

Anatomy-Derived Hemodynamic-Proxy Features Improve Image-Based Classification of Congenital Heart Disease

Yu Xiang Chong
russellchong5@gmail.com

ABSTRACT

Congenital heart disease (CHD) encompasses many types of cardiovascular abnormalities, making image-supported diagnosis particularly challenging. In this study, we evaluated whether anatomy-derived hemodynamic-proxy features could improve machine learning classification of CHD subtypes (healthy, conotruncal defects, septal defects, and complex) when physiological measurements were unavailable. Publicly available human cardiac segmentation datasets were harmonized into a single dataset containing 194 cases spread into four classes. Fifteen anatomy-derived hemodynamic-proxy features and fifteen conventional anatomical features were extracted from the segmentation masks. Four machine learning models were optimized and evaluated using repeated stratified five-fold cross-validation. Machine learning models were assessed on accuracy, class-balanced accuracy, and macro F1 score for the anatomy-derived hemodynamic-proxy feature set, conventional anatomical feature set, and combined feature set. Across all four models, the combined anatomical plus hemodynamic-proxy feature set outperformed the anatomical-only feature set, with the best combined models reaching approximately 0.64–0.66 macro F1-score. These exploratory findings suggest that anatomy-derived hemodynamic-proxy features may provide complementary information for CHD classification, but validation in larger, source-balanced external cohorts and against direct hemodynamic measurements is required before generalizability or clinical utility can be established.

Keywords: cardiac segmentation, congenital heart disease, hemodynamic-proxy features, machine learning, medical image classification.

INTRODUCTION

Congenital heart disease (CHD) is composed of a heterogeneous group of structural cardiovascular abnormalities involving cardiac chambers, septa, great vessels, and other features of the heart [1,2]. Given the three-dimensional anatomical variability of CHD, image-based classification remains challenging [2,3]. Public CHD imaging datasets, although limited, have enabled computational CHD classification

[3,4]. Clinically, CHD diagnosis is largely based on anatomical and functional assessment using imaging [1,2]. Echocardiography is commonly used as an initial diagnostic tool because it can evaluate cardiac chambers, septa, valves, and ventricular function in real time [1]. These images are interpreted by clinicians who assess the structure of the heart to diagnose patients with CHD [1,2]. Cardiac computed tomography (CT) is another tool used by clinicians to diagnose patients with CHD, providing high-resolution, three-dimensional visualization of cardiac structures and great-vessel anatomy, particularly in patients with complex CHD [2,3].

Machine learning models have been applied to numerous medical image analysis tasks, including segmentation, diagnosis, and risk stratification [3,4]. Xu et al. introduced ImageCHD, a three-dimensional cardiac CT dataset for CHD classification, showing that automated CHD diagnosis is possible using segmentation-based analysis of cardiac shape and great-vessel connections [3]. Xu et al.'s work establishes one of the first public benchmarks for image-based CHD classification, but it also demonstrates the difficulty of CHD classification as structural defects are highly variable and public datasets are small [3]. Yang et al. introduced PointCHD, a point-cloud benchmark for CHD diagnosis, segmentation, and reconstruction, showing that three-dimensional anatomical representations can be useful for CHD classification [4]. More recently, Xu et al. developed a clinically applicable CT-based AI system for CHD diagnosis, while Wang et al. showed that CHD can also be interpreted automatically from multi-view echocardiograms [5,6]. These studies show that defects within pivotal heart structures are directly relevant to CHD classification because many defects alter chamber dominance, atrial or ventricular dilation, or systemic outflow geometry [2–4].

Most machine learning CHD classification approaches focus primarily on anatomical features, such as chamber volume ratios, vessel size, shape, and spatial relationships [3,4]. Although these features are relevant to CHD classification, CHD is not purely an anatomical disorder. Defects also alter blood flow distribution, outflow resistance, and ventricular-arterial coupling [7,8]. Therefore, features that approximate hemodynamic consequences may provide useful additional information for machine-learning algorithms. True hemodynamic assessment requires flow measurements, pressure gradients, or cardiac phase information, which are rarely available in the limited public CHD datasets [7,8]. Direct hemodynamic measurements are not consistently available in public CHD datasets and may require specialized imaging, invasive assessment, or patient-specific simulation [7,8].

Because direct hemodynamic measurements are often unavailable, other methods of estimating hemodynamic impact from structural defects could help with classification. Features such as vessel narrowing, chamber-to-outflow size mismatch, and ventricular-great vessel alignment may reflect structural factors that influence circulation even when flow is not directly measured. The hypothesis of this study is that anatomy-derived hemodynamic-proxy features may act as useful additional predictors for CHD classification. To address this, we compare machine learning models trained on anatomy-derived hemodynamic-proxy features, purely anatomical features, and a feature set combining both.

METHODS

A. Dataset Creation and Harmonization

This study utilized manually segmented, publicly available datasets of healthy hearts and hearts of individuals with CHD. The final outcome of dataset harmonization was a single harmonized machine-learning table in which each row represented one subject, and each column represented either a class label, metadata, or a quantitative feature extracted from the segmentation mask. Four sources for heart models were combined: CHD68, ImageCHD, HVSMR-2.0, and MM-WHS CT training cases [3,9–11]. ImageCHD and CHD68 shared 63 case identifiers. Overlapping cases were only counted once. Five CHD68 cases were retained [3,9]. ImageCHD and HVSMR-2.0 were used as the CHD segmentation datasets [3,10]. MM-WHS CT training labels were added to increase the number of healthy whole-heart segmentation samples [11]. Because datasets were collected from different sources and used different label conventions, all segmentation masks were mapped into a shared anatomical ontology before feature extraction. Using segmentation masks to extract features for the shared dataset also allowed differences in modality-specific intensity to not directly affect feature calculation.

The universal anatomical structures used for this study were the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), aorta, and pulmonary artery (PA). All cases utilized were required to contain segmentations for all six structures. Labels unrelated to these structures were ignored.

To account for differences in voxel spacing across scans, voxel spacing was also harmonized. All volumes, surface areas, diameters, and distances were calculated in physical units using spacing in the x, y, and z directions. Pixel spacing for MM-WHS and HVSMR-2.0 was obtained from the extracted NIfTI metadata, while ImageCHD relied on the accompanying spreadsheet [3,10,11] because the metadata that was published was inaccurate.

Diagnostic labels were grouped into four machine-learning classes: Healthy, Septal defects, Conotruncal defects, and Complex CHD. Septal defects included atrial septal defect, ventricular septal defect, atrioventricular septal defect, and common atrium. Conotruncal defects included tetralogy of Fallot, transposition of the great arteries, double outlet right ventricle, common arterial trunk, and pulmonary atresia. Complex CHD included single-ventricle physiology and other complex or mixed CHD patterns. When a case contained multiple diagnoses, a predefined hierarchy was applied: Complex CHD > Conotruncal defects > Septal defects > Healthy. This hierarchy was used because the multiclass framework required one mutually exclusive outcome for each case, while the numbers of individual diagnoses and combinations of defects were too small to model as separate classes. Priority was based on the increasing extent of structural and circulatory abnormality represented by the four study categories as well as implications for patient health. This hierarchy was an operational rule for constructing the machine-learning outcome and should not be interpreted as a universally applicable scale of clinical severity.

Because specific categories (Healthy and Complex) were primarily extracted from certain source groups and disease labels were not fully independent, model performance may partly reflect dataset-specific segmentation conventions, acquisition differences, or population differences. Therefore, results were interpreted as exploratory, and improvements between feature sets were prioritized over raw accuracy. The final dataset contained 194 total cases: 41 Healthy, 44 Septal defects, 45 Conotruncal defects, and 64 Complex CHD cases. The distribution of diagnostic classes across dataset source groups is summarized in Table I.

Table I. Distribution of Cases Between Datasets

Source group	Healthy	Septal defects	Conotruncal defects	Complex CHD	Total
ImageCHD/CHD68 merged cohort	13	41	39	22	115
HVSMR-2.0	8	3	6	42	59
MM-WHS	20	0	0	0	20
Total	41	44	45	64	194

To assess whether diagnostic class was associated with dataset origin, cases were arranged in a 3×4 contingency table containing the three harmonized source groups and four diagnostic classes. A Pearson chi-square test of independence was applied to compare the observed source-by-class counts with the counts expected if dataset source and diagnostic class were independent. Cramér's V was calculated as an effect-size measure of the strength of this association, with values ranging from 0, indicating no association, to 1, indicating complete association. Statistical significance was defined as $p < 0.05$.

B. Development of Anatomy-Derived Hemodynamic-Proxy Features

The anatomy-derived hemodynamic-proxy feature set consisted of static geometric descriptors selected because they may reflect anatomical conditions relevant to cardiovascular flow. These features do not directly measure flow, pressure, resistance, or shunt magnitude. Direct hemodynamic calculations require information such as time-resolved flow, cardiac phase, blood pressure, patient-specific inflow and outflow conditions, and other physiological data that are not available in public segmentation datasets [7,8].

The anatomy-derived hemodynamic-proxy features were extracted from segmentation masks and were intended to approximate structural defects that may influence blood flow. Heart size was also incorporated into several hemodynamic-proxy features to account for differences in patient or heart size. Specifically,

centroid distance features were normalized by an estimated heart size defined as $\text{heart size} = (\text{total chamber volume in mm}^3)^{1/3}$ where $\text{total chamber volume} = \text{LV} + \text{RV} + \text{LA} + \text{RA}$.

Although this approach is a simplification and does not replace physiological measurement, it may provide complementary information in settings where only static image scans and segmentation masks are available. This is especially relevant for CHD datasets and other limited imaging datasets where detailed hemodynamic measurements remain unavailable [3,10]. Even in clinical imaging workflows, CT or anatomical MRI data may be available without matching flow MRI, catheterization-derived pressures or patient-specific CFD boundary conditions [7,8]. Anatomy-derived hemodynamic-proxy features were therefore evaluated as an exploratory method for representing flow-relevant anatomy when matched physiological measurements were unavailable.

C. Feature Extraction

Feature extraction was performed on the segmentation masks from the combined datasets. For each section, voxel counts were converted into physical volumes using case-specific voxel spacings derived from either metadata or the dataset's accompanying spreadsheet.

$$V_{\text{voxel}} = s_x \times s_y \times s_z$$
$$V_{\text{structure}} = N_{\text{voxel}} \times V_{\text{voxel}}$$

Where V_{voxel} represents the volume of a single voxel, s_x , s_y , and s_z represent the voxel spacing in the x-, y-, and z-directions, $V_{\text{structure}}$ represents the physical volume of the segmented anatomical structure, and N_{voxels} represents the number of voxels assigned to that structure in the segmentation mask.

For each anatomical structure, volume was calculated by multiplying the number of labeled voxels by the voxel volume. Surface area was estimated from exposed voxel faces, with each face area calculated using the corresponding physical voxel dimensions. Centroids were calculated as the mean physical coordinate of all voxels belonging to each structure. For vessel diameter estimation, axial cross-sectional areas were first calculated from labeled voxels, and equivalent circular diameters were then derived from those areas.

A total of 30 features were extracted and divided into two groups: conventional anatomical features and anatomy-derived hemodynamic-proxy features. This split was made to separate more conventional features used in CHD classification that primarily describe general cardiac anatomy from features that describe anatomical relationships that are often linked to possible flow pathways. All features were derived from static segmentation masks and were anatomical in origin.

Anatomical features described chamber size, global shape, and balance. Hemodynamic-proxy features described vessel caliber, chamber-outflow relationships, compactness, narrowing, and spatial alignment between major inflow and outflow structures. The conventional anatomical features and their definitions are presented in Table II.

Table II. Anatomical features calculated from segmentation masks.

Feature	Category	Formula/calculation	Structures used	Units
LV over total heart volume	Chamber size proportion	LV volume ÷ total chamber volume	LV, RV, LA, RA	Dimensionless
RV over total heart volume	Chamber size proportion	RV volume ÷ total chamber volume	LV, RV, LA, RA	Dimensionless
LA over total heart volume	Chamber size proportion	LA volume ÷ total chamber volume	LV, RV, LA, RA	Dimensionless
RA over total heart volume	Chamber size proportion	RA volume ÷ total chamber volume	LV, RV, LA, RA	Dimensionless
RV over LV volume	Ventricular balance	RV volume ÷ LV volume	RV, LV	Dimensionless
RA over LA volume	Atrial balance	RA volume ÷ LA volume	RA, LA	Dimensionless
PA over aorta volume	Great-vessel size balance	PA volume ÷ aorta volume	PA, aorta	Dimensionless
LV sphericity	Chamber shape	$\pi^{(1/3)} \times (6 \times \text{LV volume})^{(2/3)} \div \text{LV surface area}$	LV	Dimensionless
RV sphericity	Chamber shape	$\pi^{(1/3)} \times (6 \times \text{RV volume})^{(2/3)} \div \text{RV surface area}$	RV	Dimensionless
LA sphericity	Chamber shape	$\pi^{(1/3)} \times (6 \times \text{LA volume})^{(2/3)} \div \text{LA surface area}$	LA	Dimensionless
RA sphericity	Chamber shape	$\pi^{(1/3)} \times (6 \times \text{RA volume})^{(2/3)} \div \text{RA surface area}$	RA	Dimensionless

LV elongation	Chamber shape	Maximum LV bounding-box extent ÷ minimum LV bounding-box extent	LV	Dimensionless
RV elongation	Chamber shape	Maximum RV bounding-box extent ÷ minimum RV bounding-box extent	RV	Dimensionless
Aorta elongation and tortuosity proxy	Great-vessel morphology	Maximum aorta bounding-box extent ÷ minimum aorta bounding-box extent	Aorta	Dimensionless
PA elongation and tortuosity proxy	Great-vessel morphology	Maximum PA bounding-box extent ÷ minimum PA bounding-box extent	PA	Dimensionless

Abbreviations: LV, left ventricle; RV, right ventricle; RA, right atrium; LA, left atrium; PA, pulmonary artery.

LV over total heart volume, RV over total heart volume, LA over total heart volume, and RA over total heart volume describe the relative contribution of each chamber to total chamber size. RV over LV volume and RA over LA volume describe ventricular and atrial balance, respectively, and may reflect chamber dominance or asymmetry. PA over aorta volume describes the relative size balance between the pulmonary and systemic outflow vessels. LV sphericity, RV sphericity, LA sphericity, and RA sphericity describe how closely each chamber approximates a compact spherical shape, which may reflect chamber remodeling or distortion. LV elongation and RV elongation describe chamber shape distortion based on bounding-box dimensions, while aorta elongation and tortuosity proxy and PA elongation and tortuosity proxy describe elongated or distorted great-vessel morphology. The anatomy-derived hemodynamic-proxy features and their definitions are presented in Table III.

Table III. Anatomy-derived Hemodynamic-Proxy features calculated from segmentation mask.

Feature	Category	Formula/calculation	Structures used	Units
---------	----------	---------------------	-----------------	-------

PA over aorta cross-sectional area	Outflow size balance	Median PA axial cross-sectional area $\div$ median aorta axial cross-sectional area	PA, aorta	Dimensionless
Narrowest PA diameter over maximum PA diameter	Vessel narrowing	Minimum PA equivalent diameter $\div$ maximum PA equivalent diameter	PA	Dimensionless
Narrowest aorta diameter over maximum aorta diameter	Vessel narrowing	Minimum aorta equivalent diameter $\div$ maximum aorta equivalent diameter	Aorta	Dimensionless
RV volume over PA volume	Chamber-to-outflow coupling	RV volume $\div$ PA volume	RV, PA	Dimensionless
LV volume over aorta volume	Chamber-to-outflow coupling	LV volume $\div$ aorta volume	LV, aorta	Dimensionless
RA volume over RV volume	Inflow-to-ventricle balance	RA volume $\div$ RV volume	RA, RV	Dimensionless
LA volume over LV volume	Inflow-to-ventricle balance	LA volume $\div$ LV volume	LA, LV	Dimensionless
PA surface area over volume	Vessel surface complexity	PA surface area $\div$ PA volume	PA	mm ⁻¹
Aorta surface area over volume	Vessel surface complexity	Aorta surface area $\div$ aorta volume	Aorta	mm ⁻¹
PA compactness	Vessel compactness	$\pi^{(1/3)} \times (6 \times \text{PA volume})^{(2/3)} \div \text{PA surface area}$	PA	Dimensionless
Aorta compactness	Vessel compactness	$\pi^{(1/3)} \times (6 \times \text{aorta volume})^{(2/3)} \div \text{aorta surface area}$	Aorta	Dimensionless

RV to PA centroid distance over heart size	Spatial pathway relationship	Distance between RV and PA centroids $\div$ heart size	RV, PA	Dimensionless
LV to aorta centroid distance over heart size	Spatial pathway relationship	Distance between LV and aorta centroids $\div$ heart size	LV, aorta	Dimensionless
PA to aorta centroid distance over heart size	Great-vessel spatial relationship	Distance between PA and aorta centroids $\div$ heart size	PA, aorta	Dimensionless
RV to LV centroid distance over heart size	Ventricular spatial relationship	Distance between RV and LV centroids $\div$ heart size	RV, LV	Dimensionless

The anatomy-derived hemodynamic-proxy features were calculated as ratios, shape indices, and normalized spatial measurements intended to represent anatomical characteristics that may influence cardiovascular flow. PA over aorta cross-sectional area reflects relative pulmonary versus systemic outflow caliber. Relative pulmonary and aortic caliber may also reflect differential great-vessel remodeling associated with chronic pressure or volume loading, although this ratio does not directly measure pulmonary-to-systemic flow. Narrowest-to-maximum diameter ratios quantify focal vessel narrowing. A reduction in outflow caliber may increase local velocity, energy loss, resistance, and the trans-stenotic pressure gradient, although the magnitude of these effects also depends on flow rate, lesion length, vessel geometry, and downstream conditions [7,8]. RV volume over PA volume and LV volume over aorta volume describe chamber-to-outflow coupling, while RA volume over RV volume and LA volume over LV volume describe inflow-to-ventricle balance. These ratios may capture disproportionate chamber or vessel remodeling associated with chronic preload, pressure loading, or shunt-related volume loading, but they are not direct measurements of inflow, outflow, or ventricular-arterial coupling. PA surface area over volume and aorta surface area over volume describe vessel surface complexity, while PA compactness and aorta compactness describe vessel shape compactness. RV to PA centroid distance over heart size, LV to aorta centroid distance over heart size, PA to aorta centroid distance over heart size, and RV to LV centroid distance over heart size describe normalized spatial relationships between chambers and great vessels that may reflect abnormal outflow alignment, vessel arrangement, or intracardiac flow pathways. The relationships between these exact feature definitions and actual flow,

resistance, pressure gradients, or shunt physiology are not yet established so these features remain as exploratory hemodynamic-proxy features.

D. Machine Learning Models

Four supervised machine learning models were evaluated to classify cases into the four diagnostic groups. The models used for this study were k-nearest neighbors (KNN), multinomial logistic regression (MNL), support vector machine (SVM), and random forest (RF) [12]. Models were selected based on the way they represent different classification strategies that are commonly used for tabular biomedical datasets. A variety of models were selected to test the validity of the Anatomy-derived Hemodynamic Proxy across different classification strategies.

MNL was included as a linear baseline model for comparison with more flexible classifiers. KNN was included as a distance-based classifier to assess whether cases with similar segmentation-derived feature profiles clustered by diagnostic group. SVM was included as a margin-based classifier because it is commonly used for small tabular datasets with limited feature dimensions. Random forest was included as a nonlinear ensemble model to capture potential interactions between anatomical and anatomy-derived hemodynamic-proxy features.

Hyperparameter optimization was performed to identify ideal model settings for this dataset. Hyperparameter tuning was performed using cross-validation and was designed using macro F1-score as the optimization metric. Optimizing macro F1-score prevents the moderately uneven dataset from skewing accuracy results. GridSearchCV, a function inside the Python package scikit-learn, was employed to find optimal combinations of hyperparameters for each model [12,13]. All models utilized median imputation in order to replace missing feature values before classification. KNN, MNL and SVM models also used MinMax scaling because these models are sensitive to feature magnitude and distance-based associations. During cross-validation, preprocessing steps were fitted only on the training portion of each split and then applied to the corresponding held-out validation fold. This was done to reduce preprocessing-related data leakage during model evaluation. No data-driven feature selection was performed; all predefined features belonging to each feature set were retained during model training and evaluation.

Optimized parameters were fixed and used for the final comparison between the anatomical-only, hemodynamic-proxy-only, and combined feature sets. Model settings were applied homogeneously for each classifier. This was done to ensure that differences in performance reflected differences between feature sets rather than differences in model tuning.

E. Model Evaluation and Statistical Comparison

Model performance was evaluated using repeated stratified five-fold cross-validation with 20 repeats, producing 100 train-test splits for each model and feature set [12]. For each split, the model pipeline was fitted solely on the training data. Median imputation and MinMax scaling, when used, were learned from the training fold and then applied to the held-out fold. Model predictions were then generated for the held-out fold. This process was repeated across all cross-validation splits. Stratification was used to

maintain the relative class distribution within each fold. This ensured that imbalances between the numbers of Healthy, Septal defects, Conotruncal defects, and Complex CHD cases reduced the influence of class imbalance on accuracy metrics. Within the repeated stratified five-fold cross-validation, median imputation and, where applicable, MinMax scaling were fitted using only the training portion of each split and then applied to the corresponding held-out test portion. Hyperparameters had previously been selected using the complete dataset and were held fixed during the repeated cross-validation evaluation. Thus, preprocessing and model fitting were fold-specific, but hyperparameter optimization was not nested within the evaluation procedure.

Performance was calculated separately for the anatomical-only feature set, the hemodynamic-proxy-only feature set, and the combined anatomical plus hemodynamic-proxy feature set. This allowed evaluation of whether the anatomy-derived hemodynamic-proxy features performed well alone and whether they improved performance when added to conventional anatomical features. Comparing the performance difference between the combined feature set and the anatomical-only feature set was used to evaluate whether anatomy-derived hemodynamic-proxy features could be a constructive addition to future datasets.

Performance differences were calculated on a split-by-split basis. For example, the macro F1-score from the anatomical-only model would be subtracted from the macro F1-score from the combined model on the same cross-validation split. This produced paired differences across the 100 repeated cross-validation splits. Mean improvement and standard deviation were then reported for each difference. Because the same subjects appeared in multiple training and held-out folds, the 100 repeated cross-validation evaluations were not treated as independent experiments.

Formal comparisons between the combined and anatomical-only feature sets were performed using the corrected repeated-cross-validation t-test described by Nadeau and Bengio [14]. For each classifier and performance metric, a paired difference was calculated on each matched split by subtracting the anatomical-only score from the combined-feature score. The standard error was corrected for dependence between overlapping training sets using $SE_{corrected} = s_d \sqrt{(1/N + n_{test}/n_{train})}$, where s_d was the standard deviation of the paired differences, and $N=100$ was the number of evaluations. For five-fold cross-validation, $n_{test}/n_{train}=1/4$. Two-sided corrected 95% confidence intervals and p-values were calculated. Mean $\pm$ SD values across folds were retained as descriptive measures of cross-validation variability rather than interpreted as estimates from 100 independent experiments.

F. Feature Importance Analysis

Feature importance was performed using the combined feature set to identify which features contributed most strongly to CHD classification within our model. Permutation feature importance was calculated for the optimized KNN model using the held-out test set from a separate stratified 75/25 train-test split. This split provided a fixed held-out subset on which all feature permutations could be evaluated while preserving the relative distribution of the four diagnostic classes. Median imputation, MinMax scaling, and KNN model fitting were performed using only the 75% training partition and were subsequently

applied to the 25% held-out partition. However, because the KNN hyperparameters had previously been selected using the complete dataset, the held-out partition was not fully independent of model development and hyperparameter optimization. Each of the 30 features was randomly permuted 100 times. Then, the resulting decrease in model performance was measured. Macro F1-score was used as the primary scoring metric because it reduces the effect of class imbalance. Features that produced a larger decrease in macro F1-score after being permuted were interpreted as having greater importance for classification. Feature importance was reported as the permutation importance mean across repeated random permutations, with higher values indicating greater reliance of the fitted KNN model on that feature. These values were interpreted as model-specific measures of predictive contribution and not as evidence that a feature was biologically causal or independently physiologically important.

Results

A. Optimized Model Hyperparameters

Optimized model hyperparameters were used to compare the accuracy of CHD classification for models trained on the anatomy-derived hemodynamic-proxy-only feature set, anatomical-only feature set, and combined feature set. After hyperparameter optimization, the final KNN model used 31 neighbors, distance-based weighting, and the Euclidean distance metric. The optimized MNL model used elastic net regularization with $C = 0.6553$, $l1_ratio = 0.10$, balanced class weighting, the saga solver, and a maximum of 10,000 iterations. The optimized SVM model used a linear kernel with $C = 0.1052$ and balanced class weighting. The optimized random forest model used 644 trees, a maximum depth of 3, a minimum of 4 samples per leaf, a minimum of 13 samples required to split an internal node, square-root feature selection, log-loss splitting criterion, no bootstrapping, and no class weighting.

B. Repeated Cross-Validation Performance Across Models

Model performance was evaluated across hemodynamic-proxy-only, anatomical-only, and combined feature sets using repeated stratified five-fold cross-validation with 20 repeats. The results of the evaluation are shown in Table IV. Across the four models, the combined feature set had the highest mean performance for accuracy (+2.8% for RF), macro F1-score (+0.032 for RF), and balanced accuracy (+3.3% for RF). The hemodynamic-proxy-only feature set outperformed the anatomical-only set in all models in terms of accuracy: KNN (+2.7%), SVM (+0.2%), MNL models (+1.5%), and RF(+1.9%).

Table IV. Repeated cross-validation performance for each model and feature set.

Model	Feature set	Accuracy, mean $\pm$ SD (%)	Macro F1, mean $\pm$ y, SD	
KNN	Anatomical only	61.2 $\pm$ 7.1	0.606 $\pm$ 0.075	0.607 $\pm$ 0.073
KNN	Hemodynamic-proxy only	63.8 $\pm$ 6.9	0.631 $\pm$ 0.073	0.635 $\pm$ 0.070

KNN	Combined	65.9 ± 6.7	0.653 ± 0.071	0.655 ± 0.068
MNL	Anatomical only	61.3 ± 6.6	0.609 ± 0.070	0.617 ± 0.067
MNL	Hemodynamic-proxy only	62.8 ± 6.7	0.623 ± 0.071	0.630 ± 0.068
MNL	Combined	65.7 ± 6.7	0.654 ± 0.070	0.660 ± 0.067
SVM	Anatomical only	58.1 ± 6.2	0.569 ± 0.069	0.586 ± 0.065
SVM	Hemodynamic-proxy only	58.3 ± 6.5	0.574 ± 0.070	0.597 ± 0.065
SVM	Combined	65.8 ± 7.0	0.651 ± 0.076	0.655 ± 0.071
Random forest	Anatomical only	58.6 ± 6.7	0.575 ± 0.071	0.572 ± 0.069
Random forest	Hemodynamic-proxy only	60.5 ± 7.7	0.600 ± 0.079	0.597 ± 0.077
Random forest	Combined	61.4 ± 7.0	0.607 ± 0.073	0.605 ± 0.070

The highest mean accuracy was produced by KNN using the combined feature set with an accuracy of 65.9% ± 6.7%. The highest mean macro F1-score and balanced accuracy were observed for MNL using the combined feature set, with a macro F1-score of 0.6539 ± 0.0703 and balanced accuracy of 66% ± 6.71%.

The combined feature set produced positive mean paired differences relative to the anatomical-only feature set across all classifiers and evaluation metrics. For KNN, the mean improvements were 0.0470 in accuracy (95% CI: -0.0031 to 0.0971; corrected p = 0.066), 0.0466 in macro F1-score (95% CI: -0.0027 to 0.0960; corrected p = 0.063), and 0.0474 in balanced accuracy (95% CI: -0.0042 to 0.0990; corrected p = 0.071). For MNL, the corresponding improvements were 0.0442 in accuracy (95% CI: -0.0087 to 0.0972; corrected p = 0.101), 0.0454 in macro F1-score (95% CI: -0.0088 to 0.0995; corrected p = 0.099), and 0.0436 in balanced accuracy (95% CI: -0.0084 to 0.0955; corrected p = 0.099). SVM demonstrated improvements of 0.0773 in accuracy (95% CI: 0.0135 to 0.1410; corrected p = 0.018), 0.0819 in macro F1-score (95% CI: 0.0095 to 0.1543; corrected p = 0.027), and 0.0693 in balanced accuracy (95% CI: 0.0047 to 0.1339; corrected p = 0.036). For random forest, the improvements were 0.0278 in accuracy (95% CI: -0.0171 to 0.0728; corrected p = 0.222), 0.0320 in macro F1-score (95% CI: -0.0137 to 0.0777; corrected p = 0.167), and 0.0332 in balanced accuracy (95% CI: -0.0142 to 0.0805; corrected p = 0.167). Thus, the corrected confidence intervals excluded zero for all three SVM comparisons, whereas the intervals for KNN, MNL, and random forest included zero. These results represent corrected internal comparisons rather than evidence from independent validation experiments.

C. Confusion Matrix Results

Confusion matrices were generated for each model and feature set; the KNN confusion matrix is shown in Figure 1, and the RF confusion matrix is shown in Figure 2. Across models, the largest off-diagonal errors

occurred between complex CHD and septal defects, and between septal defects and conotruncal defects. Specifically for the combined feature set, complex CHD cases were most often misclassified as septal defects in KNN, MNL, and random forest.

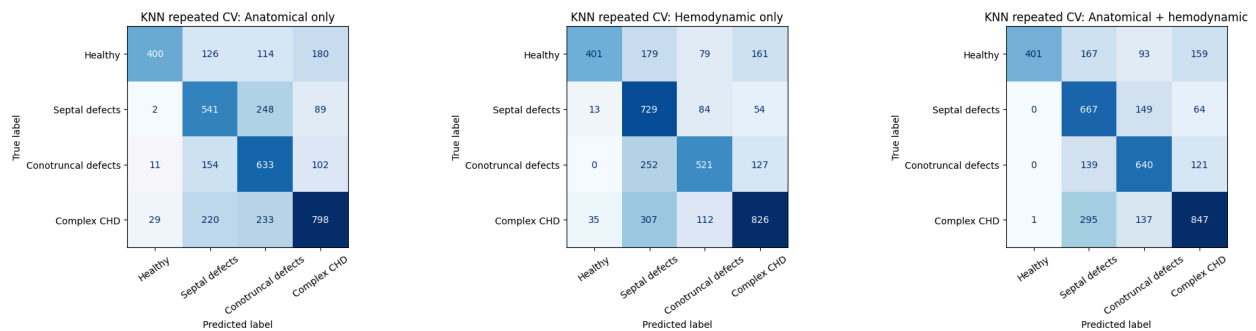


Figure 1. KNN repeated cross-validation confusion matrices. Aggregated confusion matrices from repeated stratified five-fold cross-validation with 20 repeats for KNN models trained using anatomical-only, hemodynamic-proxy-only, and combined anatomical + hemodynamic-proxy feature sets.

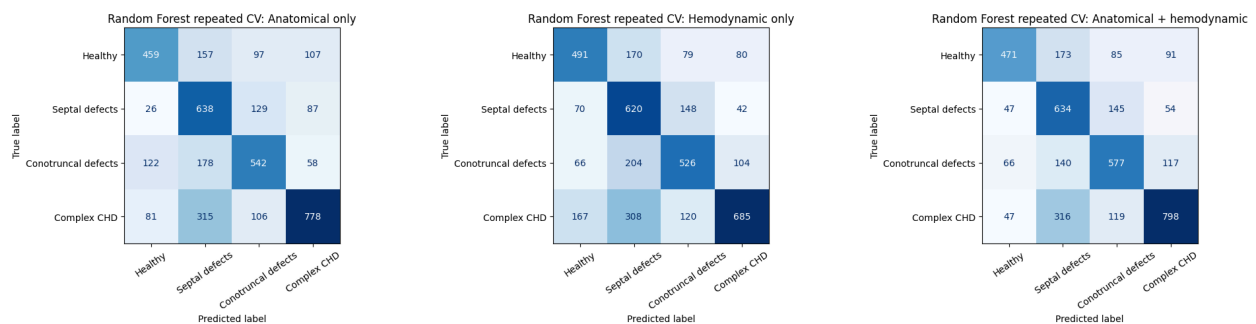


Figure 2. Random forest repeated cross-validation confusion matrices. Aggregated confusion matrices from repeated stratified five-fold cross-validation with 20 repeats for random forest models trained using anatomical-only, hemodynamic-proxy-only, and combined anatomical + hemodynamic-proxy feature sets.

D. One-vs-Rest Receiver Operating Characteristic (ROC) Analysis

One-vs-rest ROC curves were generated for the healthy, septal defects, conotruncal defects and complex CHD classes (Fig. 3).

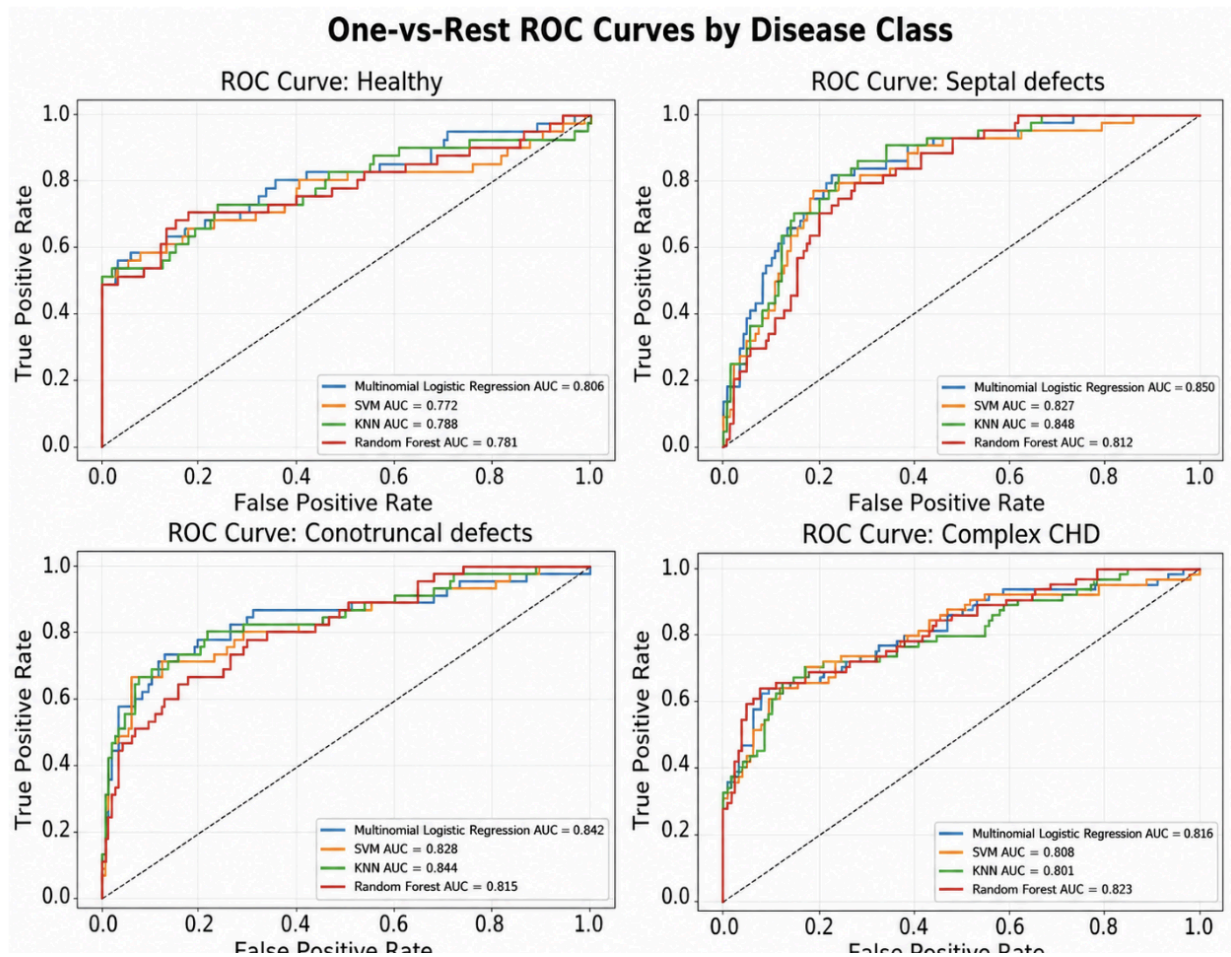


Figure 3. One-vs-rest ROC curves by disease class. ROC curves comparing MNL, SVM, KNN, and random forest models for each disease class using the combined anatomical + hemodynamic-proxy feature set. The area under the curve (AUC) values are shown in each subplot legend.

AUC values ranged from 0.772 to 0.850 across all classes and models. The highest AUC for healthy classification was observed for MNL (0.806). The highest AUC for septal defects was also observed for MNL (0.850). The highest AUC for conotruncal defects was observed for KNN. The highest AUC for complex CHD was observed for random forest (Figure 3).

E. Feature Importance

Importance for all 30 features was calculated for the optimized KNN model. The permutation importance mean (PIM) for each feature was calculated from the decrease in macro F1-score after that feature was randomly permuted in the held-out test set. Larger PIM values indicate a higher positive importance in classification.

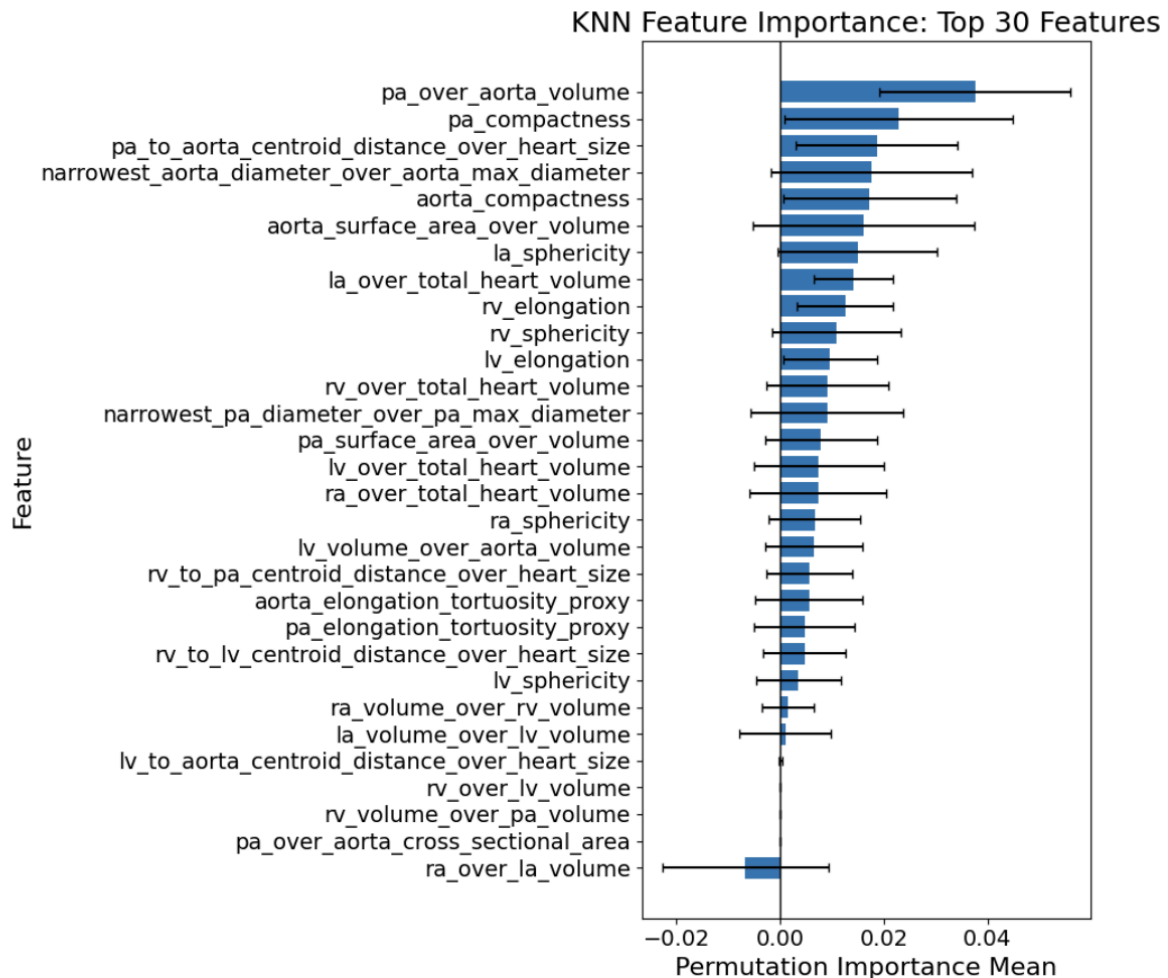


Figure 4. KNN permutation feature importance for the combined anatomical plus hemodynamic-proxy feature set. Feature importance was calculated on the held-out test set using 100 random permutations for each feature. Error bars represent the standard deviation across the 100 permutation repeats. Higher values indicate a greater contribution to the predictive performance of the fitted KNN model within this held-out split.

Permutation importance was calculated for all 30 features using the optimized KNN model with the combined feature set (Fig. 4). The feature with the highest PIM across both feature sets was Pulmonary Artery over Aorta Volume, an anatomical great-vessel size-balance feature, with a permutation importance mean of 0.038. Although the highest-ranked feature was anatomical, the next five highest-ranked features were anatomy-derived hemodynamic-proxy features. The lowest-ranked and only feature with a negative PIM was RA over LA volume with a PIM of -0.007. The average PIM for anatomical features was 0.00915, while it was 0.00860 for anatomy-derived hemodynamic-proxy

features. Feature importance was also performed on SVM, RF, and MNL, yielding similar results (not shown).

F. Source Confounding

Diagnostic class was significantly associated with dataset source ($\chi^2(6)=140.88$, $p<0.001$), with a Cramér's V of 0.603. MM-WHS contributed exclusively Healthy cases, and HVSMR-2.0 was predominantly composed of Complex CHD cases. The magnitude of Cramér's V indicated substantial source–class association. Because several diagnostic classes were absent or sparsely represented within individual sources, reliable source-specific four-class performance estimates could not be calculated.

DISCUSSION AND CONCLUSION

In this study, we evaluated whether anatomy-derived hemodynamic-proxy features could aid in machine learning classification of CHD (split into the classes Healthy, Conotruncal defects, Septal defects, and Complex CHD) when added to conventional segmentation-derived anatomical features. Across all four machine learning models, the combined feature set performed better than the anatomical-only feature set. Among the four models, the KNN model produced the highest mean accuracy with the combined feature set, while multinomial logistic regression produced the highest macro F1-score and balanced accuracy. SVM had the biggest improvement after adding the hemodynamic-proxy features. The improvements with the combined feature set suggest that anatomical measurements that have plausible relationships with flow may serve as meaningful features within a machine learning classification framework. In the feature-importance analysis, the highest-ranked feature was the anatomical PA-to-aorta volume ratio, while the next five highest-ranked features were anatomy-derived hemodynamic-proxy features. This pattern indicates that both conventional great-vessel anatomy and flow-relevant geometric descriptors contributed to classification. These model-specific associations do not establish biological causality or independent physiological importance, particularly because correlated features may share predictive information. The features ranked the highest were primarily related to the great vessels, including the pulmonary artery and aorta. This is medically consistent because blood vessels leading into and out of the heart often carry particularly useful diagnostic information for diseases that alter flow like CHD [2,16]. Although the combined feature set improved macro F1-score and balanced accuracy across the evaluated models, the absolute improvements were modest and varied by classifier. These gains should be interpreted as exploratory evidence that the proxy features may provide complementary predictive information, rather than evidence of a clinically meaningful improvement.

These findings align with those of previous studies showing that three-dimensional cardiac anatomy can be used for automated CHD classification. Like in the studies of ImageCHD [3], PointCHD [4], Xu et al. [5], and Wang et al. [6], our work demonstrates that cardiac structure contains diagnostic information that can be used by machine learning models. However, our method differs from prior image-based and point-cloud approaches by using interpretable tabular features extracted from segmentation masks rather than training directly on image voxels, echocardiogram views, or point-cloud representations. This

approach produced lower performance than some larger deep learning studies, but it allowed each model input to be connected to a specific anatomical or anatomy-derived hemodynamic-proxy measurement. This feature-based representation provides methodological interpretability by connecting model inputs to defined geometric measurements; however, its clinical utility has not yet been established. The results also differ from true hemodynamic studies because this work did not measure flow directly. Recent 4D-flow CMR studies have shown that direct flow imaging can quantify velocity, flow volumes, Qp/Qs, stenosis, shunts, and valve function in CHD [15,16]. Although these manual measurements can offer more direct quantification on blood flow, they are often difficult to obtain in practical clinical settings. In contrast to direct flow imaging, the present study used anatomy-derived hemodynamic-proxy features that may reflect flow-related structure but do not replace direct physiological measurements.

There are several limitations to this study. First, the database used to train the machine learning model was relatively small, with 194 total cases distributed across the disease classes. Although repeated stratified cross-validation allowed model performance to be estimated across multiple train-test splits, the absence of a larger, independent external dataset limits conclusions about generalizability. Second, dataset source and diagnostic class were strongly associated, indicating that source-specific characteristics may have contributed to model predictions alongside disease-related anatomy. Consequently, the reported performance may partly reflect imaging, segmentation, or population characteristics shared within each source. The source imbalance prevented reliable separation of dataset-origin effects from diagnostic effects, and external validation using an independent, source-balanced cohort containing all four diagnostic classes is necessary before generalizability or clinical applicability can be established. Third, the proposed features were derived exclusively from static anatomy and were not validated against 4D-flow CMR, catheterization, or computational fluid dynamics. Outflow-caliber and vessel-narrowing features have a plausible physiological relationship to obstruction, whereas the volume ratios, shape measurements, and centroid distances remain exploratory geometric descriptors. Their predictive performance therefore does not establish a relationship with actual flow, resistance, pressure gradients, or shunt physiology. Therefore, these features should only be interpreted as anatomy-derived proxies. Fourth, the final evaluation of the machine learning models was not performed on a fully nested cross-validation design. Hyperparameters were optimized before the final evaluation and subsequently kept fixed across feature sets. Therefore, the reported performance values may be slightly optimistic compared with performance on a fully independent external dataset. The permutation feature-importance analysis was also based on a single stratified 75/25 split. Although the held-out partition was not used to fit the KNN model or preprocessing transformations in that analysis, it was not fully independent of hyperparameter optimization, which had previously used the complete dataset. Consequently, the feature rankings may depend on the selected split and should be interpreted as exploratory, model-specific predictive associations rather than stable estimates of biological importance. Finally, individual diagnoses were grouped into four broader classes in order to increase class size. Cases with multiple diagnoses were also assigned to one class using a predefined hierarchy based on anatomical complexity. The broad categories may have made classification less granular and potentially easier than diagnosis at the individual-lesion level, while heterogeneity among defect combinations within each category may also have reduced model performance. As a result, the one-class label may limit clinical utility as specific

dual-defects are common and may be overlooked with this method. However, using the hierarchical system allowed for larger class sizes while still classifying the most anatomically significant CHD.

Future directions for this research topic include increasing dataset size, improving disease specificity, and validating models on external datasets. Future studies should compare the hemodynamic-proxy features with 4D-flow CMR-derived velocity, flow volume, and Qp/Qs; catheterization-derived pressures and pressure gradients; and CFD-derived velocity fields, pressure fields, and wall shear stress [7,8,15,16]. In addition, combining the segmentation-derived tabular features present in this study with image-based, mesh-based, or point-cloud machine learning approaches may improve classification accuracy. To address the one-type CHD hierarchy limitation, future studies should use larger and more diagnostically diverse cohorts, retain labels for each individual cardiac defect, and evaluate multilabel or hierarchical classification methods capable of identifying multiple coexisting defects in the same patient. Our results suggest that anatomy-derived hemodynamic-proxy features can improve CHD classification, but further validation is needed before this approach should be considered for clinical use. Overall, this study provides an exploratory foundation for evaluating anatomy-derived hemodynamic-proxy features in larger, independently validated CHD datasets.

REFERENCES

1. Stout, K.K.; Daniels, C.J.; Aboulhosn, J.A.; Bozkurt, B.; Broberg, C.S.; Colman, J.M.; Crumb, S.R.; Dearani, J.A.; Fuller, S.; Gurvitz, M.; et al. 2018 AHA/ACC Guideline for the Management of Adults With Congenital Heart Disease. *Circulation* 2019, 139, e698–e800. <https://doi.org/10.1161/CIR.0000000000000603>
2. Lapierre, C.; Déry, J.; Guérin, R.; Viremouneix, L.; Dubois, J.; Garel, L. Segmental Approach to Imaging of Congenital Heart Disease. *RadioGraphics* 2010, 30, 397–411. <https://doi.org/10.1148/rg.302095112>
3. Xu, X.; Wang, T.; Zhuang, J.; Yuan, H.; Huang, M.; Cen, J.; Jia, Q.; Dong, Y.; Shi, Y. ImageCHD: A 3D Computed Tomography Image Dataset for Classification of Congenital Heart Disease. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*; Springer: Cham, Switzerland, 2020; Volume 12266, pp. 77–87. https://doi.org/10.1007/978-3-030-59719-1_8
4. Yang, D.; Gao, W. PointCHD: A Point Cloud Benchmark for Congenital Heart Disease Classification and Segmentation. *IEEE Journal of Biomedical and Health Informatics* 2025, 29, 2683–2694.
5. Xu, X.; Jia, Q.; Yuan, H.; Qiu, H.; Dong, Y.; Xie, W.; Yao, Z.; Zhang, J.; Nie, Z.; Li, X.; et al. A Clinically Applicable AI System for Diagnosis of Congenital Heart Diseases Based on Computed Tomography Images. *Medical Image Analysis* 2023, 90, 102953. <https://doi.org/10.1016/j.media.2023.102953>

6. Wang, J.; Liu, X.; Wang, F.; Zheng, L.; Gao, F.; Zhang, H.; Zhang, X.; Xie, W.; Wang, B. Automated Interpretation of Congenital Heart Disease from Multi-View Echocardiograms. *Medical Image Analysis* 2021, 69, 101942. <https://doi.org/10.1016/j.media.2020.101942>
7. Gerrah, R.; Hsia, T.Y.; Marsden, A.L.; Migliavacca, F.; Pennati, G.; Corsini, C. Computational Fluid Dynamics: A Primer for Congenital Heart Disease. *Perfusion* 2021, 36, 342–348. <https://doi.org/10.1177/0218492320957163>
8. Pennati, G.; Corsini, C.; Hsia, T.Y.; Migliavacca, F.; for the Modeling of Congenital Hearts Alliance Investigators. Computational Fluid Dynamics Models and Congenital Heart Diseases. *Frontiers in Pediatrics* 2013, 1, 4. <https://doi.org/10.3389/fped.2013.00004>
9. Xu, X.; Wang, T.; Shi, Y.; Yuan, H.; Jia, Q.; Huang, M.; Zhuang, J. Whole Heart and Great Vessel Segmentation in Congenital Heart Disease Using Deep Neural Networks and Graph Matching. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*; Springer: Cham, Switzerland, 2019; Volume 11765, pp. 477–485. https://doi.org/10.1007/978-3-030-32245-8_53
10. Pace, D.F.; Contreras, H.T.M.; Romanowicz, J.; Ghelani, S.; Rahaman, I.; Zhang, Y.; Gao, P.; Jubair, M.I.; Yeh, T.; Golland, P.; et al. HVSMR-2.0: A 3D Cardiovascular MR Dataset for Whole-Heart Segmentation in Congenital Heart Disease. *Scientific Data* 2024, 11, 721. <https://doi.org/10.1038/s41597-024-03469-9>
11. Zhuang, X.; Li, L.; Payer, C.; Štern, D.; Urschler, M.; Heinrich, M.P.; Oster, J.; Wang, C.; Smedby, Ö.; Bian, C.; et al. Evaluation of Algorithms for Multi-Modality Whole Heart Segmentation: An Open-Access Grand Challenge. *Medical Image Analysis* 2019, 58, 101537. <https://doi.org/10.1016/j.media.2019.101537>
12. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011, 12, 2825–2830. Available online: <https://jmlr.org/papers/v12/pedregosa11a.html>
13. Scikit-learn Developers. GridSearchCV. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
14. Nadeau, C.; Bengio, Y. Inference for the Generalization Error. *Machine Learning* 2003, 52, 239–281. <https://doi.org/10.1023/A:1024068626366>
15. Voges, I.; Ide, Y.; Puntmann, V.; Gabbert, D.; Krupickova, S. Emerging Role of 4D Flow for Clinical Use in Adult Congenital Heart Disease. *Current Cardiovascular Imaging Reports* 2025, 18, 1. <https://doi.org/10.1007/s12410-024-09599-3>

16. Lagerstrand, K.; Svensson, P.-A.; Andersson, L.; Nyström, A.; Dangardt, F.; de Lange, C. 4D Flow Cardiac Magnetic Resonance in Pediatric Congenital Heart Disease: Insights from over Four Years of Clinical Practice. *Clinical Imaging* 2025, 119, 110399. <https://doi.org/10.1016/j.clinimag.2024.110399>