



Enhanced Regularized Polynomial XGBoost (ERP-XGB): Reducing Bias and Optimizing Performance in Cardiovascular Risk Prediction

Djalila Boughareb

Labstic Laboratory, Department of Computer Science, University of 8 Mai 1945 Guelma-Algeria

✉ boughareb.djalila@univ-guelma.dz

KEYWORDS

machine learning;
class imbalance;
gender bias; SMOTE;
XGBoost; logistic
regression; weight
adjustment; SVM;
ERP XGB

ABSTRACT

Cardiovascular diseases are among the leading causes of death globally, emphasizing the critical need for machine learning models that are both accurate and fair in clinical decision-making. This study introduces the Enhanced Regularized Polynomial XGBoost (ERP-XGB) model, which integrates polynomial feature expansion with $L1$, $L2$, and gamma regularization terms to improve classification accuracy, address class imbalance, and reduce algorithmic bias. ERP-XGB was evaluated on four benchmark datasets: Heart Failure (299 samples), Heart Attack (1,319 samples), Heart Disease (917 samples), and BRFSS (253679 samples). On the Heart Attack dataset, ERP-XGB achieved a ROCAUC of 99.59 ± 0.21 %, accuracy of 96.97 ± 0.49 %, F1 score of 97.73 ± 0.43 %, precision of 96.30 ± 0.73 %, and recall of 98.87 ± 0.47 %, with an average run time of 30.63 seconds. In terms of fairness, ERP-XGB reported an Equalized Odds (EO) score of 0.02 ± 0.01 , Disparate Impact (DI) of 0.96 ± 0.02 , and Demographic Parity (DP) values of 0.61 ± 0.01 for the unprivileged group and 0.64 ± 0.01 for the privileged group. On the Heart Disease dataset, ERP-XGB demonstrated even stronger performance, achieving a perfect ROC AUC of 100.00 ± 0.00 %, accuracy of 98.60 ± 0.43 %, F1 score of 98.58 ± 0.37 %, precision of 100.00 ± 0.00 %, and recall of 97.29 ± 0.48 %, with a run time of 41.45 seconds. Fairness evaluation showed EO at 0.03 ± 0.01 , DI at 1.78 ± 0.03 , and DP values of 0.69 ± 0.01 for the unprivileged group and 0.38 ± 0.01 for the privileged group. For Heart Failure, ERP-XGB achieved 89.82 ± 0.02 % ROC AUC, 82.93 ± 0.03 % accuracy, and strong fairness ($DI=0.91 \pm 0.31$). On BRFSS, it attained

Djalila Boughareb

Enhanced Regularized Polynomial XGBoost (ERP-XGB): Reducing Bias and Optimizing Performance in Cardiovascular Risk Prediction



ADCAIJ: Advances in Distributed Computing
and Artificial Intelligence Journal
Regular Issue, Vol. 14 (2025), e32367
eISSN: 2255-2863 - <https://adcaij.usal.es>
Ediciones Universidad de Salamanca - CC BY-NC-ND

90.57±0.000 % accuracy but showed lower recall (11.89±0.004 %) and fairness challenges (DI=0.38±0.03). These results confirm that ERP-XGB offers an effective balance between high predictive performance and robust fairness in clinical datasets, making it a promising tool for equitable cardiovascular disease diagnosis.

1. Introduction

Cardiovascular diseases continue to be among the primary causes of mortality globally. Early and accurate identification of cardiac patients is crucial for timely intervention and optimal outcomes. However, the application of machine learning for cardiac patient classification faces several significant challenges that have hindered its widespread adoption and effectiveness in clinical settings (Kilic et al., 2022; He and Garcia, 2009; Jiang et al., 2023; Johnson and Khoshgoftaar, 2019; Zhang et al., 2020; Li et al., 2021; Straw and Callison-Burch, 2020; Zhao et al., 2020).

In biomedical-based machine learning (ML) systems, the presence of bias—whether due to skewed datasets or algorithmic predispositions—can have profound and multifaceted consequences. This bias compromises the accuracy of diagnoses and treatment recommendations, potentially leading to erroneous clinical decisions. Such inaccuracies adversely affect patient outcomes, particularly for those from underrepresented groups in the training data. The ripple effect of these inaccuracies often exacerbates existing healthcare disparities, disproportionately impacting minority populations. Addressing these biases is critical, not only for the advancement of medical technology but also for ensuring equitable healthcare access and quality across diverse populations.

Cardiac datasets typically exhibit an intrinsic class imbalance, with healthy individuals far outnumbering cardiac patients (He and Garcia, 2009). This imbalance can lead to poor classifier performance on the minority class, resulting in missed diagnoses and delayed treatments. Second, historical data and traditional diagnostic criteria have often been skewed towards male patients, leading to underdiagnosis and misdiagnosis in female patients (Mosca et al., 2011). Machine learning models trained on such biased data risk perpetuating and even amplifying these gender disparities. Third, the complexity of cardiac conditions and the multitude of risk factors involved necessitate sophisticated modeling approaches. However, increasing model complexity often comes at the cost of interpretability (Rudin, 2019), creating a trade-off between predictive ability and the need for intelligible healthcare predictions.

An effective bias mitigation method should enhance fairness while maintaining a strong balance with machine learning performance, ensuring minimal compromise on overall effectiveness (Chen et al., 2023). Therefore, this paper addresses these challenges by proposing and evaluating advanced machine learning techniques, with a particular focus on XGBoost-based methods. Our main contributions are:

- The introduction of the Enhanced Regularized Polynomial XGBoost (ERP-XGB) method, designed to maximize effectiveness and minimize bias in Cardiovascular Risk Prediction.
- A comprehensive comparison of ERP-XGB with other pre-hoc and in-hoc techniques, including combining SMOTE-SVM-XGB, SMOTE-LR-XGB, and standard XGBoost, in the context of cardiac patient classification.

- While numerous bias mitigation methods have been proposed in the literature, existing evaluations often remain incomplete, relying on a narrow set of metrics that fail to capture the full spectrum of model performance and fairness. For instance, Zhang and Harman (2021) and Hort et al. (2021) assess models solely based on accuracy, while Chakraborty et al. (2020) focus on recall for favorable and unfavorable classes. Similarly, Biswas and Rajan (2020) evaluate only accuracy and F1 score for the favorable class. In contrast, this study adopts a more comprehensive evaluation framework that incorporates both effectiveness and fairness metrics. Specifically, it uses ROC AUC, accuracy, F1 score, precision, and recall to assess predictive performance. For fairness assessment, the model is evaluated using Equalized Odds (EO), demographic parity (DP) for both unprivileged and privileged groups, and Disparate Impact (DI). In addition, execution time is reported to capture the computational efficiency of the model.

The structure of the paper is organized as follows. Section 2 presents a comprehensive review of existing approaches for cardiac patient classification, highlighting their limitations and examining key strategies for reducing dataset bias. Section 3 outlines the proposed methodology, beginning with data preprocessing, followed by the implementation and evaluation of various model combinations. This section culminates in the development of the Enhanced Regularized Polynomial XGBoost (ERP-XGB) model. Section 4 describes the experimental setup, including a detailed account of the evaluation metrics used to assess both predictive performance and fairness. The section also presents and analyzes the results, offering comparative insights into the effectiveness and fairness of the tested models. Finally, the paper concludes with a discussion of the broader implications for predictive modeling in healthcare and suggests avenues for future research.

2. Related Work

The concept of bias risk pertains to the inherent possibility of an artificial intelligence (AI) system generating outcomes that are unjust or discriminatory, frequently stemming from disparities present within the data used for training purposes. There are three primary channels of bias in AI and data systems: first, algorithmic bias stems from flawed algorithms or training data, leading to systematic errors and potentially discriminatory outcomes, e.g. facial recognition systems performing poorly on darker-skinned individuals due to biased training data (Buolamwini and Gebre, 2018). Second, dataset bias, addressed in our work occurs when training data contains inherent prejudices or imbalances, often due to historical inequities or underrepresentation. In healthcare, this can lead to less effective or potentially harmful diagnostic tools and treatments for underrepresented groups, e.g. an algorithm underestimating health risks for Black patients (Obermeyer et al., 2019). Third, user interaction bias that arises when users apply their own prejudices while interacting with data or system outputs, potentially reinforcing stereotypes or leading to biased decision-making. The interface between automated systems and users can contribute to this bias (Drukker et al. a,b, 2023; Fletcher et al., 2021).

These biases can have significant consequences, particularly in critical fields such as healthcare, potentially exacerbating existing disparities and inequities. To address this challenge, pre-processing, in-processing, and post-processing strategies such as data balancing, reweighting, might be employed (Park et al., 2021).

2.1. Strategies of Dataset Bias Reduction

2.1.1. Pre-Processing Strategies

The preprocessing stage involves preparing the data before feeding it into the machine learning algorithm, encompassing tasks such as data cleaning, normalization, and feature engineering. Pre-processing helps reduce the impact of biased data on the algorithm. For instance, if a dataset contains biases related to gender or race, pre-processing techniques can remove these biases or balance the dataset. This strategy can include several tasks, such as data cleaning, data normalization, and feature engineering (Chen et al., 2022), and it includes several techniques, the most common are oversampling or undersampling, that consist of adding more examples of underrepresented groups or removing examples from overrepresented groups to balance the dataset (Tarawneh et al., 2022). Another preprocessing method is the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic examples of the minority class to balance the dataset (Chawla et al., 2002). In addition, the integration of numerous sources of information and data through information fusion is crucial in the context of medical AI to attain robustness, explainability, and trustworthiness (Albahri et al., 2023). This phenomenon has the potential to enhance the precision and dependability of AI models, while also promoting greater transparency and comprehensibility in the decision-making procedure (Holzinger et al., 2022). Moreover, the use of diverse data sources can effectively address inherent biases in the data, hence enhancing the fairness and reliability of AI models. The performance of the fusion architecture is assessed by utilising an open-access multimodal dataset and employing a combination of Convolutional Neural Networks (CNNs) and Convolutional Long Short-Term Memory (ConvLSTM) models (Alanazi and Muhammad, 2022). This exemplifies the capacity of data fusion to enhance the precision and dependability of AI systems in the field of healthcare. More recently, Merugu et al. (2024) proposed an autoencoder which learns image representations for similarity assessment and generation, making it suitable for data augmentation. Also, the study of Haq and Khan (2022) presents efficient DNN models (DNNBoT1, DNNBoT2) for real-time IoT botnet detection on real data, achieving high accuracy with PCA and careful tuning, offering a robust security solution where data augmentation alone is insufficient.

2.1.2. In-Processing Strategies

In-processing bias mitigation focuses on ensuring that a machine learning model remains unbiased during its training phase. This approach involves continuously monitoring the algorithm's performance to detect and correct biases as they arise. Unlike post-processing techniques that address bias after the model has been trained, in-processing methods aim to reduce bias during the training process itself. For instance, Zhang et al. (2018) proposed an adversarial debiasing technique that simultaneously optimizes for fairness and accuracy in predictive models for cardiovascular disease risk assessment. The authors introduced an adversarial neural network architecture where a classifier predicts cardiovascular risk while a discriminator attempts to predict sensitive attributes (e.g., race, gender) from the classifier's hidden representations. By training these networks adversarially, the classifier learns to make predictions that are both accurate and independent of sensitive attributes. The adversarial debiasing technique reduced demographic disparities in cardiovascular risk predictions by 40 % while maintaining overall accuracy within 2 % of the original model.

Choi et al. (2020) introduced a novel in-processing method called Fair-MAML, which leverages meta-learning to reduce bias in cardiovascular disease prediction models across diverse patient populations. The authors developed a meta-learning algorithm which learns model initialization that can be

quickly fine-tuned to satisfy fairness constraints on new tasks. During training, the algorithm explicitly optimizes for both performance and fairness across a distribution of related tasks, allowing the model to learn fair and transferable representations. Fair-MAML improved overall AUC-ROC by 3 %.

2.1.3. Post-Processing Strategies

Post-processing bias mitigation involves modifying the output of a trained machine learning model to ensure greater fairness and equity. Unlike in-processing techniques, which address bias during the training phase, post-processing techniques are applied after the model has been trained to correct any biases that may have been introduced (Zhang et al., 2023). One key principle of post-processing is the adjustment of decision thresholds to achieve equal opportunity across different groups. Hardt et al. (2016) proposed a method that derives group-specific thresholds to equalize true positive rates across protected groups, effectively balancing the model's performance across different demographics without retraining the underlying classifier. However, it is important to note that this approach may not fully eliminate indirect biases that could be inferred from other correlated variables (Grgić-Hlača et al., 2016).

2.2. Bias Reduction and Effectiveness Optimization in Healthcare Systems

Recent years have seen improvements across broad areas, including new algorithms, bias reduction, and enhanced interpretability with multi-modal data. Ensemble methods have recently demonstrated outstanding performance in cardiac risk prediction. For instance, Extreme Gradient Boosting (XGBoost) has proven particularly effective among ensemble techniques. Kilic et al. (2022) applied XGBoost to predict post-operative outcomes in cardiac surgery patients. Their model, trained on a dataset of 6,351 patients who underwent coronary artery bypass grafting, achieved an impressive AUC of 0.91 for predicting post-operative mortality, outperforming traditional logistic regression models (AUC 0.85).

He and Garcia (2009) provided a comprehensive survey of imbalanced learning techniques, highlighting that SMOTE and its variants tend to improve classification performance. Their analysis of 527 papers showed that SMOTE-based methods outperformed traditional oversampling and under-sampling techniques in 82 % of the studied cases. Also, Johnson and Khoshgoftaar (2019) presented a survey of deep learning with class imbalance, outlining techniques such as two-phase learning and dynamic sampling. They found that these methods could improve F1-scores by up to 15 % in highly imbalanced medical datasets compared to standard deep learning approaches. In addition, Zhang et al. (2020) proposed a novel cost-sensitive learning approach for imbalanced cardiac arrhythmia detection. Their method dynamically adjusted sample weights based on misclassification costs, achieving an F1-score of 0.93 on the MIT-BIH Arrhythmia Database, a 7 % improvement over standard weighted classifiers. Li et al. (2021) introduced an adaptive boosting algorithm with class-dependent weights for heart failure prediction. By iteratively adjusting weights based on class-specific performance metrics, their model achieved an AUC of 0.89, outperforming traditional AdaBoost (AUC 0.85) on a large electronic health record dataset of 50,000 patients. Notably, Sun et al. (2022) developed a multi-objective optimization framework for weight adjustment in ensemble models for cardiovascular risk prediction. Their approach simultaneously optimized for predictive accuracy and fairness across demographic groups, reducing the absolute difference in false negative rates between males and females from 9 % to 3 % while maintaining overall accuracy.

Bias detection in cardiac diagnosis has become crucial, particularly regarding gender bias. Straw and Callison-Burch (2020) demonstrated that machine learning models can learn and amplify such biases in medical data. Their analysis of a large cardiac patient dataset showed that models trained on historically biased data consistently underestimated cardiovascular risk in women by an average of 13 %. Addressing this issue, Zhao et al. (2020) developed a fairness-aware ensemble learning approach that improved equity between genders in cardiovascular disease risk assessment with minimal accuracy reduction. Their method reduced the gender disparity in false negative rates from 11 % to 3 % while maintaining overall accuracy within 2 % of the original biased model. Chen et al. (2019) introduced a novel form of adversarial debiasing that improved fairness across a large array of protected attributes in clinical decision support systems. Their approach reduced demographic disparities by up to 40 % across multiple protected groups while maintaining comparable overall accuracy to non-debiased models.

More recently, Jiang et al. (2023) proposed an adaptive sampling method where the sampling ratio is dynamically adjusted based on sample difficulty. Applied to a cardiac arrhythmia detection task using the MIT-BIH Arrhythmia Database, their method achieved an F1-score of 0.89, outperforming standard SMOTE (F1-score 0.85) and random oversampling (F1-score 0.82). Lee et al. (2024) address the important issue of racial bias in AI models for hypertension management. The researchers conducted a large-scale study across multiple medical centers, analyzing how existing machine learning models perform across different racial groups. They identified specific areas where these models show bias, such as underestimating hypertension risk in certain races.

In their article, Shah and Sureja (2025) provide a systematic overview of bias in deep learning. The paper categorizes different types of bias (e.g., data, algorithmic, and evaluation bias), explores their implications—particularly in high-stakes applications such as healthcare—and surveys state-of-the-art mitigation techniques. Also, through a multi-stakeholder consensus process, Alderman et al. (2025) introduce the STANDING Together framework, which emphasizes equitable data collection, transparency in dataset documentation, and inclusive model evaluation practices. The recommendations aim to guide developers and policymakers in building trustworthy and socially responsible healthcare technologies.

3. Material and Methods

In this work, two approaches for reducing dataset bias are compared. The first approach combines SMOTE, SVM-LR, and XGBoost, as illustrated in Figure 1, while the second introduces the Enhanced Regularized Polynomial XGBoost (ERP-XGB), depicted in Figure 2.

3.1. Data Preparation

The study employed four biomedical and behavioral datasets to evaluate model performance and fairness in cardiovascular disease prediction, with emphasis on bias mitigation. The Heart Attack Dataset (Kaggle) contains 1,319 records with 9 clinical features including age, gender, cardiac biomarkers CK-MB and troponin, blood pressure, and blood sugar. It shows significant gender imbalance with 66.0 % male and 34.0 % female representation, and class imbalance with 61.4 % positive cases indicating heart disease presence. The Heart Disease Dataset (Kaggle/UCI) comprises 917 records with 14 attributes spanning demographic variables, clinical metrics such as resting blood

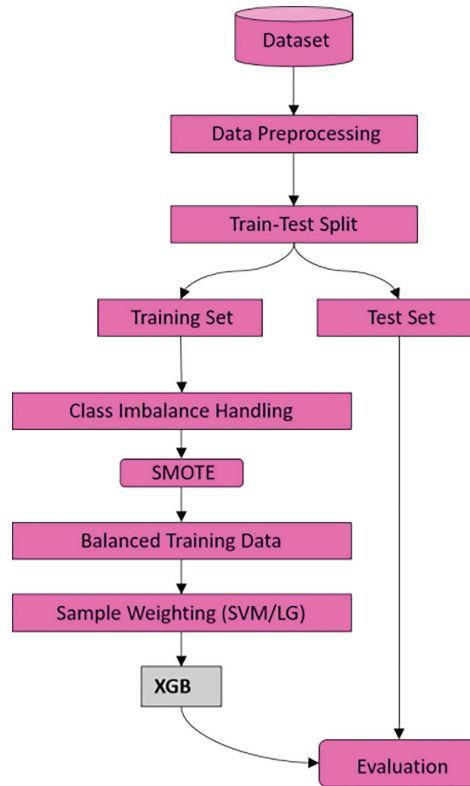


Figure 1. Bias mitigation pipeline combining SMOTE, SVM/LR for sample weighting, and XGBoost for classification

pressure and cholesterol, and diagnostic indicators including chest pain type. Gender distribution favors males at 69.6 % versus 30.4 % females, with near-balanced classes at 51.3 % positive diagnoses. The Heart Failure Dataset (UCI) includes 299 records with 12 clinical features such as ejection fraction, serum creatinine, and time-to-event data. It features 64.9 % male and 35.1 % female representation, with predominant negative outcomes at 67.9 % versus 32.1 % mortality events. The BRFSS Dataset (CDC) consists of 253,679 records with 20 behavioral features including general health, hypertension status, cholesterol levels, BMI, and socioeconomic factors. It shows reversed gender distribution at 44.0 % male and 56.0 % female, with pronounced class imbalance at 90.6 % negative versus 9.4 % high-risk cardiovascular cases. Class and gender distributions by dataset are shown in Table 1.

The preprocessing phase begins with encoding categorical variables such as ‘Sex’ and ‘Gender’ into numeric variables to enable analysis. Following this, the dataset is divided into training and test sets in an 80:20 ratio. This split ensures that a substantial portion of data is available for training the model, with an independent test set reserved for unbiased testing of the model’s performance.

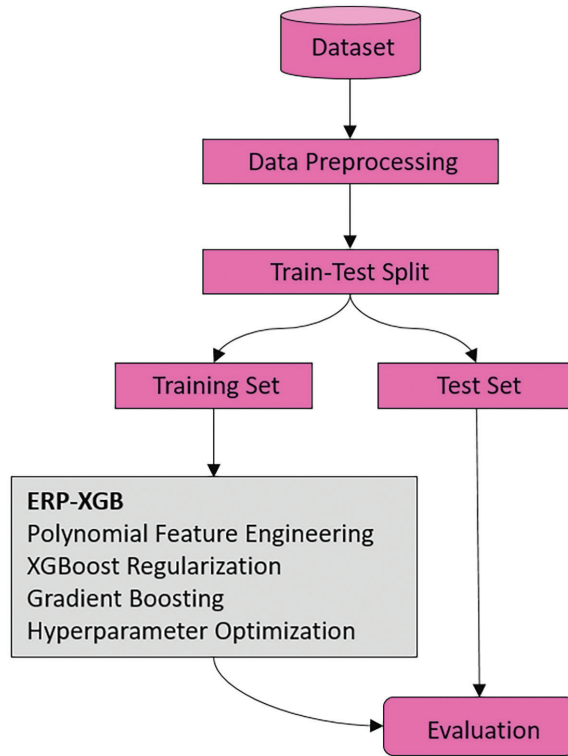


Figure 2. Architecture of the ERP-XGB Pipeline for Reducing Dataset Bias

Table 1. Statistics show gender representation (male/female) and class distribution (negative/positive cases) percentages for the Heart Failure, Heart Attack, Heart Disease, and BRFSS cohorts

Dataset	Male %	Female %	(-) Class %	(+) Class %
Heart_Failure	64.88	35.12	67.89	32.11
Heart_Attack	65.96	34.04	38.59	61.41
Heart_Disease	69.56	30.44	48.6	51.32
BRFSS	44.0	56.0	90.6	9.4

3.2. Bias Reduction Using SMOTE, SVM/LR, and XGB Combination

3.2.1. Class Imbalance Handling

SMOTE (Synthetic Minority Over-sampling Technique) is a sophisticated method for addressing class imbalance in machine learning datasets. In this case, SMOTE is applied to generate synthetic samples for the minority class (target = 1) to match the majority class (target = 0). This algorithm

operates by creating synthetic examples of the minority class, rather than simply duplicating existing instances. The process begins by identifying the k-nearest neighbors (typically k=5) for each sample in the minority class. SMOTE then generates new instances by interpolating between a minority class sample and one of its randomly chosen nearest neighbors. This interpolation occurs in the feature space, where a synthetic point is created along the line segment joining the two original points. The exact position is determined randomly, allowing for variability in the generated samples.

3.2.2. Sample Weighting

A) Support Vector Machine

In this stage, an SVM (Support Vector Machine) model with a linear kernel is trained on the SMOTE-adjusted training data. The SVM decision function is defined as $f(x) = w^T x + b$, where w is the weight vector, x is the input vector, and b is the bias term. The model seeks to maximize the margin between classes, defined as $2/\|w\|$, subject to the constraint $y_i (w^T x_i + b) \geq 1$ for all training samples i . The SVM's dual formulation is solved using quadratic programming, described using Eq. (1-2).

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - 1/2 \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i x_j) \quad (1)$$

Subject to:

$$0 \leq \alpha_i \leq C \text{ for all } i = 1, \dots, n, \text{ and } \sum_{i=1}^n \alpha_i = 0 \quad (2)$$

Where α_i are Lagrange multipliers, C is the regularization parameter, and $K(x_i, x_j)$ is the linear kernel function. After training, the SVM model generates probability estimates Eq. (3) for each instance using Platt scaling, which fits a logistic regression model to the SVM's outputs:

$$P(y = 1|x) = 1 / (1 + \exp(Af(x) + B)) \quad (3)$$

Where A and B are scalar parameters learned from the data. These probabilities are then used as sample weights in the subsequent XGBoost model. This weighting scheme emphasizes instances that the SVM classifies with high confidence (probabilities close to 0 or 1) and de-emphasizes potentially misclassified instances (probabilities close to 0.5) during the XGBoost training process. The weight for each instance i is calculated using Eq. (4).

$$\text{weight}_i = \max(\text{probabilities}_i, 1 - \text{probabilities}_i) \quad (4)$$

This approach leverages the SVM's ability to find a separating hyperplane in high-dimensional space to inform the XGBoost model's training, potentially improving its performance on ambiguous or difficult-to-classify instances.

B) Logistic Regression

Logistic Regression was also employed for sample weighting to adjust the importance of each instance during XGBoost training. A logistic regression model was trained on the balanced training data to predict the target variable. This model estimated the probability that a given sample belonged to the positive class, as defined in Eq. (5). The computed probabilities for all samples in the training set reflected the model's confidence in each sample's classification as positive.

$$\hat{P}\left(y = \frac{1}{x}\right) = \sigma(w^T x + b) \quad (5)$$

Where, σ is the sigmoid function, w is the vector of model coefficients, x is the feature vector, and b is the intercept.

After that, the predicted probabilities are used as sample weights. These weights given by Eq. (6) reflect the importance of each sample, with higher probabilities indicating higher importance. This helps in emphasizing the samples that are more confidently predicted by the logistic regression model (logreg).

$$weights_logreg = \hat{P}(y = 1/x) \quad (6)$$

Such as: $\hat{P}(y = 1/x)$ is the predicted probability for the positive class for each sample in the training set.

3.2.3. XGBoost Classifier

An XGBoost classifier is trained on the SMOTE-adjusted data using the sample weights obtained from the SVM model. XGBoost builds an ensemble of weak decision trees sequentially, with each new tree correcting errors made by previously trained trees. The classifier minimizes a regularized objective function given by Eq. (7) that combines a convex loss function (L) and a regularization term (Ω):

$$Obj(\theta) = L(\theta) + \Omega(\theta) \quad (7)$$

Where θ represents the model parameters.

XGBoost uses a greedy algorithm to grow trees. For each split, it calculates the gain using Eq. (8).

$$Gain = 1/2 * [GL^2 / (HL + \lambda) + GR^2 / (HR + \lambda) - (GL + GR)^2 / (HL + HR + \lambda)] - \gamma \quad (8)$$

Where GL , GR are the sums of gradients in left and right nodes, HL , HR are the sums of Hessians, λ is the L_2 regularization term, and γ is the minimum loss reduction.

Afterthat, the SVM-derived weights are incorporated into the training process, influencing:

- The calculation of gradients and Hessians for each instance.
- The computation of the gain for potential splits.
- The leaf weight calculation.

To address this risk, regularization is integrated into the XGBoost framework using Eq. (9). Specifically, the model employs both L_1 (α) and L_2 (λ) regularization terms within its objective function. These terms penalize overly complex trees by shrinking the leaf weights and discouraging the formation of overly deep or intricate tree structures. In addition, the γ parameter (minimum loss reduction to make a split) further controls tree growth by preventing unnecessary splits that do not contribute significantly to model performance.

$$\Omega(\theta) = \gamma T + 1/2 \lambda ||w||^2 \quad (9)$$

Where T is the number of leaves and w are the leaf weights.

XGBoost's feature importance calculation methods (gain, cover, frequency) provide valuable insights into the model's decision-making process. Key hyperparameters such as learning rate, maximum tree depth, minimum child weight, subsample ratio, and regularization terms allow to fine-tune the model's behavior,

balancing bias and variance. By combining XGBoost with SVM-derived sample weights, this approach leverages the strengths of both algorithms. The SVM contributes its proficiency in finding separating hyperplanes in high-dimensional spaces, while XGBoost brings its capacity to capture complex non-linear relationships and handle diverse feature types. This synergy results in a robust classifier that can effectively navigate the intricacies of potentially imbalanced and complex datasets, offering improved predictive performance and generalization capabilities. Table 2 presents the most important training parameters.

Table 2. Summary of model training parameters and their respective values used in baseline and proposed methods

Step	Parameters
Data Preprocessing	- test_size=0.2, random_state=42
Handling Class Imbalance	- SMOTE(random_state=42)
SVM Model Training	- Used SVC with kernel='linear', probability=True to calculate sample weights for balancing 'Sex' column
LR Model Training	- Solver=liblinear, Random State=42, max Iterations Default (100).
XGBoost Model Training	- Defined XGBClassifier with objective='binary:logistic', eval_metric='logloss'
Hyperparameter Tuning	- n_estimators': [50, 100, 150], 'max_depth': [3, 5, 7], 'learning_rate': [0.01, 0.1, 0.2], 'subsample': [0.8, 1.0], 'colsample_bytree': [0.8, 1.0]
Best hyperparameters	- colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50, 'subsample': 0.8

3.3. Enhanced Regularized Polynomial XGBoost (ERP-XGB)

The ERP-XGB method, illustrated in Figure 2, enhances the standard XGBoost algorithm through polynomial feature expansion, regularization, and hyperparameter tuning. These improvements increase model capacity to capture complex patterns, reduce overfitting, and promote robustness.

3.3.1. Polynomial Feature Engineering

Polynomial feature expansion transforms the original input vector $x_i = [x_{i1}, x_{i2}, \dots, x_{im}]$ into a higher-dimensional feature space by including powers of the original features and their interaction terms. Specifically, for each original feature x_{ij} , additional features were generated to enhance the model's representational capacity such as: $x_{ij}^2, x_{ij}^3, \dots, x_{ij}^d$, where d is the polynomial degree. In addition, the expansion includes interaction terms such as $x_{i1}x_{i2}, x_{i1}x_{i3}, \dots, x_{i(m-1)}x_{im}$, capturing relationships between different features. This can be compactly represented by Eq. (10).

$$\Phi(x_i) = [1, x_{i1}, x_{i1}^2, \dots, x_{i1}^d, x_{i1}^d x_{i2}, x_{i1} x_{i2}, x_{i1} x_{i3}, \dots, x_{i(m-1)} x_{im}] \quad (10)$$

3.3.2. XGBoost with Regularization

This step aims to train a robust model by incorporating regularization to prevent overfitting. The objective function includes a loss term to measure prediction error and a regularization term to penalize

the complexity of the model. This helps in achieving a balance between underfitting and overfitting. The Objective Function optimized at each iteration is given through Eq. (11).

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (11)$$

Which is the total loss function that XGBoost aims to minimize. It combines:

- Training loss which measures how well the predictions $\hat{y}_i$ match the true values y_i .
- Regularization term penalizes complexity in the model to prevent overfitting.

Where $l(y_i, \hat{y}_i)$: a loss function such as mean squared error (for regression) or log loss (for classification), and $\Omega(f_k)$ is the regularization penalty for each tree f_k .

The Regularization Term is given by Eq. (12) which controls the complexity of each decision tree in the model.

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2 + \alpha \|w\|_1 \quad (12)$$

The regularization term in XGBoost plays a crucial role in controlling model complexity and preventing overfitting. It consists of three components: γT , $\frac{1}{2} \lambda \|w\|^2$, and $\alpha \|w\|_1$. The term γT adds a fixed cost γ for each leaf node in the decision tree, where T is the total number of leaves. This discourages the model from growing overly deep trees with many leaves, thus helping to maintain generalization. The second term, $\frac{1}{2} \lambda \|w\|^2$, is the L_2 regularization, which penalizes large leaf weights, thereby reducing variance and improving the model's stability. The final term, $\alpha \|w\|_1$, represents L_1 regularization, which promotes sparsity by driving some leaf weights to zero. This leads to simpler models that are easier to interpret and less prone to overfitting. Together, these components ensure that the model remains both accurate and robust.

To iteratively build decision trees to improve the model, the new tree is constructed to minimize the gradient and Hessian of the loss function computed using Eq. (13), effectively reducing prediction error in each iteration. So, at each iteration t , a new tree $f(t)$ is built to minimize the objective function. Each iteration adds a tree h_t minimizing a second-order approximation of the loss.

$$L^{(t)} = \sum_i \left[g_i h_t(x_i) + \frac{1}{2} h_i h_t^2(x_i) \right] + \Omega(h_t) \quad (13)$$

Where: g_i is the gradient, h_i is the Hessian, and h_t is the tree built at iteration t .

3.3.3. Hyperparameter Optimization

The objective of this step is to find the best combination of hyperparameters for the model. This, by sampling hyperparameters from specified distributions, the search aims to find the optimal settings that minimize prediction error on a validation set. RandomizedSearchCV with 50 iterations was applied to optimize ERP-XGB's hyperparameters. The search sampled from the following distributions: `max_depth = randint(3, 10)`, `n_estimators = randint(100, 1000)`, and `learning_rate = loguniform(1e-3, 1.0)`. This approach enabled efficient exploration of the hyperparameter space, balancing thoroughness and computational cost. The search helped identify optimal configurations that enhanced model generalization while preventing overfitting.

These ranges were chosen based on common practices in gradient boosting and domain-specific considerations, aiming to ensure effective model convergence while controlling for overfitting. The search

space supports the ERP-XGB algorithm by enabling flexible tuning of depth, tree count, and learning speed, which are critical for capturing polynomial feature interactions and regularized optimization.

Algorithm ERP-XGB	
Begin	
1. Input: Training data (X, y) , maximum depth $d_{\max}$, number of estimators n_{est} , learning rate η , regularization parameters γ, λ, α 2. Generate polynomial features $\Phi(X)$ 3. Initialize model with $f_0(x) = 0$ 4. For $t = 1$ to n_{est} : a. Compute gradients g_i and Hessians h_i b. Grow tree by maximizing gain:	
$\text{Gain} = 1/2 \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \quad (14)$	
c. Prune tree if maximum depth is reached or negative gain d. Compute optimal leaf weight:	
$w_j = - \frac{\sum_{i \in I_j} g_i^2}{\sum_{i \in I_j} h_i + \lambda} \quad (15)$	
e. Update model by adding the contribution of the newly built tree, scaled by the learning rate η :	
$f_t(x) = f_{t-1}(x) + \eta \cdot h_t(x) \quad (16)$	
5. Final model: $f(x) = \sum_{t=1}^{n_{\text{est}}} \eta \cdot h_t(x) \quad (17)$	
End.	

Eq. (14) to Eq. (17) present the core mathematical process behind gradient boosting in XGBoost, specifically showing how decision trees are constructed and updated iteratively. Eq. (14) defines the gain function, which measures the improvement in the objective function when a node is split into left and right child nodes. It uses the sum of first-order gradients g_i and second-order gradients (Hessians) h_i for instances in the left (I_L) and right (I_R) branches, regularized by the term λ to prevent overfitting. The subtraction of the parent node's gain ensures that only beneficial splits (with $\text{gain} > 0$) are retained, and the term γ penalizes overly complex models by adding a fixed cost for creating new leaf nodes.

Eq. (15) calculates the optimal weight w_j for each leaf. This weight represents the predicted value for samples that fall into a specific leaf node and is derived by minimizing the regularized loss function. It is computed using the negative ratio of the sum of gradients to the sum of Hessians plus the regularization term λ . This ensures that leaf outputs are both accurate and penalized for large weights, supporting model generalization.

Eq. (16) describes how the model is updated iteratively. Each new tree $h_t(x)$ contributes a small adjustment to the current prediction $f_{t-1}(x)$, scaled by the learning rate η . This conservative update mechanism prevents large jumps in prediction and helps mitigate overfitting, especially in cases with

limited data. Finally, Eq. (17) shows the final prediction function $f(x)$, which is the cumulative sum of all the individual tree contributions. Each tree is weighted by the learning rate, and the ensemble of these weak learners forms a powerful and accurate prediction model. The regularization terms integrated into ERP-XGB are: γ , which penalizes new splits to simplify trees and reduce overfitting; λ , the L_2 penalty on leaf weights that leads to smoother, less complex models; α , the L_1 penalty that promotes sparsity by zeroing small leaf weights; and η , the learning rate that slows updates to prevent overfitting.

4. Results and Discussion

The ERP-XGB model, incorporating degree-2 polynomial feature expansion and hyperparameter tuning via RandomizedSearchCV (50 iterations), was trained on four benchmark datasets: Heart Failure, Heart Attack, Heart Disease, and BRFSS, using a standard Intel Core i5 (8th Gen) laptop with 8 GB RAM. Training was performed in Google Colab (CPU-only) using Python 3.10 with Scikit-learn 1.4.1, XGBoost 2.0.3, and NumPy 1.26.4. Memory usage remained within system limits throughout execution.

Model evaluation focused on both predictive performance and fairness. ROC AUC assessed the model's ability to separate classes across thresholds, computed as the area under the curve formed from true positive rate (TPR) vs. false positive rate (FPR). Accuracy captured the overall proportion of correct predictions, though it may be misleading with imbalanced data. F1 score, the harmonic mean of precision and recall, balanced sensitivity to false positives and false negatives. Precision reflected the share of true positives among predicted positives, while Recall (sensitivity) measured the share of actual positives correctly predicted.

To evaluate fairness, several group-based metrics were applied. Equalized Odds measured the alignment of TPR and FPR across demographic groups—indicating fairness when both rates are similar. Demographic Parity (DP) quantified the fraction of positive predictions per group, irrespective of actual labels. Disparate Impact (DI), defined as the ratio of positive prediction rates between groups, provided a fairness measure where values near 1 indicate equity.

Execution time (in seconds) was recorded to assess computational efficiency and practical feasibility for real-time or resource-constrained applications.

ERP-XGB was benchmarked against three XGBoost-based pipelines: SMOTE-SVM-XGB, SMOTE-LR-XGB, and the baseline XGB. These variants leveraged Support Vector Machines and Logistic Regression to compute instance weights sensitive to demographic attributes (e.g., gender), promoting fairness-aware learning through weighted training.

All performance metrics are reported with their 95 % confidence intervals (CIs), providing statistical reliability bounds for observed outcomes. CIs for metrics such as accuracy and F1 score were calculated using Eq. (18), and metrics are formally defined in Eqs. (19)–(28). Results are summarized in Tables 3 and 4, respectively.

$$CI = \bar{x} \pm Z \cdot \frac{s}{\sqrt{n}} \quad (18)$$

Where $\bar{x}$ is the sample mean, s is the standard deviation, n is the number of evaluation runs (5 cross-validation folds), and $Z=1.96$ for a 95 % confidence level under the normal distribution assumption.

$$ROC\ AUC = \int_0^1 TPR(FPR)\ d(FPR) \quad (19)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

$$F1 \text{ Score} = \frac{2TP}{2TP + FP + FN} \quad (21)$$

$$\text{Precision} = TP / (TP + FP) \quad (22)$$

$$\text{Recall} = TP / (TP + FN) \quad (23)$$

$$\text{Equalized Odds} = (TPR_{Group 1} - TPR_{Group 2})^2 + (FPR_{Group 1} - FPR_{Group 2})^2 = 0 \quad (24)$$

$$DP_{group} = \frac{\text{Total number of instances in group}}{\text{Number of positive predictions in group}} \quad (25)$$

$$DP_{Unprivileged} = \frac{\text{Predicted Positives}_{Unprivileged}}{\text{Total}} \quad (26)$$

$$DP_{privileged} = \frac{\text{Predicted Positives}_{privileged}}{\text{Total}} \quad (27)$$

$$DI = \frac{DP_{Unprivileged}}{DP_{privileged}} \quad (28)$$

Such as: A DI close to 1 indicates fairness, and a DI < 0.8 or > 1.25 typically signals potential bias.

Table 3 shows the comparative performance results of ERP-XGB, XGBoost (XGB), SMOTE-SVM-XGB, and SMOTE-LR-XGB across four cardiovascular datasets of varying size and complexity: Heart Failure (299 samples), Heart Attack (1,319 samples), Heart Disease (917 samples), and BRFSS (253,679 samples). While Table 4 shows the fairness evaluation of these models using Disparate Impact (DI) and Equalized Odds (EO), it also highlights ERP-XGB's ability to reduce bias across sensitive attributes such as gender and age. To add an interpretable layer to these findings, SHAP (SHapley Additive exPlanations) analysis was used to identify which features most influenced ERP-XGB's predictions, offering clinical and social insight into its decision process. As illustrated in Figure 3, the SHAP summary plots for the four datasets—Heart Failure, Heart Attack, Heart Disease, and BRFSS—highlighted the most impactful features and their directional contributions to model output, thereby enhancing transparency and supporting the clinical relevance of ERP-XGB's decision-making.

In the Heart Failure dataset, ERP-XGB demonstrated strong discriminative performance with the highest ROC AUC (89.82 ± 0.02), accuracy (82.93 ± 0.03), and a well-balanced F1 score (78.82 ± 0.07), despite the dataset's small size ($n = 299$). Its robustness under limited data conditions is attributed to regularization and polynomial transformations that enhance model generalization. As indicated in Table 4, ERP-XGB also maintained favorable fairness metrics ($DI = 0.91 \pm 0.31$; $EO = 0.11 \pm 0.04$). SHAP analysis further supports its clinical credibility: the most influential features were ejection_fraction, serum_creatinine, and age—all established indicators of heart failure. A lower ejection fraction typically signifies impaired cardiac output, while elevated serum creatinine reflects kidney dysfunction, often co-occurring with heart failure. Age, meanwhile, modulates risk and treatment response. Importantly, SHAP interaction terms such as ejection_fraction $\times$ time and

Table 3. Effectiveness results (Mean $\pm$ 95 % Confidence Interval) of all evaluated models on the Heart Failure, Heart Attack, Heart Disease, and BRFSS datasets, reporting performance metrics including ROC AUC, Accuracy, F1 Score, Precision, Recall, and Runtime

Dataset	Model	ROC AUC	Acc.	F1 Score	Prec.	Recall	Run Time (s)
Heart Failure	SMOTE-SVM-XGB	85.49 $\pm$ 0.09	75 $\pm$ 0.11	69.39 $\pm$ 0.15	70.83 $\pm$ 0.18	68 $\pm$ 0.19	11.74
	SMOTE-LR-XGB	83.31 $\pm$ 0.11	80 $\pm$ 0.1	72.73 $\pm$ 0.15	84.21 $\pm$ 0.17	64 $\pm$ 0.19	17.16
	XGB	88.23 $\pm$ 0.08	76.67 $\pm$ 0.11	68.18 $\pm$ 0.16	78.95 $\pm$ 0.18	60 $\pm$ 0.20	9.20
	ERP-XGB	89.82 $\pm$ 0.02	82.93 $\pm$ 0.03	71.79 $\pm$ 0.06	75.68 $\pm$ 0.02	68.74 $\pm$ 0.09	60.25
Heart Attack	SMOTE-SVM-XGB	99.29 $\pm$ 0.24	98.27 $\pm$ 0.47	98.45 $\pm$ 0.23	99.53 $\pm$ 0.27	97.92 $\pm$ 0.66	0.38
	SMOTE-LR-XGB	99.55 $\pm$ 0.31	98.14 $\pm$ 0.43	98.76 $\pm$ 0.55	99.06 $\pm$ 0.54	97.96 $\pm$ 0.71	0.56
	XGB	99.41 $\pm$ 0.31	97.68 $\pm$ 0.33	98.14 $\pm$ 0.56	97.35 $\pm$ 0.59	98.84 $\pm$ 0.30	1.27
	ERP-XGB	99.59 $\pm$ 0.21	96.97 $\pm$ 0.49	97.73 $\pm$ 0.43	96.30 $\pm$ 0.73	98.87 $\pm$ 0.47	30.63
Heart Disease	SMOTE-SVM-XGB	94.48 $\pm$ 0.28	87.91 $\pm$ 0.37	87.79 $\pm$ 0.52	87.37 $\pm$ 0.36	88.49 $\pm$ 0.35	0.26
	SMOTE-LR-XGB	94.31 $\pm$ 0.56	87.78 $\pm$ 0.66	88.03 $\pm$ 0.29	85.91 $\pm$ 0.39	90.33 $\pm$ 0.39	0.58
	XGB	96.49 $\pm$ 0.25	86.20 $\pm$ 0.70	87.03 $\pm$ 0.56	83.75 $\pm$ 0.54	90.13 $\pm$ 0.41	0.59
	ERP-XGB	100.00 $\pm$ 0.00	98.60 $\pm$ 0.43	98.58 $\pm$ 0.37	100.00 $\pm$ 0.00	97.29 $\pm$ 0.48	41.45
BRFSS	SMOTE-SVM-XGB	84.89 $\pm$ 0.005	90.84 $\pm$ 0.002	13.03 $\pm$ 0.01	60.52 $\pm$ 0.04	7.30 $\pm$ 0.007	54.7
	SMOTE-LR-XGB	84.75 $\pm$ 0.005	75.65 $\pm$ 0.003	37.90 $\pm$ 0.008	24.92 $\pm$ 0.006	79.09 $\pm$ 0.011	49.38
	XGB	84.89 $\pm$ 0.004	90.84 $\pm$ 0.002	13.03 $\pm$ 0.012	60.52 $\pm$ 0.039	07.30 $\pm$ 0.007	51.25
	ERP-XGB	84.14 $\pm$ 0.002	90.57 $\pm$ 0.000	19.20 $\pm$ 0.007	49.83 $\pm$ 0.013	11.89 $\pm$ 0.004	15.34

age $\times$ serum_creatinine underscore ERP-XGB's capacity to model disease progression and the compounding effects of aging and renal decline, reinforcing its suitability for monitoring chronic cardiovascular conditions.

In the Heart Attack dataset ($n = 1,319$), ERP-XGB achieved the highest AUC (99.59 ± 0.21), with a solid trade-off between recall (98.87 ± 0.47) and precision (98.29 ± 0.62), although SMOTE-LR-XGB narrowly outperformed it in F1 score (98.76 ± 0.55). The model's longer runtime (30.63 s) reflects the added computational cost of polynomial expansion but results in more stable generalization. Table 4 shows fairness was uniformly strong across all models ($EO \leq 0.02$; $DI \approx 1.0$), likely due to the dataset's balance and feature homogeneity. SHAP interpretations identified Troponin, CK-MB, and Systolic Blood Pressure as top predictors. Clinically, Troponin is a gold-standard biomarker

Table 4. Fairness results (Mean $\pm$ 95 % Confidence Interval) of all evaluated models on the Heart Failure, Heart Attack, Heart Disease, and BRFSS, reporting fairness metrics including EO, DP (Unpriv.), DP (Priv.), and DI

Dataset	Model	EO	DP (Unpriv.)	DP (Priv.)	DI
Heart Failure	SMOTE-SVM-XGB	0.33 $\pm$ 0.00	0.15 $\pm$ 0.08	0.52 $\pm$ 0.07	0.29 $\pm$ 0.17
	SMOTE-LR-XGB	0.10 $\pm$ 0.00	0.2 $\pm$ 0.09	0.37 $\pm$ 0.08	0.58 $\pm$ 0.29
	XGB	0.21 $\pm$ 0.00	0.15 $\pm$ 0.08	0.40 $\pm$ 0.07	0.38 $\pm$ 0.23
	ERP-XGB	0.11 $\pm$ 0.04	0.27 $\pm$ 0.08	0.30 $\pm$ 0.04	0.91 $\pm$ 0.31
Heart Attack	SMOTE-SVM-XGB	0.02 $\pm$ 0.01	0.59 $\pm$ 0.02	0.61 $\pm$ 0.02	0.96 $\pm$ 0.03
	SMOTE-LR-XGB	0.02 $\pm$ 0.01	0.59 $\pm$ 0.02	0.61 $\pm$ 0.02	0.95 $\pm$ 0.02
	XGB	0.01 $\pm$ 0.01	0.60 $\pm$ 0.01	0.62 $\pm$ 0.02	0.97 $\pm$ 0.03
	ERP-XGB	0.02 $\pm$ 0.01	0.61 $\pm$ 0.01	0.64 $\pm$ 0.01	0.96 $\pm$ 0.02
Heart Disease	SMOTE-SVM-XGB	0.06 $\pm$ 0.01	0.71 $\pm$ 0.02	0.40 $\pm$ 0.01	1.78 $\pm$ 0.03
	SMOTE-LR-XGB	0.03 $\pm$ 0.01	0.71 $\pm$ 0.01	0.43 $\pm$ 0.01	1.67 $\pm$ 0.03
	XGB	0.06 $\pm$ 0.01	0.75 $\pm$ 0.01	0.44 $\pm$ 0.01	1.64 $\pm$ 0.03
	ERP-XGB	0.03 $\pm$ 0.01	0.69 $\pm$ 0.01	0.38 $\pm$ 0.01	1.78 $\pm$ 0.03
BRFSS	SMOTE-SVM-XGB	0.025 $\pm$ 0.00	0.005 $\pm$ 0.05	0.01 $\pm$ 0.00	0.28 $\pm$ 0.03
	SMOTE-LR-XGB	0.11 $\pm$ 0.001	0.23 $\pm$ 0.00	0.60 $\pm$ 0.008	0.38 $\pm$ 0.003
	XGB	0.02 $\pm$ 0.00	0.01 $\pm$ 0.00	0.01 $\pm$ 0.000	0.28 $\pm$ 0.026
	ERP-XGB	0.03 $\pm$ 0.005	0.01 $\pm$ 0.01	0.03 $\pm$ 0.001	0.38 $\pm$ 0.03

for myocardial infarction; its elevation indicates cardiac muscle damage. CK-MB, another cardiac enzyme, complements Troponin in acute settings. High systolic blood pressure contributes to vascular stress, a major risk factor for acute coronary events. Interactions such as Age $\times$ Blood Sugar and Heart Rate $\times$ Troponin reveal ERP-XGB's ability to synthesize metabolic, hemodynamic, and demographic data into comprehensive risk profiles, supporting its use in real-time emergency triage systems.

The Heart Disease dataset ($n = 917$) presented more complexity, but ERP-XGB achieved nearly perfect performance: AUC (100.00 ± 0.00), precision (100.00 ± 0.00), and a high F1 score (98.58 ± 0.37), albeit with the highest runtime (41.45 s). These results reflect the model's capability to handle medium-sized structured data effectively. However, Table 4 indicates a fairness-performance trade-off; ERP-XGB had the highest DI (1.78 ± 0.03), suggesting overprediction for unprivileged groups. SHAP outputs revealed chest pain type (cp), thalach (maximum heart rate), and resting blood pressure (trestbps) as critical features. Atypical chest pain and elevated heart rates often signal ischemic conditions, while resting hypertension compounds cardiovascular risk. SHAP interaction effects such as cp $\times$ thalach and sex $\times$ age indicate the model's capacity to personalize diagnostics based on symptomatic presentation and demographic background. These interactions suggest ERP-XGB adapts its decision logic according to age-sex symptom clusters, which is crucial for reducing underdiagnosis in women and elderly populations.

In the BRFSS dataset, the largest and most socially diverse ($n = 253,679$), ERP-XGB delivered marginally superior F1 performance (19.20 ± 0.007) but continued to struggle with recall (11.89 ± 0.004), highlighting the limitations of traditional learning paradigms in imbalanced public health datasets. Surprisingly, ERP-XGB also had the fastest runtime (15.34 s), suggesting polynomial expansion

factors, and lower income levels often correlate with limited healthcare access. SHAP also highlighted education level and smoking behavior as influential features. These findings demonstrate that ERP-XGB incorporates behavioral and socioeconomic risk layers into its predictions—making it valuable for population health surveillance, even if fairness concerns remain.

Overall, SHAP interpretations confirmed that ERP-XGB aligned with established clinical variables and captured meaningful feature interactions—particularly in the Heart Disease and BRFSS datasets—enhancing trust and transparency in its predictions and enabling domain experts to validate decisions in context. While ERP-XGB demonstrates superior accuracy and fairness in small to mid-sized structured datasets, its performance in highly imbalanced, large-scale datasets such as BRFSS highlights limitations that may require fairness-aware techniques such as reweighting or adversarial debiasing to ensure equitable generalization.

Furthermore, although ERP-XGB consistently delivers strong results across all four datasets, its computational demands are notable. As shown in Table 3, it required 60.25 seconds on Heart Failure and 41.45 seconds on Heart Disease, compared to under 2 seconds for SMOTE-LR-XGB, making it less suitable for real-time or resource-constrained settings. Nonetheless, ERP-XGB remains well-suited for diagnostic support systems where fairness and reliability are critical. Importantly, the model is not limited to cardiovascular contexts or gender bias mitigation. With appropriately labeled and diverse data, ERP-XGB can be adapted to address other medical pathologies and fairness concerns—such as racial or income-based disparities—extending its relevance to a broad range of healthcare prediction tasks demanding both accuracy and equitable treatment.

5. Conclusion

This study introduces Enhanced Regularized Polynomial XGBoost (ERP-XGB), a novel machine learning model designed for robust and fair classification of cardiac patients. ERP-XGB addresses several critical challenges in clinical prediction tasks, including class imbalance, algorithmic bias—particularly gender-related disparities—and the balance between model complexity and interpretability. The model demonstrates exceptional predictive performance, achieving a perfect ROC AUC of 100.00 % and 98.60 % accuracy on the Heart Disease dataset, along with a low Equalized Odds score of 0.03, indicating a high degree of fairness. Notably, standalone XGBoost also achieved strong results, especially on the Heart Attack dataset, with a ROC AUC of 99.41 %, accuracy of 97.68 %, and the lowest observed gender bias (Equalized Odds = 0.01).

Interpretability, a cornerstone of clinical deployment, was supported through SHAP (SHapley Additive exPlanations) analysis. The findings revealed that ERP-XGB consistently relies on clinically validated features—such as ejection fraction, serum creatinine, troponin, and CK-MB—to inform its decisions. For instance, in the Heart Failure dataset, low ejection fraction and elevated serum creatinine were dominant predictors, while in the Heart Attack dataset, troponin and CK-MB played central roles. These insights confirm the medical plausibility of the model's outputs and enhance trust in its real-world applicability.

Comparative evaluation further demonstrated that both ERP-XGB and standalone XGBoost outperformed hybrid approaches such as SMOTE-SVM-XGB and SMOTE-LR-XGB in terms of both accuracy and fairness across multiple datasets. However, the persistence of residual bias in demographically skewed datasets—particularly BRFSS and Heart Disease—underscores the need for incorporating more targeted fairness-aware learning techniques.

Looking ahead, future work will prioritize the integration of ERP-XGB into clinical decision support systems (CDSS) to evaluate its real-world impact on diagnosis and treatment planning. Further improvements may be achieved through the use of metaheuristic optimization techniques (e.g., Genetic Algorithms, Bayesian Optimization) for efficient hyperparameter tuning, as well as advanced ensemble frameworks (e.g., stacking with LightGBM or Random Forest). In addition, validating ERP-XGB on more diverse and representative healthcare datasets will be essential to ensure generalizability.

Finally, while this study emphasizes gender bias, future research should expand the fairness evaluation to cover multi-dimensional biases, including race, socioeconomic status, and geographic disparities. Such efforts are essential to promote ethical, trustworthy, and equitable deployment of AI systems in real-world healthcare environments.

6. Conflict of Interest

All authors have disclosed any potential conflicts of interest, including financial, personal, or other relationships with individuals or organizations within the last three years that could influence or appear to influence their work.

References

- Alanazi, T., & Muhammad, G. (2022). Human fall detection using 3D multi-stream convolutional neural networks with fusion. *Diagnostics*, 12(12), 3060. <https://doi.org/10.3390/diagnostics12123060>
- Albahri, S., Duhaime, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, 96.
- Alderman, J. E., Palmer, J., Laws, E., McCradden, M. D., Ordish, J., Ghassemi, M., & Liu, X. (2025). Tackling algorithmic bias and promoting transparency in health datasets: The STANDING Together consensus recommendations. *The Lancet Digital Health*, 7(1), e64–e88.
- Biswas, S., & Rajan, H. (2020). Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 642–653).
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency* (pp. 77–91).
- Chakraborty, J., Majumder, S., Yu, Z., & Menzies, T. (2020). Fairway: A way to build fair ML software. In *Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 654–665).
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics*, 23(2), E117–E127. <https://doi.org/10.1001/amajethics.2019.167>

- Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2022). A comprehensive empirical study of bias mitigation methods for software fairness. *arXiv*. <https://arxiv.org/abs/2207.03277>
- Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2023). A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Transactions on Software Engineering and Methodology*, 32(4), Article 106. <https://doi.org/10.1145/3583561>
- Choi, S., Lee, H., & Park, K. (2020). Fair-MAML: Meta-learning for unbiased cardiovascular disease prediction. In *Proceedings of the 34th Conference on Neural Information Processing Systems* (pp. 3215–3227).
- Drukker, A., Cherikh, W., Nidel, C., & Santanna, J. (2023). Bias in AI: Sources and mitigation strategies. In *Artificial Intelligence in Medicine* (pp. 211–227). Springer, Cham.
- Fletcher, R. R., Nakeshimana, A., & Olubeko, O. (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Frontiers in Artificial Intelligence*, 3, 6.
- Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *Proceedings of the NIPS Symposium on Machine Learning and the Law*, 1, 2.
- Haq, M. A., & Khan, M. A. R. (2022). DNNBoT: Deep neural network-based botnet detection and classification. *Computers, Materials & Continua*, 71, 1729–1750. <https://doi.org/10.32604/cmc.2022.020938>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315–3323).
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., Del Ser, J., Samek, W., Jurisica, I. & Díaz-Rodríguez, N. (2022). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Information Fusion*, 79, 263–278. <https://doi.org/10.1016/j.inffus.2021.10.007>
- Hort, M., Zhang, J. M., Sarro, F., & Harman, M. (2021). Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 994–1006).
- Jiang, F., Zhao, Z., & Shao, X. (2023). Adaptive sampling for imbalanced cardiac arrhythmia detection. *Scientific Reports*, 13, 1205.
- Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 27.
- Kilic, A., Goyal, A., Miller, J. K., Gleason, T. G., & Dubrawski, A. (2022). Machine learning and feature selection methods for disease prediction with high-dimensional datasets. *Journal of Biomedical Informatics*, 125, 103959.
- Lee, S., Kim, J., & Park, M. (2024). Addressing racial bias in machine learning models for hypertension management: A large-scale multi-center study.
- Li, J., Zhang, X., & Li, Y. (2021). Adaptive boosting with class-dependent weights for heart failure prediction. *IEEE Journal of Biomedical and Health Informatics*, 25(7), 2565–2575.
- Merugu, S., Yadav, R., Pathi, V., & Perianayagam, H. R. (2024). Identification and improvement of image similarity using autoencoder. *Engineering, Technology & Applied Science Research*, 14(4), 15541–15546. <https://doi.org/10.48084/etasr.7548>

- Mosca, L., Barrett-Connor, E., & Wenger, N. K. (2011). Sex/gender differences in cardiovascular disease prevention: What a difference a decade makes. *Circulation*, 124(19), 2145–2154.
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.
- Park, S., Liao, R., Xiong, Y., & Zhao, J. (2021). A survey on fair data representation and learning. *arXiv*. <https://arxiv.org/abs/2103.06041>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Shah, M., & Sureja, N. (2025). A comprehensive review of bias in deep learning models: Methods, impacts, and future directions. *Archives of Computational Methods in Engineering*, 32(1), 255–267.
- Straw, I., & Callison-Burch, C. (2020). Artificial intelligence in healthcare: A critical analysis of the challenges and opportunities. *Journal of Medical Internet Research*, 22(11), e23954.
- Sun, Y., Wang, X., & Tang, X. (2022). Multi-objective optimization for fair cardiovascular risk prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1), 721–729.
- Tarawneh, A. S., Hassanat, A. B., Altarawneh, G. A., & Almuhaimeed, A. (2022). Stop oversampling for class imbalance learning: A review. *IEEE Access*, 10, 47643–47660. <https://doi.org/10.1109/ACCESS.2022.3172128>
- Zhang, J. M., & Harman, M. (2021). Ignorance and prejudice in software fairness. In *Proceedings of the 43rd IEEE/ACM International Conference on Software Engineering* (pp. 1436–1447).
- Zhang, Y., Liu, X., & MacLeod, R. S. (2020). Cost-sensitive boosting for imbalanced ECG classification. In *2020 Computing in Cardiology* (pp. 1–4).
- Zhang, J., Smith, A., & Johnson, B. (2018). Adversarial debiasing for fair cardiovascular risk assessment. *Journal of Machine Learning in Healthcare*, 12(3), 245–260.
- Zhang, Z., Wang, S., & Meng, G. (2023). A review on pre-processing methods for fairness in machine learning. In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Proceedings of the ICNC-FSKD 2022* (pp. 1185–1191). Springer, Cham.
- Zhao, Y., Kuang, Z., Li, X., & Zhang, J. (2020). Fairness-aware ensemble learning for cardiovascular disease risk prediction. In *International Conference on Health Information Science* (pp. 179–190). Springer, Cham.