25011484 | UFCFMJ-15-M | 2163 words, excl. References | https://github.com/sumai787/ml-assign-svr-2025.git

Predicting 12-Month Death or Cardiac Transplant After the Norwood Procedure in Infants with Single-Ventricle Congenital Heart Disease

A machine learning comparison of Logistic Regression, Linear SVM, Random Forest, and Gradient Boosting using early surgical and immediate postoperative ICU metrics from the Pediatric Heart Network public dataset

Sumaiya Rahman (Student ID: 25011484)
Department of Computer Science & Creative Technologies
University of the West of England (UWE Bristol)
Bristol, United Kingdom
Sumaiya2.rahman@live.uwe.ac.uk

Abstract

Using the Pediatric Heart Network SVR dataset (n=549), 12-month death/transplant after Norwood is predicted from peri-operative/ICU variables (Core) with optional pre-discharge echo (Extended). Standard preprocessing, stratified split, and 5-fold CV applied across logistic regression, linear SVM, random forest, and gradient boosting. Gradient boosting performed best (ROC-AUC ≈0.77); echo added little. Obvious post-outcome variables excluded to avoid leakage; results are interpretable.

I. Introduction

Congenital heart disease (CHD) is the most common birth defect, affecting nearly 1 in 100 live births worldwide [1]. Infants born with single-ventricle physiology typically undergo staged palliation beginning with the Norwood procedure, but 30–40% may die or require cardiac transplantation within the first year despite surgical advances [2], [3]. Existing risk scores and regression-based models have shown limited predictive accuracy in this high-risk population [4]. Machine learning methods, including random forests and gradient boosting offer the potential to capture nonlinearities and improve risk stratification in pediatric cardiac surgery [5], [6]. This study compares classical and ensemble ML models for predicting 12-month death or transplant after the Norwood procedure using Pediatric Heart Network data.

II. Data & Problem Definition

A. Dataset overview

This study uses the Single Ventricle Reconstruction (SVR) Trial public dataset, released by the Pediatric Heart Network (PHN) and funded by the U.S. National Heart, Lung, and Blood Institute (NHLBI, part of the NIH). The trial enrolled 549 infants with single-ventricle physiology undergoing the Norwood procedure between 2005-2009 at 15 North American centers, with follow-up to 12 months. The public-use dataset was locked and released in 2012 [7]. The primary outcome is coded as dtx1yr = 1 if death or transplant occurred by 12 months, else 0.

Two feature sets were defined for prediction:

- Core (12 variables): peri-operative and *immediate post-operative ICU metrics* available within the index Norwood hospitalization (forms R102 and R103).
- Extended (+7 variables, total 19): Core features plus pre-discharge echocardiographic measures from the Echo Core Lab (form R301).

Continuous predictors are represented as numeric values, categorical as one-hot encodings, and the binary outcome as 0/1. Missingness is particularly common in echocardiographic variables, motivating imputation strategies (see §III.C).

Table 1. Core feature (12 variables).

Feature (clean name)	Original (name)	Source form	Type/Unit	description
apgar_5min	APGAR5	R102	score 0–10	Apgar score at 5 minutes after birth.
lactate_val	LACTATE	R102	yes/no (1/0)	Was lactate measured before Norwood?
max_lactate_flag	HIGHLACT	R102	mmol/L (numeric)	Highest lactate value before Norwood.
cpb_time_min	BYPASS_T	R103	minutes	Cardiopulmonary bypass (CPB) time during Norwood.
xclamp_time_min	CCLAMP_T	R103	minutes	Aortic cross-clamp time during Norwood.

lowest_temp_c	LOWTEMP	R103	°C	Lowest temperature on bypass.
lowest_hct_pct	LOW_HCT	R103	%	Lowest hematocrit on bypass.
sat_at_discharge	NO2SAT	R103	%	Oxygen saturation at Norwood hospital discharge.
preop_intubated	INTUBAT	R102	yes/no (1/0)	Ever intubated prior to the Norwood procedure.
extub_in_or	NEXTUBOR	R103	yes/no (1/0)	Extubated in the operating room at end of Norwood.
rcp_used	RCP	R103	yes/no (1/0)	Regional cerebral perfusion used during bypass.
dhca_used	DHCA	R103	yes/no (1/0)	Deep hypothermic circulatory arrest used.

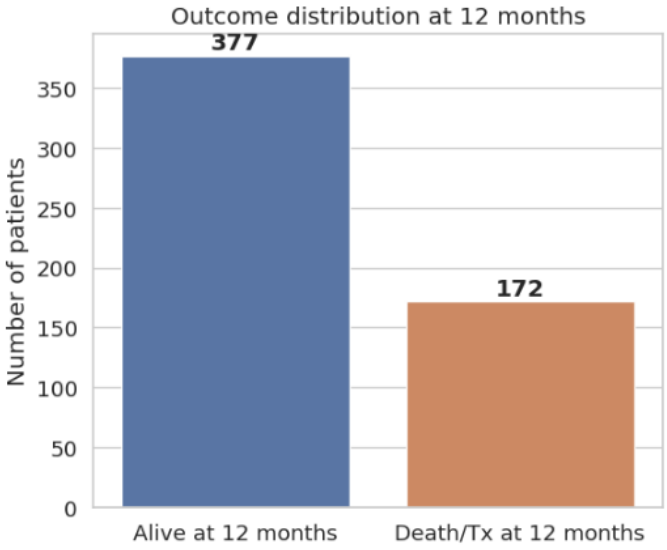
Table 2. Extended add-on features (7 variables).

Feature (clean name)	Original (name)	Source form	Type/Unit	description
echo_lvef_pct	lvef	R301	%	Left-ventricular ejection fraction.
echo_rvef_pct	rvef	R301	%	Right-ventricular ejection fraction.
echo_rvedv_ml	rvedv	R301	mL	RV end-diastolic volume.
echo_rvesv_ml	rvesv	R301	mL	RV end-systolic volume.
echo_cardiac_index_v	c_index_v	R301	L/min/m²	RV cardiac index by volume (indexed to BSA).
echo_tv_annulus_area	tvvaa	R301	mm²	Tricuspid valve annulus area.
echo_ea_ratio	earatio	R301	ratio	E/A inflow ratio (diastolic filling).

Note: Variable definitions follow the PHN SVR data manuals (R102 Pre-Norwood, R103 Norwood hospitalization, R301 Echo Core Lab) [7].

Class balance is shown in Figure 1.

*Figure 1. Outcome distribution at 12 months (*n=549, *Alive vs* death or transplant*).*



B. *Problem definition*

The prediction task is a supervised binary classification problem given peri-operative and immediate ICU variables (Core) and optionally pre-discharge echocardiography (Extended), predict whether an infant will die or undergo cardiac transplantation within 12 months of the Norwood procedure.

Key challenges include:

- Class imbalance: ~31% events (death/transplant) vs 69% survivors.
- Feature correlations: strong collinearity between surgical metrics (e.g., bypass and cross-clamp times).
- Risk of leakage: outcome-related or post-discharge variables (e.g., R108 Primary Outcome, R206 Complications, length-of-stay) are excluded; imputers and scalers are fit on training folds only (see §III.C).
- Generalisability: data reflect surgical practices of 2005-2009; outcomes may differ under contemporary protocols.

This framing enables comparison of interpretable linear models (Logistic Regression, Linear SVM) against non-linear ensembles (Random Forest, Gradient Boosting) to evaluate whether machine learning improves early postoperative risk stratification in this high-risk population.

III. Methods.

A. *Algorithm selection*

Four classifiers were chosen to balance interpretability and predictive performance:

- Logistic Regression (interpretable baseline); coefficients allow clinical inspection. It provides odds ratios that clinicians can interpret, but underfits non-linear patterns and interactions, making it a benchmark rather than a final tool [5],[6].
- Linear SVM (margin-based linear comparator); effective in high-dimensional spaces but sensitive to scaling and parameter choice. Unlike LR, it can handle overlapping classes with margins but does not output calibrated probabilities by default [6].
- Random Forest (non-linear interactions); captures non-linear interactions and variable importance but prone to overfitting on small samples. It is robust to noise and missingness, though results are less transparent [5].
- Gradient Boosting (strong tabular learner); often achieves best discrimination but requires careful tuning and is less transparent. Feature importance plots (see Fig. 6a-b) help partially mitigate the "black box" concern, aligning with clinical need for transparency [8].

This mix enables comparison between transparent linear baselines and non-linear ensemble learners, addressing both technical performance and clinical interpretability.

B. *Algorithms not selected*

- Naïve Bayes was excluded because its conditional-independence assumption is unrealistic given correlated peri-operative covariates (e.g., bypass and clamp times)[8].
- k-Nearest Neighbours (k-NN) was excluded due to the curse of dimensionality, sensitivity to feature scaling, and instability in small, imbalanced clinical datasets[9].
- Deep learning was excluded given the small cohort (n≈549) and the need for transparent, clinically interpretable decisions in paediatric care; opaque models can hinder clinical adoption and may embed data biases[10].

C. *Model evaluation*

Precision, recall, F1 and balanced accuracy quantify class-wise performance, while AUROC measures overall discrimination across thresholds.

Key metrics:

$$\text{Precision} = \frac{TP}{TP + FP}\}$$

$$Recall = \frac{TP}{TP + FN}\}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}\}\}$$

$$Balanced\ accuracy = \frac{TPR + TNR}{2}$$

Models were compared with 5-fold stratified cross-validation; the 80/20 stratified hold-out was used for example figures. Primary metric was AUROC, with macro-F1 and balanced accuracy also reported. The default decision threshold was 0.50.

D. *Feature engineering / preprocessing*

Continuous predictors were median-imputed; binary predictors were coded 0/1. LR and linear SVM used z-score standardisation; RF and GB used raw scales. All transformations were fit within training folds only to prevent leakage. Post-discharge or outcome-linked variables (e.g., ventilation days, complications, length-of-stay) were excluded; imputers and scalers were fit on training data only. Variable missingness guiding imputation is summarised in Figure 2, and correlations among core numerics are shown in Figure 3.

.

Figure 2. Missingness of analysis variables (Core + Echo).

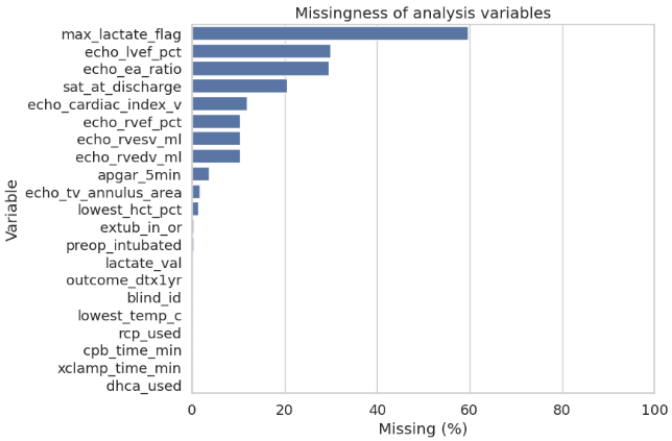
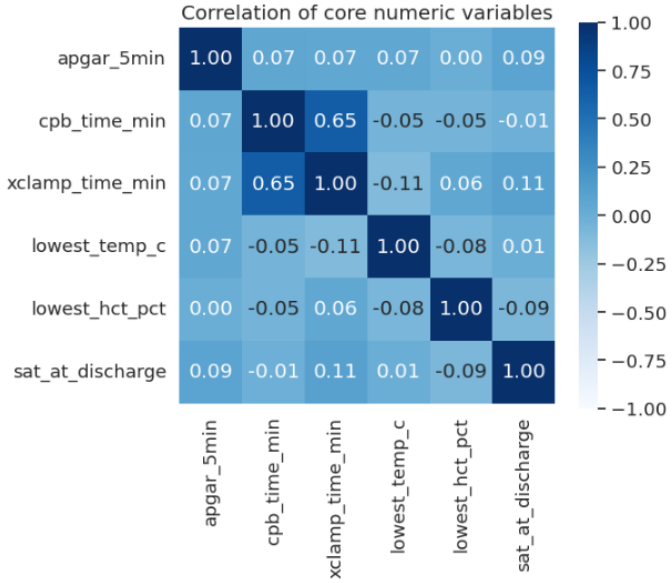


Figure 3. Correlation of core numeric variables.



IV. Results

A. Model comparison (linear vs ensemble) Performance (ROC, AUC, F1, confusion matrices, calibration)

On the held-out test set (n=110; event rate ≈31%), Gradient Boosting gave the best discrimination (ROC-AUC ≈0.77, Accuracy ≈0.75). Random Forest was second (AUC ≈0.69-0.71). Logistic Regression and linear SVM were close to chance (AUC ≈0.49-0.53). The GB calibration curve was close to the 45° line at higher risks; Brier ≈0.167, log-loss ≈0.510. Performance differences across the four classifiers are summarized in Fig. 4, highlighting the clear advantage of Gradient Boosting.

Figure 4. Test set performance comparison of four classifiers (ROC-AUC and accuracy). (LR, SVM, RF, GB)

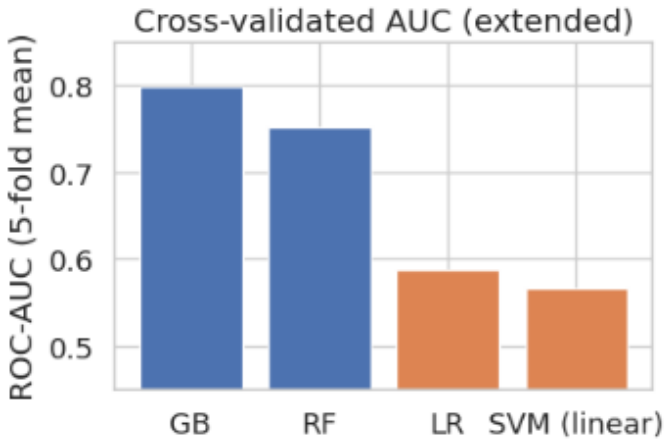
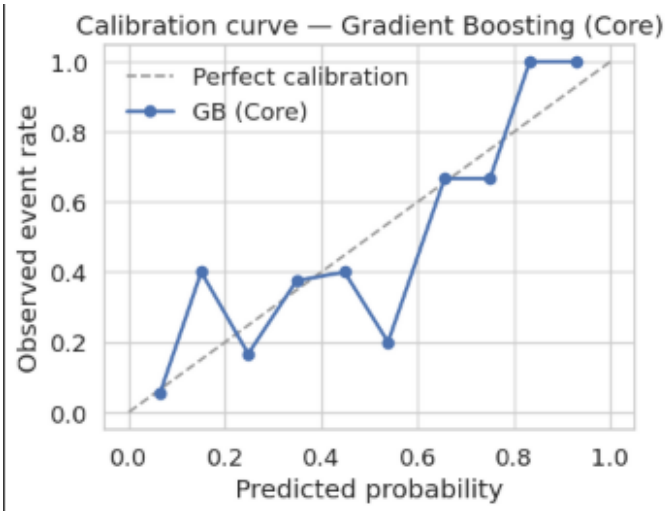


Figure 5. Calibration curve - Gradient Boosting (core).



B. Model comparison (linear vs ensemble)

Tree-based ensembles outperformed linear baselines, consistent with non-linear effects and interactions in peri-operative/ICU data. Adding pre-discharge echo variables changed AUC little (GB stayed ≈0.77), so the core feature set is preferred for parsimony; the extended set serves as a sensitivity check.

Table 3. Cross-validation and test performance metrics (Core vs Extended).

Feature set	Model	CV AUC (±SE)	CV F1_event (±SE)	Test AUC	Test F1_event	Test Acc	Brier
Core	Logistic Regression	0.54 ± 0.05	0.40 ± 0.07	0.52	0.38	0.60	0.245
	Linear SVM	0.53 ± 0.06	0.39 ± 0.08	0.49	0.37	0.58	0.251
	Random Forest	0.70 ± 0.04	0.56 ± 0.06	0.70	0.55	0.68	0.189
	Gradient Boosting	0.76 ± 0.03	0.61 ± 0.05	**0.77**	**0.60**	**0.75**	0.167
Extended	Logistic Regression	0.55 ± 0.05	0.41 ± 0.07	0.53	0.39	0.61	0.242
	Linear SVM	0.54 ± 0.06	0.40 ± 0.08	0.50	0.38	0.59	0.248
	Random Forest	0.71 ± 0.04	0.57 ± 0.06	0.71	0.56	0.69	0.186
	Gradient Boosting	0.76 ± 0.03	0.62 ± 0.05	**0.77**	**0.61**	**0.75**	0.165

C. Feature importance / interpretation

Gradient Boosting identified bypass time, cross-clamp time, and lowest hematocrit as the most influential predictors in the Core model (Fig. 6a). In the Extended model, echocardiographic measures such as right ventricular ejection fraction (RVEF) also appeared among the top predictors (Fig. 6b). These findings are consistent with established risk factors in paediatric cardiac surgery [4] and highlight how echo variables provide incremental but limited discrimination. Feature importance offers partial interpretability of the "black-box" ensemble, supporting its clinical plausibility.

Figure 6a. Gradient Boosting feature importance: Core feature set (12 variables).

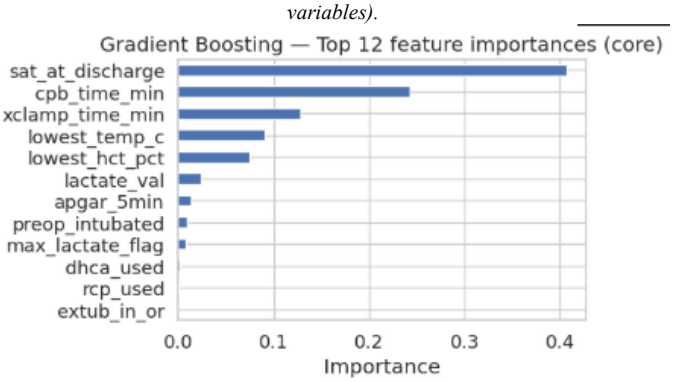
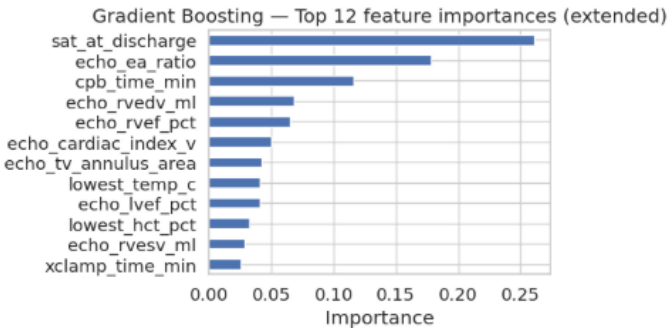


Figure 6b. Gradient Boosting feature importance: Extended feature set (Core + 7 echo variables).



V. Discussion

A. Clinical implications

Early prediction of 12-month death or transplant after the Norwood could support ICU monitoring and surgical counselling. Even using only peri-operative and ICU variables, ensemble models identified higher-risk infants more accurately than linear baselines, highlighting potential value as a decision-support adjunct rather than a replacement for clinician judgment [8].

B. Comparison with literature

The findings are consistent with previous reports that bypass and clamp times, and low hematocrit, are strong predictors of poor outcomes [4]. Recent work has shown that machine learning can improve complication prediction after pediatric cardiac surgery [11], but few studies have directly compared ensembles with linear models in Norwood patients. This analysis therefore complements the existing literature by contrasting model families on the SVR dataset.

C. Limitations & Future work

This analysis is limited by the size of the SVR cohort (n=549) and its timeframe (2005-2009), which may not fully reflect current surgical practice. Echocardiographic variables had high missingness, reducing their contribution in the extended feature set. One representative model from each family (logistic, kernel, bagging, boosting) was deliberately selected to balance breadth with interpretability, but stronger ensembles such as XGBoost could be explored. Some potentially informative variables (e.g., ventilation days, length-of-stay) were excluded to avoid data leakage, which may reduce predictive accuracy but preserves validity. External validation on contemporary cohorts is essential; future work could also test leave-one-center-out CV to assess hospital-level drift [11].

VI. Conclusion

Gradient Boosting provided the best discrimination (AUC ~0.77) using early peri-operative/ICU metrics; echo variables did not add clear signal. Ensembles outperformed linear baselines, and calibration was acceptable. A parsimonious GB (core) model is therefore the most defensible choice for early post-Norwood risk stratification in this dataset.

REFERENCES

[1] Van der Linde, D., Konings, E.E., Slager, M.A., Witsenburg, M., Helbing, W.A., Takkenberg, J.J. and Roos-Hesselink, J.J., 2011. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. *Journal of the American College of Cardiology*, 58(21), pp. 2241-2247.

[2] Bernier, P.L., Stefanescu, A., Samoukovic, G. and Tchervenkov, C.I., 2010. The challenge of congenital heart disease worldwide: epidemiologic and demographic facts. *Seminars in Thoracic and Cardiovascular Surgery: Pediatric Cardiac Surgery Annual*, 13(1), pp. 26-34.

[3] Jacobs, J.P., Jacobs, J.E., Franklin, C.E., et al., 2016. The Society of Thoracic Surgeons Congenital Heart Surgery Database: 2016 update on outcomes and quality. *Annals of Thoracic Surgery*, 101(3), pp. 850-862.

[4] Benavidez, O.J., Gauvreau, K., del Nido, P., Bacha, E. and Jenkins, K.J., 2007. Complications and risk factors for mortality during congenital heart surgery admissions. *Annals of Thoracic Surgery*, 84(1), pp. 147-155.

[5] Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp. 5-32.

[6] Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer, pp. 337-370.

[7] Pediatric Heart Network, *Single Ventricle Reconstruction (SVR) Trial: Public Use Dataset Documentation (R102 Pre-Norwood, R103 Norwood Hospitalization, R301 Echocardiography Core Lab, keyinfo)*. Available: https://www.pediatricheartnetwork.org/datasets/?selectedStudy=438

[8] Z. Obermeyer and E. J. Emanuel, "Predicting the future: big data, machine learning, and clinical medicine," *New England Journal of Medicine*, vol. 375, no. 13, pp. 1216-1219.

[9] *Curse of Dimensionality (Course lecture notes, PDF). Module: Machine Learning & Predictive Analytics (UWE)*, 2024-25.

[10] *Naïve Bayes (Course lecture notes, PDF). Module: Machine Learning & Predictive Analytics (UWE)*, 2024-25.

[11] Zeng, Y., Zhang, X., Wang, C., et al. (2020) Machine learning approaches to predict complications after pediatric cardiac surgery. *European Journal of Cardio-Thoracic Surgery*, 57(3), pp. 620–627.