JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org ISSN 2988-7003 (online)

Machine Learning-Based Outcome Prediction in Isolated Ventricular Septal Defects

Nurdan Erol[1], Çiğdem Erol[2], and Ilkim Ecem Emre[3]*

[1] *Health Sciences University Zeynep Kamil Gynecology and Pediatrics Training and Research Hospital, Türkiye*
[2] *Istanbul University, Türkiye*
[2] *Marmara University, Türkiye*

ARTICLE HISTORY
Received: 10 June 26
Final Revision: 26 June 26
Accepted: 28 June 26
Online Publication: 30 June 26

KEYWORDS
Congenital Heart Disease, Machine Learning, Risk Stratification, XGBoost, Ventricular Septal Defect

CORRESPONDING AUTHOR
ecem.emre@marmara.edu.tr

DOI
10.37034/medinftech.v4i2.151

A B S T R A C T

Ventricular Septal Defect (VSD) is one of the most common congenital heart defects. Predicting whether isolated VSD will close spontaneously, require surgical intervention, or remain unclosed is essential for optimizing patient management and avoiding unnecessary treatment. This study aimed to develop and evaluate machine learning (ML) models for predicting VSD outcomes using maternal and neonatal clinical characteristics. A retrospective dataset of 382 patients with isolated VSD was analyzed and categorized into spontaneous closure, surgical closure, and non-closure outcomes. Data preprocessing included duplicate removal and listwise deletion of records with missing values. To address class imbalance, random undersampling and oversampling were applied exclusively to the training set (80%), while the independent test set (20%) remained unchanged. Five ML algorithms-Decision Tree, Random Forest, K-Nearest Neighbor, Naive Bayes, and XGBoost-were evaluated using accuracy, macro-average area under the receiver operating characteristic curve (AUC), and class-specific F1-scores. XGBoost achieved the best overall performance with an accuracy of 65.8% and a macro-average AUC of 0.81, demonstrating balanced classification across all outcome groups. Although Decision Tree and Random Forest produced the highest F1-score (92.3%) for the minority surgical closure class, their overall multiclass performance was inferior to XGBoost. Sampling strategies had minimal impact on overall predictive performance, although ensemble-based methods showed greater robustness to class imbalance. These findings suggest that ML, particularly XGBoost, provides a promising approach for early risk stratification of isolated VSD, supporting personalized clinical decision-making and improving identification of patients requiring surgical intervention.

1. Introduction

A ventricular septal defect (VSD) is the presence of one or more holes in the interventricular septum. It accounts for 40% of all congenital heart anomalies [1], [2]. The prevalence of VSD is reported to be as high as 5.7% per 1,000 patients [1]. VSDs may indicate major congenital heart disease but are characterized primarily as single lesions. The clinical and symptomatic profiles of isolated VSDs vary depending on their type, location, number, and size. VSDs are generally considered to be benign, with reports suggesting that 88.9% of muscular lesions close spontaneously within the first year [3]. Among the 799 patients with isolated VSD, spontaneous closure occurred in 42.7% of muscular defects, 13.1% of perimembranous defects, and 25% of cases involving multiple defects [4]. However, the long-term clinical trajectory of isolated VSDs remains highly unpredictable and is not limited strictly to spontaneous

closure. Machine learning (ML) technologies have recently emerged as a tool for use in medical applications. For instance, a study conducted across multiple centers in China attempted to use artificial intelligence (AI) to predict the spontaneous closure of perimembranous VSDs [5]. In their study, at a median of 31 months, 12·1% of patients experienced spontaneous closure. However, follow-up of these cases revealed an increased risk of ventricular dysfunction, and a higher susceptibility to endocarditis, pulmonary arterial hypertension, and arrhythmia [6]. It has been reported that the Danish group has a shorter life expectancy than the general population, irrespective of whether they have undergone surgery for an isolated VSD [6]. Reports have also suggested that these cases lead to a lifelong burden of cardiovascular morbidity, irrespective of whether the defects are closed [6]. Once Eisenmenger syndrome is excluded, the rate of heart failure development is reportedly 2.5 times higher in cases of unclosed isolated VSD and 1.5 times higher in cases that have undergone surgery, compared to the general population [7]. Recent studies reveal that the history of VSDs requires further investigation. These studies emphasize the importance of conducting an in-depth follow-up investigation into VSDs and their implications for the progression of the condition, regardless of treatment.

The rate at which VSDs close spontaneously varies depending on the type of defect and the patient's age. For example, 65–75% of muscular VSDs close spontaneously, whereas the rate is lower for perimembranous VSDs [5]. Critical to clinical decision-making, predicting spontaneous closure facilitates the avoidance of unnecessary surgical operations and reduces the risk of delayed treatment. Traditionally, the prediction of spontaneous closure is based on echocardiographic parameters and clinical experience. However, these methods are characterized by limited accuracy and a lack of standardization [5]. This lack of objective tools often leads to either delayed surgical referrals or premature, unnecessary interventions.

To overcome these limitations, AI and ML techniques have lately been adopted for the analysis of medical data [8]. These techniques facilitate the early diagnosis and risk prediction of cardiovascular diseases. For instance, classification algorithms such as Naive Bayes, the Decision Tree, K-Nearest Neighbour (K-NN), and Random Forest were tested on the Cleveland dataset. The K-NN algorithm delivered the best performance, achieving an accuracy rate of 90.78% [9]. While studies utilizing data from non-VSD datasets (such as the Cleveland coronary dataset or general heart disease registries) do not share our exact clinical objective, they are highly relevant as they demonstrate the overarching methodological validity of ML in extracting non-linear patterns from multi-variable, tabular clinical data where human interpretation remains highly subjective. According to [10], artificial intelligence can improve the accuracy of diagnostic methods such as echocardiography, magnetic resonance imaging (MRI), and electrocardiography (ECG) when used to detect congenital heart diseases (CHDs). This conclusion is based on their review of artificial intelligence applications in pediatric cardiology conducted between 2002 and 2022. Innovative approaches such as the early detection of fetal heart anomalies using deep learning models [11] and the classification of pathological murmurs using artificial neural networks support clinical decision-making processes [12]. Natural language processing (NLP) methods, in particular, allow predictive models to be developed by automatically extracting data from unstructured clinical reports.

Although the literature on ML models for predicting VSD closure remains limited, the field has attracted growing interest in recent years. Several studies have used clinical, echocardiographic, and demographic data to predict the likelihood of VSD closure occurring spontaneously or the need for surgical or percutaneous intervention.

Study [13] concluded that the spontaneous closure rate was 42.8% after two years of monitoring 257 fetuses diagnosed with an isolated VSD. The logistic regression model was able to predict the probability of spontaneous closure with 82.6% accuracy. Nevertheless, this study relied on a basic linear model that cannot capture complex feature interactions, and its scope was restricted solely to fetal cases without considering post-natal maternal-neonatal risk combinations.

Deep learning has also been explored in congenital heart disease. Study [14] developed a deep learning model based on electrocardiograms (ECGs) for the diagnosis of atrial septal defects (ASDs) with 89% accuracy, yet their model was structural-specific to ASDs and did not address the dynamic closure trajectories found in ventricular anomalies. Study [15] employed five ML algorithms (logistic regression, Lasso, Random Forest, Gradient Boosting Machine [GBM], and Support Vector Machine [SVM]) to predict the risk of late-onset and prolonged conduction block following transcatheter occlusion in the treatment of perimembranous VSDs (AUC: 0.82). However, their focus was strictly restricted to post-procedural complications rather than the initial, pre-interventional clinical pathway.

More recently, Sun et al. [5] achieved high accuracy (95% AUC) in predicting the spontaneous closure of perimembranous VSDs using a natural language processing (NLP) and ML-based model. Despite its impressive sample size, a major limitation of this multi-center study is that it treated VSD tracking as a binary classification problem, ignoring patients who fall into the stable, persistent non-closure group that requires long-term conservative monitoring.

Following the preliminary studies of the 1950s, the introduction of AI and ML to the field of medicine has

marked a significant advance [8]. Techniques including deep learning (DL) and convolutional neural networks (CNNs) have achieved a high level of accuracy in the analysis of medical images and clinical data. Based on these developments, there is evidence to suggest that a similar approach could lead to the successful management of complex congenital diseases such as VSD. ML algorithms are used in this article to develop an understanding of the natural history of VSDs and predict spontaneous closure. In this respect, the aim of this study was to determine the natural history of cases of isolated VSD followed up at our center to verify the necessity of intervention and evaluate the potential of ML to assess the factors affecting VSD cases.

Existing ML applications in pediatric cardiology almost exclusively employ binary classification (e.g., spontaneous closure vs. intervention), grouping immediate surgical candidates and benign, stable non-closure patients into the same negative category. Clinically, treating a stable non-closure patient as a surgical candidate leads to overtreatment and severe parental anxiety, whereas treating a surgical candidate as a simple non-closure risks delayed intervention. Predicting a three-class trajectory (Spontaneous Closure, Surgical Intervention, and Persistent Non-Closure) provides far superior clinical value by preventing both clinical pitfalls.

To address this gap, this study introduces a novel hybrid approach that details specific clinical and maternal-neonatal variables as the basis for prediction. Our feature space combines Maternal Obstetric Factors (including maternal age, systemic diseases, smoking, and pregnancy type) and Neonatal Baseline Characteristics (such as birth weight, gestational age, gender, defect location, and associated minor anomalies) to construct a comprehensive predictive profile.

To powerfully capture these interactions, five machine learning algorithms with different mathematical foundations have been deployed: Naive Bayes, which provides a simple probabilistic basis; K-NN, which evaluates distance-based similarities; Decision Tree, which offers interpretable rules for clinicians; Random Forest, which reduces variance; and XGBoost, which maximizes predictive power by combating class imbalances.

The primary objective of this study is to develop and validate a multi-class machine learning framework capable of predicting the three distinct clinical trajectories of isolated VSDs (Spontaneous Closure, Surgical Intervention, or Persistent Non-Closure) using baseline maternal and neonatal variables. Specifically, this study aims to identify the most critical maternal and neonatal predictive factors influencing VSD progression, critically compare the performance of standard baseline algorithms against advanced tree-based ensembles under class imbalance conditions, and establish a clinically viable prediction model that supports personalized risk stratification in pediatric cardiology.

2. Research Method

2.1. Dataset & Pre-processing

Dataset used in the study is obtained from Health Sciences University Zeynep Kamil Gynecology and Pediatrics Training and Research Hospital. Approval for the dataset was granted by the Health Sciences University Zeynep Kamil Gynaecology and Pediatrics Training and Research Hospital's Ethics Committee in Clinical Research on 20/12/2023 (Ethics approval number: 175).

The basic steps of pre-processing were followed to clean and prepare the dataset for ML analysis. All of the variables included in the dataset are classified as categorical. The initial dataset contained data from 602 patients (rows/instances). During pre-processing, columns with missing values (Birth_Week_Group and Birth_Weight_Group) were identified (n = 207) and the rows with missing values in these columns removed from the dataset. These missing values were examined before model development. Since these variables contained a substantial proportion of missing values (34.4%), imputation was not preferred because it could introduce additional uncertainty and potential bias into the predictive models. Duplicate records (n = 13) were also identified, by comparing all available clinical attributes and outcome labels. Repeated instances were removed, and only unique patient records (n = 382) were retained in the final dataset for model development. The number of instances could be summarized as follows:

1. Initial number of patients = 602 patients
2. Removed due to missing values = 207 patients
3. Removed due to duplicated values = 13 patients
4. Final version of the dataset = 382 patients

The dataset used consisted of 382 instances (rows) and 16 variables (columns), 15 categorical predictor variables and one target variable. Predictor variables cover the health status of the mother, the type of pregnancy, birth characteristics, and various health conditions observed in the infant. The dataset is particularly suited to analyses focusing on VSD. A breakdown of the classes of each variable is presented in Table 1. Prior to model training, categorical predictors were converted into factor variables in R to enable processing by machine learning models.

The target variable in the dataset is the outcome to be classified. It represents the closure process of VSD and includes three classes/categories: Spontaneous closure (closed without medical intervention), surgical closure (closed through surgical intervention), and non-closure (remains open). The goal is to predict one of these three

outcomes based on various health factors and the mother's medical conditions. The distribution of the data in each class of the target variable is imbalanced. The distribution of the classes is also given in Table 1 in the Appendix.

No additional feature selection was performed before model training because the number of predictors was limited and all variables were selected based on clinical relevance. Instead, mode-based feature importance analysis was performed to evaluate the contribution of individual predictors. Feature importance analysis was also conducted using the best-performing XGBoost model. Relative importance scores were calculated and normalized, with the most influential predictor assigned a score of 100. Since categorical variables were converted into dummy variables during XGBoost training, feature importance values represent category-level predictor contributions.

Although the number of instances in Class 1 (spontaneous closure) and Class 3 (non-closure) of the target variable closely matches, the number of instances in Class 2 (surgical closure) causes an imbalance across the entire variable. Due to the imbalanced nature of the original dataset, undersampling and oversampling methods were employed to address the imbalance between the target variable's classes. Balancing methods were only applied to the training dataset (80%). The test dataset (20%) remains unchanged. According to the undersampling method, the number of instances in all classes is reduced to match the number of instances in the minority class (Class 2). In contrast, the oversampling method increases the number of instances in the other classes to match the number of instances in the majority class (Class 3). To apply sampling methods, upSample and downSample functions in caret package was used [16]. After balancing process, the distribution of the classes is given in the Table 1. All analyses were conducted with the original dataset, the undersampled dataset (Downsampled) and the oversampled dataset (Upsampled).

Table 1. Distribution of The Instances in The Target Variable in Datasets

Class	Original	Original (training dataset)	Upsampled	Downsampled
1: Spontaneous closure	170 (44.50 %)	136 (44.44 %)	141	29
2: Surgical closure	36 (9.42 %)	29 (9.48 %)	141	29
3: Non-closure	176 (46.07 %)	141 (46.08%)	141	29
Total number of instances	382 (100%)	306 (100%)	423 (100%)	87 (100%)

The dataset is divided into a training and a testing set using a stratified hold-out method [17]. Accordingly, 80% of the dataset was used for training and 20% was reserved for testing model performance, while preserving the class distribution of the target variable across both sets. To prevent data leakage, the test dataset was separated before applying any sampling methods. Class balancing was performed only within the training dataset for model development and cross validation. The 5-fold cross-validation method [18], repeating three times in the 80% training dataset, was used and this also provided hyperparameter optimization to be performed. The independent test set remained untouched throughout preprocessing and model optimization. To ensure reproducibility, fixed random seeds were used during data partitioning and model training. Specifically, set.seed(8) was used for the stratified train-test split, whereas set.seed(5) was used during cross-validation and model training.

2.2. Model Performance Evaluation Criteria

The results of the classification models were evaluated based on accuracy and the area under the curve (AUC) metric. Multiclass AUC was calculated as the macro-average of one-vs-rest AUC values across all outcome classes. As the study focuses on multiclass classification, the metrics for each class (i.e. the classes of the target variable) were also evaluated. Due to the imbalanced nature of the original dataset, the F-score was also used to compare the performance of the models in each class [19], [20].

2.3. Tools Used

In the first stage, the dataset was cleaned using a Microsoft Excel spreadsheet. Then, pre-processing procedures and model construction were applied using the R language in R-Studio (version 2024.12.1). Different packages for R-Studio were utilized. The dataset was imported using the readxl package [21]. Data preprocessing, transformation, data partitioning, class balancing procedures, hyperparameter tuning, model training, model evaluation, and variable importance analyses were performed using caret package [22]. Result organization was employed using dplyr [23] and tidyr packages [24]. ROC-AUC analyses were conducted using the pRoc package [25]. Visualization of model performance metrics was performed using ggplot2 [26] and reshape2 packages [27]. Final results were exported using writexl package [28].

2.4. Statistical Analysis

Five different ML methods were used to develop a classification model for VSD in the study. Decision trees (rpart), Random Forest (rf), K-Nearest Neighbour (knn), Naïve Bayes (nb) and XGBoost (xgb) algorithms were used as part of this process;

Naive Bayes (NB): A probabilistic classifier based on Bayes' Theorem, serving as a baseline model that

assumes conditional independence among clinical features.

K-Nearest Neighbors (KNN): An instance-based, non-parametric algorithm that classifies cases based on distance metrics (Euclidean distance) in the multi-dimensional feature space.

Decision Tree (DT - CART): A non-linear, rule-based algorithm that recursively splits data, providing a transparent, white-box visual logic highly intuitive for clinicians.

Random Forest (RF): An ensemble bagging classifier that aggregates multiple un-pruned decision trees trained on bootstrap samples to minimize variance and avoid overfitting.

eXtreme Gradient Boosting (XGBoost): An advanced gradient-boosting framework designed for high speed and predictive efficiency. It sequentially builds trees to minimize a multi-class loss function via gradient descent, inherently handling tabular class imbalances through regularization.

Hyperparameter optimization was executed using the caret package's tuning procedure, coupled with a 5-fold cross-validation repeated 3 times on the training dataset (80% split). The exact ranges of the hyperparameters tested and their final optimized values are detailed in Table 2.

Table 2. Optimized Hyperparameter Values of The Models

Model	Sampling	Optimized Hyperparameters
rpart	Original	cp = 0.042
	Upsampled	cp = 0.025
	Downsampled	cp = 0.011
rf	Original	mtry = 23
	Upsampled	mtry = 23
	Downsampled	mtry = 45
knn	Original	k = 5
	Upsampled	k = 7
	Downsampled	k = 9
nb	Original	laplace = 0,usekernel = TRUE,adjust = 1
	Upsampled	laplace = 0,usekernel = FALSE,adjust = 1
	Downsampled	laplace = 0,usekernel = FALSE,adjust = 1
xgb	Original	nrounds = 150, max_depth = 1, eta = 0.3, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1, subsample = 0.5
	Upsampled	nrounds = 100, max_depth = 1, eta = 0.4, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1, subsample = 0.5
	Downsampled	nrounds = 50, max_depth = 1, eta = 0.4, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1, subsample = 0.75

3. Result and Discussion

3.1. Descriptive Results

Following preprocessing, the final dataset comprised 382 patient records and was subsequently used for model development and evaluation. The class distribution of the target variable, described previously in Table 1, showed a substantial imbalance, particularly for the surgical closure class, thereby justifying the application of sampling strategies before model training.

3.2. Machine Learning Results

All analyses were conducted with the original dataset, undersampling dataset (Downsampled) and oversampling dataset (Upsampled). The performances of five (Rpart, Random Forest, KNN, Naïve Bayes, XGBoost) different models were calculated and evaluated based on the metrics of confusion matrix. These metrics indicated the models' predictive performance, i.e. their ability to classify patients based on their given features. The accuracy and AUC metrics of the models were used to evaluate their performance (Table 3). Performance metrics were reported with their corresponding 95% confidence intervals.

Table 3. Accuracy and AUC Results

Model	Original	Upsampled	Downsampled
Accuracy results			
rpart	0.605	0.605	0.605
rf	0.539	0.539	0.578
knn	0.565	0.539	0.526
nb	0.460	0.421	0.421
xgb	0.657	0.657	0.657
Macro-average ROC-AUC results			
rpart	0.757	0.757	0.757
rf	0.758	0.747	0.769
knn	0.701	0.704	0.736
nb	0.730	0.717	0.722
xgb	0.809	0.810	0.804

Based on the accuracy metric (Table 3), the *xgb* model provided the best performance across all datasets. Regardless of the sampling technique used — original, upsampled or downsampled — the *xgb* model achieved an accuracy of approximately 65%. The *rpart* model followed with 60.5% accuracy, then rf with 57.9%, *knn* with 56.6%, and *nb* with 46.1%. *nb* achieved the lowest accuracy in all sampling types. Sampling methods did not significantly affect model performance. Most models showed minimal or no change across different datasets. The accuracy results are visualized in Figure 1.

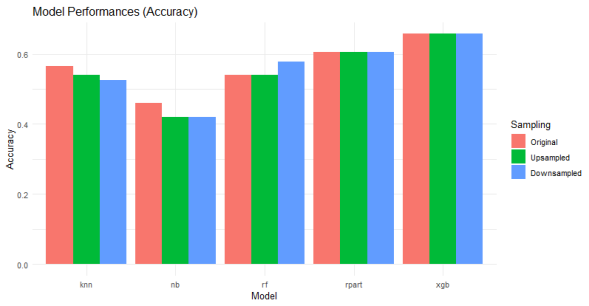

Figure 1. Model Performances (Accuracy)

Although the XGBoost models trained on the original, upsampled, and downsampled datasets achieved similar accuracy values, the model trained on the upsampled

dataset demonstrated the highest macro-average AUC. Therefore, this model was selected for further evaluation, and its confusion matrix is presented in Table 4. The confusion matrix demonstrated that 22/34 cases in Class 1 (Spontaneous closure), 6/7 cases in Class 2 (Surgical closure), and 22/35 cases in Class 3 (Non-closure) were correctly classified. This model achieved balances classification performance across all three classes, yielding precision score of 0.647, 0.857, and 0.629 and identical recall scores of 0.647, 0.857, and 0.629 for classes 1,2, and 3 respectively.

Table 4. Confusion Matrix of The Best-Performing XGBoost Model

Predicted	Spontaneous Closure	Actual Surgical closure	Non-Closure
Spontaneous Closure	22	0	12
Surgical Closure	0	6	1
Non-Closure	12	1	22

Based on the macro-average ROC-AUC results presented in Table 3, the *xgb* model achieved the best performance across all the datasets with the upsampled dataset, with an AUC of 81%. This was followed by *rf* (76.9%), *rpart* (75.7%), *knn* (73.6%) and *nb* (73%). The sampling methods affected the AUC values slightly in all models, with the *rf* and *knn* models demonstrating an increase in AUC under the downsampled condition. The AUC results are visualized in Figure 2.

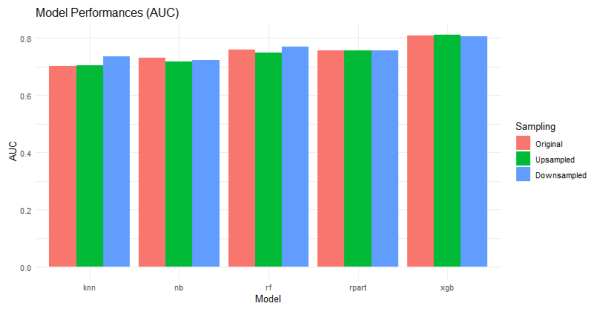

Figure 2. Model Performances (AUC)

Class-based metrics were also calculated to evaluate performance based on each of the three classes. To further evaluate model performance across individual outcome classes, class-specific F1-scores were calculated (Table 5). F1 scores higher than 80% are highlighted and reported. The highest F1-score equivalent to 92.3%, was achieved by both the *rpart* and *rf* models. While the *rpart* model delivered the same value for all datasets, the rf model achieved this value in the downsampled dataset. These results indicate the discriminatory potential of both models for the "Surgical closure" class. *xgb* exhibited relatively high F1 scores across the datasets. This model reached 85.7% for class 2 in both the upsampled and downsampled datasets, and 83.3% in the original dataset. Other models achieved lower F1 scores, indicating poor class discrimination capability. The class-specific F1-score comparisons are illustrated in Figure 3.

Table 5. Class Based F1-Score Metrics

Metric	Class: 1	Class: 2	Class: 3	Model	Sampling
F1	0.674	0.923	0.400	rpart	Original
F1	0.674	0.923	0.400	rpart	Upsampled
F1	0.674	0.923	0.400	rpart	Downsampled
F1	0.575	0.923	0.515	rf	Downsampled
F1	0.647	0.857	0.628	xgb	Upsampled
F1	0.647	0.857	0.628	xgb	Downsampled
F1	0.657	0.833	0.628	xgb	Original
F1	0.555	0.727	0.492	rf	Original
F1	0.542	0.727	0.507	rf	Upsampled
F1	0.600	0.705	0.363	knn	Downsampled
F1	0.613	0.444	0.529	knn	Original
F1	0.623	0.444	0.456	knn	Upsampled
F1	0.594	0.368	0.150	nb	Downsampled
F1	0.649	0.350	-	nb	Upsampled
F1	-	-	0.630	nb	Original

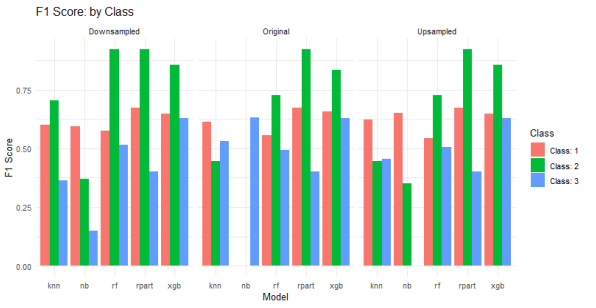

Figure 3. Class Based F1-Score Metrics

Furthermore, recall and precision values were also examined. The highest recall values for Class1, 2, and 3 were achieved by rpart (0.882), nb (1), and xgb (0.629).

3.3. Discussion

Five ML algorithms (Decision Tree, Random Forest, K-Nearest Neighbour, Naive Bayes, and XGBoost) were used to predict the closure status of patients with VSD in this study. Among the evaluated models, XGBoost achieved the highest overall performance with AUC rate of 0.810 in terms of both accuracy and macro-average AUC. Naïve Bayes (nb) showed the weakest accuracy rate of 0.421, whereas decision tree (rpart) demonstrated stable but moderate performance across sampling methods. The random forest model followed XGBoost by achieving a ROC AUC value of 0.769. Analysis of the results reveals that ensemble methods (XGBoost and Random Forest) demonstrate superior predictive capabilities in complex clinical scenarios, such as VSD. Significant class imbalance in the dataset adversely affected the performance of all models, particularly in the surgical closure group. Although undersampling and oversampling techniques are used to eliminate imbalance and ensure consistency in XGBoost, overfitting presents a risk in simple models such as Naive Bayes. These findings provide further evidence that ensemble methods are more robust in imbalanced datasets. Likewise, the literature suggests that XGBoost performs well in medical prediction models [15].

An interesting finding of this study was that neither undersampling nor oversampling substantially

improved model performance. Across most algorithms, accuracy and AUC values remained relatively stable regardless of the sampling strategy applied. This observation suggests that predictive performance may have been driven primarily by the discriminatory power of the available clinical features rather than by class distribution alone. Therefore, while class imbalance likely contributed to classification difficulty, it does not appear to be the sole factor limiting predictive performance. Additional predictors such as detailed echocardiographic measurements, longitudinal follow-up variables, and genetic information may provide greater improvements than further class balancing techniques.

The surgical closure class, typically underrepresented, reached a peak F1-score of 92.3% with both decision tree and random forest models. Although this finding suggests strong discrimination of clinically important surgical cases, it should be interpreted cautiously because the surgical closure group represented only 9.4% of the study population. Performance estimates derived from small minority classes may be sensitive to sampling variability and may overestimate real-world performance. Therefore, external validation using larger multicenter cohorts is required to confirm the robustness and generalizability of these results. A comparative review of the available literature corroborates the potential of AI in pediatric cardiology, as demonstrated by the findings of this study. Although [5] reported substantially higher predictive performance (AUC=0.95), their study addressed a binary classification problem in a very large multicenter cohort. Therefore, direct performance comparison should be interpreted cautiously.

The model's clinical benefit is that it may prevent unnecessary surgical interventions in cases with a high probability of spontaneous closure while enabling the early identification of cases necessitating surgery. Analysis of the confusion matrix revealed that most misclassifications occurred between the spontaneous closure and persistent non-closure groups, whereas the surgical closure group was identified with relatively high accuracy. This finding may reflect overlapping clinical characteristics between patients who ultimately experience spontaneous closure and those who remain stable without closure, making discrimination between these two outcomes more challenging than identification of patients requiring surgical intervention.

Although XGBoost achieved the highest overall performance, the observed accuracy of 65.7% indicates that predicting the clinical course of isolated VSD remains a challenging task. Several factors may explain this moderate predictive performance. First, after exclusion of records with missing values, the final dataset consisted of only 382 patients, which may have limited the model's ability to learn complex clinical patterns. Second, the prediction problem addressed in this study involved three clinically distinct outcome classes rather than a simpler binary classification task, increasing the complexity of model learning. Third, the available predictors were limited to baseline maternal and neonatal categorical variables. Important prognostic information such as detailed echocardiographic measurements, hemodynamic parameters, serial follow-up findings, and molecular or genetic biomarkers were not available for model development. Finally, the natural history of isolated VSD is inherently heterogeneous, and outcomes may be influenced by biological and environmental factors that are not routinely captured in retrospective clinical datasets. Therefore, the moderate accuracy observed in this study likely reflects both data-related limitations and the intrinsic complexity of predicting long-term VSD outcomes.

From a clinical perspective, the model is not intended to replace physician judgment or guide treatment decisions independently. Rather, its value lies in supporting preliminary risk stratification. Despite the moderate accuracy, the model achieved a macro-average AUC of 0.81 and demonstrated strong discrimination of the clinically important surgical closure group, suggesting that machine learning may assist clinicians in identifying patients who require closer monitoring or earlier specialist evaluation. Future studies incorporating larger multicenter cohorts, longitudinal follow-up data, and detailed echocardiographic variables may further improve predictive performance and enhance clinical applicability.

Beyond its predictive performance, the proposed model has several potential implications for pediatric cardiology practice. Accurate prediction of VSD outcomes remains clinically important because management strategies differ substantially among patients who experience spontaneous closure, require surgical intervention, or remain clinically stable despite persistent defects. Early identification of patients with a higher likelihood of surgical closure may facilitate closer monitoring, timely referral to specialized centers, and improved treatment planning. Conversely, identifying patients with a high probability of spontaneous closure may reduce unnecessary follow-up examinations and alleviate parental anxiety. Therefore, machine learning-based prediction models may contribute to more personalized and resource-efficient care pathways.

Comparison with previous studies further highlights the clinical relevance of our findings. Li et al. [13] reported an accuracy of 82.6% using a logistic regression model to predict spontaneous closure of isolated VSDs. However, their study was limited to fetal cases and evaluated only spontaneous closure as a binary outcome. Similarly, Sun et al. [5] achieved an AUC of 0.95 in a multicenter cohort of 29,142 patients using AI-based methods to predict spontaneous closure of

perimembranous VSDs. Although their predictive performance exceeded that observed in our study, their model addressed a binary classification problem and did not include patients with persistent non-closure as a distinct clinical category. In contrast, our study evaluated a more challenging multiclass prediction framework by simultaneously distinguishing spontaneous closure, surgical closure, and persistent non-closure. This approach more closely reflects real-world clinical decision-making, where persistent non-closure represents a separate management pathway rather than merely a negative outcome. Furthermore, Li et al. [15] demonstrated the usefulness of machine learning in predicting post-procedural conduction block after transcatheter closure of perimembranous VSDs (AUC=0.82). Unlike their study, which focused on post-interventional outcomes, our model aimed to predict the natural clinical course before any intervention was performed. Collectively, these findings support the growing role of machine learning in pediatric cardiology and suggest that future multicenter studies integrating longitudinal clinical data and detailed echocardiographic parameters may further improve predictive performance and clinical applicability.

3.4. Limitations

Although this study an insight into the application of machine learning models for predicting VSD closure outcomes, several limitations should be noted. First, the dataset was obtained from a single healthcare center and included a relatively limited sample size, which may restrict the generalizability of the findings. Second, the presence of a large amount of missing data represents an important limitation of this study because exclusion of incomplete records may have resulted in the loss of potentially informative clinical features and introduced selection bias. Third, despite the use of cross-validation and hyperparameter tuning strategies, the possibility of overfitting cannot be completely excluded due to the limited size of data. Additionally, the absence of external validation using a different dataset limits the assessment of the models.

4. Conclusion

The model may contribute to preliminary risk classification; however, its clinical utility should be validated in larger, multicenter, and prospective cohorts before using to support treatment decisions. In conclusion, XGBoost is the most promising algorithm for predicting VSD closure. However, class imbalance and limitations due to retrospective data impede its clinical integration. To overcome these challenges, future studies are recommended to optimize ensemble methods using large-scale datasets and integrate deep learning models. Such developments could pave the way for the routine use of AI-supported decision systems in pediatric cardiology.

This study has several limitations. First, it was based on a retrospective, single-center dataset, which may limit the generalizability of the findings. Second, although the initial dataset included 602 patients, 207 patients' record with missing values were excluded, resulting a final dataset of 382 instances. Third, the surgical closure class included a relatively small number of cases, which may have affected the stability of the class-specific performance metrics such as F1-score. Therefore, the high F1-scores observed for this minority class should be interpreted cautiously, as performance estimates derived from small sample sizes may be sensitive to sampling variability and may not fully generalize to external populations. Fourth, although multiple machine learning algorithms were compared, formal statistical significance testing of performance differences between models was not performed. Consequently, the observed superiority of XGBoost should be interpreted as descriptive rather than inferential. As the last limitation, no external and prospective validation was performed. Therefore, the proposed models should be considered as preliminary and should not be used for clinical decision-making without further investigation. Further analyses with a larger or multi-centered dataset could provide deeper insights into predictions and enhance the clinical interpretability of the developed models. Although the proposed models demonstrated predictive performances, due to the limitations of the study, the results cannot be generalized or directly used by the clinicians. Rather this study could be considered as an exemplary study, which show the possible implementations of machine learning methods in VSD prediction.

References

[1] K. Cox, C. Algaze-Yojay, R. Punn, and N. Silverman, "The Natural and Unnatural History of Ventricular Septal Defects Presenting in Infancy: An Echocardiography-Based Review," Journal of the American Society of Echocardiography, vol. 33, no. 6, pp. 763–770, Jun. 2020, doi: 10.1016/j.echo.2020.01.013.

[2] D. J. Penny and G. W. Vick, "Ventricular septal defect," in The Lancet, Lancet, 2011, pp. 1103–1112. doi: 10.1016/S0140-6736(10)61339-6.

[3] N. Roguin, Z. D. Du, M. Barak, N. Nasser, S. Hershkowitz, and E. Milgram, "High prevalence of muscular ventricular septal defect in neonates.," J. Am. Coll. Cardiol., vol. 26, no. 6, pp. 1545–1548, Nov. 1995, doi: 10.1016/0735-1097(95)00358-4.

[4] S. Erdem, N. Özbarlas, O. Küçükosmanoğlu, H. Poyrazoğlu, and O. K. Salih, "Long-term follow-up of 799 children with isolated ventricular septal defects," Turk Kardiyol Dern Ars, vol. 40, no. 1, pp. 22–25, 2012, doi: 10.5543/tkda.2012.01679.

[5] J. Sun et al., "Leveraging artificial intelligence for predicting spontaneous closure of perimembranous ventricular septal defect in children: a multicentre, retrospective study in China," Lancet Digit. Health, vol. 7, no. 1, pp. e44–e53, 2025, doi: https://doi.org/10.1016/S2589-7500(24)00245-0.

[6] F. Eckerström, C. Nyboe, A. Redington, and V. E. Hjortdal, "Lifetime Burden of Morbidity in Patients With Isolated Congenital Ventricular Septal Defect," J. Am. Heart Assoc., vol. 12, no. 1, p. e027477, Jan. 2023, doi: 10.1161/JAHA.122.027477.

[7] F. Eckerström, C. Nyboe, M. Maagaard, A. Redington, and V. E. Hjortdal, "Survival of patients with congenital ventricular septal defect.," Eur. Heart J., vol. 44, no. 1, pp. 54–61, Jan. 2023, doi: 10.1093/eurheartj/ehac618.

[8] V. Kaul, S. Enslin, and S. A. Gross, "History of artificial intelligence in medicine," Gastrointest. Endosc., vol. 92, no. 4, pp. 807–812, Oct. 2020, doi: 10.1016/j.gie.2020.06.040.

[9] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Comput. Sci., vol. 1, no. 6, p. 345, 2020, doi: 10.1007/s42979-020-00365-y.

[10] Y. Sethi et al., "Artificial Intelligence in Pediatric Cardiology: A Scoping Review," J. Clin. Med., vol. 11, no. 23, Nov. 2022, doi: 10.3390/jcm11237072.

[11] S. Nurmaini et al., "Deep Learning-Based Computer-Aided Fetal Echocardiography: Application to Heart Standard View Segmentation for Congenital Heart Defects Detection," Sensors (Basel), vol. 21, no. 23, Nov. 2021, doi: 10.3390/s21238007.

[12] W. R. Thompson, A. J. Reinisch, M. J. Unterberger, and A. J. Schriefl, "Artificial Intelligence-Assisted Auscultation of Heart Murmurs: Validation by Virtual Clinical Trial," Pediatr. Cardiol., vol. 40, no. 3, pp. 623–629, Mar. 2019, doi: 10.1007/s00246-018-2036-z.

[13] X. Li et al., "Prediction of spontaneous closure of isolated ventricular septal defects in utero and postnatal life," BMC Pediatr., vol. 16, no. 1, p. 207, Dec. 2016, doi: 10.1186/s12887-016-0735-2.

[14] H. Mori, K. Inai, H. Sugiyama, and Y. Muragaki, "Diagnosing Atrial Septal Defect from Electrocardiogram with Deep Learning," Pediatr. Cardiol., vol. 42, no. 6, pp. 1379–1387, Aug. 2021, doi: 10.1007/s00246-021-02622-0.

[15] Z. F. Li et al., "Machine learning prediction for prognosis and long-term effectiveness for transcatheter ventricular septal defect closure: a 5-year single center experience," Eur. Heart J., vol. 45, no. Supplement_1, p. ehae666.2132, Oct. 2024, doi: 10.1093/eurheartj/ehae666.2132.

[16] M. Kuhn, "caret: Classification and Regression Training," CRAN: Contributed Packages. Accessed: Apr. 07, 2025. [Online]. Available: https://cran.r-project.org/package=caret

[17] R. Kohavi, "A Study of Cross-Validation and Bootstrapfor Accuracy Estimation and Model Selection," in International Joint Conference on Artificial Intelligence, May 1995, pp. 1137–1145.

[18] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," Journal of the Royal Statistical Society. Series B (Methodological), vol. 36, no. 2, pp. 111–147, Jun. 1974, doi: 10.1111/j.2517-6161.1974.tb00994.x.

[19] T. R. Hoens and N. V. Chawla, "Imbalanced Datasets: From Sampling to Classifiers," in Imbalanced Learning: Foundations, Algorithms, and Applications, H. Haibo and Y. Ma, Eds., Wiley-IEEE Press, 2013, pp. 43–59. doi: 10.1002/9781118646106.

[20] N. Japkowicz, "Assessment Metrics for Imbalanced Learning," in Imbalanced Learning: Foundations, Algorithms, and Applications, H. Haibo and Y. Ma, Eds., Wiley-IEEE Press, 2013, pp. 187–206. doi: 10.1002/9781118646106.

[21] H. Wickham and J. Bryan, "readxl: Read Excel Files," 2025. doi: 10.32614/CRAN.package.readxl.

[22] Kuhn and Max, "Building Predictive Models in R Using the caret Package," J. Stat. Softw., vol. 28, no. 5, pp. 1–26, 2008, doi: 10.18637/jss.v028.i05.

[23] H. Wickham, R. Francois, L. Henry, K. Muller, and D. Vaughan, "dplyr: A Grammar of Data Manipulation," 2026. doi: 10.32614/CRAN.package.dplyr.

[24] H. Wickham, D. Vaughan, and M. Girlich, "tidyr: Tidy Messy Data," 2025. doi: 10.32614/CRAN.package.tidyr.

[25] X. Robin et al., "pROC: an open-source package for R and S+ to analyze and compare ROC curves," BMC Bioinformatics, vol. 12, p. 77, 2011.

[26] H. Wickham, ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[27] H. Wickham, "Reshaping Data with the reshape Package," J. Stat. Softw., vol. 21, no. 12, pp. 1–20, 2007, [Online]. Available: https://www.jstatsoft.org/v21/i12/

[28] J. Ooms, "writexl: Export Data Frames to Excel 'xlsx' Format," 2025. doi: 10.32614/CRAN.package.writexl.