i]}, \quad \eta_{\mathbf{b}}[i] = \sum_{c=1}^{N_d} \mathrm{ReLU}(\mathbf{d}_{\mathbf{b},c}[i]) \tag{9}$$

We tuned TabNet using Bayesian Optimization over architectural and regularization parameters.

TabPFN (Bayesian Few-Shot Learner) :
TabPFN approximates the posterior predictive distribution:

$$p(y \mid x, D) \propto \int_{\Phi} p(y \mid x, \phi)\, p(D \mid \phi)\, p(\phi)\, d\phi \tag{10}$$

It excels on small datasets like UCI Heart Disease and requires no tuning, being pretrained for generalization [19].

FT-Transformer (Attention on Columns) :
FT-Transformer [20] uses feature tokenization and row-wise self-attention:

$$\begin{aligned} \boldsymbol{T}_j &= \boldsymbol{b}_j + f_j(x_j), \\ \boldsymbol{T}_0 &= \text{stack}([\text{CLS}], \boldsymbol{T}), \\ \boldsymbol{T}_i &= F_i(\boldsymbol{T}_{i-1}) \end{aligned} \tag{11}$$

The [CLS] token's final embedding is used for prediction:

$$\hat{y} = \texttt{Linear}(\texttt{ReLU}(\texttt{LayerNorm}(\boldsymbol{T}_L^{\texttt{[CLS]}}))) \tag{12}$$

We used default hyperparameter settings, as tuning on small datasets may lead to overfitting.

These three tabular neural models capture complex interactions and provide strong baselines for tabular healthcare data.

3.4.2 Bayesian Neural Network (BNN)

The UCI Heart Disease dataset is relatively small and involves high-stakes predictions, where understanding uncertainty is essential. Bayesian Neural Networks (BNNs) with Monte Carlo (MC) Dropout provide a principled way to estimate predictive uncertainty and prevent overfitting.

Bayesian Inference in Neural Networks :
A BNN learns a distribution over weights $\mathbf{W}$, instead of point estimates, capturing epistemic uncertainty [21]. Before training, we observe the joint distribution between weights and data:

$$p(\boldsymbol{\omega}, \mathcal{D}) = p(\boldsymbol{\omega}) p(\mathcal{D}|\boldsymbol{\omega}). \tag{13}$$

For a 1-D homoscedastic regression with MSE loss:

$$p(\mathcal{D}|\boldsymbol{\omega}) = \mathcal{N}\big(\mathbf{f}^{\boldsymbol{\omega}}(\mathcal{D}), \sigma^2\big). \tag{14}$$

Assuming i.i.d. samples:

$$p(\mathcal{D}|\boldsymbol{\omega}) = \prod_{i=1}^{N} \mathcal{N}\big(\mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x}_i), \sigma^2\big). \tag{15}$$

Bayes theorem gives the posterior:

$$\pi(\boldsymbol{\omega}|\mathcal{D}) = \frac{p(\boldsymbol{\omega}) p(\mathcal{D}|\boldsymbol{\omega})}{p(\mathcal{D})}, \quad \text{where} \quad p(\mathcal{D}) = \int p(\boldsymbol{\omega}) p(\mathcal{D}|\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{16}$$

Predictions are made via posterior expectation:

$$\mathbb{E}_{\pi}[f] = \int f(\boldsymbol{\omega}) \pi(\boldsymbol{\omega}|\mathcal{D}) d\boldsymbol{\omega}. \tag{17}$$

MC Dropout Approximation :
MC Dropout approximates Bayesian inference by performing multiple stochastic forward passes with dropout at test time:

$$\hat{p}(y \mid \mathbf{x}) \approx \frac{1}{T} \sum_{t=1}^{T} f^{(t)}(\mathbf{x}) \tag{18}$$

where $f^{(t)}(\mathbf{x})$ is the t-th forward pass with different dropout masks, and T is the number of passes.

There are several reasons behind why it would be a good fit for predicting CVD. They are mentioned below.

- **Relatively Smaller Dataset:** Regular dropout helps reduce overfitting; MC Dropout also quantifies uncertainty.
- **Clinical Use:** Epistemic uncertainty aids in identifying cases where the model is unsure and may need human oversight.
- **Interpretability:** Predictive variance can be visualized to support clinical decision-making.

BNNs with MC Dropout are a strong choice for the UCI Heart Disease dataset due to their robustness in low-data regimes and ability to provide trustworthy, uncertainty-aware predictions.

3.4.3 Causal-Guided Ensemble Voting Classifier (CGEVC)

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

1. Identifying Potential Confounders: To estimate the causal effect of each feature (treatment variable) on the target (CVD diagnosis), we first identify the potential confounders using a data-driven approach. For each feature T_i:

1. Remove the target variable Y from the dataset.
2. Use a Random Forest Classifier (RFC) to predict T_i using the remaining features $\mathbf{X} \setminus T_i$.
3. Extract the feature importances from RFC and select the subset $\mathbf{Z}_i \subset \mathbf{X} \setminus T_i$ such that the cumulative importance satisfies:

$$\sum_{x \in \mathbf{Z}_i} \text{Importance}(x) \geq 0.80 \tag{19}$$

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

2. Estimating Average Treatment Effect (ATE): We use the backdoor criterion from causal inference to estimate the Average Treatment Effect (ATE) of each treatment T_i on the outcome Y, conditioned on the identified confounders $\mathbf{Z}_i$. According to the backdoor adjustment formula:

$$\text{ATE}(T_i \to Y) = \mathbb{E}_{\mathbf{z} \sim \mathbf{Z}_i} \left[\mathbb{E}[Y \mid T_i = 1, \mathbf{Z}_i = \mathbf{z}] - \mathbb{E}[Y \mid T_i = 0, \mathbf{Z}_i = \mathbf{z}] \right] \tag{20}$$

We implement this using the `DoWhy` framework with the `backdoor.linear_regression` estimator, which assumes the linear model:

$$Y = \beta_0 + \beta_1 T_i + \boldsymbol{\gamma}^\top \mathbf{Z}_i + \epsilon \tag{21}$$

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

3. Causal-Guided Feature Selection and Ensemble Classification: We select features with statistically significant (0.1) ATE values and use them to train a soft voting ensemble classifier composed of:

- Random Forest Classifier (RFC)
- XGBoost
- LightGBM
- k-Nearest Neighbors (KNN)
- Logistic Regression (LR)

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

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{m=1}^{M} w_m \cdot \hat{p}_m(c \mid \mathbf{x}) \tag{22}$$

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

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

4 Results and Discussion

Performance metrics of XGBoost, LGBoost, and RF model trained on Cleveland (renewed) clinical dataset, tested on Combined dataset is given in Table 2. Accuracy, AUC score, and F1-score are reported with 95% CI.

Trained on Clinical Dataset (Cleveland renewed) and tested on Clinical Combined dataset

Table 2: Performance metrics of XGBoost, LGBoost, and RF models on clinical dataset

Model	**Accuracy**	**AUC**	**F1 Score**	**Precision**	**Recall**
XGBoost	0.8028	0.8858	0.8297	0.79	0.87
LGBoost	0.8159	0.8891	0.8425	0.80	0.89
RF	0.8292±0.0016	0.9102±0.0008	0.8499±0.0016	0.84	0.85

Random Forest comparatively performed better than the other ensemble models in terms of all metrics except recall, but it has small uncertainty associated as well. LGBoost performed well in terms of recall. XGBoost could not overwhelm the other models in terms of any metrics. However,

all the metrics are optimal and satisfies the condition of generalization. It is easy to infer that, with hyperparameter tuning and a relatively larger dataset, the models would provide robust predictions.

Trained on Non-Clinical Dataset (Cleveland renewed) tested on Combined non-clinical dataset

Table 3: Performance metrics of XGBoost, LGBoost, and RF models on non-clinical dataset

Model	**Accuracy**	**AUC**	**F1 Score**	**Precision**	**Recall**
XGBoost	0.8137	0.8821	0.8452	0.78	0.92
LGBoost	0.8082	0.8729	0.8365	0.77	0.91
RF	0.8038±0.0009	0.8905±0.0006	0.8381±0.0008	0.76	0.92

This time, XGBoost outperformed other models in terms of accuracy, F1 score, precision, and recall. However, these metrics also satisfy the initial condition of OOD generalization.

Trained and Tested on Clinical Combined dataset (SMOTE)

Table 4: Performance metrics with hyperparameters tuned on clinical combined data

Model	**Accuracy**	**AUC**	**F1 Score**	**Precision**	**Recall**
XGBoost	0.9069	0.9578	0.9082	0.90	0.92
LGBoost	0.9020	0.9537	0.9029	0.89	0.91
RF	0.8995±0.0007	0.9509±0.0003	0.9012±0.0006	0.89	0.91

XGBoost performed relatively better in all performance metrics than the other models.

Trained and Tested on Non-Clinical Combined dataset (SMOTE)

Table 5: Performance metrics with hyperparameters tuned on non-clinical combined data

Model	**Accuracy**	**AUC**	**F1 Score**	**Precision**	**Recall**
XGBoost	0.84	0.9055	0.84	0.82	0.86
LGBoost	0.8480	0.9028	0.9029	0.85	0.84
RF	0.85±0.001	0.9085±0.0004	0.8536±0.0009	0.83	0.88

Random Forest Classifier seems to be comparatively better than the other model when it's about non-clinical dataset. It surpassed other models in terms of accuracy, AUC, and recall, which makes it a best fit for developing ML-systems to predict CVD with non-clinical features.

4.1 Fairness Metrics

The report on fairness metrics of the models is given in Table 6 and Table 7.

Table 6: Sex-wise performance metrics for XGBoost

Sex	Accuracy	Selection Rate	Precision	Recall
Female	0.9074	0.22	0.75	0.81
Male	0.9066	0.62	0.95	0.91

Table 7: Fairness metrics for XGBoost

Metric	Value
Demographic Parity Difference	0.40
Demographic Parity Ratio	0.35
Equalized Odds Difference	0.11
Equalized Odds Ratio	0.51

Selection rates indicate the proportion of individuals from each group (Male and Female) that were assigned a positive prediction by the model (classified as having heart disease).

- Male Selection Rate $= 0.62 :$ 62% of males were predicted as positive.
- Female Selection Rate $= 0.22 :$ 22% of females were predicted as positive.

Therefore, demographic parity difference $= 0.62 - 0.22 = 0.40$. This 0.40 difference implies that the model may be biased toward males unless this difference is justifiable by clinical or contextual factors. Also, the demographic parity ratio is 0.35, this means the minority group's selection rate is only 35% of the majority group's rate. Ideally, this ratio should be close to 1 (perfect fairness), and a common fairness threshold is ≥ 0.8 (the "80% rule"). According to the EEOC's Uniform Guidelines, a selection rate for a minority group that is less than 80% of that of the majority group may indicate potential bias or adverse impact, commonly referred to as the "80% rule" or "four-fifths rule" [26]. Equalized odds difference indicates 11% difference between groups in terms of true positive and false positive rates. A value closer to 0 implies the model performs equally well across groups, which is the goal of equalized odds. Conversely, equalized odds ratio denotes that the model's true/false positive rates for one group are only 51% of those for the other group. Like demographic parity ratio, this should ideally be close to 1. However, all these can be a result of heart disease prevalence by sex in the dataset which is shown in Figure 4.

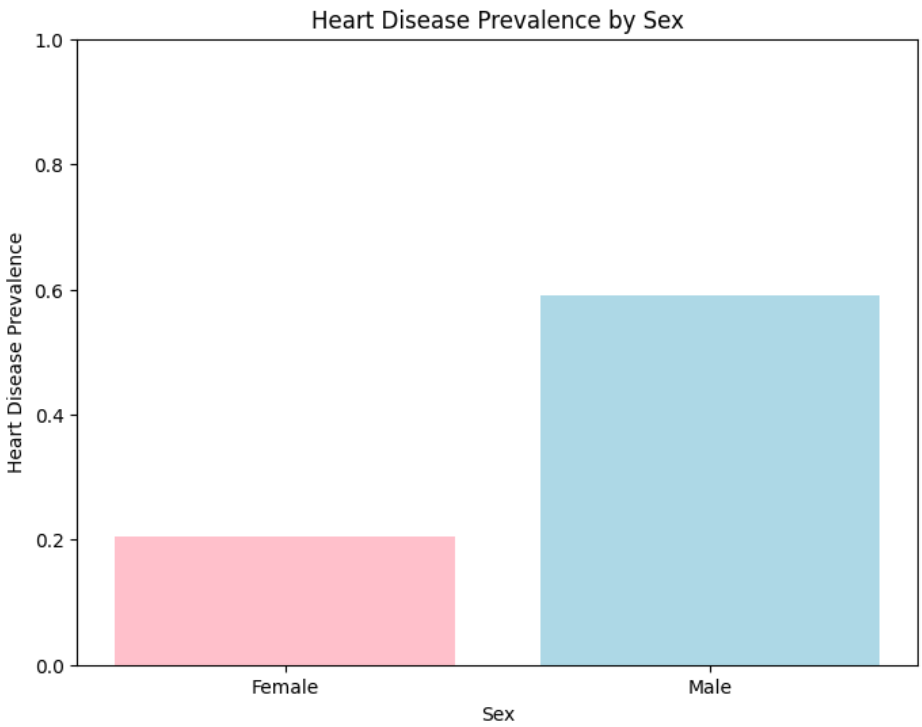

Figure 4: Heart Disease Prevalence by Sex

Sex-wise Fairness Metrics for Non-Clinical Model (RF) is presented in Table 8 and Table 9.

Table 8: Sex-wise performance metrics for RF

Sex	Accuracy	Selection Rate	Precision	Recall
Female	0.9242	0.2272	0.80	0.85
Male	0.8617	0.6329	0.86	0.91

Table 9: Fairness metrics for RF

Metric	Value
Demographic Parity Difference	0.40
Demographic Parity Ratio	0.35
Equalized Odds Difference	0.15
Equalized Odds Ratio	0.27

The fairness metrics are almost the same as what is noticed earlier in XGBoost. Now, before discussing bias mitigation, it is better to shed light on regional biases. The selection rate and recall of cluster 2 (theoretically Cleveland) is found significantly lower than the other regions for both clinical and non-clinical models. For the clinical model it is 0.13, and for the non-clinical model it is 0.082. It can be a consequence of heart disease prevalence by region as shown in Figure 5, or maybe the theoretical regions itself are not constructed well. So, we tried agglomerative clustering instead of PCA K-Means and found out that with a silhouette score of 0.44, the algorithm extracted four

clusters, but one cluster (cluster 3) has only 8 instances. It actually indicates that there is one cluster (region) that shows a similar distribution with some other cluster (region). Hence, we considered three clusters instead of four and retrained the models. Then, the models showed relatively balanced performance and fairness metrics across the regions. The selection rate for regions 0, 1, and 2 is 0.52, 0.92, and 0.42 respectively. This portion of our research supports the model generalization with the results of external validation.

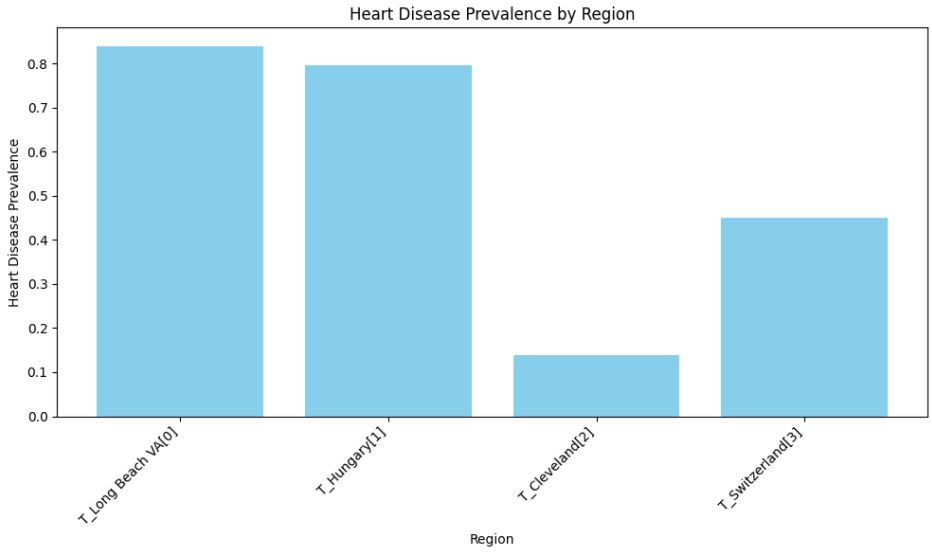

Figure 5: Heart Disease Prevalence by Region

4.1.1 Bias Mitigation Results

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

For the clinical model XGBoost:

- Demographic parity difference = 0.05
- Demographic parity ratio = 0.91
- Equalized odds difference = 0.09
- Equalized odds ratio = 0.55

For the non-clinical model Random Forest Classifier:

- Demographic parity difference = 0.07
- Demographic parity ratio = 0.892
- Equalized odds difference = 0.089
- Equalized odds ratio = 0.715

The fairness evaluation of both models shows acceptable performance in terms of demographic parity than before. However, it reveals some concerns regarding equalized odds. The general model

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

4.2 Ensemble Model Interpretability and Counterfactuals

We added SHAP interpretations and counterfactuals to both clinical and non-clinical models. This allowed us to gain both global and local interpretability for each prediction, revealing which features most significantly influenced model decisions. In clinical models, SHAP highlighted variables like chest pain type, exercise-induced angina, and ST slope as critical contributors, while in non-clinical models, it emphasized behavioral and symptom-related features such as chest pain and heart rate response. Counterfactual explanations, generated using DiCE, further complemented SHAP by demonstrating minimal actionable changes that could alter a prediction from diseased to non-diseased. These interpretability tools not only increased transparency and model trustworthiness but also provided clinicians and users with meaningful insights into potential lifestyle or symptom-based interventions, aligning with ethical AI principles and making the models more suitable for real-world healthcare deployment.

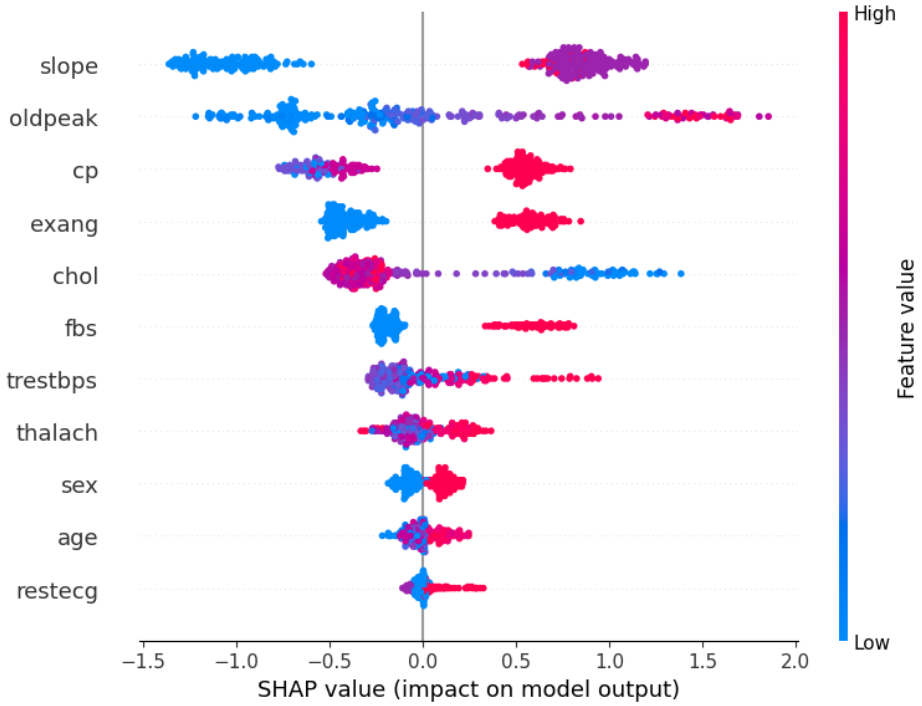

Figure 6: SHAP value (impact on clinical model output)

The SHAP summary plot of clinical model in Figure 6 shows the global feature importance and impact on the XGBoost model's predictions. Features like slope, oldpeak, and chest pain type have the highest influence, with wider distributions indicating stronger effects. Positive SHAP values (right side) push predictions toward higher CVD risk, while negative values (left side) reduce risk estimates. The plot reveals that age and sex have relatively smaller but consistent impacts compared to clinical features like max heart rate or serum cholesterol.

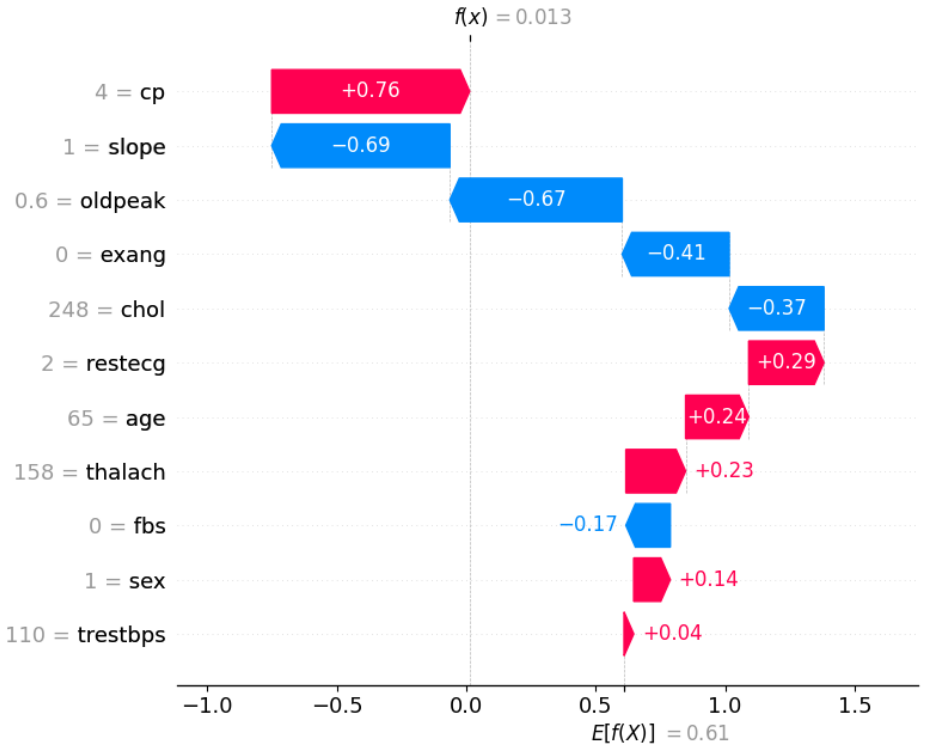

Figure 7: Local Interpretation of a CVD Patient

To understand the rationale behind the model's prediction for a patient classified as **diseased** (class 1), we employed SHAP (SHapley Additive exPlanations) analysis. The model's base value, as shown in Figure 7, or average prediction across the dataset, was 0.61, and the prediction for this specific instance rose to approximately 0.623, crossing the classification threshold. The strongest contributor to the positive prediction was the chest pain type (cp = 4), typically associated with asymptomatic or severe angina, contributing +0.76 to the final score. Additional risk factors included abnormal resting ECG results (restecg = 2), older age (65), male sex, and a high maximum heart rate (thalach = 158). Despite several features that reduced the risk, such as a flat ST slope (slope = 1), low ST depression (oldpeak = 0.6), absence of exercise-induced angina (exang = 0), and near-normal cholesterol levels- the cumulative weight of high-risk indicators led the model to classify the patient as having heart disease. This example highlights the model's ability to weigh clinical features in a nuanced manner, with significant emphasis on symptom severity and ECG abnormalities, aligning well with medical understanding.

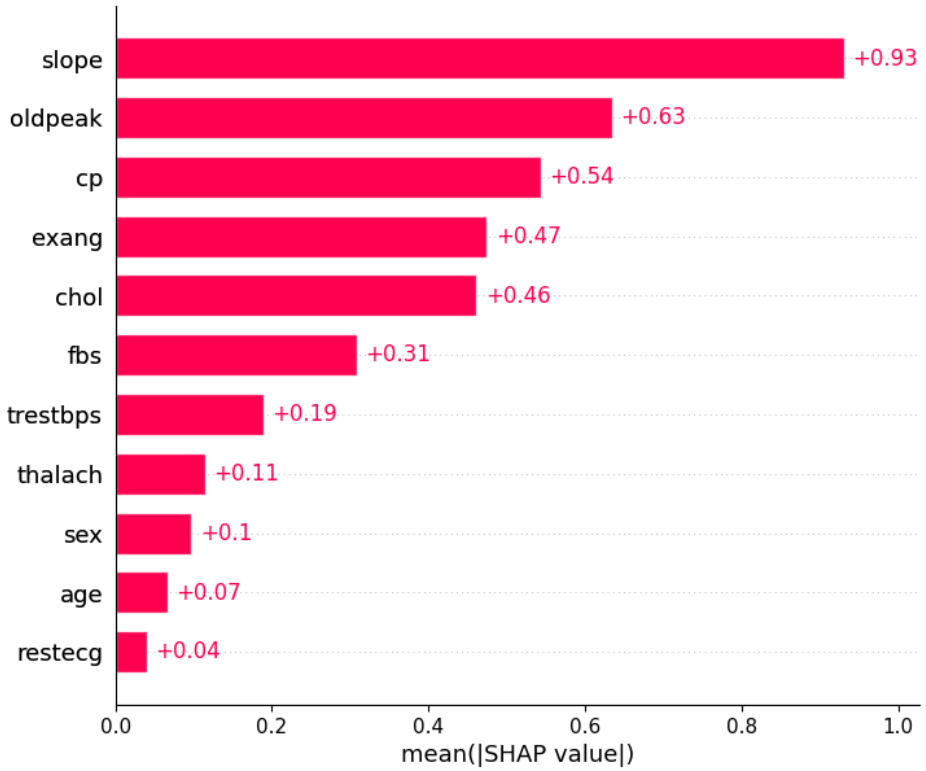

Figure 8: Mean absolute SHAP values of XGBoost

The mean SHAP value plot in Figure 8 reveals the global feature importance for the XGBoost cardiovascular disease (CVD) prediction model:

- **Top Influencers**:
  - `slope` (ST-segment slope, +0.93) and `oldpeak` (ST depression, +0.63) are the strongest predictors, indicating ECG-related features dominate CVD risk.
  - `cp` (chest pain type, +0.54) and `exang` (exercise-induced angina, +0.47) follow closely, highlighting the importance of clinical symptoms.
- **Moderate Contributors**:
  - `chol` (cholesterol, +0.46) and `fbs` (fasting blood sugar, +0.31) show measurable but smaller impacts.
- **Minimal Influence**:
  - Demographic factors (`age` +0.07, `sex` +0.10) and `restecg` (resting ECG, +0.04) have relatively low importance.

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

The local interpretation and mean SHAP values of **non-clinical model (RF)** is shown in Figure 9 and Figure 10.

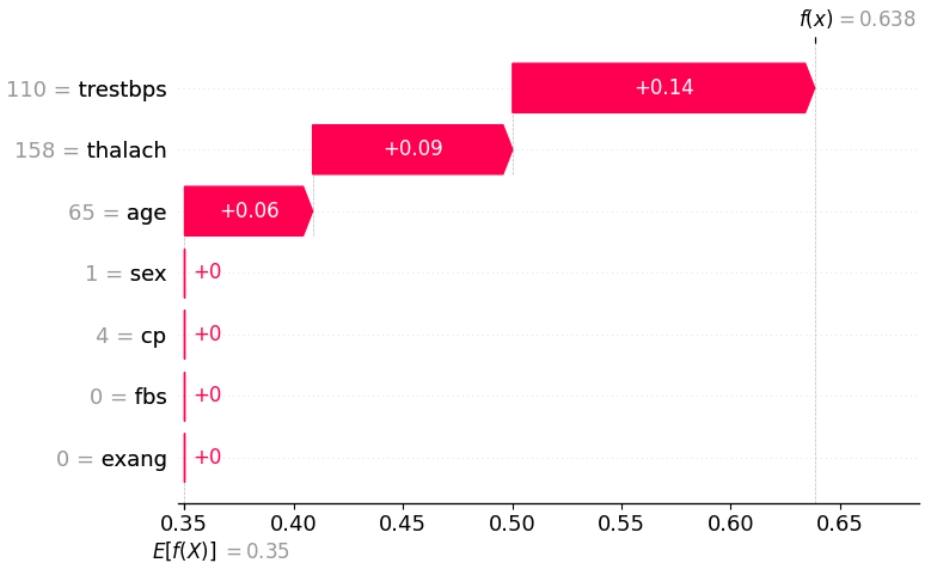

Figure 9: Local Interpretation of CVD Patient

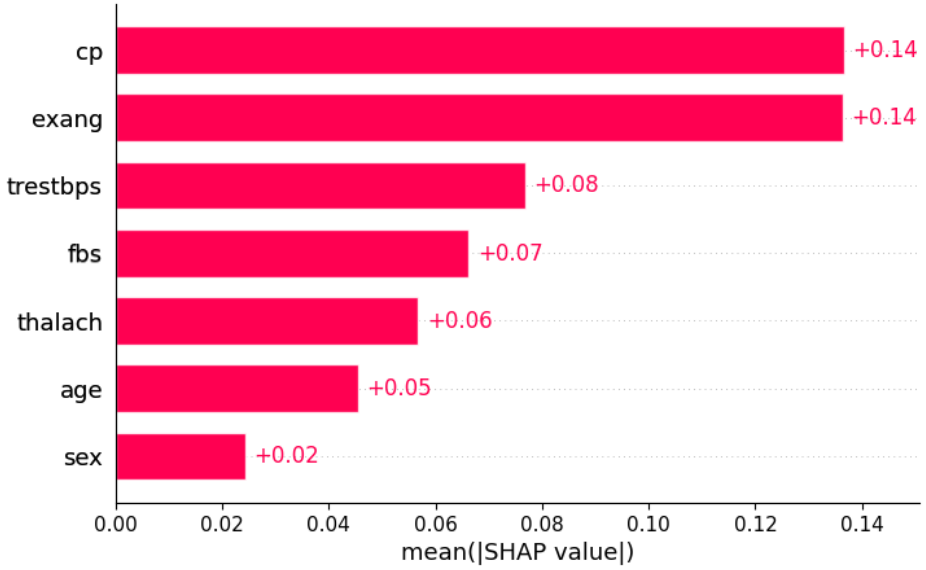

Figure 10: Mean Absolute SHAP values of RF

The interpretation method for these plots are similar as XGBoost. Figure 9 shows that the individual's age, max heart rate, and resting blood pressure are the major factors behind getting diagnosed as CVD.

4.3 Counterfactual Analysis

Counterfactual analysis was used to explore how slight changes in input features could alter the model's prediction from **diseased** to **healthy**. By identifying minimal, plausible modifications such as reducing chest pain type or increasing ST slope- we revealed actionable insights that may inform personalized interventions. This approach enhances interpretability by showing not only why a prediction was made, but also how it could change.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	target
0	38	1	4	110	289	0	0	105	1	1.5	3	1

Figure 11: Features of a Diseased Patient

Figure 11 shows the features of a diseased patient, and Figure 12 shows several counterfactuals regarding what changes are needed to be made in order to stay healthy. However, some counterfactuals often show values out of the domain. Still, that is interpretable with general knowledge.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	target
0	38	1	1	110	289	0	0	105	0	1.5	3	0
1	38	1	4	110	289	0	0	105	0	1.5	0	0
2	38	1	4	110	289	0	0	105	1	-1.5	0	0
3	38	1	4	110	289	0	0	105	1	-1.3	1	0
4	38	1	3	110	289	0	0	105	1	1.5	0	0
5	38	1	4	110	289	0	0	105	0	1.5	1	0

Figure 12: Counterfactuals of the Diseased Patient

Counterfactual explanations for clinical XGBOOST generated by DiCE revealed that reducing chest pain severity, exercise-induced angina, and improving ECG-related markers (slope, oldpeak) could reverse the model's prediction from diseased to non-diseased. Most changes are aligned with clinical expectations, although domain constraints (e.g., non-negative ST depression) are necessary to avoid physiologically implausible scenarios. The absence of changes in cholesterol and fasting blood sugar suggests these features held less influence in the model's decision boundary.

Figure 13 shows the record of an individual in early warning phase, and Figure 14 shows some counterfactuals, specially on how the indiviudal can get healthy again.

	age	sex	cp	trestbps	fbs	thalach	exang	target
0	38	1	4	110	0	105	1	1

Figure 13: Features of a Warned Individual

	age	sex	cp	trestbps	fbs	thalach	exang	target
0	38	1	1	110	0	105	0	0
1	29	1	4	110	0	105	0	0
2	28	1	4	110	0	105	0	0
3	38	1	4	110	0	199	0	0
4	38	1	2	110	0	105	0	0
5	38	1	4	110	0	183	0	0

Figure 14: Counterfactuals of the Warned Individual

Counterfactual explanations from the non-clinical model revealed that reducing exercise-induced angina and chest pain severity, increasing exercise heart rate, and lowering age were sufficient to flip predictions from diseased to healthy. All generated counterfactuals were physiologically plausible, confirming the model's medical alignment and reliability in early-warning scenarios where only non-clinical features are available.

4.4 Neural Networks Performance, Interpretation, and Uncertainty

This section focuses on the performance of neural network architectures. The ROC-AUC curve of all the neural models are illustrated in Figure 15.

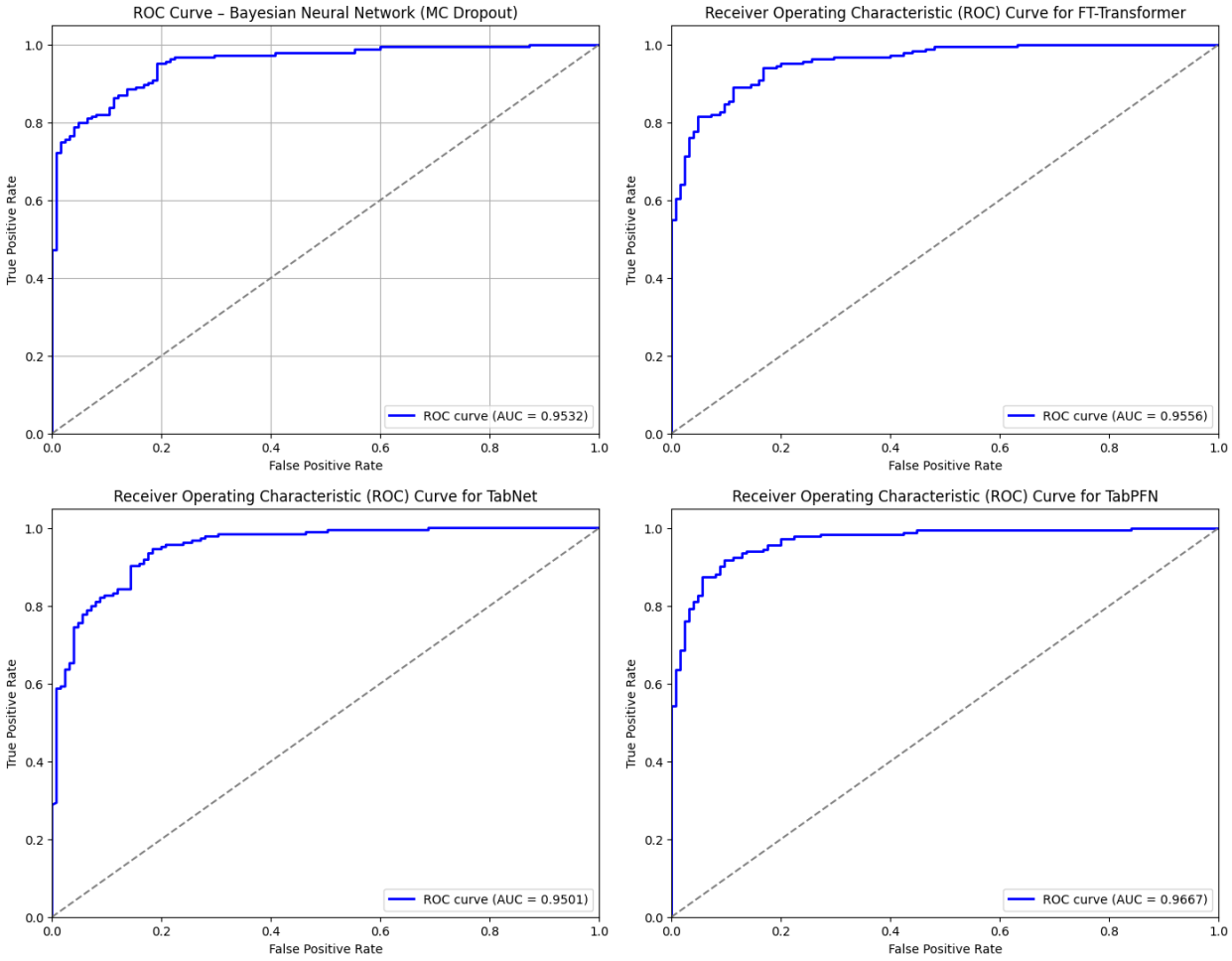

Figure 15: ROC-AUC of neural models

Table 10: Performance Comparison of Neural Models and CGEVC on Cardiovascular Disease Prediction

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

Table 10 shows that among all the evaluated neural models, TabPFN outperforms others with the highest accuracy (90%), macro F1-score (0.90), and weighted F1-score (0.90). This consistency across all key metrics indicates its superior balance in handling both positive and negative classes effectively. The model's performance reflects not only strong predictive ability but also generalizability across patient groups. CGEVC's balanced performance similar to other models indicate that the causal factors were truly important features for CVD diagnosis. Also, due to filtering out features that have an ATE of less than 0.1, CGEVC successfully reduced the dimensionality while preserving diagnosis

performance.

4.4.1 Tabular Models' Drawback On Synthetic Data

To explore the real-world scalability and robustness of neural models in large-scale settings, we simulated a high-data environment by augmenting the dataset using the Gaussian Copula method. Our assumption was that the neural models would perform better with the augmented data than the prevision one. However, the results show that, even though the Gaussian Copula method preserves the marginal and joint distributions of the original data, the performance of all tabular neural network models- TabNet, FT-Transformer, TabPFN, and Bayesian Neural Network (BNN) declined noticeably compared to prior results on smaller, high-quality datasets. It can also occur that the augmentation method itself failed to capture the pattern of the original data. Since Gaussian Copula assumes that the transformed variables follow a multivariate normal distribution, it works well when the dependencies are monotonic, and relationships are linear or weakly non-linear. Real world clinical data often does not satisfy this assumption. It can have non-monotonic relationships (e.g., cholesterol and age vs risk), and conditional interactions (e.g., chest pain affects prediction only if blood pressure is high).

As summarized in Table 11, all four models experienced difficulty in effectively separating the two classes. TabNet, for example, achieved an overall accuracy of 66% with a macro F1-score of only 0.63. Similarly, TabPFN and FT-Transformer performed comparably, with accuracies around 69% and macro F1-scores of 0.66 and 0.67 respectively. The Bayesian Neural Network (BNN) performed slightly better, achieving 70% accuracy and a macro F1 of 0.68, but still fell short of its performance on non-synthetic data. Notably, across all models, the recall for class 0 (non-diseased) remained consistently lower than that of class 1 (diseased), suggesting a class-specific bias or dilution of decision boundaries caused by the synthetic samples.

Table 11: Performance of Neural Models Trained on Gaussian Copula–Generated Synthetic Data

Model	**Accuracy**	**Macro F1-Score**	**Weighted F1-Score**
TabNet	0.66	0.63	0.65
FT-Transformer	0.69	0.67	0.68
TabPFN	0.69	0.66	0.68
BNN	**0.70**	**0.68**	**0.69**

These findings underscore an important limitation: while Gaussian Copula–based augmentation expands dataset size and preserves statistical dependencies, it may fail to retain discriminative class-level structures essential for deep learning. Tabular neural networks, which rely heavily on complex, non-linear feature interactions and class-conditioned distributions, appear particularly sensitive to such subtle shifts. This supports the hypothesis that not all synthetic data is equally useful, and the quality of augmentation plays a crucial role in model generalization, especially for high-capacity models.

Such phenomena has concerns regarding trustworthy AI and model generalizability. Many studies show impressive results after augmenting small datasets with synthetic data (CTGAN, SMOTE, Copula, etc.), but fail to validate whether the augmentation preserved the real-world decision boundary. Synthetic data augmentation, while often improving reported model performance, introduces significant ethical and generalization concerns if not handled rigorously. In some studies, augmenting datasets with synthetic instances yields models with remarkably high accuracy; however, without validating whether the synthetic data preserves true feature-label relationships, these gains can be

illusory. There is a risk that augmentation introduces artificial patterns or smooths out decision boundaries in ways that make the classification task easier, essentially allowing the model to excel on a synthetic distribution that no longer reflects the real world. This misalignment can result in models that underperform on genuine patient data, violating the principles of Trustworthy AI, particularly fairness, transparency, and generalizability. For high-stakes applications like cardiovascular disease diagnosis, it is therefore essential to critically evaluate not only model performance but also the integrity and ethical implications of the data generation process itself.

4.4.2 Uncertainty Quantification

The histogram in Figure 16 illustrates the uncertainty distribution (measured as the standard deviation of predictive probabilities) for both cardiovascular disease (CVD) and non-CVD predictions using the Bayesian Neural Network (BNN).

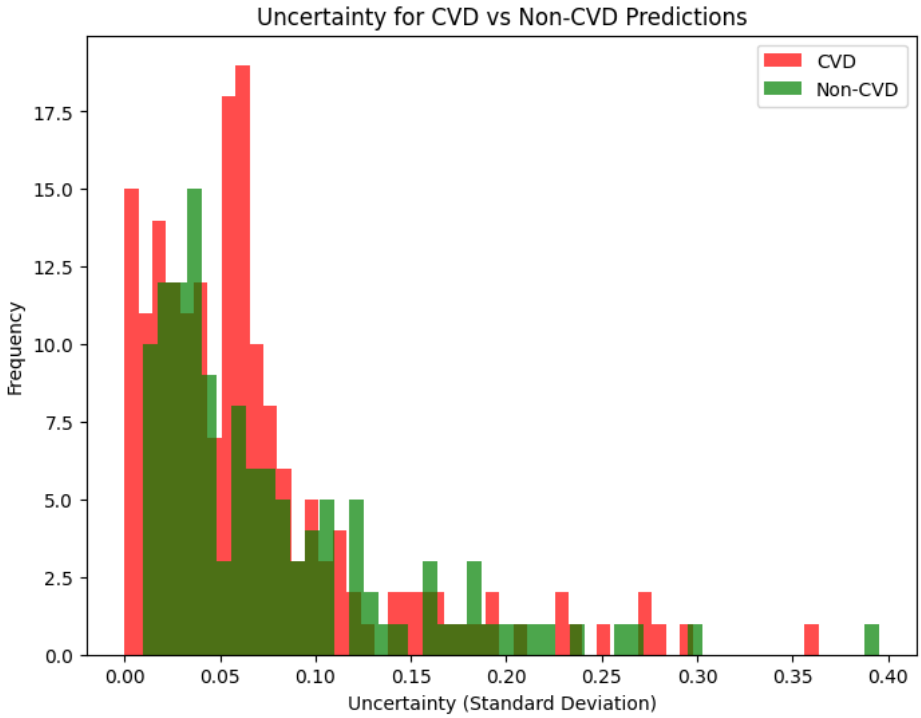

Figure 16: BNN Prediction Uncertainty Histogram

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

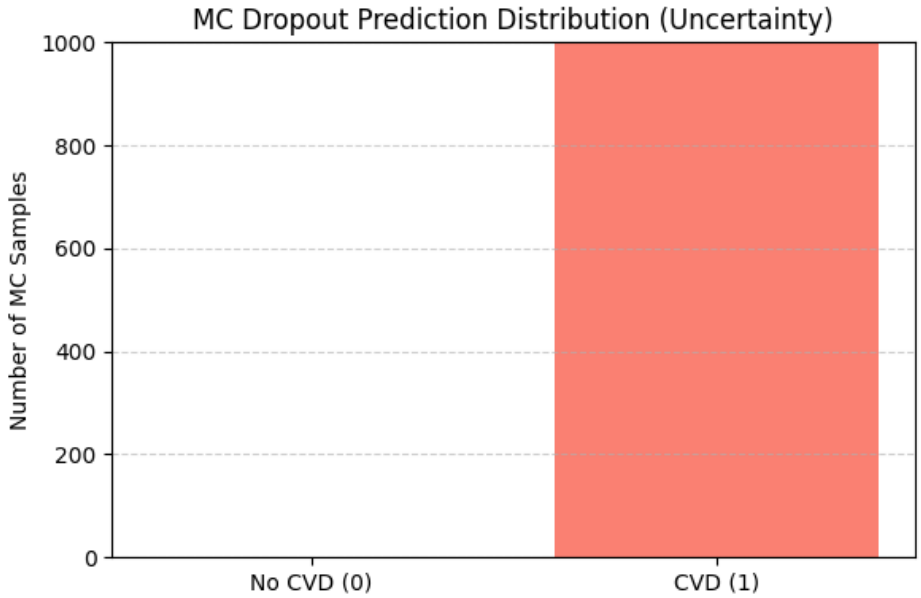

Figure 17: Predicted as Diseased by BNN

Figure 17 illustrates the prediction characteristics of the Bayesian Neural Network (BNN). Owing to its probabilistic nature, the BNN samples multiple weight configurations during inference, resulting in a distribution of predictions for each instance. A sample is classified as class 1 (diseased) if the mean of these predictions exceeds 0.5. A prediction count concentrated near the total number of Monte Carlo (MC) samples indicates high confidence, while broader distributions reflect greater uncertainty. This approach offers a principled and robust quantification of uncertainty, a distinctive capability of Bayesian models.

The uncertainty analysis for TabPFN in Figure 18 is presented through two key metrics: entropy (left plot) and prediction confidence (right plot). In the entropy histogram, we observe that most predictions are concentrated at very low entropy values (close to 0), indicating that the model often outputs highly confident and sharp class probabilities. However, there's a notable tail extending up to 0.7, suggesting that a minority of predictions remain highly uncertain, potentially due to ambiguous or borderline cases.

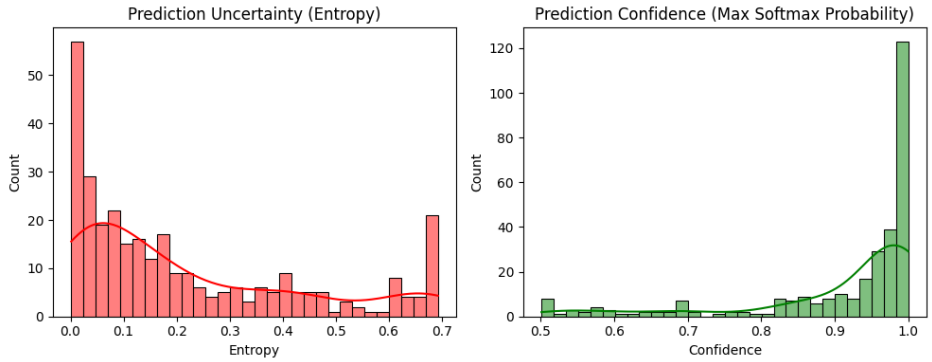

Figure 18: TabPFN Uncertainty

The confidence histogram complements this by showing that the majority of predictions have a softmax maximum probability above 0.95, further confirming the model's decisiveness on most samples. The sharp peak near 1.0 in the confidence plot supports the low-entropy pattern observed earlier. Together, these visualizations imply that TabPFN tends to be highly confident in its predictions, but retains the ability to flag uncertainty where necessary, which is essential for trustworthy AI in clinical settings.

Figure 19 presents the global feature importance derived from TabNet, highlighting the overall contribution of each variable to the model's predictions. Figures 20 and Figure 21 offer local interpretability insights. Figure 20 corresponds to a correctly predicted healthy individual, while Figure 21 represents a diseased case. These interpretations are made possible through TabNet's intrinsic sparse attention mechanism, which enables the model to selectively focus on the most relevant features for each individual prediction.

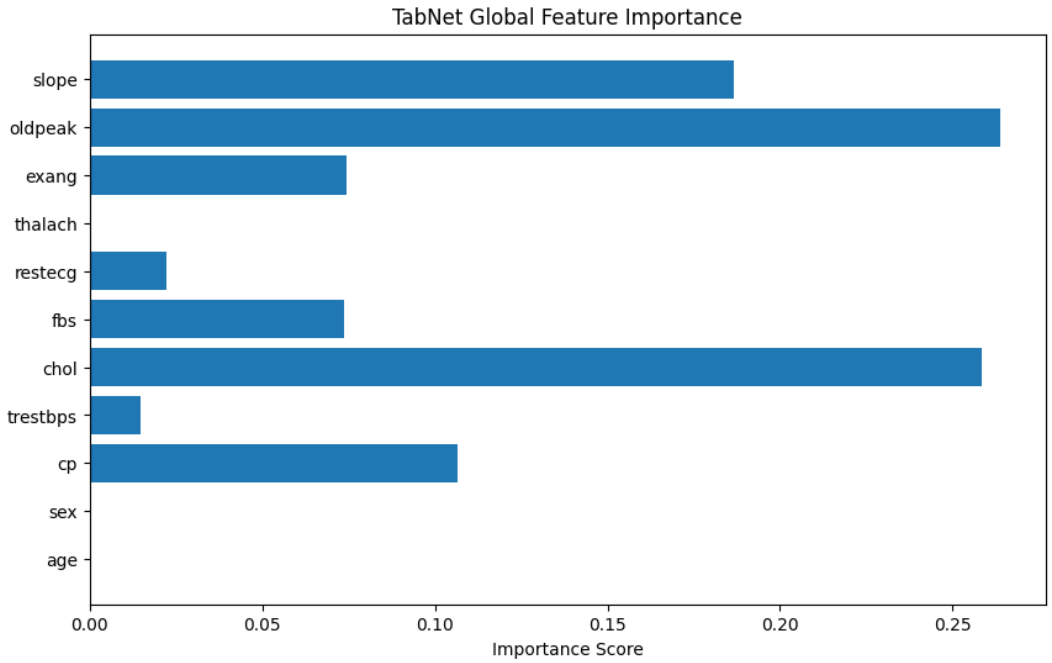

Figure 19: TabNet Global Feature Importance

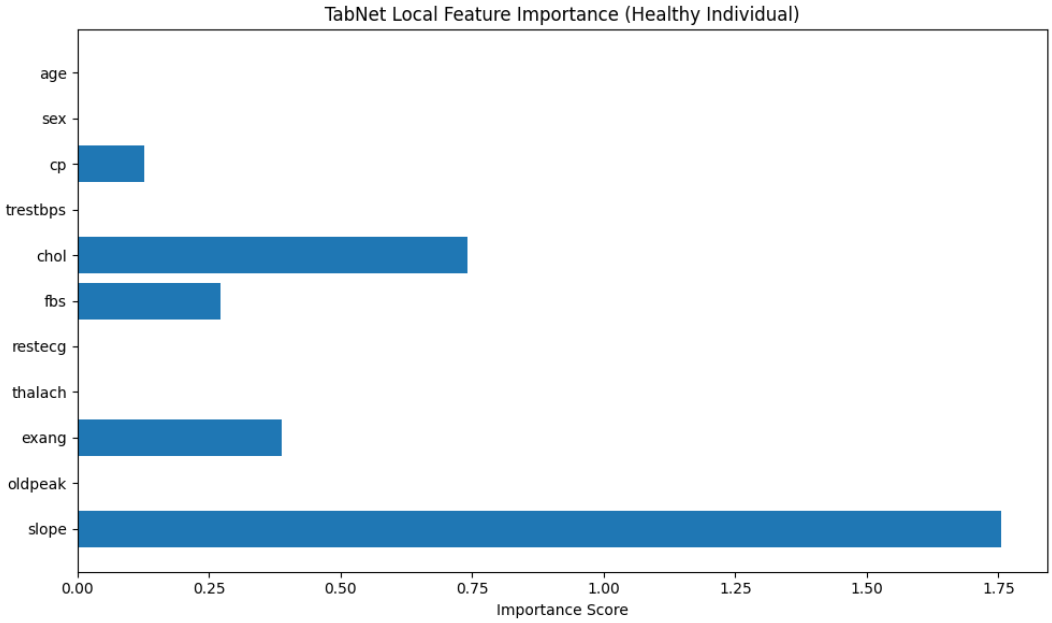

Figure 20: TabNet Local Interpretation (Healthy Individual)

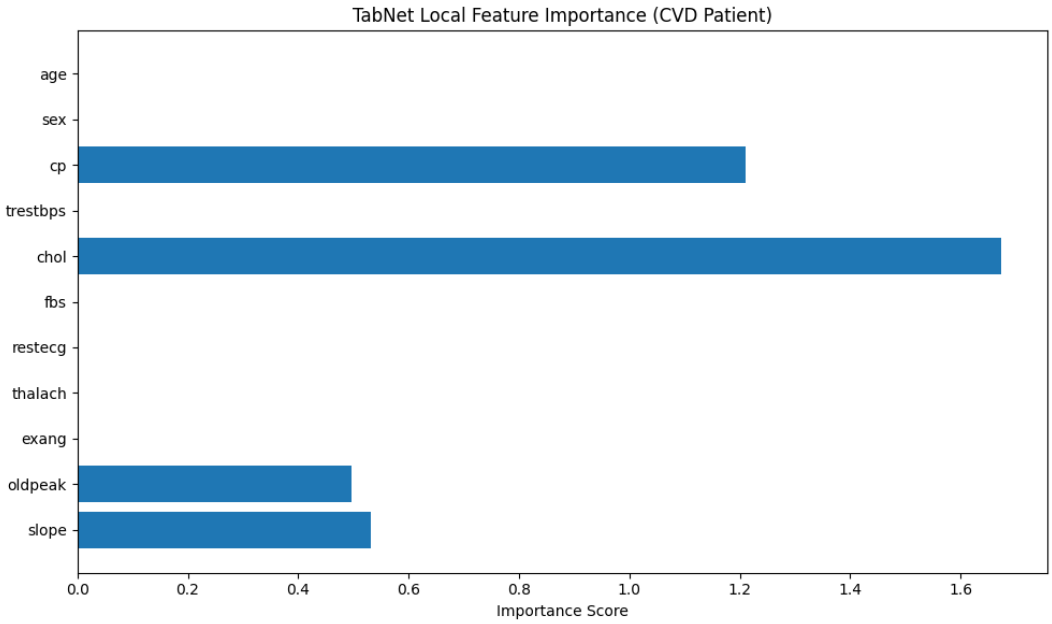

Figure 21: TabNet Local Interpretation (Diseased Individual)

TabNet also supports SHAP interpretation, as shown in Figure 22.

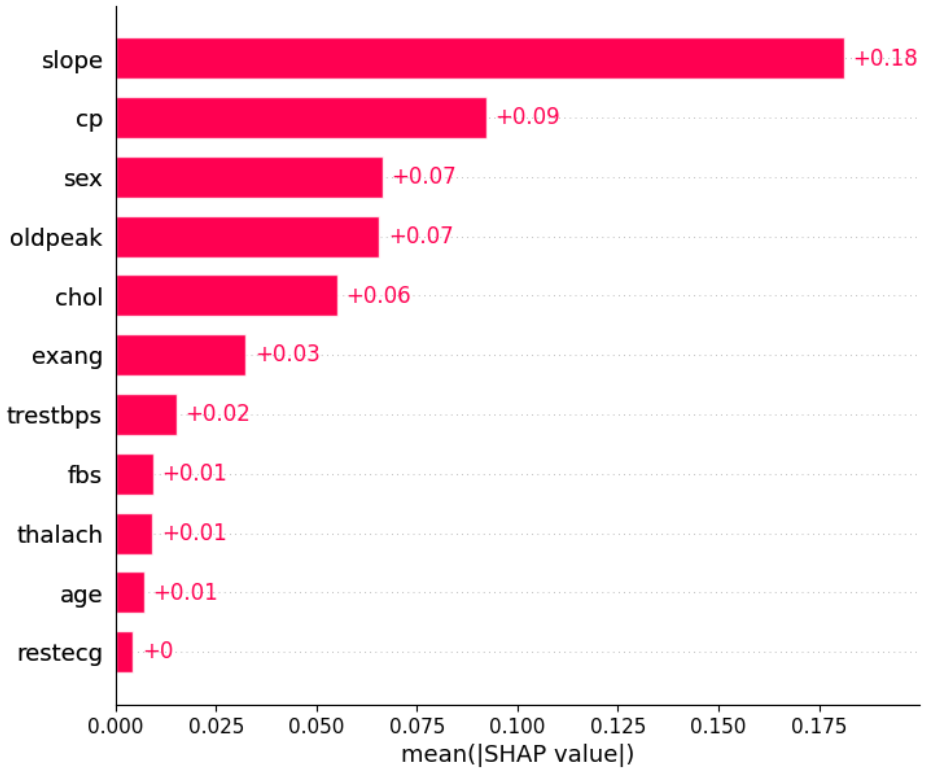

Figure 22: TabNet mean SHAP values

4.5 Causal Inference

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

4.5.1 Estimated Causal Effects (Average Treatment Effect - ATE)

Each feature is treated as a potential treatment; its causal impact on the likelihood of heart disease is estimated while controlling for relevant confounders. A bar plot of the features (treatments) and their average treatment effect (ATE) is shown in Figure 23.

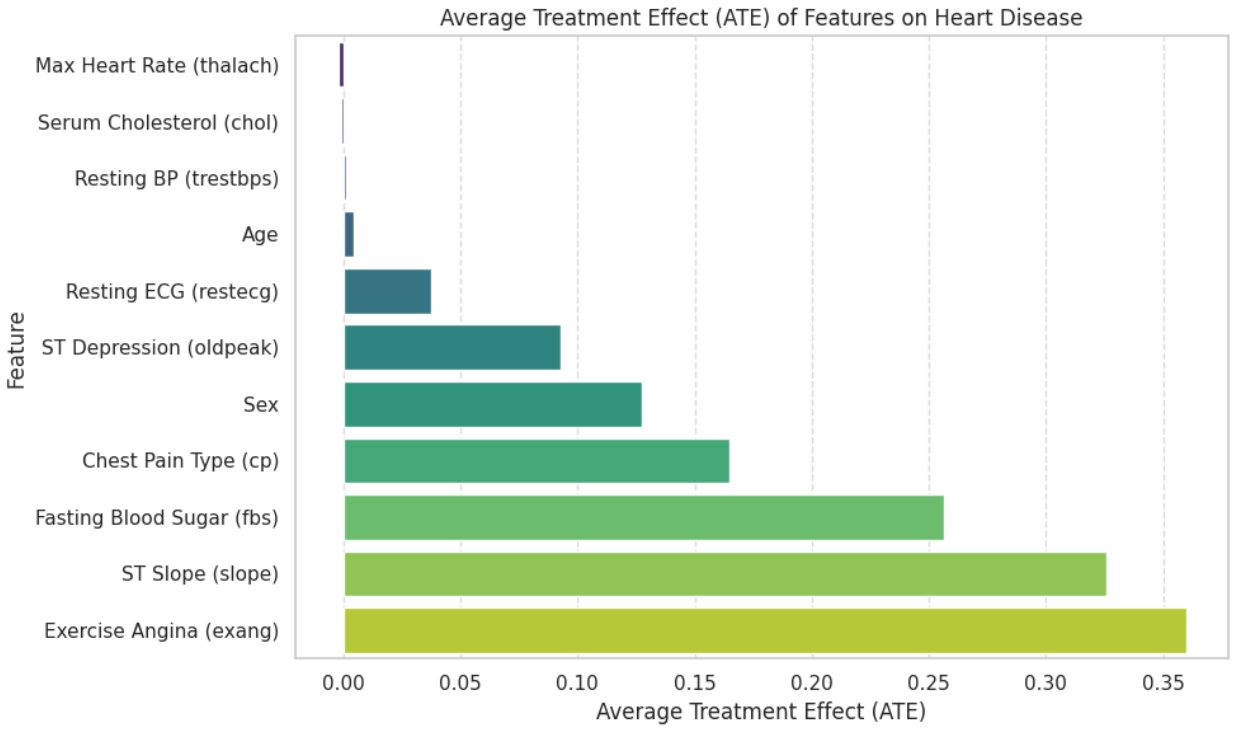

Figure 23: Average Treatment Effect (ATE) of Features on Heart Disease

The ATEs in Figure 23 are presented in ascending order so that it seems easier to understand which features (treatment) function as causal indicators of CVD.

1. Age

- ATE: 0.0045
- Confounders: oldpeak, thalach, trestbps, chol

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

2. Sex

- ATE: 0.1274
- Confounders: oldpeak, thalach, trestbps, age

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

3. Chest Pain Type (cp)

- 0.1651
- Confounders: oldpeak, thalach, trestbps, chol, age

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

4. Resting Blood Pressure (trestbps)

- ATE: 0.0011
- Confounders: oldpeak, thalach, chol, age

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

5. Serum Cholesterol (chol)

- ATE: -0.00089
- Confounders: oldpeak, thalach, trestbps, age

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

6. Fasting Blood Sugar (fbs)

- ATE: 0.2566
- Confounders: thalach, trestbps, chol, age

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

7. Resting ECG (restecg)

- ATE: 0.0375
- Confounders: thalach, trestbps, chol, age

Interpretation: Resting ECG abnormalities contribute moderately to heart disease probability. While not the most significant factor, it remains relevant in clinical contexts. A one-unit change in ECG findings leads to a 3.75% point increase in the likelihood of disease, assuming confounders are constant.

8. Max Heart Rate Achieved (thalach)

- ATE: -0.0023
- Confounders: oldpeak, trestbps, chol, age

Interpretation: Maximum heart rate achieved is negatively correlated with heart disease risk, suggesting that better cardiovascular fitness may be protective. Statistically, a one-unit increase in maximum heart rate results in a 0.23% point decrease in heart disease probability, given fixed confounding variables.

9. Exercise-Induced Angina (exang)

- ATE: 0.3598
- Confounders: oldpeak, thalach, trestbps, chol, age

Interpretation: This is one of the most influential features. Patients experiencing angina during exercise are significantly more likely to develop heart disease. The ATE suggests that a one-unit increase in exercise induced angina leads to a 35.98% point increase in heart disease probability, when other confounders are controlled.

10. ST Depression Induced by Exercise (oldpeak)

- ATE: 0.0925
- Confounders: thalach, slope, trestbps, chol, age

Interpretation: ST depression during stress testing moderately increases the risk of heart disease. On average, each unit increase in oldpeak corresponds to a 9.25% point rise in heart disease probability, assuming other variables are fixed.

11. ST Slope (slope)

- ATE: 0.3260
- Confounders: oldpeak, thalach, trestbps, chol, age

Interpretation: The slope of the ST segment is a strong causal indicator of heart disease, particularly in exercise ECG tests. A one-unit increase in the slope of the ST segment leads to a 32.60% point increase in disease probability, controlling for related confounders.

Overall, exercise-induced angina (exang), ST slope, fasting blood sugar, and chest pain type demonstrate the strongest causal effects. Variables traditionally considered important, like cholesterol or resting BP may not have strong causal impact in this dataset, potentially due to mediation or correlation with stronger signals. This causal analysis not only supports known clinical relationships but also justifies the feature selection used in the voting classifier.

5 Conclusion

In conlusion, our research presents a comprehensive and ethically aligned AI ecosystem for cardiovascular disease (CVD) diagnosis, addressing critical gaps in early detection, clinical applicability, and Responsible AI (RAI) practices. By integrating advanced tabular neural networks (TabNet, TabPFN, FT-Transformer), ensemble models (XGBoost, Random Forest, LightGBM), Bayesian Neural Network (BNN), and causal model (CGEVC), we achieved robust diagnostic performance with 0.90 accuracy while ensuring interpretability, uncertainty quantification, and bias mitigation. We also shed light on potential harm of solely relying on augmented data with the help of Gaussian Copula. Our work stands out by developing non-clinical models for accessible, early risk assessment, enabling proactive interventions before symptoms manifest. Incorporating SHAP, DiCE counter-

factuals, FairLearn, and Bayesian uncertainty, we aimed to enhance transparency, fairness, and trustworthiness in clinical decision-making. We addressed demographic and regional biases through synthetic data augmentation (CTGAN, SMOTE) and unsupervised clustering, ensuring equitable performance across diverse populations. We also presented a novel ML-system (CGEVC) guided by causal inference for diagnosing purposes, obtaining an accuracy and precision of 0.89, while reducing dataset dimensionality. Compared to prior works that focus narrowly on accuracy, our approach aligns with FDA Good ML Practices and EU AI ethics, offering a holistic solution for CVD diagnosis. Future directions include expanding datasets for rare CVD subtypes, integrating multimodal data (e.g., ECG, imaging), validating data augmentation procedures in clinical sectors, and refining causal inference for personalized treatment insights. Future research should also focus on reducing disparities in equalized odds while preserving overall model performance. As our focus was on research than development, we did not develop and maintain CI/CD pipeline and work on data drifts, but our detailed RAI framework and reproducibility guidelines presented in the paper can guide the industry specialists to develop AI systems that are ethically and clinically safe. By combining technological innovation with ethical rigor, this research advances AI-driven healthcare toward safer, more equitable, and actionable outcomes.

References

[1] Centers for Disease Control and Prevention. (2025). *National Center for Health Statistics mortality data on CDC WONDER*. CDC WONDER. https://wonder.cdc.gov/mcd.html

[2] Martin, S. S., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., Baker-Smith, C. M., Barone Gibbs, B., Beaton, A. Z., Boehme, A. K., et al. (2024). 2024 heart disease and stroke statistics: A report of US and global data from the American Heart Association. *Circulation, 149*(8). https://doi.org/10.1161/CIR.0000000000001209

[3] Centers for Disease Control and Prevention. (2024). *Heart disease facts*. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[4] Cleveland Clinic. (2022). *Cardiovascular disease*. https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease

[5] World Health Organization. (n.d.). *Cardiovascular diseases*. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_2

[6] Karatza, A. A., Fouzas, S., Gkentzi, D., Kostopoulou, E., Loukopoulou, C., Dimitriou, G., & Sinopidis, X. (2025). Missed or delayed diagnosis of heart disease by the general pediatrician. *Children, 12*(3), 366. https://doi.org/10.3390/children12030366

[7] Mandal, A. (2023). *Cardiovascular disease diagnosis*. News-Medical.net. https://www.news-medical.net/health/Cardiovascular-Disease-Diagnosis.aspx

[8] Qadri, A. M., Raza, A., Munir, K., & Almutairi, M. S. (2023). Effective feature engineering technique for heart disease prediction with machine learning. *IEEE Access, 11*, 56214–56223. https://doi.org/10.1109/ACCESS.2023.3281484

[9] Al-Alshaikh, H. A., Prabu, P., Poonia, R. C., Saudagar, A. K. J., Yadav, M., AlSagri, H. S., & AlSanad, A. A. (2024). Comprehensive evaluation and performance analysis of machine learning in heart disease prediction. *Scientific Reports, 14*, 7819. https://doi.org/10.1038/s41598-024-58489-7

[10] Almazroi, A. A., Aldhahri, E. A., Bashir, S., & Ashfaq, S. (2023). A clinical decision support system for heart disease prediction using deep learning. *IEEE Access, 11*, 61646–61659. https://doi.org/10.1109/ACCESS.2023.3285247

[11] Subramani, S., Varshney, N., Vijay Anand, M., Soudagar, M. E. M. S., Al-keridis, L. A., Upadhyay, T. K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K., & Rohini, K. (2023). Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in Medicine, 10*, 1150933. https://doi.org/10.3389/fmed.2023.1150933

[12] Saeed, M. H., & Hama, J. I. (2023). Cardiac disease prediction using AI algorithms with SelectKBest. *Medical & Biological Engineering & Computing, 61*, 3397–3408. https://doi.org/10.1007/s11517-023-02918-8

[13] Eleyan, A., AlBoghbaish, E., AlShatti, A., AlSultan, A., & AlDarbi, D. (2024). RHYTHMI: A deep learning-based mobile ECG device for heart disease prediction. *Applied System Innovation, 7*(5), 77. https://doi.org/10.3390/asi7050077

[14] Rohan, D., Pradeep Reddy, G., Pavan Kumar, Y. V., Purna Prakash, K., & Pradeep Reddy, Ch. (2025). RHYTHMI: An extensive experimental analysis for heart disease prediction using artificial intelligence techniques. *Scientific Reports, 15*, 6132. https://doi.org/10.1038/s41598-025-90530-1

[15] Kumar, A., Singh, K. U., & Kumar, M. (2023). A clinical data analysis based diagnostic system for heart disease prediction using ensemble method. *Big Data Mining and Analytics, 6*(4), 513–525. https://doi.org/10.26599/BDMA.2022.9020052

[16] Nandy, S., Adhikari, M., Balasubramanian, V., Menon, V. G., Li, X., & Zakarya, M. (2023). An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Computing and Applications, 35*, 14723–14737. https://doi.org/10.1007/s00521-021-06124-1

[17] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*. https://arxiv.org/abs/1603.02754

[18] Arık, S. Ö., & Pfister, T. (2020). TabNet: Attentive interpretable tabular learning. *arXiv preprint arXiv:1908.07442*. https://arxiv.org/abs/1908.07442

[19] Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023). TABPFN: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*. https://arxiv.org/abs/2207.01848

[20] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *arXiv preprint arXiv:2106.11959*. https://arxiv.org/abs/2106.11959

[21] Goan, E., & Fookes, C. (2020). Bayesian neural networks: An introduction and survey. *arXiv preprint arXiv:2006.12024*. https://arxiv.org/abs/2006.12024

[22] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access, 7*, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

[23] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine, 136*, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

[24] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms, 16*(2), 88. https://doi.org/10.3390/a16020088

[25] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *arXiv preprint arXiv:1907.00503*. https://arxiv.org/pdf/1907.00503

[26] U.S. Equal Employment Opportunity Commission, U.S. Department of Labor, U.S. Department of Justice, & U.S. Civil Service Commission. (1978). *Uniform guidelines on employee selection procedures*. https://www.eeoc.gov/laws/guidance/questions-and-answers-about-uniform-guidelines-employee-selection-procedures

[27] Hasan, K. S. (2025). *CVD-AI-Research: A diverse ML ecosystem for cardiovascular disease diagnosis* [Computer software]. GitHub. https://github.com/SakibHasanSimanto/CVD-AI-Research

[28] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). *Heart Disease* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X