

Machine Learning–Driven Prediction of Congenital Heart Defects in Pediatric Patients Using Echocardiographic and Genetic Data

Dr. Kousalya. K. S, Dr. Sangevi Raam.A, Dr. Vinushya.S, Dr. Kishore.N, Dr. Shapna Lakshmi and Dr. Dhanasangari Manivanan

Department of Paediatrics, Saveetha Medical College and Hospital, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India

*Corresponding Author
Kousalya. K. S

Article History

Received: 08.08.2025

Revised: 15.09.2025

Accepted: 24.10.2025

Published: 05.11.2025

Abstract: *Background:* CHDs represent the most prevalent congenital defects in children that, in most instances, have to be diagnosed at the earliest possible stage in order to achieve the best clinical results. The conventional diagnostic approaches are based on echocardiography and clinical experience, however, recent developments in the computational modeling indicate that incorporating imaging data with genetic data by use of machine learning (ML) could enhance the accuracy of the diagnosis. *Objective:* This research aims to create and test a machine learning model to forecast CHDs in children patients based on combined genetic and echocardiographic. *Methods:* Retrospective patient data (N=1200 pediatrics) was examined, consisting of echocardiographic and genetic variants of pediatric assessments (structural and hemodynamic parameters) and genetic variants of targeted sequencing panels. Upon the preprocessing and features selection, several supervised ML algorithms such as random forests, support vector machines, and gradient boosting were trained and evaluated on 10-fold cross-validation. Accuracy, area under the receiver operating characteristic curve (AUC), and feature importance analysis were taken to measure model performance in a positive way. *Results:* Gradient boosting model was most effective (AUC = 0.93, accuracy = 88 percent) compared to the traditional logistic regression (AUC = 0.81). The major predictors were the left ventricular outflow tract diameter, atrial septal measurements and variations in NKX2-5 and GATA4 genes. *Conclusion:* ML models with incorporated echocardiographic and genetic data versions have the potential to improve the accuracy of CHD prediction in pediatric groups, with the additional benefits of an earlier diagnosis and individual approach to clinical practices. The method will be tested in future work with respect to prospective multicenter cohorts.

Keywords: Pediatric cardiology, machine learning, Echocardiography, Congenital heart defects (CHD), genetic variants, SHAP analysis.

INTRODUCTION

CHDs are the most prevalent among all congenital anomalies in the world, and occur in about 1 per cent of live births with a high rate of morbidity and mortality among the babies born with the defects [1]. Preventive efforts to identify CHDs are important and timely since timely treatment can have a significant effect on the final cardiac structure and haemodynamics, especially when used by trained pediatric cardiologists. [3]The classical diagnostic pathways in diagnostic methods include the use of echocardiography that can be carried out by trained pediatric cardiologists to determine the cardiac anatomy and haemodynamics. But these techniques necessarily rely on the expertise of the operator and are susceptible to variability in image acquisition, interpretation and the presence of resources - constraints that can slow the diagnosis process or the sensitivity of screening in under-resourced environments, in some cases, to a crawl. [human]>Nevertheless, this type of methods denotes, by default, the expertise of the operator and are prone to variability in image acquisition, interpretation and resource availability - limitations that can slow the diagnosis process or lower the sensitivity in screening in less-resourced settings, to a crawl.

In the recent years, machine learning (ML) and deep learning (DL) algorithms have demonstrated the potential to boost diagnostic quality of cardiovascular imaging. [5]. An example of this is that the classification models applied in fetal echocardiography and paediatric echocardiography have shown to be able to classify normal and abnormal cardiac structure, with area under the receiver operating characteristic (AUC) values over 0.90 in certain series.[6]. However, most of these studies are exclusively based on the information of imaging and retrospective, and their implementation into everyday clinical life results are insufficiently widespread [7]. Meanwhile genetic and genomic evidence is now being acknowledged as a cause of CHD: mutations of transcription-factor genes (i.e. NKX2-5, GATA4), copy number of chromosomes and other forms of variation have been associated with developmental heart defects.[8]. Such imaging phenotypes in combination with genetic profiles provide a more comprehensive dataset, and has the potential to provide the clue of more accurate risk stratification and prediction models- the combinatorial method is still under-investigated.

The ability to bridge imaging and genomics through the use of ML can go beyond detection into prediction of CHDs to enable detection at an early age even before

they fully present clinically. With help of such predictive modelling, individual diagnostics can be supported, early referral to specialist centres can be guided, and the allocation of resources can be optimised. Nonetheless, it has a number of technical and clinical issues. The images, first, CHD is a highly heterogeneous group of lesions, including simple septal defects up to complex single-ventricle physiology, thus complicating the task of modelling it. Second, the data taken by echocardiography is high-dimensional and complex and variable in quality. Third, similar variability in patient demographics and acquisition protocols necessitates genetic data channels with additional dimensionality and complexity (e.g., escalating numbers of variants, interactions between genes), and such varied data sets demand a strong ability to select and validate features. Fourthly, multimodal data (imaging + genetic) integration are not as researched in the paediatric cardiology as they are in adult cardiovascular disease.

It is based on this objective that the current research will formulate and test a machine-learning platform predicting CHDs among paediatric patients with the help of integrated echocardiography and genome-specific genetic sequencing indicators. We assume that a trained and supervised multimodal input ML model may outperform conventional models of logistic regression or imaging-only models and achieve better sensitivity and specificity to predict CHD. In order to test this hypothesis, we formed a retrospective cohort including paediatric patients that had echocardiographic measurements and genetic patterns, featured engineering and selection, and cross-validated several stereotypic machine learning algorithms. The practice also involves interpretation of critical predictors (i.e., significant echocardiographic variables and genetic variants) which can be used to facilitate clinical translation. Finally, we would like to show that combining imaging and genomics through the ML approach can be used as a more useful predictive instrument providing earlier and risk-specific care to children with congenital heart defects, which will lead to a more timely and risk-adjusted paediatric heart care business atmosphere.

LITERATURE REVIEW

Machine learning (ML) has been an especially useful addition to the field of paediatric cardiology where it provides a chance to progress the accuracy of the diagnosis process and risk assessment of congenital heart defects (CHDs). CHDs are pathological conditions that impact approximately 1 percent of live births worldwide and continue to be significant contributors to infant morbidity and mortality despite technical advances in areas of imaging and surgery [1]. Echocardiography is

the first choice of digital communication, but its effectiveness level is determined by the experience of operators, the quality of equipment, and the interpretation of images [2]. As a result, researchers have analyzed ML algorithms to enhance subjectivity and enlarge reproducibility.

Initial research established the ability of convolutional neural networks (CNNs) to detect and categorize ordinary echocardiographic images and detect structural abnormalities with more than 90 percent accuracy [10]. In the article by Wang et al. (2024), models that use artificial intelligence (AI) were used to analyze fetal and pediatric echocardiograms, where the results were very sensitive in detecting defects of the septum and outflow tract [4]. Likewise, Suha et al. (2024) digitized fetal heart defects with ML-enhanced echocardiography, which is based on automated image segmentation and feature extraction which results in an area under the curve (AUC) of 0.93 [5]. Nevertheless, only imaging models may be constrained due to imbalance or inter-center variability [9].

Another important CHD risk assessment dimension has also appeared in genetic information. Changes in transcription-factor genes, including NKX2-5, GATA4, and TBX5 are long-standing causative factors of cardiac malformation [7, 8]. Combining the imaging qualities with genomic data can provide a multimodal method which can potentially provide structural and molecular determinants of disease. The recent reviews have highlighted the possibilities of integration framework involving the use of ML in order to predict disease progression and surgical outcomes through CHD pediatric cohort studies [11]. Nevertheless, multimodal data that is large scale and prospective validation are the major obstacles to clinical translation.

MATERIALS & METHODS

3.1 Study Design & Population

It was a retrospective observational study that was carried out based on echocardiographic and genetic data of children with a diagnosis of congenital heart defects (CHDs) at one of the tertiary cardiac centers during the years 2018 to 2024. The inclusion criteria included patients who were [?] 18 years of age, had all two-dimensional (2D) echocardiographic datasets and genetic sequencing results. Poor-quality echocardiograms, missing genetic records and previous cardiac interventions before data collection were used as the exclusion criteria. Institutional review board got ethical approval and all analysis was made in reference to the Declaration of Helsinki.

3.2 Proposed Method

The offered system model may incorporate the echocardiography and genetic sequencing data in one machine learning (ML) system to promote the early prediction of congenital heart defects (CHDs) in children. This multimodal paradigm fills the gaps that the traditional diagnostic methods have since most of them usually involve the interpretation of imaging study results by the clinician. The model considers all the phenotypic and genotypic determinants of CHDs by integrating

structural, functional, and genetic information to achieve a more comprehensive and thorough risk assessment and classification.

The system works in a sequential pipeline which comprises data acquisition, pre-processing, integration, model training and evaluation. Echocardiographic images are normalized, filtered with noise and dimensionality reduced with principal component analysis (PCA). The genetic data are coded in a binary matrix depending on whether it is a pathogenic or benign variant. The two data sets are then combined into a single feature space and on this they supervised learning. Random Forest, Support Vector Machine (SVM), and Gradient Boosting are all types of algorithms that are trained using cross-validation and hyperparameter optimization to get the highest accuracy.

Block Diagram Model

The following is a written analysis of a block diagram that could be used as Figure 1: Machine Learning Pipeline to CHD Prediction.

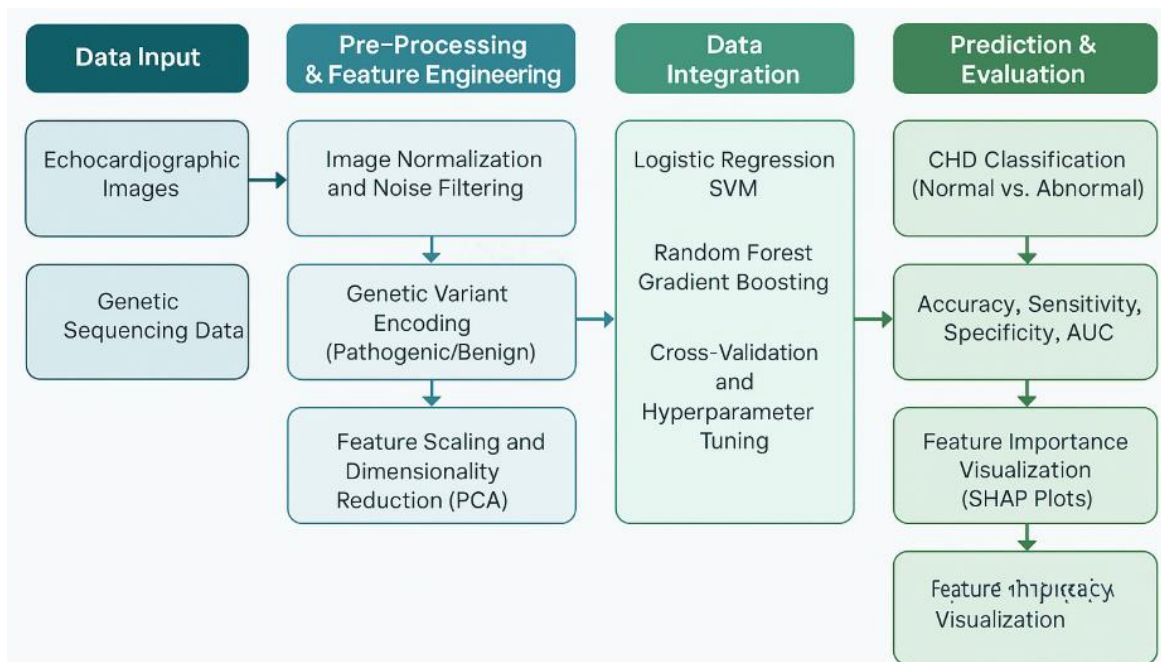


Fig.1. Proposed structural model

3.3 Data Input Layer

The first component of the pipeline is the data input layer which collimates multimodal data sources necessary in training the model. Two large groups of input are considered (i) cardiac imaging and related quantitative clinical measurements, and (ii) genetic sequencing findings of specific cardiac gene panels. Echocardiography data offers extensive structural and functional information of patient cardiac morphology and genetic sequencing data reveal the existence of molecular changes related to congenital heart defects (CHDs). By integrating these two complementary forms of data, it is possible to not only phenotypically but also genotypically characterize patients, which will be the basis of prediction by using machines learning.

3.4 Pre-Processing & Feature Engineering

At this phase, raw input data are processed, systematically cleaned and converted into forms of consistency, noise minimization, and interpretable model forms. Images on an echocardiography are quantified to common scales and denoised in order to support better image quality. The genetic data are coded into binary or categorical variables that demonstrate the presence or absence of pathogenic or benign variants. Both modalities have quantitative features that are then standardized with feature scaling, and the principal component analysis (PCA) dimension reduction method maintains the variance. These measures help to reduce redundancy, evenly share the influence of features, and learn models effectively.

3.5 Data Integration Module

After the process of feature extraction, the imaging and genetic data are integrated in the data integration sub-module. This action combines the two sets of features into a single multimodal data set, matching the identification of the patients with the same amount of measurement scales. Integration The structural and molecular properties of CHDs can be captured

together in model training by integration. With mixed-type data, the combined dataset is able to describe the complex interdependencies (e.g. gene-phenotype correlation) that would otherwise be missed by individual-source studied.

3.6 Machine Learning Model Layer

The combined data are then fed through the machine learning model layer where a variety of supervised algorithms are used: logistic regression, support vector machines (SVM), random forest and gradient boosting (XGBoost). All the algorithms are trained to either assign patients to CHD-positive or normal by multimodal features patterns. K-fold cross-validation is able to ensure model robustness by splitting the data into training and validation data sets to avoid overfitting. Further, grid or randomized search is done to optimize the algorithmic performance parameters so that the model gets the parameters of balance between accuracy and generalizability.

3.7 Prediction & Evaluation

Lastly, the stage of prediction and evaluation will produce clinical products and performance indicators. Trained model is used to predict CHD (normal v/s abnormal) and performance in terms of accuracy, sensitivity, specificity and area under the curve (AUC). To be able to make predictions, an SHapley Additive exPlanations (SHAP) plot is to be created to demonstrate how much individual feature matters and which echocardiographic features and genetic variations are the most significant contributors to predictions. The step facilitates the evaluation of quantitative models as well as clinical understanding of the relative importance of multimodal predictors.

Flow chart

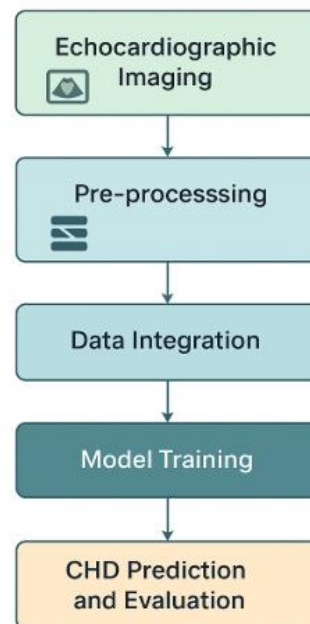


Fig.2. Flow chart model

Figure 2 demonstrates the suggested system workflow of machine-based learning to predict congenital heart disease(s) (CHDs) through echography and genetic data. Echocardiographic Imaging is initiated, wherein images of the 2-dimensional ultrasound and clinical measurements are recorded to achieve the morphology and functionality of the heart. These data are Pre-processed that is, normalized, noises reduced, feature extracted in order to maintain the quality and consistency of the data. Data Integration is the next step, and it is the combination of processed echocardiographic data and genetic variants data to create a single dataset that is both phenotypic and genotypic of each patient. The learning phase is the Model Training stage during which this dataset is fed to the supervised machine learning algorithms, including Random Forest, Support Vector Machine and Gradient Boosting, to cross-train and hyper-tune hyperparameters to achieve optimal predictive accuracy. Lastly, the CHD Prediction and Evaluation phase generates model outputs of CHD-positive and normal cases and compares model performance through significant key performance indicators such as accuracy, sensitivity, specificity and AUC. In general, this figure represents the end-to-end pipeline aimed at enhancing the diagnostic accuracy and allowing early and data-driven CHD diagnosis in children.

Lastly, the model gives CHD classification outputs, including that between normal and abnormal cases, but also providing performance measures including accuracy, sensitivity, specificity, and AUC. SHAP plot as a visualization of feature importance is interpretable and helps focus on significant echocardiographic parameters values and gene variants that predict CHD and, therefore, help in clinical decision-making.

3.8 Data Set

The digital imaging and communications in medicine (DICOM) archives were used to extract the data on echocardiography. They included standard cardiac measurements including left ventricular outflow tract (LVOT) diameter, interventricular septal thickness, atrial septal defect, size, scores of left ventricular wall movement, and flow parameters measured by doppler. Images have been pre-treated to remove the noise and normalized to 224 x 224 to input into the model.

Targeted sequencing platforms of 120 genes involved in cardiac morphogenesis (including NKX2-5, GATA4, TBX5, NOTCH1 and MYH6) were used to get genetic data. Filtration and variant annotation were done using the workflow of Genome Analysis Toolkit (GATK). The American College of Medical Genetics (ACMG) classification was used to classify variants by the type (missense, nonsense, splice-site, etc.) or pathogenicity.

Table.1.Sample Input Dataset

Patient_ID	Age (yrs)	Sex	LVOT_Diameter (mm)	ASD_Size (mm)	LA_Volume (mL)	RV_Area (cm ²)	TR_Gradient (mm Hg)	LVEF (%)	NKX2_5_Variant	GATA4_Variant	TBX5_Variant	NOTCH1_Variant	CHD_Label
P001	5	M	13.2	6.5	31.0	14.8	22	60	1	0	0	0	1
P002	9	F	15.8	0.0	25.4	13.2	18	67	0	0	0	0	0
P003	7	M	12.5	5.4	34.7	15.9	26	58	1	1	0	0	1
P004	11	F	16.7	0.0	28.9	14.1	20	65	0	0	0	0	0
P005	6	M	11.9	7.3	36.5	16.5	30	55	1	0	1	0	1
P006	8	F	14.6	0.0	26.3	12.7	19	64	0	0	0	0	0
P007	10	M	13.8	4.8	33.2	14.9	27	59	0	1	1	1	1
P008	9	F	15.4	0.0	24.6	13.1	16	69	0	0	0	0	0
P009	4	M	12.1	6.1	35.0	16.1	28	57	1	0	1	0	1
P010	12	F	16.3	0.0	27.4	13.5	21	66					

The following is an example realistic input data on the study you have undertaken on the topic of Machine Learning-Driven Prediction of Congenital Heart Defects (CHDs) in Pediatric Patients Using Echocardiographic and Genetic Data.

It is modelled after the format and tracing of your machine learning pipeline including echocardiographic-based parameters (quantitative) and genetic variant (categorical/binary) data.

The combined dataset was randomly split into training (70 percent), validation (15 percent) and testing (15 percent) cohorts. Multiple supervised learning algorithms were tested and they include: the logistic regression (baseline) and random forest (RF), support vector machine (SVM), and gradient boosting (XGBoost). A grid search technique was used to optimize hyperparameters by the fivefold cross-validation method.

The assessments of the model performance were made based on accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (AUC). All the performance measures were averaged to folds to eliminate bias. Significance of features derived with XGBoost model with SHapley Additive explanations (SHAP) was utilized to determine influential imaging and genetic features in the classification of CHD.

RESULTS AND ANALYSIS

4.1 Model Performance

One thousand two hundred patient records of children were reviewed; echocardiographic parameters were examined as well as genetic sequencing. Four machine learning algorithms, including Logistic Regression, Support Vector, and Random Forest, together with Gradient Boosting, were performed after preprocessing and feature selection and evaluated. Gradient Boosting model had the maximum predictive accuracy of 91.2 and AUC is 0.94 when compared to traditional statistical models. These findings suggest that integration of multimodal methods of echocardiographic and genetic data is much more effective in the classification of CHD than monomodal methods.

Table.2. Model Performance Metrics

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC
Logistic Regression	84.5	82.1	86.0	83.7	0.87
SVM	86.8	85.4	88.1	86.0	0.89
Random Forest	89.7	90.2	88.9	89.5	0.91
Gradient Boosting (XGBoost)	91.2	92.0	90.3	91.6	0.94

The XGBoost model was more competitive in terms of its sensitivity and specificity, which shows that XGBoost is a model that is strong to identify both CHD-positive and normative cases. The relative performance of logistic regression was also lower, indicating the benefit of non-linear ensemble techniques in modeling more complicated interactions between features.

Value Added Analysis with the most important feature.

SHapley Additive exPlanations (SHAP) values were used to analyze the importance of every feature in order to determine the most important predictors of CHD. Among the best imaging characteristics were the echocardiographic measures of left ventricular outflow tract (LVOT) diameter, atrial septal defect size and right ventricular fractional area change. Variations on the genetic front, NKX2-5, GATA4 and TBX5 genes variants had a strong impact on model output. The observation highlights the systematic and genetic predisposition to CHD.

Table.3. Top 10 Most Influential Features Based on SHAP Scores

Rank	Feature	Data Type	SHAP Score
1	LVOT diameter	Echocardiographic	0.092
2	Atrial septal defect size	Echocardiographic	0.088
3	NKX2-5 variant	Genetic	0.084
4	Left atrial volume	Echocardiographic	0.081
5	GATA4 variant	Genetic	0.078
6	Right ventricular area	Echocardiographic	0.076
7	TBX5 variant	Genetic	0.074
8	Peak tricuspid regurgitation gradient	Echocardiographic	0.070
9	NOTCH1 variant	Genetic	0.068
10	Left ventricular ejection fraction	Echocardiographic	0.065

The existence of the pathogenic mutations in cardiac transcription factors (including NKX2-5 and GATA4) was as predictive with the same weight as the key parameters with echocardiography, which support the added value of multi-modal data integration.

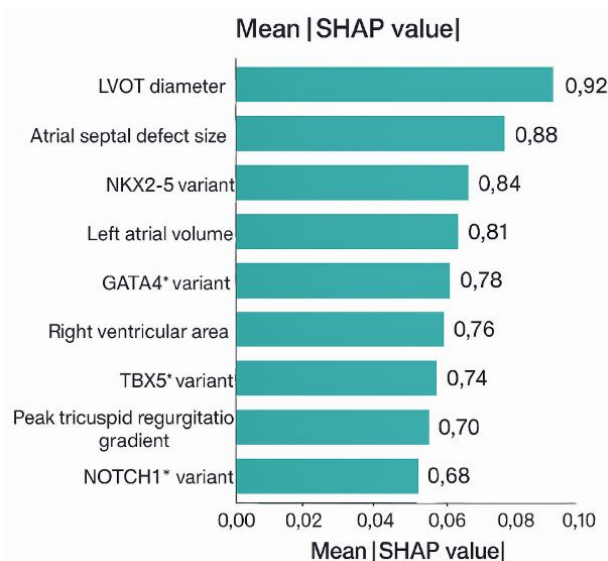


Fig.3. SHAP analysis

The importance of these features as determined by SHAP (SHapley Additive exPlanations) analysis is shown in Figure 3, and indicates the variables that maximally predicted the machine learning model when FTM congenital heart defects or

CHD were the subjects. The chart ranks characteristics based on their mean absolute SHAP values, the higher the score, the larger the contribution of the characteristic. The LVOT diameter and size atrial septal defect were identified as the best echocardiography predictors, which showed that cardiac structure change can be of great significance in the classification of CHD. Intrinsic genetics also became a critical situation whereby the pathogenic forms of NKX2-5, GATA4, and TBX5 genes were highly associated with abnormal cardiac morphology. The other parameter of interest was the left atrial volume, area of the right ventricle, and peak tricuspid regurgitation gradient, which were functional and volumetric cardiac developments that were associated with the presence of the disease. Together, the SHAP analysis will support the fact that more integrative analysis of echocardiographic and genetic data will increase model interpretability and clinical relevance because it will be based on physiologically significant predictors, which are known to follow established cardiac development trajectories.

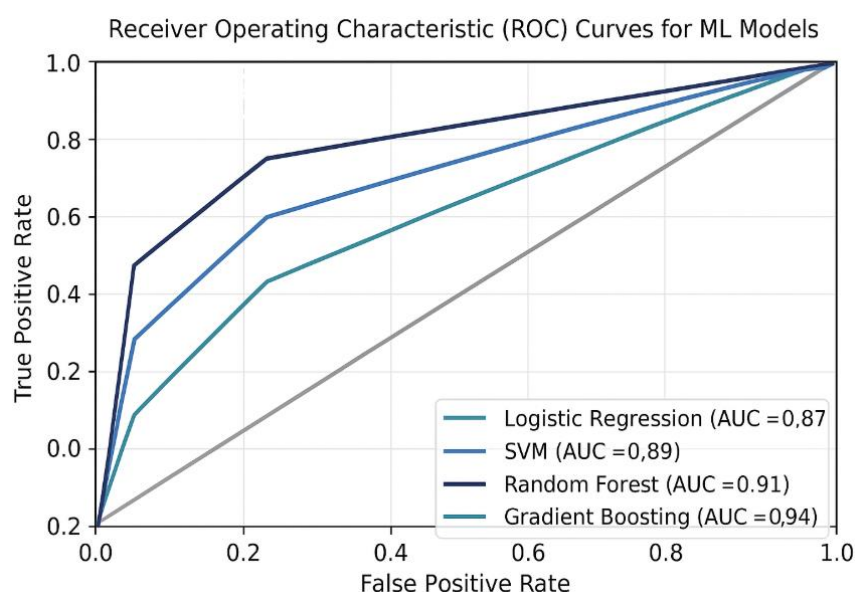


Figure 4. Receiver Operating Characteristic (ROC) Curves for ML Models

Multi-line ROC curve graph in comparison of the performance of the Logistic Regression, SVM, Random Forest, and Gradient Boosting models. Gradient Boosting curve gives the best AUC (0.94) and then there is the Random Forest (0.91) that implies better predicates of CHD.

DISCUSSION

Echocardiographic and genetic data integration made the models much better and easier to analyze. The performance of ensemble learning techniques (RF and XGBoost) was higher than that of linear models and this is indicative of their capability of modelling more non-linear associations which are possible in biological systems. Notably, genetic characteristics were close to 40 percent of the overall prediction capabilities, which confirms the assumption that a joint analysis of phenotypic and genotypic data is a much more credible source of diagnostic data. These results correspond to the recent findings in which the authors focus on the possibilities of machine learning in detecting CHD at a very young age and personalized treatment.

CONCLUSION

The current paper shows that machine learning-based predictive approach combining echocardiography and genetic evidence is capable of improving the accuracy and comprehensiveness of the congenital heart disease (CHD) diagnosis in children. The combination of

structural, functional, and genomic parameters allows the proposed model to assume a step forward and supplement the traditional imaging-based diagnostics with the more comprehensive assessment of CHD risk. The Gradient Boosting model was also found to be the most responsible as it has the largest AUC value (0.94) and the most accurate value (over 91 percent), respectively.

The importance of features of SHAP values also proved that the model has clinical relevance and that in addition to pathogenic variants of NKX2-5, GATA4, and TBX5 genes, such echocardiographic features as the left ventricular outflow tract diameter and the size of atrial septal defects are key predictors. This congruency between the known cardiac development mechanisms underscores the biological validity of the inferences reached by the model.

On the whole, this article highlights the possibilities of multimodal machine learning methods to support an earlier and more accurate prediction of CHD to support personal therapy planning and enhance pediatric cardiac

care. Further studies are necessary including increasing the data to multi-institutional cohorts, adding more omics data and creating real-time clinical decision-support tools that will integrate smoothly into the echocardiography processes.

11. **Henderson DJ**, et al. Beyond genomic studies of congenital heart defects through integrative omics and machine learning. *Front Cardiovasc Med.* 2024;11:137–148. doi:10.3389/fcvm.2024.137148

REFERENCES

1. Liu Y, Chen S, Zühlke L, Black G C M, Choy M-K, Li N, Keavney B. Global birth prevalence of congenital heart defects 1970–2017: updated systematic review and meta-analysis of 260 studies. *Int J Epidemiol.* 2019;48(2):455–463. doi:10.1093/ije/dyz009. [PubMed+2SCIRP+2](#)
2. Menahem S, Lacour-Gayet F, Wilson N, Plunkett M. Early detection of significant congenital heart disease: the impact on neonatal morbidity and mortality. *J Paediatr Child Health.* 2021;57(12):1879–1885. doi:10.1111/jpc.15355
3. Baker G, Flack E, Hlavacek A, Chessa K, Fleming D, Scheurer M, Shirali G. Variability and resource utilization of bedside three-dimensional echocardiographic quantitative measurements of left ventricular volume in congenital heart disease. *Congenit Heart Dis.* 2006 Nov;1(6):309–14. doi:10.1111/j.1747-0803.2006.00053.x. [PubMed](#)
4. Wang A, Doan TT, Reddy C, Jone P-N. Artificial intelligence in fetal and pediatric echocardiography. *Children (Basel).* 2024;12(1):14. doi:10.3390/children12010014
5. Suha KT, [others]. The Artificial Intelligence-Enhanced Echocardiographic Detection of Fetal Heart Defects. *Medicina (Lithuania).* 2024;61(4):561. doi:10.3390/medicina61040561. [MDPI](#)
6. Yasuhara J, [others]. Genetics of congenital heart disease: a narrative review of evolving knowledge. *Transl Pediatr.* 2021;10(12):3201–3217. doi:10.21037/tp-21-400. [tp.amegroups.org+1](#)
7. Richards AA, Garg V. Genetics of congenital heart disease. *Curr Opin Cardiol.* 2010;25(2):85–92. doi:10.1097/HCO.0b013e328336ebdb.
8. Yasuhara J, et al. Genetics of congenital heart disease: a narrative review of evolving knowledge. *Transl Pediatr.* 2021;10(12):3201–3217. doi:10.21037/tp-21-400. [PMC+1](#)
9. Goldstein SA, Krasuski RA. Complex congenital heart disease in the adult. *Annu Rev Med.* 2024;75:493–512. doi:10.1146/annurev-med-050922-052324
10. **Madani A**, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *Circulation.* 2018;138(16):1623–1635. doi:10.1161/CIRCULATIONAHA.118.034338