SCHD
Structural & Congenital Heart Disease

Tech Science Press

http://doi.org/10.32604/schd.2026.079590

ARTICLE

Predicting Congenital Heart Disease Using Maternal Risk Factors: A Machine Learning Study from an Indian Tertiary Cardiac Care Centre

Shruthi S* and D. Hanumanth Rao Naidu

Department of Mathematical and Computational Sciences, Sri Sathya Sai University for Human Excellence, Sathya Sai Grama, Muddenahalli, Chikkaballapura, Karnataka, India
*Corresponding Author: Shruthi S. Email: shruthi.s@sssuhe.ac.in

Received: 24 January 2026; Accepted: 08 June 2026; Published: 11 June 2026

ABSTRACT: Background: Congenital Heart Disease (CHD) is an abnormality of the heart arising before birth. CHD diagnosis poses a critical challenge, particularly in resource-constrained settings where access to doctors and skilled radiologists is limited. The maternal risk factors contributing to CHD include modifiable and non-modifiable causes. Very few studies mention about these maternal risk factors for the Indian population to build predictive machine learning models for disease forecasting. The aim is to explore the feasibility of predicting CHD occurrence using maternal risk factor data and machine learning models in an Indian context. **Methods:** This research utilizes Indian-origin retrospective case and control data from a hospital, harnessing maternal risk factors for CHD prediction. The study analyzed 1040 preprocessed records with 56 features, which were organized into thematic categories. A diverse set of machine learning algorithms were employed for CHD prediction, including Weighted Support Vector Machine, Weighted Random Forests, Naive Bayes, Artificial Neural Network, Logistic Regression, Decision Tree, Extreme Gradient Boosting, Adaptive Boosting, and Gradient Boost Machine. The dataset was evaluated using stratified train (80%)-test (20%) split, repeated 5-fold cross-validation, and bootstrap-based evaluation. **Results:** The study reveals that tree-based models demonstrated comparatively stronger performance for the identification of CHD, with sensitivity values exceeding 93%, a maximum specificity of 87.5%, and Area Under the Precision–Recall Curve (AUPRC) values greater than 94% for all models. XGBoost achieved the best performance among the evaluated models, with a balanced accuracy of 81.7% and F1 score of 93.4%. Data insights generated by the Shapley Additive Explanations (SHAP) explainable AI framework provide further insight into the features contributing to CHD. **Conclusions:** The maternal risk factor "supplement score" emerged as a significant variable, indicating that the maternal nutritional status plays a critical role in influencing CHD risk. The findings indicate that non-invasive CHD screening based on maternal risk-factor data can be effectively performed using tree-based machine-learning models, with ensemble methods such as gradient boosting and random forest.

KEYWORDS: Congenital heart disease; screening; maternal risk factors; data balancing; machine learning; SHAP

1 Introduction

One of the biggest challenges in terms of neonatal and pediatric morbidity that India and the world at large are facing is congenital malformations. The global prevalence of Congenital Heart Disease (CHD) is 8–12 per 1000 live births [1], and in India, 9 out of 1000 live births account for this disease.

CHD is a multifaceted gross structural abnormality of the heart arising before birth. Though CHD is considered a rare disease that affects less than 1% of the population, more than 0.25 million children are born with this disease in India every year, and one-fifth of these cases are said to have serious defects that

require treatment in the first year of their lives [2]. The issue is serious because of significant constraints in accessing essential diagnostic and treatment facilities. Currently, advanced cardiac care is available only to a minority of such children, considering the cost incurred for the treatment and the scarcity of resources available for the timely intervention. Also, in a recent report published [1], India accounts for the largest absolute number of CHD cases and has consequently emerged as a significant global public health challenge.

CHD etiology of the phenotype can be complex and multimodal [3], having non-modifiable and modifiable causes. The modifiable risk factors are the attributes that, when controlled, can help reduce the CHD risk and are hence considered of prime importance in this study. Modifiable risk factors include nutrition and folic acid supplementation, environmental exposures, and unhealthy lifestyle practices. Non-modifiable risk factors primarily consist of genetic influences such as family history, as well as metabolic disorders, infections, medications used, certain pregnancy-related complications, etc.

With the help of machine learning (ML) algorithms, many studies now use these advanced techniques to improve the diagnosis, prediction, and management of CHD [4]. In accordance with this data-driven approach to healthcare, the proposed study uses Electronic Health Record (EHR) data collected from Sri Sathya Sai Sanjeevani Hospital (SSSSH), Kharghar, Navi Mumbai, India to evaluate different ML techniques and identify the model that performs best for predicting CHD. The primary aim of this study was to develop and internally validate machine-learning models to predict CHD in an Indian context. The secondary aim was to identify important maternal predictors of CHD. The EHR data was classified using ML models: Probabilistic Naive Bayes model (NB), Kernel-based Weighted Support Vector Machine (WSVM), Logistic regression (LR), Tree-based models like Decision Tree (DT), Weighted Random Forest (WRF), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost).

This study makes the following important contributions:

1. Domain-driven feature engineering: A clinically validated dataset was collected from a tertiary cardiac care centre. Instead of handling features one by one, maternal risk factors were grouped into 10 medically approved, easy-to-understand categories, motivated by [5].
2. Comprehensive multi-model evaluation: The study compared nine ML models, including linear, probabilistic, neural, and combined methods, all tested under the same conditions.
3. Explainable AI integration: The study used SHapley Additive exPlanations (SHAP) to show the most important maternal risk factors that contributed to CHD occurrence, making the model's decisions easier for doctors to understand and trust.
4. Clinically oriented evaluation: The evaluation of the models in this work does not rely solely on the commonly used accuracy metric but uses sensitivity, False Negative Rate, and AUPRC to better reflect real-world screening requirements.

The rest of the paper is structured as follows. Section 2 provides a comprehensive review of existing literature. The proposed methodology of Section 3 includes an explanation of the used dataset, the clinically validated feature grouping, the preprocessing and transformation techniques, the data balancing, model selection, evaluation metrics, robustness analysis and explainability methods. Section 4 presents the experimental results together with bootstrap-based uncertainty estimation, statistical validation and model performance comparisons. The explainability analysis of SHAP, together with the importance of Indian origin data and deployment factors, study limitations and future research directions are discussed in Section 5. Section 6 presents the main findings of the study, concluding the paper.

2 Literature Review

Various studies in the past have addressed the problem of identifying risk factors contributing to CHD and the prediction of CHD using the risk factors. The proposed study investigates antenatal maternal risk factors recorded in pediatric CHD cases to predict the occurrence of the disease in offspring. Prior research have had examined the identification of such risk factors, and few of them have used them in CHD prediction. This section provides a review of such studies. This section provides the details of some of those relevant studies.

In a meta-analysis of risk factors for CHD, L Wu et al. [6] included all studies published from 1991 to May 2021 in PubMed, EMBASE, and the Cochrane libraries. Most of the studies reviewed in the paper were observational case-control trials, in which maternal factors, CHD and non-CHD case counts, and study quality were assessed using the Newcastle-Ottawa Scale (NOS). The work also aimed to evaluate the importance of risk factors by statistically estimating odds ratios, providing valuable insights into the associations between maternal risk factors and CHD occurrence. This review study identifies four major risk factors revealing the relationship between the features, such as maternal obesity during pregnancy, smoking during pregnancy, maternal diabetes, and exposure to organic solvents, that lead to increased CHD risk.

In a hospital-based case-control study [7], subjects were selected from birth defect surveillance hospitals in Hunan Province, China, including 119 CHD cases and 239 controls, between 2013 July and 2014 June. The data was collected for 36 CHD-related variables like family history of CHD, pregnancy history (gravidity, history of abnormal reproduction, parity), lifestyle and dietary behavior (smoking, alcohol drinking, intake of milk, protein), and environmental risk factors (chronic disease, pet keeping, folic acid intake). Univariate logistic regression was used to select 15 impacting features for CHD, and a Back-Propagation Neural Network (BPNN) model was used for the prediction of CHD with a training and test data ratio of 85:15. The study aimed to evaluate the neural network model and concluded that the BPNN model used was able to predict CHD risk effectively.

The work by Y Luo et al. [5] deals with non-clinical data and 9 grouped features: mother's age at delivery, annual income, family history, maternal illness history before and during pregnancy, insufficient diet, medication use in pregnancy, unhealthy environment, and unhealthy lifestyle during pregnancy. The study aimed to develop a predictive model using birth defect data collected from 2006 to 2008 across 6 counties in the Shanxi domain, China. Out of 33,831 live births, 78 cases were CHD. This data was used to predict congenital malformations using three methods: WSVM, Logistic Regression (logit), and WRF. The best median Area Under the Receiver Operating Curve (AUC) of 0.817 and weighted accuracy of 0.7681 were achieved by the WSVM method.

Samta Rani et al. [8] used non-clinical data from expecting mothers to predict congenital malformations using various ML Algorithms. The research utilized data from [5] and employed new algorithms to improve accuracy. The authors recommended the use of ANN-based models for CHD prediction on yet-to-be-born babies. NB, ANN, WSVM, and WRF models were used, with ANN reporting the best accuracy of 0.996 but a very low sensitivity of up to 0.16.

Amirreza Salehi et al. [9] used the same dataset as [5] and employed the multi-attribute decision-making (MADM) technique to cluster high-risk patients and improve diagnosis. The authors used both supervised and unsupervised methods to demonstrate an 8% improvement in recall (sensitivity). The supervised Balanced Random Forest Classifier was the optimal model used in this study. Since it was identified as the best-performing classifier overall, the researchers extracted feature importance from this model. These feature importance scores indicated the influence of each feature in predicting CHD risk. The importance

values were then used as weights in an MADM ranking method to evaluate and rank the different clusters or individual cases based on the weighted features.

Zahra Hoodbhoy et al. in [10] estimated the diagnostic accuracy of predicting CHD without the need for trained personnel. This review demonstrates the potential of ML models, particularly neural networks, as effective decision-support tools for CHD diagnosis. The authors extracted the data from various studies with different modalities. However, the number of eligible studies was limited to generalize a common best model for various data modalities selected for the review. Also, substantial heterogeneity existed in both the diagnostic modalities used to train the models and the spectrum of CHD diagnoses considered, ranging from critical to minor defects.

In another study [11], the authors used data from 1389 patients of Iranian origin with 399 features and a hybrid methodology combining Support Vector Machine (SVM) and Particle Swarm Optimisation (PSO) to predict CHD, achieving around 81.57% accuracy. They have also listed the factors affecting the CHD with their importance as given by Random Forest, SVM, and PSO-SVM.

Ariane J. Marelli et al. [12] harnessed Quebec claims and hospitalization databases from 1983 to 2000, where data consisting of 19,187 patients with 3784 CHD cases was used for the study. They compared ML methods and reported that Gradient Boosting Decision Trees (GBDT) demonstrated superior performance with an Area Under the Precision-Recall Curve (AUPRC) of approximately 0.99, sensitivity of 98.0%, and specificity of 99.7%.

The proposed study is similar to several methods used in previous research. Similar to Luo et al. [5], Rani and Masood [8], and Salehi and Khedmati [9], the study uses combined maternal risk factor scores grouped into medically approved categories to simplify the data and address overlapping information. The study also adjusts class weights to address uneven class distributions, following the method used by Luo et al. [5].

The study of literature shows that researchers have experimented with a wide range of ML models for CHD prediction. Each identified a different best-performing model, which depended on the characteristics of the dataset (size, noise, the number of features), and the hyperparameter optimization strategies used. On observing this variability, the present study evaluates a broad spectrum of ML models to ensure comprehensive coverage of linear, probabilistic, neural, and ensemble-based approaches, enabling a fair comparison and identifying the model class best suited to the structure and complexity of the Indian CHD dataset.

The present study proposes a comprehensive machine learning framework using an Indian hospital-based dataset, incorporating clinically validated feature grouping, class imbalance handling, extensive multi-model evaluation, and explainable AI techniques for improved clinical interpretability.

The present study makes the following contributions relative to the cited literature:

- Unlike [5], who use Chinese population survey data, the proposed study used hospital-collected EHR data from an Indian tertiary cardiac care center (SSSSH, Kharghar, Navi Mumbai) which deals with early antenatal screening using questionnaire-derived maternal data. This makes it one of the very few ML-based CHD prediction studies grounded in Indian clinical data, with domain driven feature engineering validated by clinicians, addressing the significant translational gap in the world's largest CHD-burden population.
- Unlike the studies considered from the literature, the proposed study evaluated nine models from major ML algorithm families (linear, probabilistic, neural, kernel, bagging, and multiple boosting variants) on the same dataset under identical evaluation conditions with hyperparameter optimization via 5-fold

cross-validation. This comparison across algorithms reduces the bias and provides a more reliable identification of the best performing model for this specific classification problem.

- Explainability: The present study applies SHAP to the best-performing XGBoost model and further conducts sub feature level analysis.

The proposed study used the data from the Sri Sathya Sai Sanjeevani Centre for Child Heart Care & Training in Pediatric Cardiac Skills (SCCHC) at Kharghar, Navi Mumbai, India. The Sri Sathya Sai Sanjeevani Hospitals constitute a network of pediatric cardiac care and maternal–child health centresthat provide free treatment for CHD and specialized care for mothers and children [13].

3 Materials and Methods

The end-to-end developmental pipeline used for the study is depicted in Fig. 1. This study follows an eight-stage analytical pipeline, beginning with data acquisition and preprocessing, feature grouping, data balancing, and model development. The workflow concludes with performance evaluation, robustness analysis, and explainability to ensure reliable and clinically interpretable CHD risk prediction. The detailed explanation of each stage is given below.

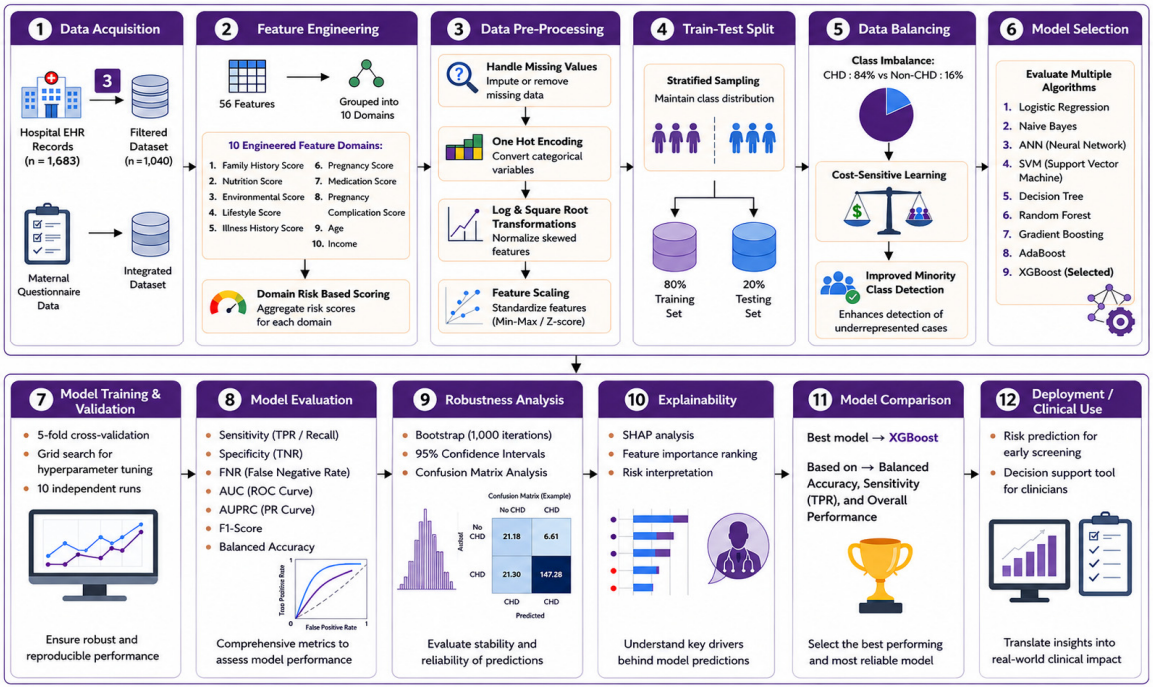

Figure 1: Complete machine learning workflow for CHD prediction.

This is a hospital-based retrospective case-control study, in which the cases are children reporting to the SCCHC hospital with complaints such as very low birth weight, delayed developmental milestones, dyspnea (difficulty breathing), difficulty in feeding, and chronic infections. Data was collected at the hospital through in-person parental interviews when the proband reported to the center. Information about the proband's mother was collected through questionnaires, which included maternal age at the time of delivery, monthly family income, family history, mother's illness history, unhealthy lifestyle of the mother, maternal illness, etc., which were the input variables for model development.

The inclusion criteria for the data collected for the study were cases with CHD and controls without heart disease or other diseases that indirectly affect the heart, who reported to the center during the period from January 2024 to June 2025. Data with incomplete information were excluded from the study. This retrospective study was approved by the Institutional Ethics Committee (IEC) of Sri Sathya Sai University for Human Excellence (Approval No. IEC/Certificate/01/2025, dated 20 January 2025). This approval covered the overall study design, data analysis plan, and the use of de-identified data for machine learning modelling. Meanwhile, the IEC of Sri Sathya Sai Sanjeevani Centre for Child Heart Care & Training in Pediatric Cardiac Skills at Kharghar, Navi Mumbai Kharghar (Approval No. -SSK0051/V3/PR/2024/IEC-11, approved 2 August 2025) was obtained specifically for the collection and retrospective use of anonymized electronic health records (EHR) and maternal questionnaire data. The requirement for individual informed consent was waived due to the retrospective use of de-identified data. The study was conducted in accordance with the Declaration of Helsinki and followed TRIPOD guidelines.

As the study was retrospective, the CHD outcome labels were assigned by clinicians using the echocardiographic modality, and clinical diagnoses were already documented in the institutional database before the model was built and the analysis began. Since this work was retrospective and relied on pre-existing anonymized records, investigator blinding during data extraction or during model training was not applicable.

3.1 Feature Grouping of Risk Factors

The data collected from SCCHC consisted of 1229 cases, 454 controls, and 254 features. An exploratory data and feature analysis was conducted to study the distribution of the data. The extensive list of features collected at the hospital included “Date of Questionnaire”, probands’ information like blood group, symptoms (like rapid breathing, fast heartbeat), ICD coding, paternal risk factors, pre-operation lab report etc., To identify relevant maternal risk factors features associated with CHD cases diagnosed by pediatric cardiologists using echocardiography at the hospital, this study reviewed contributing factors reported in the existing literature [5,6,14,15] and compared them with patterns observed in the hospital dataset. Based on this comparison, 55 independent maternal risk variables and one dependent variable (target variable) were considered from the feature set given by the hospital. The 55 risk factors (independent variables) were grouped into feature categories/indicators by categorizing them into ten distinct domains, based on their nature and the type of exposure or characteristic they represent, which helped to reduce the dimensionality of the data.

The grouping of individual risk factors into ten clinically interpretable domains was not solely driven by dimensionality reduction considerations. The grouping strategy was developed based on clinical relevance, literature-supported maternal risk categories, and in consultation with domain experts involved in CHD care. In particular, the domain structure received validation through consultation with a congenital pediatric cardiology expert and a practicing clinician. The clinician substantiated that each grouped variable represented a meaningful risk category rather than a random combination of features. Also, the approach to feature grouping for dimensionality reduction was motivated by Luo et al. [5,15] to address high dimensionality, noise, and multicollinearity in the data and to make the model generalizable. The grouping process was clinically guided rather than fully data-driven and, therefore, may contain subjective components. The grouped variables were evaluated using downstream model performance and explainability analyses.

Guided by these considerations, clinically related variables were grouped as detailed below: Family History: A positive family history [6,16,17], which includes consanguinity and affected relatives,

suggests that genetic factors contribute to the condition. The risk of CHD recurrence among children who have affected parents or parents with past congenital anomalies in their family history shows a significant increase.

- Nutrition and Supplementation: Research shows that people who lack essential micronutrients, especially folic acid and iron [6,18,19], encounter a higher chance of developing CHD. Environmental Exposure: The body can lose its ability to process environmental toxins from pollutants, radiation, or contaminated water during the first three months of pregnancy [6,20], which disrupts the normal process of organ development.
- Unhealthy Lifestyle: Researchers categorized alcohol use, tobacco use, areca nut consumption, drug use, and passive smoking as modifiable behaviors that impact fetal oxygen supply and organ development [6,21–23].
- Maternal Illness History: The study combined all pre-existing maternal conditions, which include hypertension, thyroid disorders, anemia, tuberculosis, psychiatric illness, cancer, prior miscarriages, and assisted conception, because these conditions disrupt blood flow, metabolism, and immune function to create potential risks for fetal heart development [6,24,25].
- Maternal Illness During Pregnancy: The study classified acute pregnancy-related complications [6,26], which included gestational diabetes and infections, fever, bleeding, polyhydramnios, and seizures, as separate entities from chronic conditions because these complications function as temporary stressors during pregnancy, which disrupt the critical windows required for heart development.
- Medications Used: Fertility drugs, psychoactive medications, anesthesia exposure, and insulin-dependent diabetes treatment were grouped due to their potential pharmacological and teratogenic effects [6]. Complications in Pregnancy: Psychological stress, accidents, and other pregnancy complications were grouped as acute systemic stressors that may indirectly influence fetal development through inflammatory or hemodynamic pathways.

Maternal Delivery Age and Monthly Family Income are the categorical variables independent of the grouping. Refs. [6,27] describe maternal age as a risk factor for congenital anomalies. The income level of a family determines their socioeconomic status [28], which subsequently affects their access to proper nutrition and medical services, their exposure to different environmental factors, and their use of prenatal screening services.

Each grouped variable contains multiple specific risk factors, which were summed to create a total risk score for that particular grouped variable.

The ten grouped variables and their associated risk factors are described in Table 1, where maternal delivery age is encoded as a categorical variable (<30 years:0, ≥30 years:1), and the monthly family income (expressed in Indian Rupees, INR) is categorized as (≤9307:0, 9308–27882:1, 27883–46474:2, 46475–69534:3, 69535–92950:4, 92951–185894:5, ≥185895:6).

The remaining eight features are the variables in numeric form, and the risk factors are represented in binary form, where 1 is the value given when a risk factor is present. The grouped variable created from these factors has a minimum of 0 and a maximum equal to the sum of the total number of risk factors, with each coded as 1.

Table 1: Grouped variables, associated risk factors, and ranges.

Grouped Variable	Risk Factor	Min	Max
Maternal delivery age	(<30 years = 0, ≥30 years = 1)	0	1
Monthly family income	≤9307 INR	0	6
	9308–27,882 INR		
	27883–46,474 INR		
	46475–69,534 INR		
	69535–92,950 INR		
	92951–185,894 INR		
	≥185,895 INR		
Family history	Consanguinity	0	3
	Any other relative affected with CHD		
	Children affected with CHD		
Nutrition and Supplementations (Not taken)	Iron	0	4
	Folic acid		
	Calcium		
	Multivitamins		
Environmental Exposure	Presence of cell tower in vicinity	0	4
	Pollution source in vicinity (<1 km)		
	Potable water quality		
	Exposure to radiations		
Unhealthy Lifestyle	Alcohol addiction	0	7
	Chewable tobacco addiction		
	Pan masala/chewable betel nut addiction		
	Areca nut addiction		
	Drug addiction		
	Smoking (bidi/cigarette)		
	Passive smoking		
Maternal illness history (Chronic conditions)	Hypertension	0	8
	Thyroid disorders		
	Anemia		
	Tuberculosis (TB)		
	Psychiatric illness		
	Malignancies		
	Miscarriages/abortions		
	Conceived after treatment		
Maternal illness during pregnancy	Polyhydramnios	0	17
	Oligohydramnios		
	Hyperemesis gravidarum		
	Gestational diabetes		
	Viral illness		
	Rubella (rashes)		
	Persistent fever (>102°F)		
	Bleeding		
	Upper respiratory tract infection		
	Urinary tract infection		
	Dizziness		
	Headache		
	Edema		
	Swelling of feet		
	Seizures/fainting		
	Other infections		
	Weakness		

Table 1: *Cont.*

Grouped Variable	Risk Factor	Min	Max
Medications used	Fertility drugs	0	4
	Psychoactive drugs		
	General anesthesia		
	Insulin-dependent diabetes medication (IDDM)		
Complications in pregnancy	Mental stress	0	4
	Major accident		
	Minor accident		
	Piles/fissures		

3.2 Data Preprocessing and Data Partitioning

The preprocessing pipeline is depicted in Fig. 2.

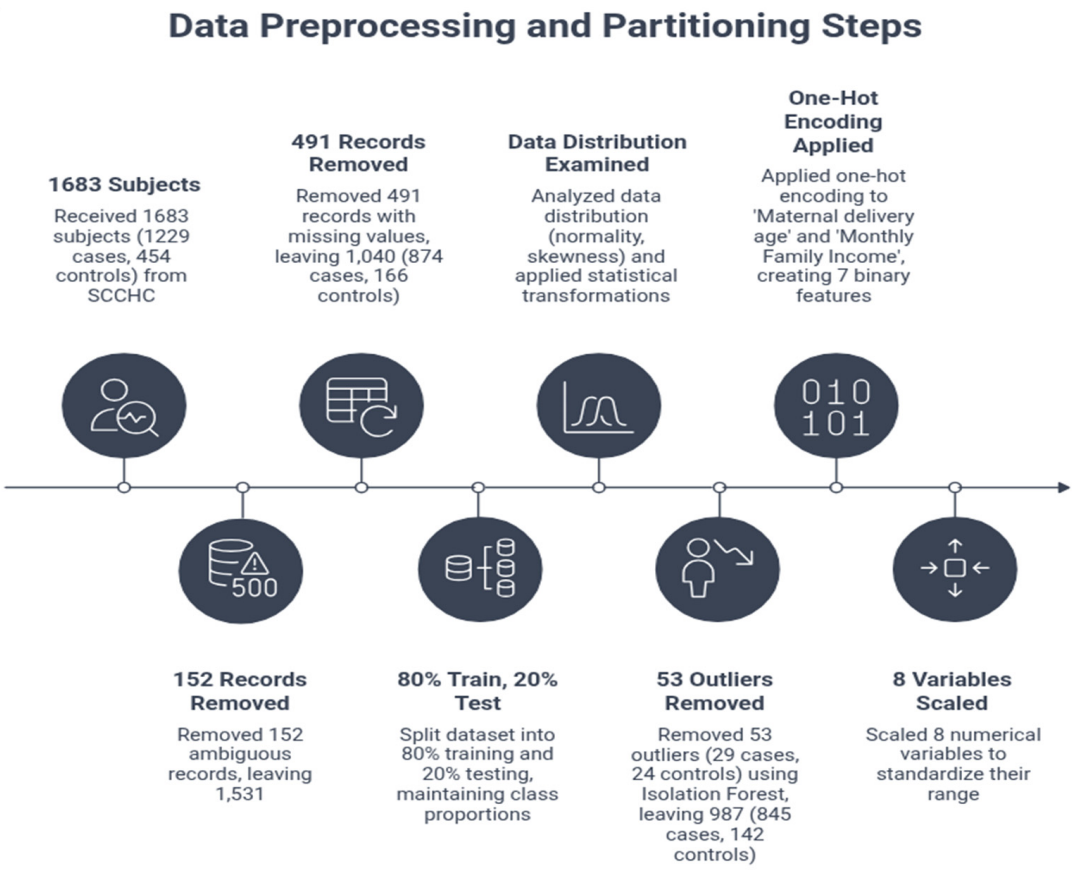

Figure 2: Data preprocessing pipeline.

The data received from SCCHC consisted of 1683 subjects, of which 1229 were cases and 454 controls. After preprocessing, 643 records were eliminated, and the remaining 1040 records were used for model development, validation, and testing as described in this section. The first step in preprocessing was to clean the data, which included:

(a) Handling ambiguously labelled data: The dataset initially consisted of 1683 records, of which 152 were excluded since the children were initially suspected of having CHD; however, subsequent echocardiographic evaluation determined that they did not require further assessment or clinical

intervention. Some children had already undergone intervention, while others had self-correcting defects (e.g., septal defects with holes smaller than 4 mm). The remaining 1531 records were retained for further analysis.

(b) Handling data that was not captured or was missing: A total of 491 incomplete records were removed because at least one of the 55 selected features contained missing data. This preprocessing step resulted in a final dataset of 1040 subjects, including 874 CHD cases and 166 non-CHD controls as shown in Fig. 3. Subject-Level Train–Test Separation: The data were split at the subject level into 80% training and 20% testing sets, while preserving the original CHD to non-CHD ratio in both subsets. The test set remained completely untouched during model training and hyperparameter optimization. Within the training data, an additional stratified 80:20 split was performed to create training and validation subsets.

(c) Data distribution: Statistical transformations were applied to the data after analyzing the normality and skewness of the data using Shapiro-Wilk normality test. Table 2 summarizes the statistical characteristics of each grouped maternal risk score and indicates the transformation applied to address distributional asymmetry before modelling.

(d) Encoding of categorical features: Maternal delivery age and Monthly Family Income were two categorical variables. These two features were encoded using one-hot encoding method to get seven binary features.

(e) Scaling of the numerical features: The eight numerical features: Family History, Nutrition and Supplementation (Not Taken), Environmental Exposure, Unhealthy Lifestyle, Maternal Illness History (Chronic Conditions), Maternal Illness During Pregnancy, Medications Used, and Pregnancy Complications were scaled to bring them onto a comparable range.

These preprocessing steps improved data quality and ensured reliable predictive modeling.

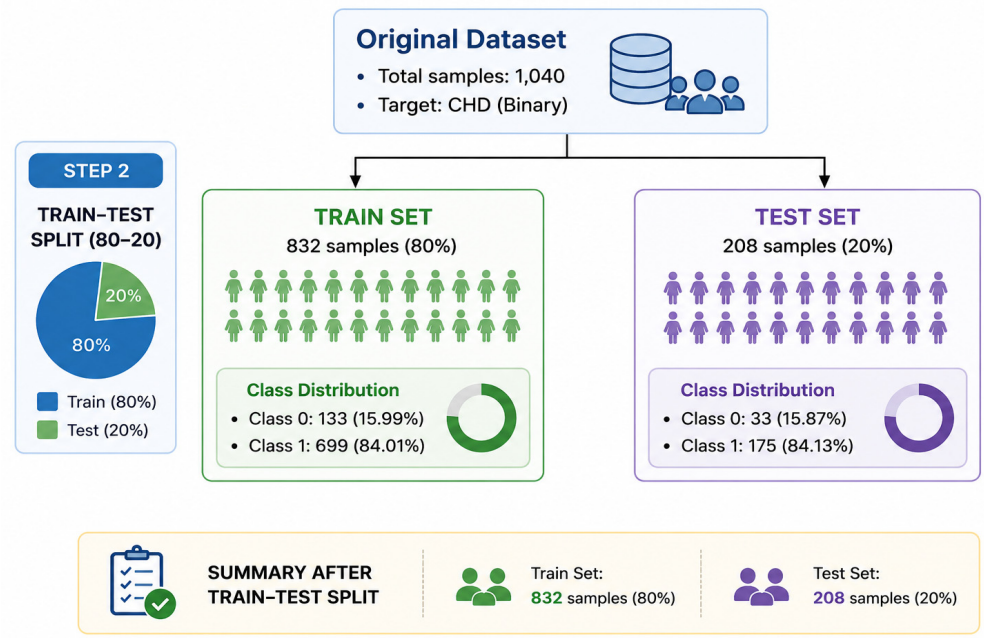

Figure 3: Data partitioning.

Table 2: Distributional characteristics of grouped features and the corresponding transformation applied prior to modeling.

Feature	Mean	Std	Skewness	Is_Normal	Transform Method
Family history	0.190	0.439	2.371	False	Log Transform (Right-skewed)
Nutrition and Supplementation	2.406	1.277	−0.984	False	Square Root Transform (Moderate skew)
Environmental Exposure	0.925	0.749	0.137	True	None (Already normal)
Unhealthy Lifestyle	0.138	0.418	3.382	False	Log Transform (Right-skewed)
Maternal Illness History	0.438	0.804	2.203	False	Log Transform (Right-skewed)
Maternal Illness During Pregnancy	0.468	0.869	2.266	False	Log Transform (Right-skewed)
Medications Used	0.012	0.111	8.776	False	Log Transform (Right-skewed)
Complications in Pregnancy	0.003	0.054	18.538	False	Log Transform (Right-skewed)

3.3 Data Balancing

In the CHD dataset received from SSSSH, after preprocessing, the final dataset contained 1040 records (874 CHD cases, 166 controls). The corresponding class distribution for majority class was 84.0% CHD and the minority class was 16.0% non-CHD. Such an imbalanced distribution can bias ML models toward the majority class, resulting in high overall accuracy but poor sensitivity for the minority class, a clinically undesirable outcome. To address dataset imbalance, several class-balancing techniques have been proposed in the literature [29].

A non-structural balancing technique using Class-weight adjustment was used in this study to deal with the imbalance problem while keeping the dataset unchanged in terms of both quantity and structure. This method does not rely on resampling but instead applies class-specific weights during model training based on class proportions to render the classifier more aware of the minority class. Thus, Class-Weight is calculated using:

$$\text{Weight for classifier} = \frac{\text{Total samples}}{\text{Number of classes} \times \text{Samples in class}}$$

The non-CHD (minority) class is assigned a greater weight, and the CHD (majority) class is assigned a lesser weight. For a binary classification problem, the dataset used in the study has the following class imbalance as per the Class Weight formula mentioned above,

- CHD class weight = $1040/(2 \times 874) = 0.595$
- Control class weight = $1040/(2 \times 166) = 3.13$

This results in the establishment of a higher misclassification penalty for minority-class errors, thus reducing the model bias towards the dominant class and increasing the sensitivity to non-CHD cases.

3.4 Prediction Model Selection

Supervised learning algorithms provide the capability to learn discriminating patterns from complex structured datasets having non-linear relations between socio-demographic, environmental, and lifestyle-related risk factors. To ensure a rigorous and unbiased comparison, nine supervised models spanning all major ML families—linear, probabilistic, neural, SVM-based, tree-based, and boosting-based methods were selected. This spectrum of models selected for the study was based on the volume and nature of EHR dataset used in the study, which can handle noise in the data and still capture different learning

behaviors, reduce algorithmic bias, test robustness under class imbalance, and enable reliable identification of the best performing model for CHD risk prediction.

Linear models such as LR provide an interpretable baseline aligned with epidemiological analysis. Probabilistic models (NB) assume simplified feature likelihoods, serving as a benchmark for non-linear and ensemble methods. Neural models (ANN) capture complex non-linear interactions between maternal risk domains and automatically learn feature representations. Kernel-based methods (WSVM) provide robust margin-based classification with weighted penalties that effectively handle class imbalance and high-dimensional risk-factor spaces. Tree-based (DT, WRF: Bagging) models learn the hierarchical decision rules that capture feature interactions naturally and remain robust to heterogeneous inputs. Boosting-based models (GBM, AdaBoost, XGBoost) iteratively correct misclassification errors, handle imbalance through gradient-based reweighting, and typically deliver state-of-the-art performance even for relatively smaller tabular clinical data.

Various studies from the past have highlighted the strength of the above models for tabular clinical data. WSVM was reported as the strongest performer in terms of True Positive Rate (TPR) and AUC [5]. ANN achieved the best results in [8] in terms of accuracy, and Ref. [11] demonstrated improved recall using WSVM. Likewise, Ref. [12] found that GBM and GBM-enhanced DT approaches performed particularly well in terms of AUPRC. As these models exhibit complementary learning behaviors and distinct advantages, a unified approach was employed in the present study to evaluate all the models used in prior literature, thereby reducing algorithmic bias and rigorously identifying the most reliable CHD prediction model.

In this study, nine machine learning models were evaluated, such as WSVM, WRF, NB, ANN, LR, GBM, DT, AdaBoost and XGBoost. Hyperparameter optimization was performed using a grid search with repeated stratified cross-validation, all on the training data. And since class imbalance was observed, class-weight balancing was used for WSVM, WRF DT, and LR. The details on grid search and best optimized hyperparameters for each model are presented in Table 3. The model parameters evaluated included regularization strength, kernel type, gamma values, tree depth, number of estimators, and learning rate. In addition, there were node splitting criteria, the hidden layer configuration, and subsampling parameters, which depended on the classifier architecture.

Table 3: Best hyperparameters selected during final model training.

Model	Best Hyperparameters
WSVM	C = **1**, γ = scale, kernel = rbf
WRF	max_depth = 10, min_samples_leaf = 4, min_samples_split = 2, n_estimators = 200
NB	Default parameters
ANN	activation = relu, alpha = 0.001, hidden_layer_sizes = (128,64), learning_rate_init = 0.01, solver = adam
LR	C = **0.1**, penalty = l2, solver = lbfgs
GBM	learning_rate = 0.01, max_depth = 3, n_estimators = 200
Decision Tree	criterion = entropy, max_depth = 5, min_samples_leaf = 2, min_samples_split = 10
AdaBoost	learning_rate = 0.1, n_estimators = 200
XGBoost	colsample_bytree = 0.8, learning_rate = 0.01, max_depth = 4, n_estimators = 200, subsample = 1.0

3.5 Performance Evaluation Metrics

To evaluate the performance of the classification models multiple metrics were implemented, considering the imbalanced nature of the dataset, to capture different quality aspects of the prediction. Below is a brief description of the metrics used in this study.

The actual positive cases correctly classified by the model as positive are the true positives (TP), while misclassified positives are false negatives (FN). Actual negatives (no CHD) correctly classified as negative are true negatives (TN), and negative instances incorrectly classified as positive are false positives (FP).

True Positive Rate (TPR), also referred to as Sensitivity or Recall, which is the primary metric for screening, assesses how well the model detects CHD cases correctly. A higher TPR will lower missed diagnoses and thus improve clinical safety. The False Negative Rate (FNR) is the next most important metric, as it shows the percentage of missed CHD cases. High FNR prevents the detection of CHD cases that need medical intervention right away. Negative Predicted Value (NPV) estimates the proportion of negative predictions that are correct. In the case of CHD prediction, where a true case is being lost, a high NPV is absolutely necessary to guarantee that healthy infants are not miscategorized as disease-free. The Positive Predictive Value (PPV), also known as Precision, measures the credibility of the positive predictions. It shows the percentage of predicted CHD cases that are really affected, out of all the predicted positives. The True Negative Rate (TNR) or Specificity examines how correctly the model distinguishes controls/healthy infants. The False Positive Rate (FPR) estimates the number of healthy individuals who have been wrongly identified as positive by the testing procedure. Balanced accuracy, F1 score, and G-Mean are all designed to evaluate classification performance under class imbalance, but they emphasize different aspects of model behavior. The metric balanced accuracy (Balanced Acc) assesses model performance through its evaluation of both majority and minority classes, which prevents false score enhancements that occur with imbalanced datasets. The system ensures that medical decision-making processes receive equal representation of a classifier's ability to identify both positive and negative test results. Geometric Mean (GMean), which merges sensitivity and specificity, is utilized to rate even performance across both categories and is handy when there is an uneven distribution of the classes by penalizing the model more strongly if it performs poorly on either class. Similarly, F1 score provides a balanced measure of a model's performance by combining precision and recall into a single metric, making it especially useful when class distributions are imbalanced, focusing only on the positive class. The Area Under Receiver Operating Curve (AUC) is an index that captures the ability of the model to differentiate patients with CHD from non-CHD/normal ones at different threshold settings. The higher AUC corresponds to better global separability. The AUPRC (Area Under the Precision-Recall Curve) illustrates the model's capability concerning the positive (CHD) class by considering the precision-recall tradeoffs rather than the overall accuracy. In case of class imbalance, it gives a more realistic view. Weighted Accuracy (Balanced Accuracy) WAcc, denotes the weighted combination of TPR and TNR and is used as a measure to detect any bias that might result from class imbalance.

The performance metrics were calculated using below formulae:

True Positive Rate: $TPR_{\text{(Sensitivity or Recall)}} = \frac{TP}{TP+FN}$

True Negative Rate: $TNR_{\text{(Specificity)}} = \frac{TN}{TN+FP}$

Geometric Mean: $GMean = \sqrt{TPR * TNR}$

Balanced Accuracy: $BalancedAcc = \frac{TPR+TNR}{2}$

F1 Score = $\frac{2TP}{2TP+FP+FN}$

Positive Predicted Value (PPV) or precision: $PPV = \frac{TP}{TP+FP}$

False Positive Rate: $FPR = \frac{FP}{FP+TN}$

False Negative Rate: $FNR = \frac{FN}{FN+TP}$

Negative Predicted Value: $NPV = \frac{TN}{TN+FN}$

It is desirable to minimize false negatives (FNR), maximize sensitivity (TPR) and NPV for clinical safety, to avoid false alarms, and using Balanced Accuracy, AUC, GMean, and F1 Score to assess the model performance under real-world class imbalance.

3.6 Robustness and Uncertainty Analysis

The study followed a carefully designed statistical framework to ensure reliable and unbiased results. The dataset was first split at the subject level using a stratified train (80% split as 65% training and 15% validation)–test (20%) approach to prevent data leakage and preserve class distribution in both sets. Within the training set, 5-fold cross-validation was performed for systematic hyperparameter tuning (the detailed hyperparameter grid is provided in Table 3 of the Experiments section). To assess model robustness, the entire training and validation process was repeated across 10 independent runs with (random state = 42). The optimal classification threshold was determined using the Youden Index, defined as Sensitivity + Specificity − 1. The threshold maximizing the Youden Index on the validation set was selected through the Youden Index calculated from the receiver operating characteristic (ROC) to achieve the best trade-off between sensitivity and specificity. The probability based predictions were then converted to binary class labels by using these best possible cutoffs.

Final model performance was then evaluated on an independent test set included discrimination, calibration, and clinical utility metrics such as balanced accuracy, AUPRC, Brier score, mean calibration error, and decision curve analysis to measure generalization ability. Performance uncertainty was quantified using 1000 bootstrap resampling iterations on the test predictions, generating the median, quartiles, and interquartile range. Models were compared using balanced accuracy with corresponding 95% confidence intervals. The best performing models were identified based on this metric. A detailed error analysis using confusion matrices was conducted to understand misclassification patterns. A statistical tests were applied to evaluate whether statistically significant differences existed among the best-performing models.

Libraries and Software Environment: All analyses were conducted in Python 3.7.16 through a conda environment. The following libraries and packages were used in implementation: NumPy 1.21.6, Pandas 1.1.5, SciPy 1.7.3, Statsmodels 0.13.5, Scikit-learn 1.0.2, XGBoost 1.6.2, Imbalanced-learn 0.8.1, SHAP 0.42.1, Matplotlib 3.1.3, Seaborn 0.12.2, OpenCV-python 4.2.0.32, Scikit-image 0.16.2, Pillow 9.5.0, TensorFlow 2.1.4, Keras 2.3.0, Joblib 1.3.2, NetworkX 2.6.3, OpenPyXL 3.1.3, and TQDM 4.64.0, as specified in the exported Conda environment configuration. The model development and validation pipeline is as depicted in the Fig. 4.

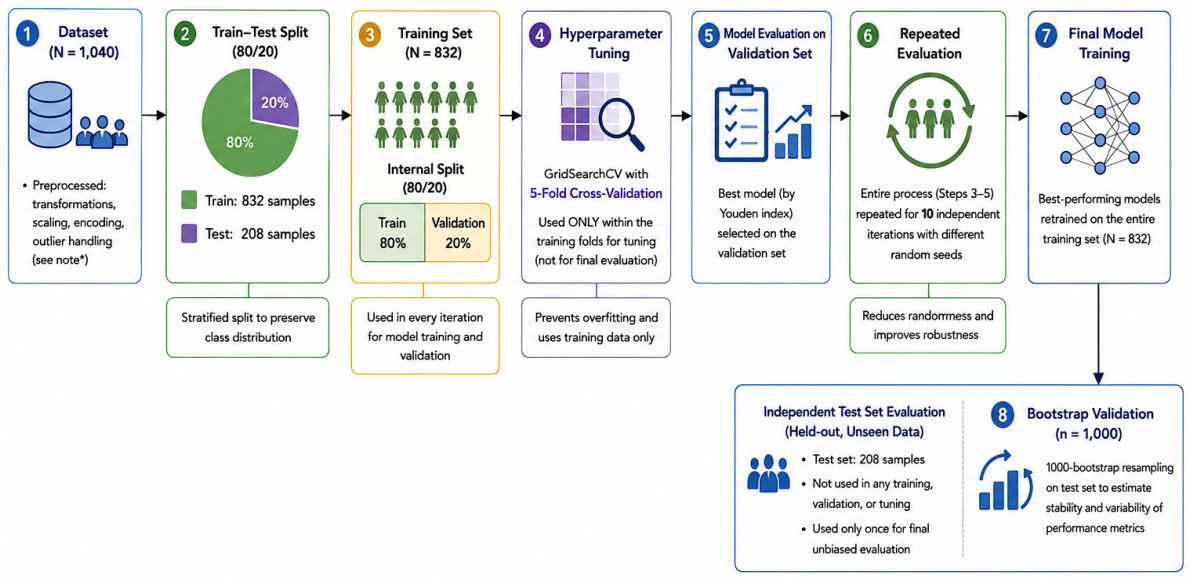

Figure 4: Model development and validation pipeline for CHD prediction.

3.7 Explainability and Risk Stratification

Calibration analysis was done using calibration plots, the Brier score and mean calibration error, to see how well the predicted probabilities line up with the observed CHD outcomes. At the same time, Decision Curve Analysis (DCA) was applied to judge the possible real-world usefulness of the models over different threshold probabilities. The net benefit curves were then checked against treat-all and treat-none strategies, basically to find if the models might actually support clinical decisions with better utility, especially within those threshold ranges that clinicians consider relevant.

SHAP (SHapley Additive exPlanations) explainable AI framework was used for interpreting the results from the model to provide the global feature importance explanations. The study calculated risk scores for each patient through the analysis. This process establishes clinical transparency which helps doctors understand the results and makes it easier to make choices during actual patient screening procedures.

4 Results

In this section, we present the experimental results of the nine machine learning models, along with a detailed evaluation of their performance metrics under the proposed experimental design.

4.1 Predictive Performance of Machine Learning Models

The data, as described in Section 3.3, was split into training and test sets in the ratio of 80:20, respectively, without any overlap between them to avoid any data leakage. 80% of the data was used for training-validation of the model development, and 20% of the data was used for testing or evaluating the developed model.

To ensure a fair and methodologically rigorous comparison across these heterogeneous algorithms, a standardized hyperparameter optimization framework was implemented for all linear, probabilistic, kernel-based, and tree-based models under identical validation settings. Each classifier in the experiment was tuned through an exhaustive grid search combined with 5-fold cross-validation on the training set. Rather than applying a one-size-fits-all tuning strategy, the search spaces were designed to reflect each algorithm's known sensitivities.

The WSVM model was tuned using three parameters, which included the regularization parameter (C = 0.01–100), the kernel type (RBF, linear, and sigmoid), and the kernel coefficient (gamma = scale, auto, and 0.0001–0.1) to achieve the correct balance between flexible decision boundary solutions and non-linear boundary solutions. The WRF model tested various tree-count settings, ranging from 100 to 500 trees, and tree-depth settings, allowing depths between 10 and 30 or complete access. Additionally, node-splitting settings were tested, ranging from 2 to 10 for minimum samples and from 1 to 4 for minimum leaf samples, to manage model growth while preventing overfitting. The Naïve Bayes (NB) classifier required no hyperparameter tuning. The ANN assessed two different hidden layer designs, which included (64) and (128), as well as (128, 64) with ReLU and tanh activation functions and the Adam optimizer together with L2 regularization settings, which ranged from 0.0001 to 0.001, and different learning rates, operated from 0.001 to 0.01. LR used L2 penalty with the lbfgs solver to perform optimization over inverse regularization strength, which ranged from C = 0.01 to C = 10. The tuning experiments with the GBM model were conducted across multiple ensemble sizes (100–500), learning rates (0.01–0.1), and shallow tree depths (3–5). DT classifiers were optimized by varying the depth (5–20 or unrestricted), the splitting criterion (Gini or entropy), and the minimum node sample threshold. The AdaBoost tuning was done by testing different weak learner counts, which ranged from 50 to 300, and different learning rates, which ranged from 0.01 to 1.0. For XGBoost optimization, four parameters, including boosting rounds (100–300), learning rate (0.01–0.1), tree

depth (3–5), and stochastic regularization parameters (subsample and colsample_bytree = 0.8–1.0) were used. The final optimized hyperparameters for each model are presented in Table 3.

The development workflow described in Section 3 was repeated across 10 independent iterations to ensure that the observed performance was not dependent on a single data partition, thereby making the model stable.

The model performance was evaluated using bootstrap resampling with 1000 test-set iterations, and the results are reported as the median with interquartile range (Q1–Q3), capturing robustness to population-level case variation, which evaluated reproducibility, robustness, and stability. Table 4 summarizes the bootstrap performance of all models, and the corresponding ROC curve is depicted in Fig. 5.

Table 4: Consolidated performance metrics across 1000 iterations. Values are reported as median with interquartile range (Q1–Q3) in parentheses.

Model	AUPRC	AUC	TPR	TNR	FNR	PPV	NPV	F1 Score	Balanced Acc
DT	0.963 (0.951–0.974)	0.857 (0.828–0.888)	0.889 (0.746–0.939)	0.792 (0.704–0.875)	0.111 (0.061–0.254)	0.962 (0.949–0.976)	0.514 (0.361–0.655)	0.917 (0.844–0.945)	0.816 (0.792–0.840)
XGBoost	0.959 (0.946–0.972)	0.854 (0.822–0.887)	0.921 (0.788–0.948)	0.758 (0.686–0.845)	0.080 (0.052–0.212)	0.958 (0.946–0.971)	0.583 (0.397–0.697)	0.934 (0.867–0.950)	0.817 (0.792–0.842)
WRF	0.957 (0.943–0.969)	0.843 (0.809–0.875)	0.913 (0.869–0.943)	0.727 (0.665–0.800)	0.087 (0.057–0.131)	0.953 (0.942–0.966)	0.571 (0.469–0.667)	0.931 (0.907–0.945)	0.807 (0.779–0.834)
AdaBoost	0.957 (0.943–0.968)	0.832 (0.798–0.866)	0.937 (0.750–0.954)	0.720 (0.652–0.828)	0.063 (0.046–0.250)	0.954 (0.940–0.968)	0.630 (0.373–0.714)	0.938 (0.846–0.952)	0.804 (0.778–0.831)
GBM	0.961 (0.946–0.973)	0.845 (0.818–0.875)	0.916 (0.888–0.942)	0.700 (0.643–0.773)	0.084 (0.058–0.112)	0.950 (0.939–0.963)	0.583 (0.500–0.667)	0.931 (0.915–0.946)	0.780 (0.752–0.808)
WSVM	0.947 (0.931–0.961)	0.791 (0.758–0.824)	0.749 (0.698–0.782)	0.827 (0.774–0.880)	0.251 (0.218–0.302)	0.963 (0.953–0.974)	0.343 (0.296–0.390)	0.842 (0.807–0.862)	0.781 (0.755–0.804)
ANN	0.947 (0.936–0.957)	0.766 (0.730–0.800)	0.836 (0.809–0.864)	0.680 (0.619–0.741)	0.164 (0.136–0.191)	0.941 (0.928–0.954)	0.405 (0.356–0.463)	0.885 (0.869–0.902)	0.758 (0.725–0.786)
LR	0.941 (0.928–0.953)	0.745 (0.711–0.777)	0.710 (0.679–0.749)	0.758 (0.690–0.818)	0.290 (0.251–0.321)	0.947 (0.933–0.960)	0.304 (0.263–0.343)	0.812 (0.789–0.837)	0.733 (0.703–0.758)
NB	0.940 (0.926–0.952)	0.733 (0.696–0.770)	0.758 (0.589–0.869)	0.714 (0.591–0.833)	0.242 (0.131–0.411)	0.942 (0.926–0.959)	0.318 (0.250–0.417)	0.838 (0.729–0.897)	0.723 (0.690–0.750)

The hospital-based case-control dataset used in this study had a class imbalance, which was addressed as mentioned in Section 3.4. The predictive performance of the models was evaluated using metrics detailed in Section 3.6.

The results indicate that the highest sensitivity (TPR) was recorded for AdaBoost (0.954), followed by XGBoost (0.948). The lowest FNR (0.046) was with AdaBoost, followed by XGBoost (0.052) and WRF (0.057), which denoted the superiority of the tree-based algorithms in the detection of CHD cases that are critical for screening applications. XGBoost achieved a good specificity of 0.845 while maintaining a high sensitivity (0.921). In comparison, WSVM had the highest overall specificity (0.880), but at the cost of considerably lower sensitivity (0.749), making it less suitable for screening where missed CHD cases are critical. In the case of precision (PPV), DT had the highest median value of 0.962 (Q1-Q3 as 0.949–0.976), slightly exceeding XGBoost (0.958), AdaBoost (0.954), WRF (0.953), and GBM (0.95), the reason being that it gives the most reliable positive predictions, thus helping in reducing unnecessary referrals.

The highest balanced accuracy was achieved by XGBoost with a score of 0.817 and DT with a score of 0.816, which demonstrated superior class discrimination abilities on the imbalanced dataset. LR (0.733) and NB (0.723) showed the weakest balance between sensitivity and specificity. WRF (0.807) and AdaBoost (0.804) demonstrated strong performance, which they maintained throughout their testing period. ANN (0.758) performance fell short of tree-based methods. F1-scores were generally high across most models, with GBM (0.931), AdaBoost (0.938), and XGBoost (0.934) achieving the highest scores, which demonstrated their ability to perform well on both minority and majority classes.

The DT model showed the highest median AUC of 0.857 (Q1–Q3: 0.828–0.888), followed by XGBoost at 0.854 (0.822–0.887), with observable overlap in their bootstrap IQR ranges. From all the models selected for the study, the DT and XGBoost classifiers demonstrated the highest median GMean (Table 8, last column) of 0.809, with IQR of 0.785–0.835 and 0.782–0.837, respectively, indicating the best balance between TPR and TNR. The WRF model was close with a median GMean of 0.798 (0.766–0.830). The highest median AUPRC was for DT with 0.963 (Q1–Q3: 0.951–0.974), followed by XGBoost with 0.959 (0.946–0.972), WRF, AdaBoost, and GBM (all approximately 0.961), indicating a slight, yet consistent advantage in precision–recall discrimination despite overlapping interquartile ranges. AUPRC and PPV may appear inflated due to CHD cases and should be interpreted in the context of prevalence imbalance.

WSVM achieved high specificity and precision but also had high FNR, limiting its suitability for screening where missed CHD cases are critical. ANN demonstrated moderate and stable performance, but with lower discriminative and balanced metrics, indicating limited robustness in highly imbalanced settings. LR exhibited reduced sensitivity and AUPRC due to its linear modelling constraints. Naive Bayes demonstrated the weakest and most variable performance, reflecting sensitivity to distributional assumptions and reduced reliability for CHD screening.

Upon evaluating the outcomes, it was found that among all the methodologies assessed, the tree-based ensemble models exhibited the best discrimination. The median TPR of XGBoost was notably high at 0.921, its median TNR was 0.758, resulting in a powerful 0.854 AUC and 0.809 GMean, which translates to an adequate detection of CHD cases with not too many false positives. DT and WRF were very close in performance with GMean values of over 0.79, which suggests that they were both good in classifying CHD and non-CHD cases. AdaBoost showed high TPR, PPV, NPV, and reported the lowest FNR. Besides, it had lower TNR than XGBoost and WRF. GBM showed a similar trend, demonstrating strong detection capability for CHD cases. The tree-based ensemble methods in general were found to be superior in dealing with the imbalance between classes and in capturing the complex nonlinear relationships that are naturally formed in CHD risk prediction.

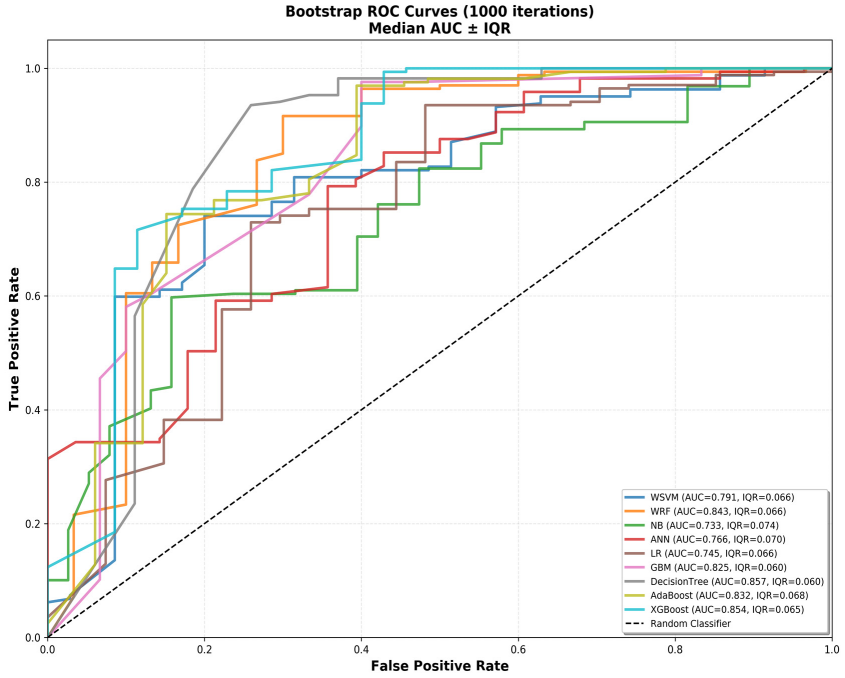

Figure 5: Area under receiver operating curve.

In the present study, LR was included as the baseline classifier because it is the most widely adopted conventional clinical risk prediction approach in healthcare modelling. The performance comparison shows that the suggested tree-based ensemble methods substantially outperform LR. To assess statistical significance between predictive models, DeLong's test was used to compare AUC values for LR and the other models. Table 5 details the DeLong AUC Statistical Comparison.

Table 5: DeLong AUC statistical comparison with logistic regression baseline.

Model	AUC	ΔAUC vs. LR	Significance ($p < 0.01$)
Logistic Regression	0.745	–	(Baseline)
Naive Bayes	0.733	−0.012	No improvement
ANN	0.766	+0.021	Small improvement
Weighted SVM	0.791	+0.046	Moderate improvement
AdaBoost	0.832	+0.087	Strong improvement
GBM	0.845	+0.100	Strong improvement
Weighted RF	0.843	+0.098	Strong improvement
XGBoost	**0.854**	**+0.109**	**Very strong improvement**
Decision Tree	**0.857**	**+0.112**	**Very strong improvement**

The bold value indicates the models with significant p-value.

4.2 Predictive Performance with 95% Confidence Intervals

To assess the model stability at 95% confidence intervals and uncertainty of performance estimates, balanced accuracy was used as a comparison metric across the models. The metric balanced accuracy was chosen as it assigns equal importance to both minority and majority classes, preventing inflated performance in imbalanced datasets and providing a more reliable measure of a model's true discriminative ability, independent of prevalence. In rare conditions such as CHD, it balances sensitivity and specificity so that the model detects affected cases while avoiding excessive false diagnoses. The Table 6 and Fig. 6 present the results.

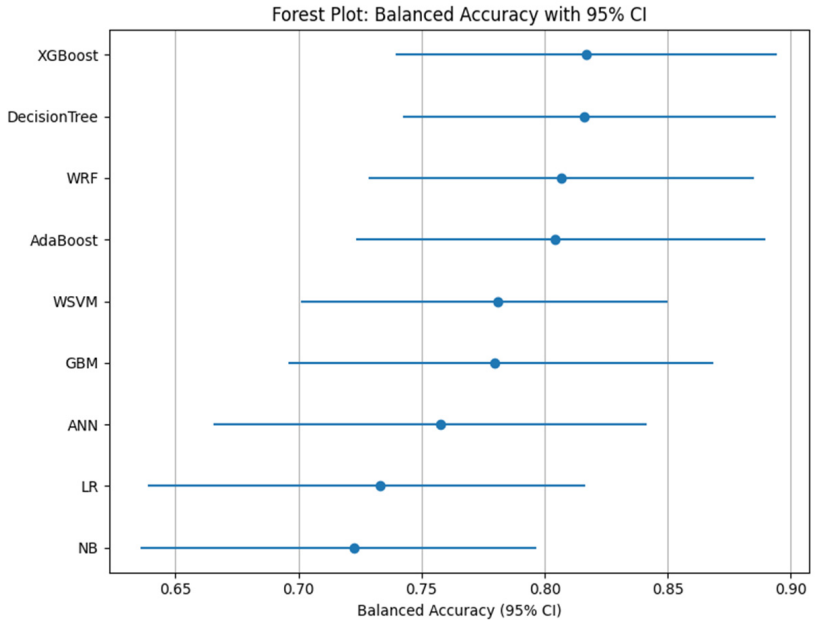

Figure 6: Balanced accuracy with 95% confidence interval.

Table 6: Balanced accuracy (median with 95% percentile confidence interval) across 1000 bootstrap iterations.

Model	Balanced Accuracy
XGBoost	**0.817 (0.739–0.894)***
DT	0.816 (0.742–0.894)
WRF	0.807 (0.728–0.885)
AdaBoost	0.804 (0.723–0.890)
GBM	0.799 (0.695–0.868)
WSVM	0.781 (0.701–0.850)
ANN	0.757 (0.665–0.841)
LR	0.733 (0.639–0.816)
NB	0.723 (0.635–0.797)

*The bold value indicates the highest median balanced accuracy achieved among all evaluated models.

The XGBoost model recorded the highest Balanced Accuracy of 0.817, which had a 95% CI between 0.739 and 0.894, and the DT followed closely with 0.816, which had a CI between 0.742 and 0.894. The two models showed identical median scores which produced overlapping CI that proved no significant statistical difference existed between them. WRF recorded 0.807, while AdaBoost scored 0.804 performance above the critical 0.80 threshold. The confidence interval of AdaBoost showed a lower bound at 0.723 which demonstrated more variability than the confidence interval of WRF that showed a lower bound at 0.728. The two models GBM and WSVM scored below 0.80 while GBM showed the most extensive CI which ranged from 0.695 to 0.868 indicating GBM exhibited high sensitivity to sample variation which raised doubts about its clinical deployment reliability. The group of lower-performing models which included ANN at 0.757, LR at 0.733 and NB at 0.723 proved that neural and linear and probabilistic methods failed to handle the complex non-linear nature of CHD prediction, which suffered from limited data and restricted model capabilities. The approximate 0.094 performance difference between the best and worst models shows that ensemble methods and tree-based approaches achieve better results than linear and probabilistic methods for this particular classification problem.

The tree-based models applied in this research, which were XGBoost, DT, Adaboost, WRF, and GBM, outperformed the models that were used in the earlier CHD prediction studies, especially in sensitivity and precision-recall behavior, demonstrating the strong discriminative power of boosted ensemble algorithms. The analysis of repeated runs showed that the model had good reproducibility, which in turn implies that the model's performance is not reliant on a specific data split and is hence robust enough for hospital-based applications.

Although the DT achieved the best median AUPRC and AUC, and AdaBoost had the least FNR, XGBoost showed up as the most balanced and trustworthy model in the view of all clinically relevant metrics combined. XGBoost is ranked the best model across AUPRC, AUC, and GMean, at the same time having a high sensitivity (TPR = 0.921) and low FNR, therefore, favorable for CHD screening in case of an unbalanced data setting. The inter-quartile ranges over key metrics were narrow and overlapped with the best models, which is an indication of strong and consistent performance of the bootstrapping resampling method. Further, to evaluate the models and assess whether the predicted probabilities match the observed probabilities in the study dataset, Brier score analysis, reliability analysis through calibration plots, and DCA were performed using patient-level predicted probabilities from all models.

Brier Score Analysis: It is used to assess a model's calibration by measuring the mean squared difference between predicted probabilities and actual clinical outcomes.

The XGBoost model achieved a Brier Score of 0.081, which was interpreted relative to the observed dataset prevalence rather than using an absolute threshold. Since the dataset contained

approximately 84% CHD-positive samples, a naive prevalence-based predictor assigning the same probability ($p = 0.84$) to all patients would produce an expected baseline Brier Score of 0.134 using the equation: $BrierScore_{baseline} = p(1-p)^2 + (1-p)p^2$. In comparison, the proposed XGBoost model achieved a Brier score of 0.081, which is substantially lower than the prevalence-based baseline prediction error. Numerically, the model reduced the probabilistic prediction error from approximately 0.134 to 0.081, corresponding to an improvement of nearly 39.7% relative to naive prevalence prediction as shown in Fig. 7. The results imply that the XGBoost model had the lowest calibration error and calibrated well on the studied dataset.

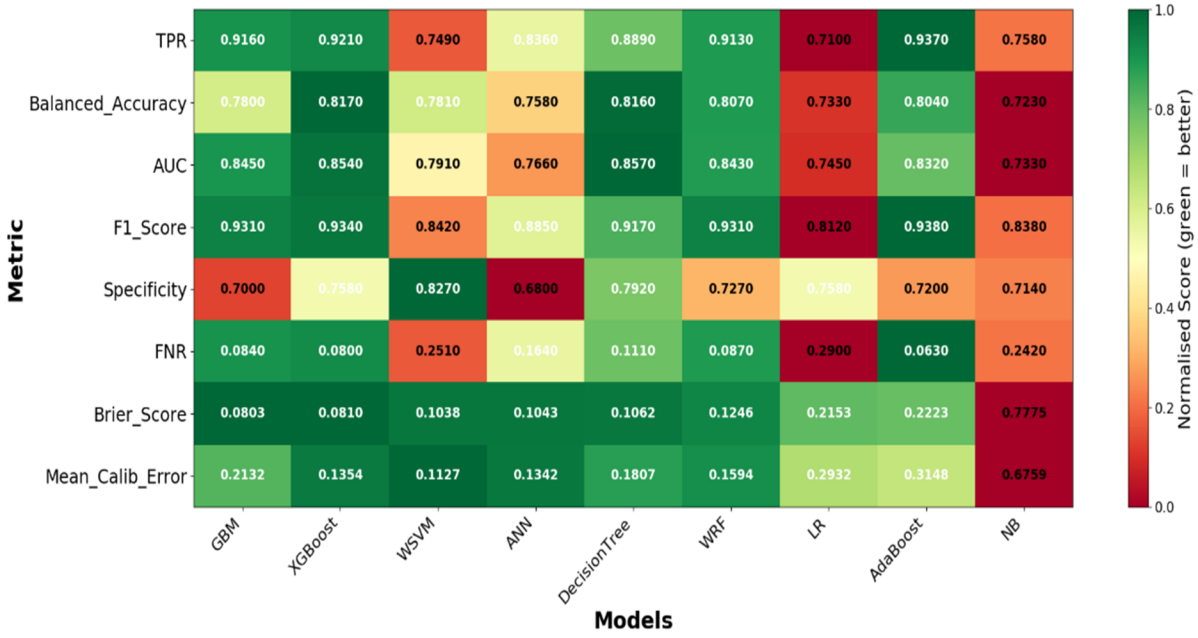

Figure 7: Model comparison heatmap with Brier score.

Further, the consolidated calibration (reliability) diagram shown in Fig. 8 for all the models evaluated how closely the predicted probabilities from each ML model align with the actual observed CHD outcomes.

To evaluate clinical utility beyond conventional discrimination metrics such as AUC and accuracy, DCA was performed using patient-level prediction probabilities as shown in Fig. 9.

DCA: It quantifies the net clinical benefit obtained when applying a predictive model across varying threshold probabilities. The net benefit was computed using:

$$Net\ Benefit = \frac{TP}{N} - \frac{FP}{N} \times \frac{p_t}{1-p_t}$$

where N represents the total patient population, p_t refers to the selected threshold probability.

The DCA curve was generated by the steps: a. Note the predicted probabilities from the trained model. b. Vary the threshold probability (0.01 to 0.99). c. Calculate TP and FP for every threshold. d. Calculate the net benefit for every threshold probability. The references, Treat-all (assumes all patients are classified as CHD-positive) and Treat-None (assumes no patients are classified as CHD-positive) were used. The results showed that XGBoost consistently provided better overall clinical benefit across many threshold levels. It showed good balance for Treat-all and Treat-none during classification by scoring a maximum benefit of 0.8564.

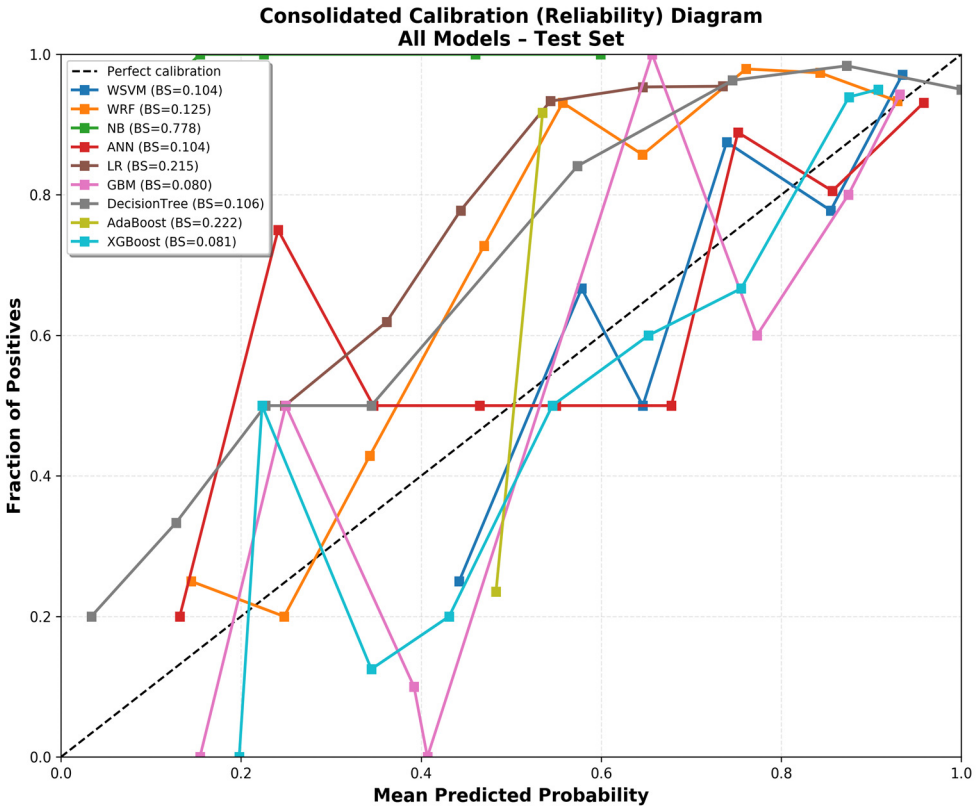

Figure 8: Calibration Diagram for all the models.

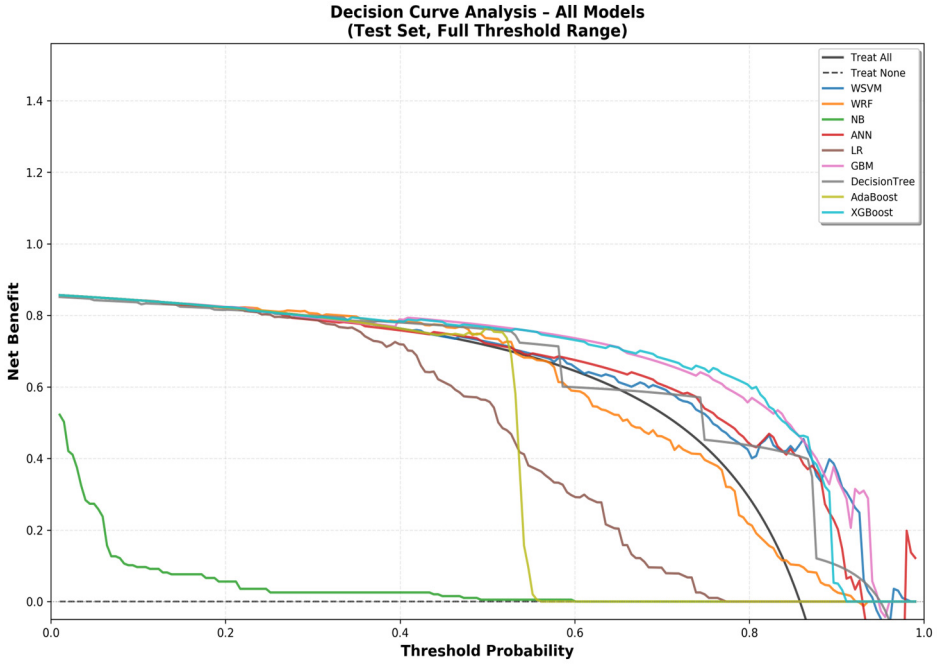

Figure 9: DCA analysis for the models.

4.3 Comparative Analysis of the Best Performing Models

To have a deeper and more transparent look at the error types, confusion matrices were used for the best-performing models: XGBoost, DT, GBM and AdaBoost. These models were selected based on the metrics: Balanced accuracy, sensitivity, specificity, F1 score, FNR, and Brier score, as shown in Fig. 7. These confusion matrices were generated by averaging the confusion matrix values obtained across all bootstrap iterations. The mean values of TP, TN, FP, and FN obtained from the bootstrap run for the best-performing models are depicted in Fig. 10.

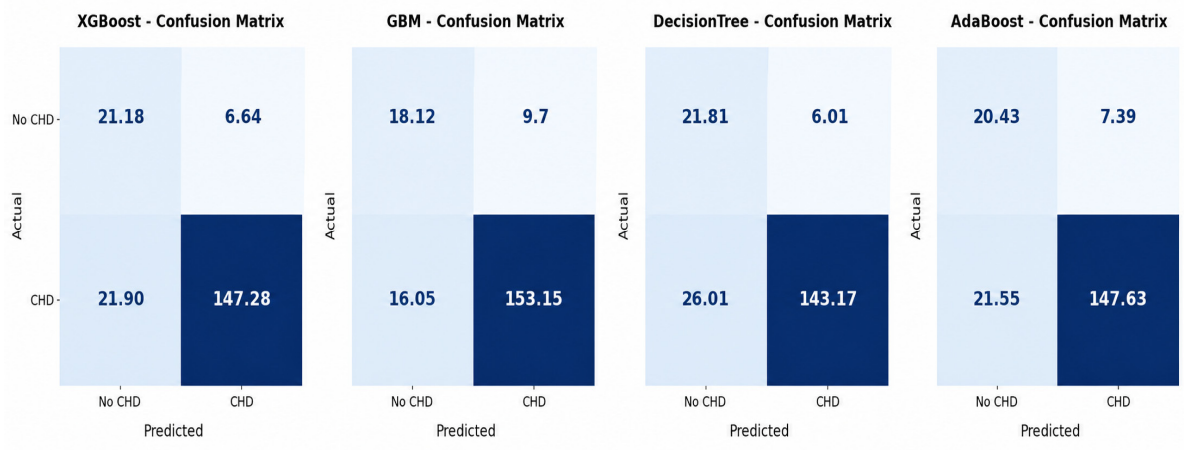

Figure 10: Bootstrap-averaged Confusion matrix for the best-performing models.

XGBoost was the most effective model among the competing models. The system achieves its best performance through its ability to correctly identify approximately 147.28 actual cases and 21.18 control cases while making 6.64 false positives and 21.90 false negatives. DT had the highest true negatives (21.81) and the fewest false positives (6.01), but false negatives amounted to 26.01, missing actual CHD cases, which is the highest among the models. AdaBoost comes close to XGBoost in true positive count (147.63 vs. 147.28) and false negatives (21.55 vs. 21.90), but it generates the most false positives (7.39) and the fewest true negatives (20.43), meaning it over-predicts CHD more than XGBoost, leading to unnecessary follow-ups or interventions for healthy patients. Similarly, GBM demonstrated highly competitive predictive performance. It recorded a true positive of approximately 152.50, which is the highest count among the models, while identifying 18.58 true negatives, being the lowest and the highest FP of 8.18 and 17.75 false negatives. The specificity of GBM is the lowest among the three models.

The performance of the models can be further quantified through sensitivity, specificity, and their combined measure, balanced accuracy and F1 score as shown in Table 4. To statistically evaluate the model's performance under a class imbalance setting, the study equally weighted sensitivity (CHD detection) and specificity (correct identification of non-CHD), using the balanced accuracy metric. This metric provides a prevalence-independent assessment of diagnostic discrimination and prevents inflated performance estimates that may arise from conventional accuracy in skewed clinical datasets.

In this study, GBM showed the highest sensitivity, but the lowest specificity, which reduced balanced accuracy. DT had the highest specificity but the lowest sensitivity. Adaboost was very close to XGBoost, but XGBoost was better in terms of sensitivity and specificity. Thus, XGBoost provides the best trade-off between detecting CHD and correctly ruling out non-CHD, resulting in the highest balanced accuracy overall. The F1 score analysis shows that AdaBoost achieves the highest F1 score, closely followed by

XGBoost and GBM, whereas the DT shows slightly lower performance due to reduced sensitivity despite high precision.

Further, a paired Wilcoxon signed-rank test was conducted on the generated results to compare the performance of XGBoost, Adaboost, GBM, and DT. XGBoost demonstrated statistically significant improvement over others in this study, as listed in Table 7.

Table 7: Paired wilcoxon signed-rank test comparing model performance across metrics. statistically significant differences ($p < 0.05$) are shown in bold.

Metric	XGBoost vs. DT	XGBoost vs. AdaBoost	XGBoost vs. GBM	AdaBoost vs. GBM	AdaBoost vs. DT
AUPRC	$\mathbf{2.56 \times 10^{-26}}$	$\mathbf{2.43 \times 10^{-29}}$	1.32×10^{-1}	$\mathbf{3.84 \times 10^{-8}}$	$\mathbf{1.20 \times 10^{-51}}$
AUC	$\mathbf{4.36 \times 10^{-11}}$	$\mathbf{2.07 \times 10^{-98}}$	$\mathbf{1.14 \times 10^{-41}}$	$\mathbf{3.55 \times 10^{-37}}$	$\mathbf{1.06 \times 10^{-80}}$
TPR	$\mathbf{7.51 \times 10^{-19}}$	5.95×10^{-2}	$\mathbf{1.39 \times 10^{-2}}$	$\mathbf{2.68 \times 10^{-7}}$	$\mathbf{1.75 \times 10^{-13}}$
TNR	$\mathbf{2.21 \times 10^{-15}}$	$\mathbf{5.89 \times 10^{-11}}$	$\mathbf{1.93 \times 10^{-27}}$	$\mathbf{4.12 \times 10^{-19}}$	$\mathbf{3.75 \times 10^{-28}}$
FPR	$\mathbf{2.47 \times 10^{-15}}$	$\mathbf{5.35 \times 10^{-11}}$	$\mathbf{1.98 \times 10^{-27}}$	$\mathbf{4.31 \times 10^{-19}}$	$\mathbf{3.77 \times 10^{-28}}$
FNR	$\mathbf{6.40 \times 10^{-19}}$	6.20×10^{-2}	$\mathbf{1.33 \times 10^{-2}}$	$\mathbf{2.81 \times 10^{-7}}$	$\mathbf{1.50 \times 10^{-13}}$
PPV	$\mathbf{2.02 \times 10^{-15}}$	$\mathbf{5.00 \times 10^{-17}}$	$\mathbf{2.35 \times 10^{-52}}$	$\mathbf{7.61 \times 10^{-11}}$	$\mathbf{1.07 \times 10^{-31}}$
NPV	$\mathbf{3.58 \times 10^{-24}}$	2.54×10^{-2}	$\mathbf{4.30 \times 10^{-7}}$	$\mathbf{6.88 \times 10^{-5}}$	$\mathbf{4.60 \times 10^{-22}}$
Balanced Accuracy	$\mathbf{3.51 \times 10^{-2}}$	$\mathbf{1.04 \times 10^{-64}}$	$\mathbf{2.44 \times 10^{-55}}$	$\mathbf{5.92 \times 10^{-29}}$	$\mathbf{3.91 \times 10^{-40}}$

The bold value indicates the significant p-value difference observed among the evaluated models. **Abbreviations**: AUPRC, Area Under the Precision–Recall Curve; AUC, Area Under the Receiver Operating Characteristic Curve; TPR, True Positive Rate (Sensitivity/Recall); TNR, True Negative Rate (Specificity); FPR, False Positive Rate; FNR, False Negative Rate; PPV, Positive Predictive Value (Precision); and NPV, Negative Predictive Value.

The results demonstrate that XGBoost gives better performance than AdaBoost for AUC, AUPRC, specificity, and balanced accuracy, though TPR and FNR show only low statistical difference. XGBoost and GBM seem similar in terms of AUPRC, but XGBoost had better overall discrimination, too, sensitivity, specificity and steadier balanced accuracy. GBM showed better class separation and reduced error rates when compared to Adaboost. XGBoost shows better results than DT for sensitivity, FNR, and balanced accuracy metrics. AdaBoost performs better than DT on all the metrics. Thus, XGBoost demonstrated the best balanced performance in this study as its advantage lies in its ensemble gradient boosting mechanism, which iteratively corrects errors from weak learners, resulting in better generalization and a more optimal trade-off between false positives and false negatives—which is exactly what is needed in a binary medical classification task like CHD prediction.

4.4 Comparison with Existing Studies

The performance of the present study was compared with [5], and the comparison results are presented in Table 8.

Prior work [5] demonstrated high specificity (TNR) for WSVM and logistic regression models (TNR > 0.94), but at the cost of relatively lower sensitivity (TPR ≈ 0.65), indicating a tendency to miss a substantial proportion of CHD cases. In contrast, models developed in the present study showed a more balanced trade-off between sensitivity and specificity, which is critical for screening applications. Thus, the present study demonstrates improved sensitivity and competitive AUC performance while maintaining acceptable specificity.

Table 8: Comparison of CHD prediction performance with prior literature [5] and the present study.

Paper/Model	TPR	TNR	AUC	GMean
WSVM [5]	0.692 (0.615–0.731)	**0.948** (0.946–0.949)	**0.819** (0.782–0.839)	**0.809** (0.764–0.832)
WSVM (Present study)	**0.749** (0.698–0.782)	0.827 (0.774–0.880)	0.791 (0.758–0.824)	0.775 (0.748–0.801)
LR [5]	0.654 (0.615–0.692)	**0.981** (0.970–0.987)	**0.815** (0.784–0.839)	**0.799** (0.757–0.826)
LR (Present study)	**0.710** (0.679–0.749)	0.758 (0.690–0.818)	0.745 (0.711–0.777)	0.731 (0.698–0.755)
WRF [5]	0.654 (0.615–0.731)	**0.930** (0.921–0.940)	0.799 (0.771–0.828)	0.786 (0.756–0.823)
WRF (Present study)	**0.913** (0.869–0.943)	0.727 (0.665–0.800)	**0.843** (0.809–0.875)	**0.798** (0.766–0.830)
Decision Tree (Present study)	0.889 (0.746–0.939)	0.792 (0.704–0.875)	0.857 (0.828–0.888)	0.809 (0.785–0.835)
XGBoost (Present study)	0.921 (0.788–0.948)	0.758 (0.686–0.844)	0.854 (0.822–0.887)	0.809 (0.782–0.837)
AdaBoost (Present study)	0.937 (0.750–0.954)	0.720 (0.652–0.828)	0.832 (0.798–0.866)	0.794 (0.765–0.823)
GBM (Present study)	0.916 (0.888–0.942)	0.700 (0.643–0.773)	0.845 (0.818–0.875)	0.791 (0.758–0.821)
ANN (Present study)	0.836 (0.809–0.864)	0.680 (0.619–0.741)	0.766 (0.730–0.800)	0.752 (0.715–0.784)
NB (Present study)	0.758 (0.589–0.869)	0.714 (0.591–0.833)	0.733 (0.696–0.770)	0.709 (0.677–0.738)

The bold value indicates the highest metric values achieved among the compared models.

Results were further compared and validated with [11], which reported AUC values of 0.84–0.92, accuracy of 0.75–0.86, and sensitivities of 0.70–0.88; the present study demonstrates AUC in the overlapping range and sensitivities ranging from 0.788–0.948, showing that the models XGBoost matches the TPR range and WRF exceeds with performance. Our study findings are also consistent with [9], which reported AUC values of 0.80–0.95 and F1-scores of 0.70–0.86.

The study was also compared with [12], and the results are presented in Table 9. The AUPRC value observed in the present study was 0.96 for the DT model, which is consistent with the ensemble-boosted CHD models reported in [12], where risk-aggregated feature representations enhanced the separability between CHD and non-CHD populations. The median AUPRC values summarized in Table 9 confirm that the tree-based methods show tight IQRs and low differences across multiple experiment runs.

Table 9: AUPRC comparison with [12].

AUPRC Comparison with [12]	Regularized Logistic Regression	Decision Tree	SVM	GBDT
Median [12]	0.935	0.942	**0.957**	**0.994**
Present study	**0.941**	**0.963**	0.947	0.961

The bold value indicates the highest AUPRC values achieved among the compared models.

The current research uses a different data type when compared with earlier investigations. Marelli et al. [12] used administrative claims data, which contains the yearly average number of CHD-related information billed by cardiovascular specialists for 24 subtype ICD-9 disease codes and shows documented medical events instead of actual patient assessment results. Luo et al. [5] investigated maternal interview survey records. Our research uses clinical data that was collected formally at a hospital setup at SSSSH, which also includes confirmed diagnostic results, and covers CHD as the majority class. The study by Rani et al. [8] reported very high accuracy (99%) but showed low sensitivity of 16%, which limits their use in medical settings.

5 Discussion

This study harnessed nine ML models and evaluated their CHD prediction capability on a hospital-based case–control dataset. Models like XGBoost and WRF capture these complex, non-linear interactions and intricate patterns, but are difficult to interpret. To address this interpretability challenge, one of the explainable artificial intelligence frameworks, SHAP [30,31], was employed to explain the predictions of

these high-performing complex models by quantifying the contribution and direction of influence of each risk factor on CHD prediction.

5.1 SHAP

SHAP functions as a game-theory-based method that demonstrates how each feature of a model affects its predictions. The system provides a method that enables users to understand black-box models through medical applications. Traditional ML models produce a single risk score or class label, which fails to explain the reasons behind a patient's classification as high or low risk and the specific elements that contributed to that choice, which reduces clinical confidence and makes it difficult to take action while eliminating the possibility of delivering customized patient guidance. SHAP provides an interpretable representation of feature contributions to model predictions by indicating whether specific variables are associated with higher or lower model-predicted CHD risk. The system provides a risk score together with a patient-specific risk explanation, which enables customized patient management.

5.1.1 SHAP-Based Feature Importance Analysis

The SHAP-based feature importance was analyzed using the best-performing XGBoost model from the study, and the results are presented in Fig. 11 and Table 10.

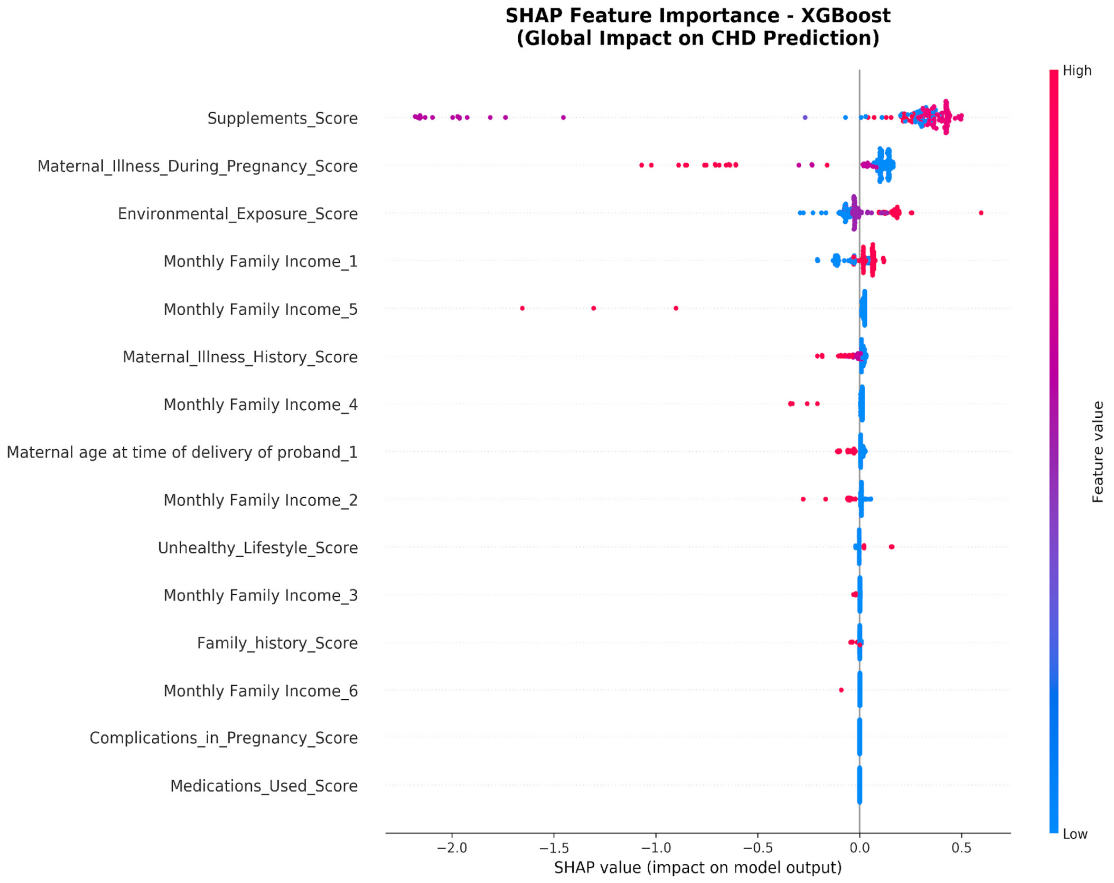

Figure 11: The SHAP summary plots from XGBoost model.

Table 10: SHAP-based feature importance, standardised effect magnitude, and clinical interpretation for CHD prediction in the present study using XGBoost.

Feature	Mean SHAP	Std. Effect	Clinical Interpretation
Supplements_Score	0.473	3.375	High
Maternal_Illness_During_Pregnancy_Score	0.164	0.846	High
Environmental_Exposure_Score	0.082	0.174	High
Monthly Family Income_1	0.060	−0.003	Low
Monthly Family Income_5	0.040	−0.171	Low
Maternal_Illness_History_Score	0.023	−0.310	Low
Monthly Family Income_4	0.019	−0.340	Low
Maternal age at time of delivery of proband_1	0.018	−0.349	Low
Monthly Family Income_2	0.017	−0.355	Low
Unhealthy_Lifestyle_Score	0.006	−0.442	Low
Monthly Family Income_3	0.002	−0.477	Low
Family_history_Score	0.002	−0.478	Low
Monthly Family Income_6	0.001	−0.484	Low
Medications_Used_Score	0.000	−0.494	Low
Complications_in_Pregnancy_Score	0.000	−0.494	Low

The SHAP summary plots for XGBoost models provide a global explanation of how individual features contribute to CHD risk prediction. Each plot shows the distribution of SHAP values for every feature across all samples. A SHAP value represents how much a feature pushes the prediction toward CHD (positive) or Non-CHD (negative) for each data point. Points are colored by the actual feature value (blue = low, red = high), enabling joint interpretation of magnitude and direction of impact.

In the models considered, 'Supplements Score' is the topmost of the feature list, which indicates that it has the highest impact on the model outputs. High values/red points shift the SHAP value toward the positive direction, which shows that lower supplement intake is strongly associated with a higher model-predicted probability of CHD risk, and that is consistent with known maternal nutrition risk factors. Secondly, 'Maternal Illness During Pregnancy' consistently shows high SHAP magnitudes. The top third feature is 'Environmental Exposure', which shows strong positive SHAP contributions when exposure is high, aligning with its biological relevance in congenital disorders. The fourth is set of 'Monthly Family Income' indicators (Income 1 to Income 6), which demonstrate a more nuanced pattern. The monthly family income distribution reflects a concentration of CHD high-risk correlation in the lower to middle socioeconomic brackets.

Around 67.40% (the highest proportion) of the mothers had family monthly income in the range of 9308–27,882 (in Rupees), indicating that most families belonged to middle-income groups. The income distribution is concentrated in lower socioeconomic groups, with limited representation at the upper end of the income range. In the plot, lower income categories (blue) generally push the prediction positively toward CHD, while higher categories (red) move the prediction in the negative direction. This aligns well with socioeconomic literature indicating that lower income is associated with limited access to antenatal care and higher congenital anomaly burden. These income encoded variables, therefore, behave predictably in the SHAP space.

The fifth important feature is 'Maternal Illness History score', which showed low SHAP variability. The analysis of 'maternal age at delivery', which is the top sixth listed risk factor from the given data, shows that the majority of mothers (79.62%) were young under 30 years, while 20.38% were 30 years or more. This shows that most deliveries occurred in younger maternal age groups, with only one-fifth of mothers in the advanced maternal age category. Although advanced maternal age is an established risk

factor for congenital anomalies [6], the predominance of younger mothers in the present dataset suggests that maternal age did not have a major impact on CHD risk in this study for the Indian population.

For the features 'Family history', 'Unhealthy Lifestyle', 'Medications Used', and 'Complications in Pregnancy Score', the SHAP distributions' magnitudes are low in the current study.

XGBoost yields larger SHAP value spreads for modifiable risk factors, such as supplement intake and maternal illness during pregnancy, highlighting that these features are the primary drivers of the model's risk-stratification capability.

5.1.2 SHAP Sub-Analysis

The "Supplements Score" and "Maternal illness during pregnancy" were identified as the most significant features using the SHAP analysis as described in Section 5.1.1. To further understand the individual feature contributions, these grouped variables were disaggregated into individual components, and the experimental pipeline was executed. The sub-analysis of the individual features is presented below.

(a) Supplements Score: It is composed of Folic Acid (SS_F), Iron (SS_I), Calcium (SS_C), and Multivitamins (SS_V). The Fig. 12 presents the ranked importance of individual supplements SHAP values. The results show a clear hierarchy of importance among individual supplements, with important clinical implications for prenatal care interventions. Folic Acid exemplifies dominance with a mean SHAP value of 0.8126, indicating 65.7% of the total supplement group contribution. This finding indicates that Folic Acid deficiency is the primary modifiable risk factor among all nutritional supplements for CHD prediction. The visualization substantiates that continuous, clinician-recommended folic acid intake during pregnancy emerges as the most influential and protective nutritional factor among the supplements analyzed. Iron supplementation is a moderate contributor, with a SHAP value of 0.3295 (26.6% of group contribution), while Calcium has a value of 0.0906 (7.3%), and Multivitamins (0.0037, 0.3%) demonstrate minimal independent predictive power. Based on these quantitative findings, we classify Folic Acid as a "Primary Driver" of CHD risk, Iron as a "Moderate Contributor", Calcium as a "Secondary Factor", and Multivitamins as having a "Weak Signal". These classifications directly inform clinical prioritization strategies.

(b) Maternal Illness During Pregnancy: The right subplot depicts the contribution of the top 15 maternal illness related features during pregnancy out of 17 maternal complications. It shows that the importance of Hyperemesis Gravidum (severe nausea/vomiting) ranks higher than other factors and its mean SHAP value is approximately 0.078. Other complications include hypertensive disorders of pregnancy, placental abnormalities, autoimmune conditions, and miscellaneous medical or obstetric complications not categorized individually contributed to mean SHAP value of 0.061. Weakness showed SHAP value of 0.032, and Swelling on Feet having 0.024. The 12 remaining complications even though they are clinically important as per conventional risk assessment methods, which include Gestational Diabetes, Oligohydramnios, Polyhydramnios, and Seizures/Fainting, Edema show near-zero SHAP values in the dataset considered in this study. Hyperemesis Gravidum, the most severe complication, has a 10-fold lesser impact than Folic Acid deficiency which has a 0.813 value while the entire complication group contributes 0.194, which is 6.4 times smaller than the supplement group total of 1.236, showing that nutritional factors which can be changed through Folic Acid supplementation have a greater impact on CHD risk prediction, so prenatal care protocols for CHD prevention should focus on Folic Acid supplementation.

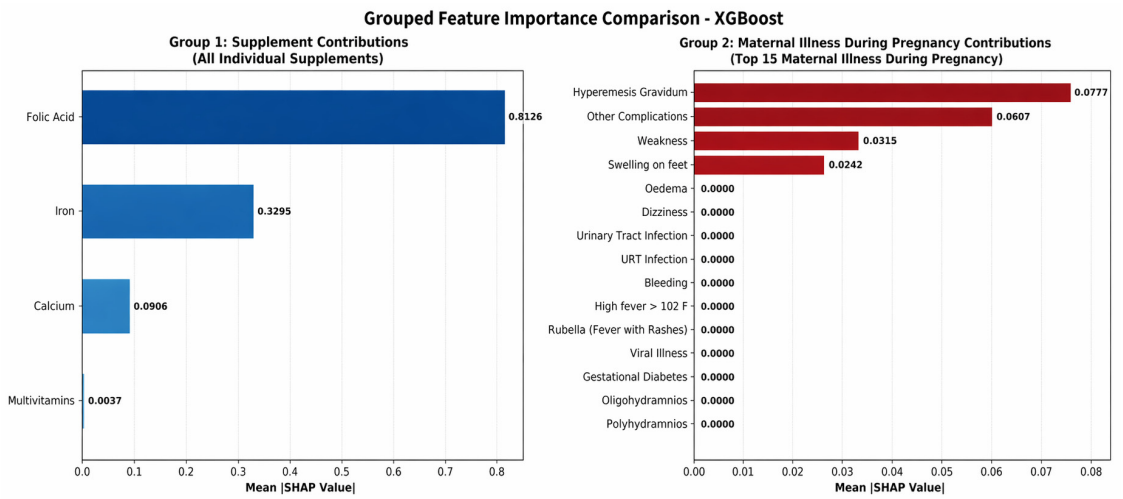

Figure 12: Group 1 (Supplement score): Four independent features and Group 2 (Maternal illness during pregnancy): Top 15 features.

The study [9] utilized maternal risk factor data from six counties in China, and the top features that contributed to the prediction included unhealthy lifestyle, annual per capita income, nutrition and folic acid supplementation, and maternal illness (with reference to Fig. 2 of [9]). When these features are compared with our SHAP results, it can be inferred that despite differences in geographical location, socioeconomic conditions, nutritional status, and maternal illness, continue to be major contributing factors influencing CHD.

5.2 Significance of Using an Indian Maternal Health Dataset

The study utilizes an Indian clinical dataset reflecting demographic, socioeconomic, environmental, and nutritional characteristics unique to Indian populations. Unlike many prior works conducted in Western or East Asian populations, Indian maternal health profiles are influenced by distinct factors such as variable nutritional status, heterogeneous socioeconomic gradients, region-specific environmental exposures, and diverse genetic backgrounds. These factors limit the direct generalizability of international CHD prediction models to Indian settings. The present work, therefore, addresses a critical translational gap by developing and validating a CHD risk model grounded in Indian maternal data.

CHD is a relatively rare condition, with a global birth prevalence of approximately 8–12 per 1000 live births (about 0.8–1.2%). In India, the reported prevalence is around 9 per 1000 live births. The hospital-collected dataset in this study contains a majority of CHD cases (84%), which is artificially enriched compared to the general population. This constitutes a key strength for modelling the feature spectrum associated with CHD, but it also means that raw model outputs will overestimate absolute risk.

While many statistical studies in the Indian context have examined the prevalence and influence of individual maternal risk factors on CHD, this is one of the first studies to harness predictive models that use tabular clinical data with an extensive set of features to assess multiple risk factors simultaneously. In this study, 55 independent maternal risk features were incorporated to enable a comprehensive, multivariate risk assessment.

5.3 Deployment Considerations for Screening

5.3.1 Prevalence Adjustment for Real-World Screening

CHD is a rare disease in the general population, with a birth prevalence of approximately 0.9% (9 per 1000 live births) in India. The probability estimates obtained in this study were derived from a hospital-based case-control study with a substantially higher CHD prevalence (84%) than in the general population. As a result, raw model outputs may overestimate true disease risk and drastically inflate PPV, whereas sensitivity remains unchanged. PPV is the metric that depends directly on disease prevalence. If we rewrite PPV ($PPV = \frac{TP}{TP+FP}$) using probability terms for the prevalence setting,

$$PPV = \frac{Sensitivity \times Prevalence}{Sensitivity \times Prevalence + (1 - Specificity) \times (1 - Prevalence)}.$$

This gap can be addressed by employing prevalence adjustment on the dataset before deploying the models for screening. An incremental prevalence-adjustment strategy can be adopted, whereby the proportion of control cases is progressively increased to better approximate population-level prevalence. In this context, probability adjustments of predicted values to reported prevalence [32], threshold optimization (to redefine the decision cut-off for the new prevalence setting) [33], and an Expectation Maximization (EM) based adjustment procedure can be applied to iteratively refine the predicted probabilities until the overall predicted prevalence converges to the true target prevalence. This stepwise readjustment should continue until false-positive rates begin to decline and the PPV stabilizes at an operationally optimal level, thereby aligning model performance with real-world screening conditions.

After prevalence adjustments, the models can be used as a rule-out or triage tool, rather than a diagnostic confirmation tool, supporting early referral decisions for fetal echocardiography while maintaining an appropriate balance between false positive burden and negative predictive confidence in large-scale screening programs. These deployment-ready models give a probability risk score for CHD prediction by flagging elevated maternal risk patterns for closer monitoring. Final diagnosis remains dependent on fetal echocardiography and specialist evaluation. The system would strictly act as a decision-support tool that does not recommend termination, intervention, or treatment decisions and is not a replacement for a clinician.

5.3.2 Clinical Deployment Workflow

The proposed system can be used for screening high-risk pregnancies for CHD from a particular geographic area by following the steps below:

1. Information about the mothers should be gathered using organized questionnaires during pregnancy checkups for a specific geography.
2. Risk factors have to be sorted into important medical categories.
3. The trained ML system should be calibrated for the new setting to analyze the maternal information and produce a risk score for CHD.
4. SHAP-based explanations to be provided at a granular level to help give doctors clear information about which risk factors have to be targeted for improving maternal health.
5. High-risk cases should be flagged for further diagnostic evaluation (e.g., fetal echocardiography).

This enables early and simple screening in places with limited resources.

5.4 Limitations and Future Work

The main limitations identified from the study are as follows.

Spectrum bias: The retrospective case and control data used for the study were collected from a single hospital in Kharghar, Mumbai, India, with a majority of CHD cases accounting for enrichment in the disease spectrum of subjects. Though imbalance handling techniques were efficiently used in the study, these artificially enriched CHD cases and fewer controls in the study dataset present a limitation for generalizing the models for deployment.

- Selection Bias: India is a large and diverse country, with big differences in diet, genetics, and mothers' health that can change the risk of CHD. The maternal risk factor profiles used in the case-control design may not fully represent the real-world population data, which limits generalization. So, the models developed in this study need to be tested further with external data from other hospitals and regions (multi-center cohorts), and adjusted for different regions, before they can be used to screen large groups of people across India.
- Recall Bias: For this retrospective study, data was collected through questionnaire. So, there may be a possibility that the person providing the information might not clearly remember past exposures, because of which the information might be missing, recorded differently, or selected in a particular way, which can affect model's performance and can lead to bias.
- In the case-control design, the artificially enriched cases do not reflect the actual prevalence screening setting. Although balanced accuracy, sensitivity, specificity, bootstrap confidence intervals, and repeated evaluation strategies were incorporated for interpretation, performance metrics such as AUPRC, accuracy may appear inflated due to class prevalence effects and should be interpreted in the context of prevalence imbalance.

Future work could be on expanding the study from a single-center study to a multi-center study design to enhance scalability across diverse populations and clinical settings [34], with Leave-One-Centre-Out Validation (LOCO) validation to ensure model robustness across diverse clinical settings. Further, federated learning–based, privacy-preserving framework can be effectively employed, wherein hospitals are not required to share patient data, but model training occurs locally at each institution in a collaborative manner for generalizability, and lightweight models can be explored for real-world deployment settings and adaptation to other domains.

6 Conclusion

In this work, a hospital data-based CHD prediction was performed using Machine Learning algorithms. Among the models used, tree-based methods showed better performance, and XGBoost consistently demonstrated the best overall balanced performance. Building on this result, the present work contributes to one of the first ML–based maternal risk prediction frameworks for CHD in the Indian population. The proposed models demonstrate potential for CHD risk prediction; however, further external validation and prospective evaluation are required before clinical implementation.

In addition to the investigation of predictive performance, a SHAP-based interpretability analysis showed that modifiable maternal factors comprise a clinically relevant risk profile. Continuous intake of recommended supplements during pregnancy is seen as a protective factor, whereas maternal illness during pregnancy and environmental exposures are regarded as the most significant contributors to increased CHD risk. The maternal factors, like age and family predisposition, have been found to have a lesser role according to this study.

Acknowledgement: Our heartfelt acknowledgment to Dr. C. Sreenivas, Chairman, Sri Sathya Sai Sanjeevani Research Foundation & Sri Sathya Sai Health and Education Trust, India and Dr. Radha Joshi and Ms. Manasi Bhoite and the clinicians from Sri Sathya Sai Sanjeevani Centre for Child Heart Care & Training in Pediatric Cardiac Skills at Kharghar, Navi Mumbai for their invaluable guidance, support in procuring the data and expertise in case classification. We are extremely grateful to Dr. Annie Arvind and Dr. Harsha, and all the clinicians from the Pediatrics Department of Sri Madhusudan Sai Institute of Medical Sciences and Research, Muddenahalli, Karnataka, for their valuable guidance and support.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Shruthi S: Conceptualisation; Methodology; Investigation; Paper curation; Formal analysis; Writing—original draft, Visualization, Validation, Project administration. D. Hanumanth Rao Naidu: Conceptualization; Supervision; Writing—review and editing, Visualization, Validation, Project administration. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data not available due to ethical restrictions.

Ethics Approval: This retrospective study was approved by the Institutional Ethics Committee (IEC) of Sri Sathya Sai University for Human Excellence (Approval No. IEC/Certificate/01/2025, dated 20 January 2025). This approval covered the overall study design, data analysis plan, and the use of de-identified data for machine learning modelling. Meanwhile, the IEC of Sri Sathya Sai Sanjeevani Centre for Child Heart Care & Training in Pediatric Cardiac Skills at Kharghar, Navi Mumbai Kharghar (Approval No. -SSK0051/V3/PR/2024/IEC-11, approved 2 August 2025) was obtained specifically for the collection and retrospective use of anonymized electronic health records (EHR) and maternal questionnaire data. The requirement for individual informed consent was waived due to the retrospective use of de-identified data. The study was conducted in accordance with the Declaration of Helsinki and followed TRIPOD guidelines.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AdaBoost	Adaptive Boosting
ANN	Artificial Neural Network
AUC	Area Under Receiver Operating Curve
AUPRC	Area Under the Precision-Recall Curve
BPNN	Back-Propagation Neural Network
CHD	Congenital Heart Disease
CTG	Cardiotocography
CVD	Cardiovascular Disease
DT	Decision Tree
ECG	Electrocardiography
EF	Ejection Fraction
EHR	Electronic Health Record
FNR	False Negative Rate
FPR	False Positive Rate
GBM	Gradient Boosting Machine
GMean	Geometric Mean
IoMT	Internet of Medical Things
LIME	Local Interpretable Model-Agnostic Explanations
LR	Logistic Regression
ML	Machine Learning
NB	Naive Bayes

NPV	Negative Predictive Value
PPV	Positive Predictive Value (Precision)
RF	Random Forest
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
TNR	True Negative Rate (Specificity)
TPR	True Positive Rate (Sensitivity/Recall)
WRF	Weighted Random Forest
WSVM	Weighted Support Vector Machine
XGBoost	Extreme Gradient Boosting

References

1. Xu J, Li Q, Deng L, Xiong J, Cheng Z, Ye C. Global burden and epidemiological trends of congenital heart disease in children under five years: 1990–2021. Front Cardiovasc Med. 2025;12:1522644. [CrossRef].
2. Saxena A. Congenital heart disease in India: A status report. Indian Pediatr. 2018;55(12):1075–82. [CrossRef].
3. Ottaviani G, Buja LM. Congenital heart disease: Pathology, natural history, and interventions. In: Buja LM, Butany J, editors. Cardiovascular pathology. 4th ed. Waltham, MA, USA: Elsevier Academic Press; 2016. p. 611–47. [CrossRef].
4. Khan K, Ullah F, Syed I, Ullah I. The role of machine learning in congenital heart disease diagnosis: Datasets, algorithms, and insights. arXiv:2501.04493. 2025.
5. Luo Y, Li Z, Guo H, Cao H, Song C, Guo X, et al. Predicting congenital heart defects: A comparison of three data mining methods. PLoS One. 2017;12(5):e0177811. [CrossRef].
6. Wu L, Li N, Liu Y. Association between maternal factors and risk of congenital heart disease in offspring: A systematic review and meta-analysis. Matern Child Health J. 2023;27(1):29–48. [CrossRef].
7. Li H, Luo M, Zheng J, Luo J, Zeng R, Feng N, et al. An artificial neural network prediction model of congenital heart disease based on risk factors: A hospital-based case-control study. Medicine. 2017;96(6):e6090. [CrossRef].
8. Rani S, Masood S. Predicting congenital heart disease using machine learning techniques. J Discrete Math Sci Cryptogr. 2020;23(1):293–303. [CrossRef].
9. Salehi A, Khedmati M. Identifying at-risk patients for congenital heart disease using integrated predictive models and fuzzy clustering analysis: A cross-sectional study. Heliyon. 2024;10(20):e39609. [CrossRef].
10. Hoodbhoy Z, Jiwani U, Sattar S, Salam R, Hasan B, Das JK. Diagnostic accuracy of machine learning models to identify congenital heart disease: A meta-analysis. Front Artif Intell. 2021;4:708365. [CrossRef].
11. Dehghan B, Sabri MR, Ahmadi A, Ghaderian M, Mahdavi C, Ramezani Nejad D, et al. Identifying the factors affecting the incidence of congenital heart disease using support vector machine and particle swarm optimization. Adv Biomed Res. 2023;12:130. [CrossRef].
12. Marelli AJ, Li C, Liu A, Nguyen H, Moroz H, Brophy JM, et al. Machine learning informed diagnosis for congenital heart disease in large claims data source. JACC Adv. 2024;3(2):100801. [CrossRef].
13. Iyer KS, Sivalingam S. Congenital heart disease in low- and middle-income countries: Can India show the way? Indian J Thorac Cardiovasc Surg. 2025;41(6):657–63. [CrossRef].
14. Dolk H, McCullough N, Callaghan S, Casey F, Craig B, Given J, et al. Risk factors for congenital heart disease: The Baby Hearts Study, a population-based case-control study. PLoS One. 2020;15(2):e0227908. [CrossRef].
15. Cao H, Wei X, Guo X, Song C, Luo Y, Cui Y, et al. Screening high-risk clusters for developing birth defects in mothers in Shanxi Province, China: Application of latent class cluster analysis. BMC Pregnancy Childbirth. 2015;15:343. [CrossRef].
16. Ramegowda S, Ramachandra NB. Parental consanguinity increases congenital heart diseases in South India. Ann Hum Biol. 2006;33(5–6):519–28. [CrossRef].
17. Dev D, Sharma R, Sharma M, Agrawal K, Garg M. Evaluation of consanguinity as a risk factor for congenital heart diseases. Int J Contemp Pediatr. 2016;3(3):868–71. [CrossRef].
18. Santander Ballestín S, Giménez Campos MI, Ballestín Ballestín J, Luesma Bartolomé MJ. Is supplementation with micronutrients still necessary during pregnancy? A review. Nutrients. 2021;13(9):3134. [CrossRef].

19. Joshi R, Bhoite M, Mandhare P, Nath S, Kapoor S, Wadke R, et al. Population-based comparative nutritional status of unoperated congenital heart defects patients from a tertiary pediatric cardiac centre in India. Preprint. 2025. [CrossRef].
20. Joshi RO, Kukshal P, Kumar A, Murthy PR, Chellappan S, Manohar K, et al. Congenital heart defects and environmental factors: A snapshot of CHD cohort from a tertiary cardiac care centre in India. Ann Pediatr. 2023;6(2):1123.
21. Wang T, Chen L, Ni B, Sheng X, Huang P, Zhang S, et al. Maternal pre-pregnancy/early-pregnancy smoking and risk of congenital heart diseases in offspring: A prospective cohort study in Central China. J Glob Health. 2022;12:11009. [CrossRef].
22. Alhindal M, Janahi J, D'Angelo EC, Lisignoli V, Palmieri R, Cutrì A, et al. Impact of smoking on cardiovascular health: Mechanisms, epidemiology and specific concerns regarding congenital heart disease. Int J Cardiol Congenit Heart Dis. 2025;20:100581. [CrossRef].
23. Deng C, Pu J, Deng Y, Xie L, Yu L, Liu L, et al. Association between maternal smoke exposure and congenital heart defects from a case-control study in China. Sci Rep. 2022;12:14973. [CrossRef].
24. Helle E, Priest JR. Maternal obesity and diabetes mellitus as risk factors for congenital heart disease in the offspring. J Am Heart Assoc. 2020;9(8):e011541. [CrossRef].
25. Liu S, Joseph KS, Lisonkova S, Rouleau J, Van den Hof M, Sauve R, et al. Association between maternal chronic conditions and congenital heart defects: A population-based cohort study. Circulation. 2013;128(6):583–9. [CrossRef].
26. Su H, Guo E, Woodward M, He JR, Waterboer T, Schuermans A, et al. First trimester maternal infections and offspring congenital heart defects: A meta-analysis. Eur Heart J. 2026;47(7):794–812. [CrossRef].
27. Hashim ST, Alamri RA, Bakraa R, Rawas R, Farahat F, Waggass R. The association between maternal age and the prevalence of congenital heart disease in newborns from 2016 to 2018 in a single cardiac center in Jeddah, Saudi Arabia. Cureus. 2020;12:e7463. [CrossRef].
28. Miao Q, Dunn S, Wen SW, Lougheed J, Yang P, Davies M, et al. Association between maternal marginalization and infants born with congenital heart disease in Ontario Canada. BMC Public Health. 2023;23(1):790. [CrossRef].
29. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. Berkeley, CA, USA: University of California, Department of Statistics; 2004.
30. Li J, Liu S, Hu Y, Zhu L, Mao Y, Liu J. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: Retrospective cohort study. J Med Internet Res. 2022;24(8):e38082. [CrossRef].
31. Feretzakis G, Sakagianni A, Anastasiou A, Kapogianni I, Bazakidou E, Koufopoulos P, et al. Integrating shapley values into machine learning techniques for enhanced predictions of hospital admissions. Appl Sci. 2024;14(13):5925. [CrossRef].
32. Gorgels KMF, van Iersel SCJL, Keijser SFA, Hoebe CJPA, Wallinga J, van Hoek AJ. Estimating infection prevalence using the positive predictive value of self-administered rapid antigen diagnostic tests: An exploration of SARS-CoV-2 surveillance data in the Netherlands from May 2021 to April 2022. PLoS One. 2024;19(2):e0298218. [CrossRef].
33. Balayla J. Prevalence threshold and the geometry of screening curves. PLoS One. 2020;15(10):e0240215. [CrossRef].
34. Reed M, Rampono B, Turner W, Harsanyi A, Lim A, Paramalingam S, et al. A multicentre validation study of a smartphone application to screen hand arthritis. BMC Musculoskelet Disord. 2022;23(1):433. [CrossRef].