Iowa Initiative for Artificial Intelligence
Final Report

Project title:	Machine Learning Model to Predict Extubation Success in Neonates
Principal Investigator:	Lindsey Knake MD, MS, Danielle Rios MD, Patrick McNamara MD, James Blum MD
Prepared by (IIAI):	Avinash Mudireddy
Other investigators:	IIAI analytics team and NICU clinical collaborators
Date:	04/07/2026

Were specific aims fulfilled:	Aim 1 was completed and the others are current in process
Readiness for extramural proposal?	Yes.
If yes … Planned submission date	Submitted October 2025
Funding agency	NIH/NHLBI
Grant mechanism	K23
If no … Why not? What went wrong?	

Brief summary of accomplished results:

This final report extends the earlier extubation prediction proposal into a complete three-phase study. Phase 1 established the strongest benchmark with engineered physiologic plus clinical features using Logistic Regression (Test AUC 0.875). Phase 2 evaluated deep representation learning from time-series signals and reached AUC 0.735 (5-fold CV) and 0.703 (held-out split). Phase 3 introduced feature-wise low-pass filtering and hybrid classical modeling, improving to AUC 0.7848. Overall, the best final-performing and most interpretable approach on this dataset remains the Phase 1 engineered-feature model.

Research report:

Aims (provided by PI):

Note: The original proposal aims remain valid for the final report. Extubation timing in preterm neonates is clinically high-risk, and prediction support is needed to reduce both failed and delayed extubation harms. The foundational aims were to build a predictive model from clinical plus high-resolution physiologic data, evaluate generalizability under lower-resolution settings, and assess subgroup effects. As the project progressed, these aims were expanded into a three-phase program: classical engineered-feature benchmarking, deep representation learning, and hybrid low-pass-filtering pipelines.

Here are the actual aims provided by the PI at the beginning of Project:

Mechanical ventilation in the Neonatal Intensive Care Unit (NICU) is a life-saving therapy. However, prolonged mechanical ventilation in preterm infants is associated with increased mortality, neurodevelopmental impairment, structural changes in the central nervous system, and bronchopulmonary dysplasia (BPD). Additionally, extubation failure is also associated with an increased risk of death, BPD and prolonged time on mechanical ventilation. Frequently, clinicians are weighing the risks between deciding to extubate too early and risk severe clinical decompensation that may result in brain or pulmonary hemorrhage versus waiting too long to extubate and risk lung damage or maldevelopment. Either decision may result in life-long morbidities. Thus, creating prediction models to help determine the appropriate timing for a trial of extubation in neonates is critically needed.

Previous models have attempted to predict extubation success in neonates without successful clinical adoption. Logistic regression and machine learning models have been created using static clinical and ventilator variables (with low resolution) to predict extubation success. However, there has been limited success with these models when evaluated at external institutions. The Heart Rate Characteristics index (HRCi) model has combined both static and dynamic variables in the model to predict extubation readiness. The HRCi model evaluates continuous heart rate wave form data and patient clinical characteristics but has not incorporated dynamic ventilator data into the model. We aim to incorporate high resolution ventilator data into our predictive model to enhance the accuracy and clinical utility of the model. We hypothesize that using this model to aid in the clinical decision of the best timing for a trial of extubation will result in more successful extubations with decreased morbidity in neonates.

Specific Aims:

1. To develop a predictive model for extubation success in neonates on mechanical ventilation with the primary outcome defined as the probability the patient will remain extubated for at least five days. The input variables would include static data such as laboratory data (blood gases) and patient factors (gestational age, birth weight, day of life, and weight at extubation) and high resolution dynamic clinical data (heart rate, mean airway pressure, ventilator measured lung compliance, oxygen trends, and patient desaturations).
2. To evaluate the algorithm developed in Aim 1 and for generalizability by determining performance characteristics based on hourly ventilator data that is extracted from the electronic health record (EHR).
3. To perform a subgroup analysis on different modes of mechanical ventilation or different post-extubation respiratory support to determine if these subgroups will have different predictors for extubation success.

Data:

We will plan to utilize SickbayTM (Medical Informatics Corp. Houston, TX) for automated data capture and real-time data analysis which is already implemented into the University of Iowa NICU. This cluster-based platform interfaces with all patient-monitoring devices in the NICU and automatically records physiologic data without intervention or configuration by research or clinical personnel. It passively collects physiologic monitor data at exceptionally high resolution. These data include waveforms, vital signs, alarms, and alarm settings. SickbayTM also provides research interfaces for self-service analytics across subject cohorts, which have been used in a variety of research projects analyzing pediatric physiologic data and has been used to support data recording and analysis in large-scale NIH-funded research projects.

SickbayTM captures and stores both static and dynamic ventilator data parameters at 0.25 second intervals while the patient is on the ventilator and vital sign monitors. The implementation of SickbayTM at the University of Iowa, has over 520 patients with continuous ventilator data collected and stored for research purposes.

The data framework from the prior proposal was preserved and expanded. Sources included cohort labels, high-frequency physiologic streams, ventilator summaries, and curated clinical variables. Core files included ga_36week_lable_df.pickle, capsule_df.pickle, and Combined_merit_Extubation_failure_key_Final_SBID_avi_upload.xls. Signals included heart rate, oxygen saturation, and respiratory rate over 24-hour windows. Final sample sizes were n=160 in Phase 1, n=145 in Phase 2, and n=144 in Phase 3. Preprocessing included missingness filtering, interpolation/imputation, scaling, and class-imbalance controls.

AI/ML Approach:

Input data:

Clinical data- This data includes the patient(neonate) clinical parameters like weight, time, procedures, status, given medications, lab results. Philips and capsule data- These datasets are timeseries of patient's vitals like ECG and ventilator readings.

The timeseries data is converted to the (Batch, #of seconds of history in the past, features)

The non-timeseries data is converted to (Batch, features)

Phase 1 (engineered-feature classical modeling): Physiologic summaries (min, max, median, variance), intermittent hypoxia count, RR category features, and clinical covariates were combined. Models included Logistic Regression, Random Forest, XGBoost, and Gaussian Naive Bayes with RandomizedSearchCV, SMOTE balancing, threshold optimization, and SHAP interpretation.

Phase 2 (time-series representation learning): Signals were segmented into fixed-length 1 Hz blocks with missingness thresholds and interpolation strategy. ConvAE, U-Net AE, and Masked AE variants were evaluated. The enhanced downstream model used U-Net latent sequences fused with static covariates via Conv1D plus dense branches, dropout, class weighting, and early stopping.

Phase 3 (hybrid LPF plus classical): Feature-wise low-pass filtering (Butterworth and Savitzky-Golay) was applied before statistical flattening (mean, std, min, max, IQR, slope). Tuned classical models were re-evaluated under unified 5-fold CV with out-of-fold probability aggregation, threshold analysis, and SHAP.

This progression was designed to test whether added temporal representation depth and preprocessing complexity could outperform the interpretable baseline.

Experimental methods, validation approach:

Phase 1 used stratified hold-out evaluation with internal cross-validation and optimized thresholding for failure-class F1; an additional 5-fold subgroup CV analysis was run for GA <28 weeks.

Phase 2 used both full-cohort 5-fold cross-validation and an 80/20 stratified split with inner 5-fold CV on the training set.

Phase 3 used a consistent 5-fold cross-validation pipeline with out-of-fold aggregation under each LPF configuration. Across phases, outputs included AUROC, threshold-specific confusion matrices, PPV/NPV, and SHAP-based interpretation.

Results:

Phase 1 remained the top benchmark. Logistic Regression achieved Test AUC 0.875 and failure-class F1 0.545 at optimized threshold. Cross-validated Random Forest for GA <28 weeks yielded AUROC 0.848.

Phase 2 deep-learning representations did not exceed this benchmark. U-Net embeddings plus the enhanced CNN head achieved AUC 0.735 in 5-fold CV and 0.703 on held-out split.

Phase 3 hybrid LPF modeling improved over Phase 2. The best Savitzky-Golay plus XGBoost configuration reached out-of-fold AUC 0.7848.

Overall, advanced pipelines improved methodological understanding and narrowed the gap, but engineered-feature classical modeling remained strongest on current data volume.

Phase 1 primary benchmark visuals:

Figure 1. Logistic Regression ROC curve (AUC 0.875).

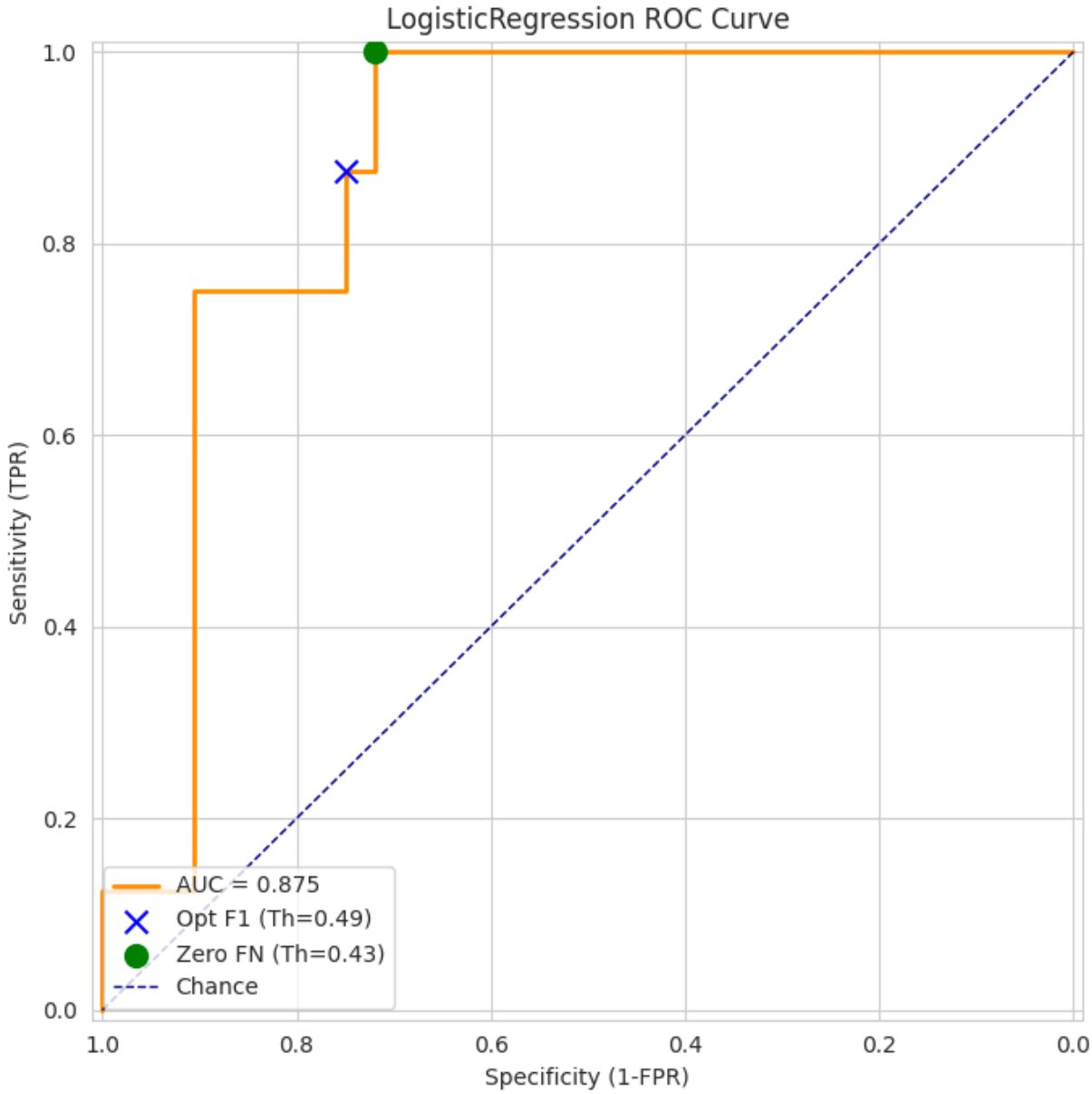


Figure 2. Logistic Regression confusion matrix at optimized threshold.

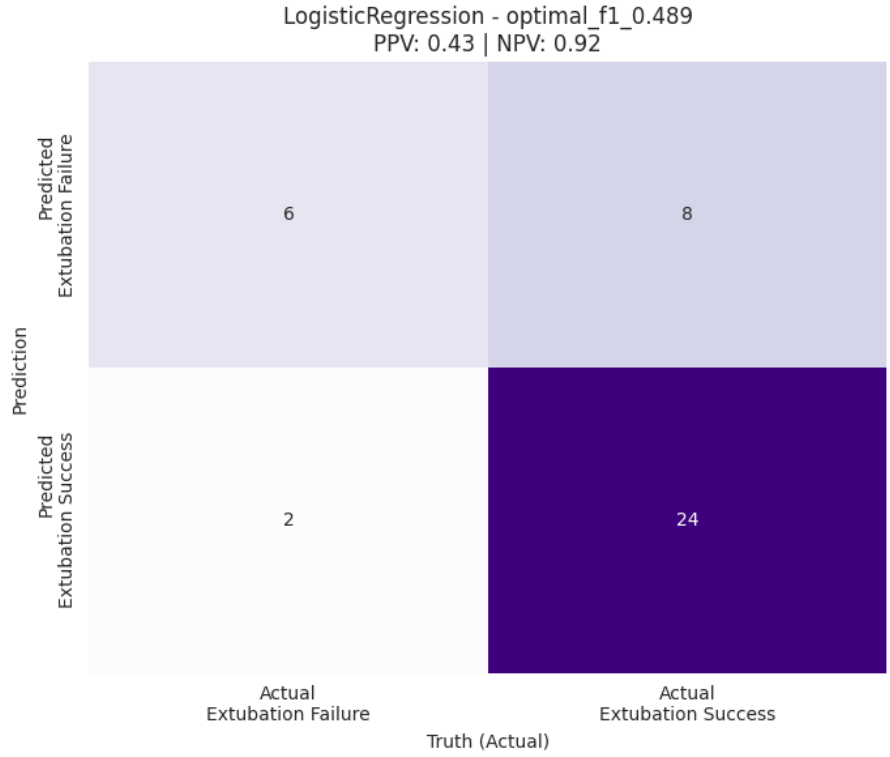


Figure 3. SHAP summary (dot) for Logistic Regression.

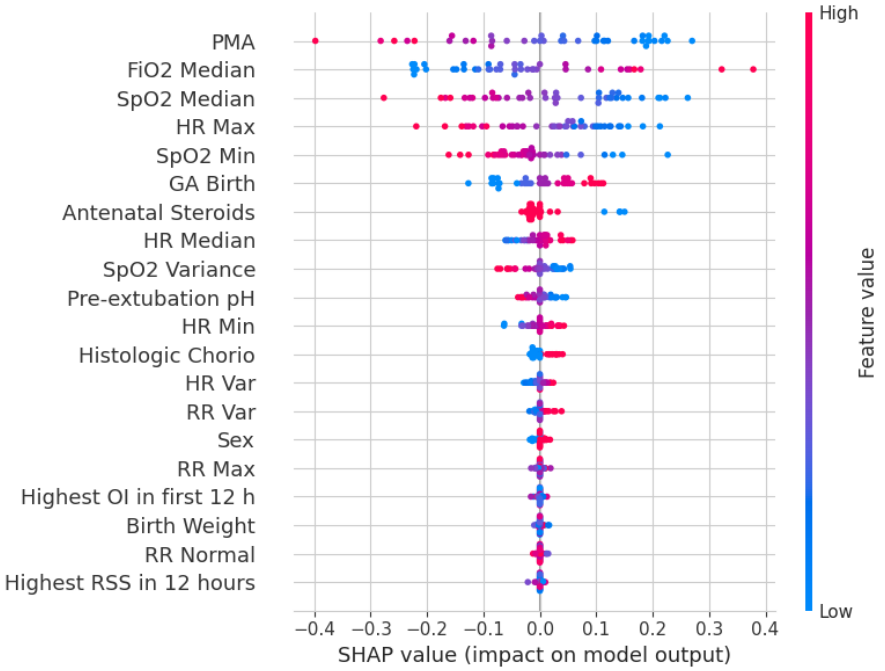
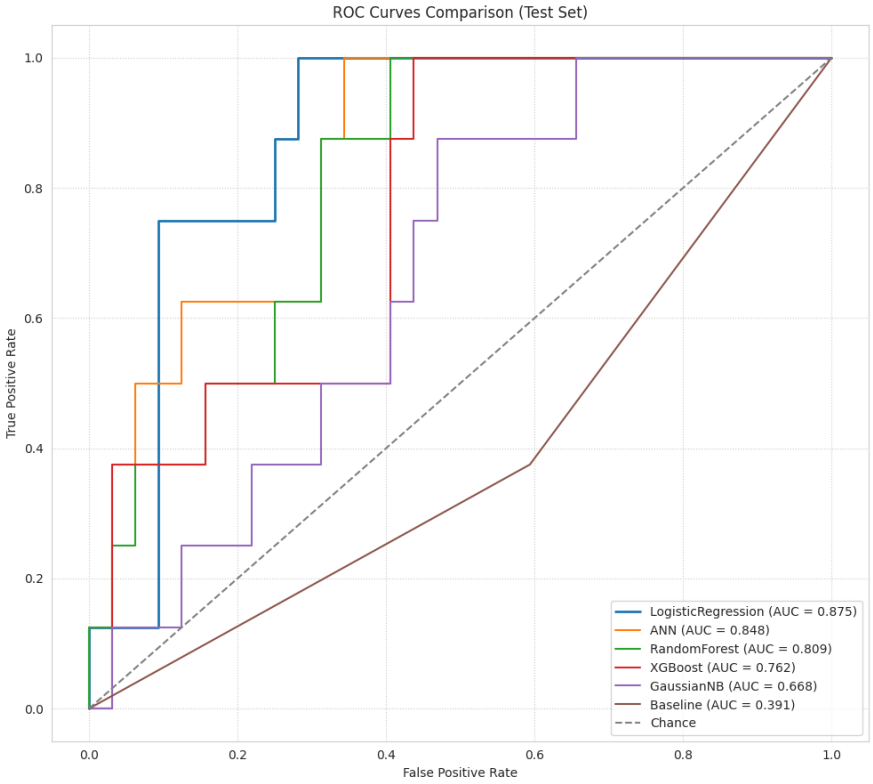


Figure 4. Combined ROC comparison across Phase 1 models.



Phase 2 representation-learning visuals:

Figure 6. Experiment 1 (5-fold CV) AUROC.

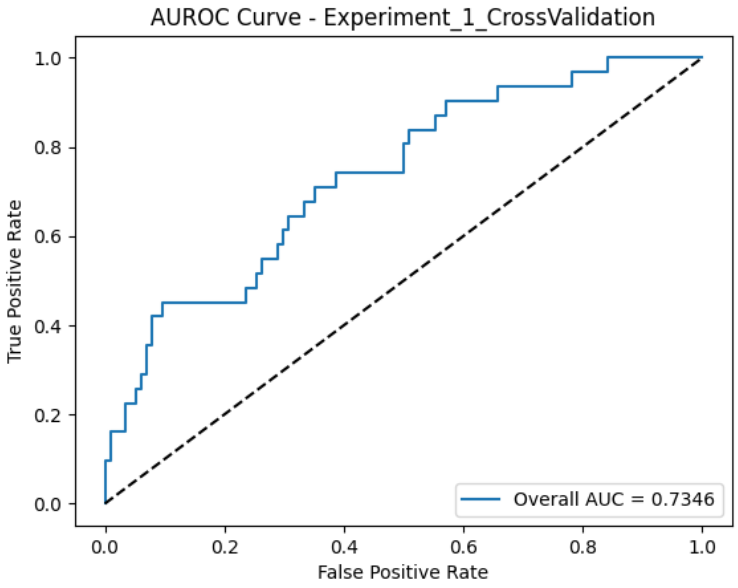


Figure 7. Experiment 1 confusion matrix at threshold 0.5.

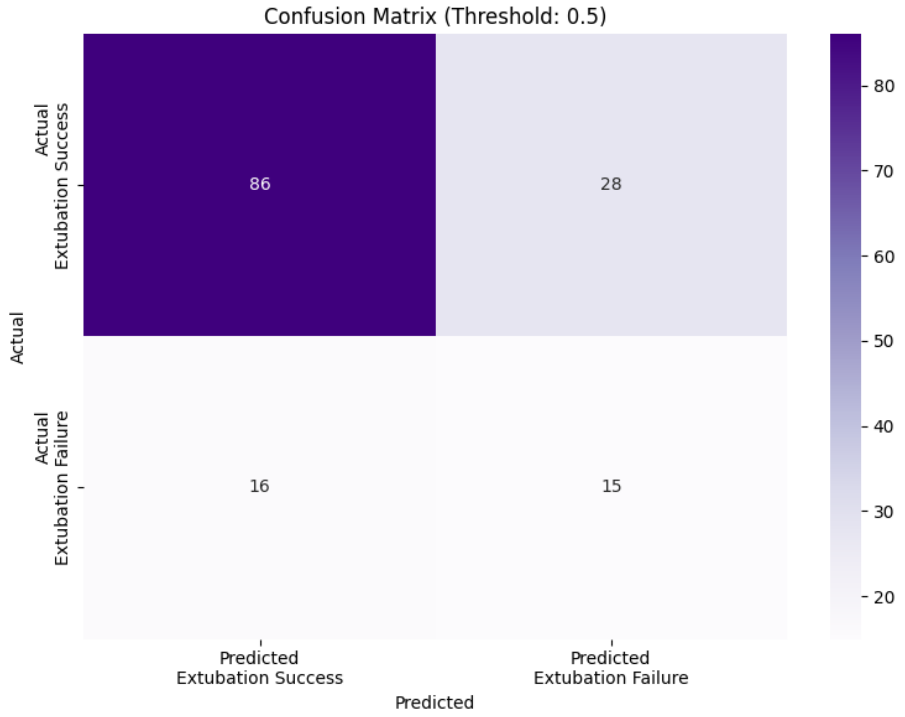


Figure 8. Experiment 2 (train/test split) AUROC.

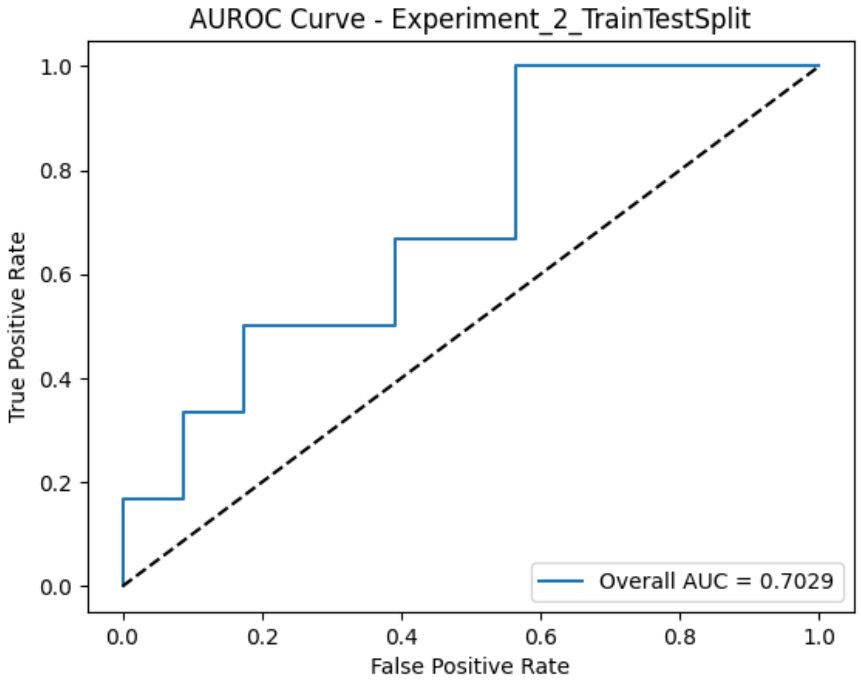
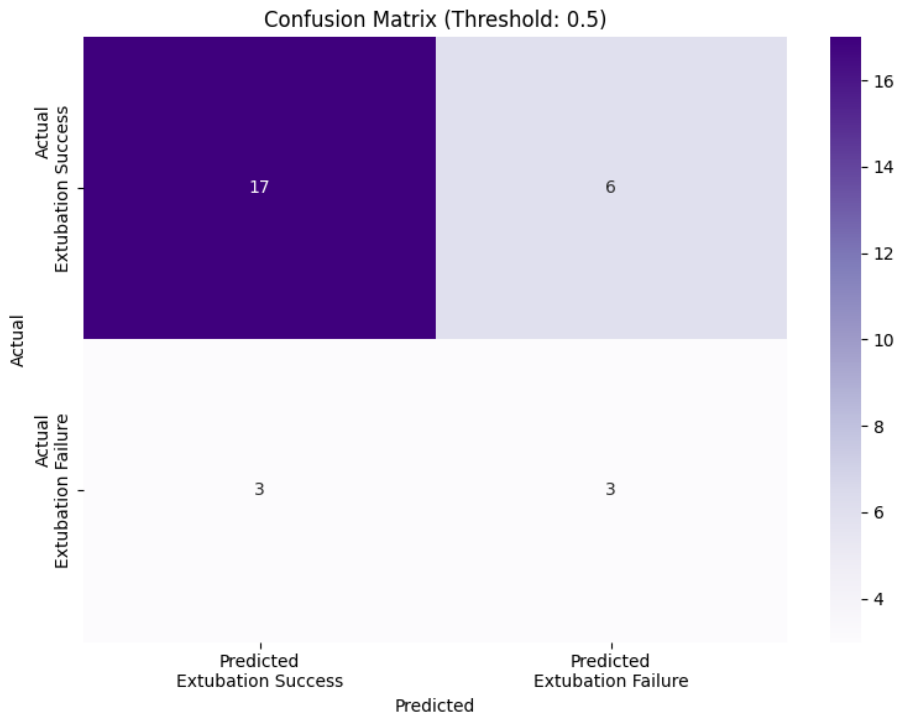


Figure 9. Experiment 2 confusion matrix at threshold 0.5.



Phase 3 hybrid LPF visuals:

Figure 10. Savitzky-Golay plus XGBoost AUROC.

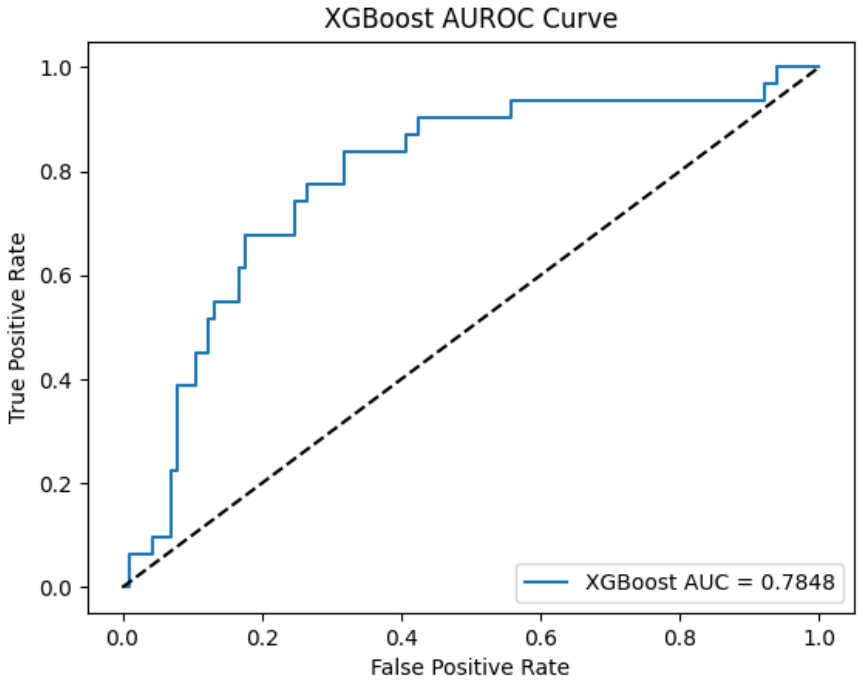
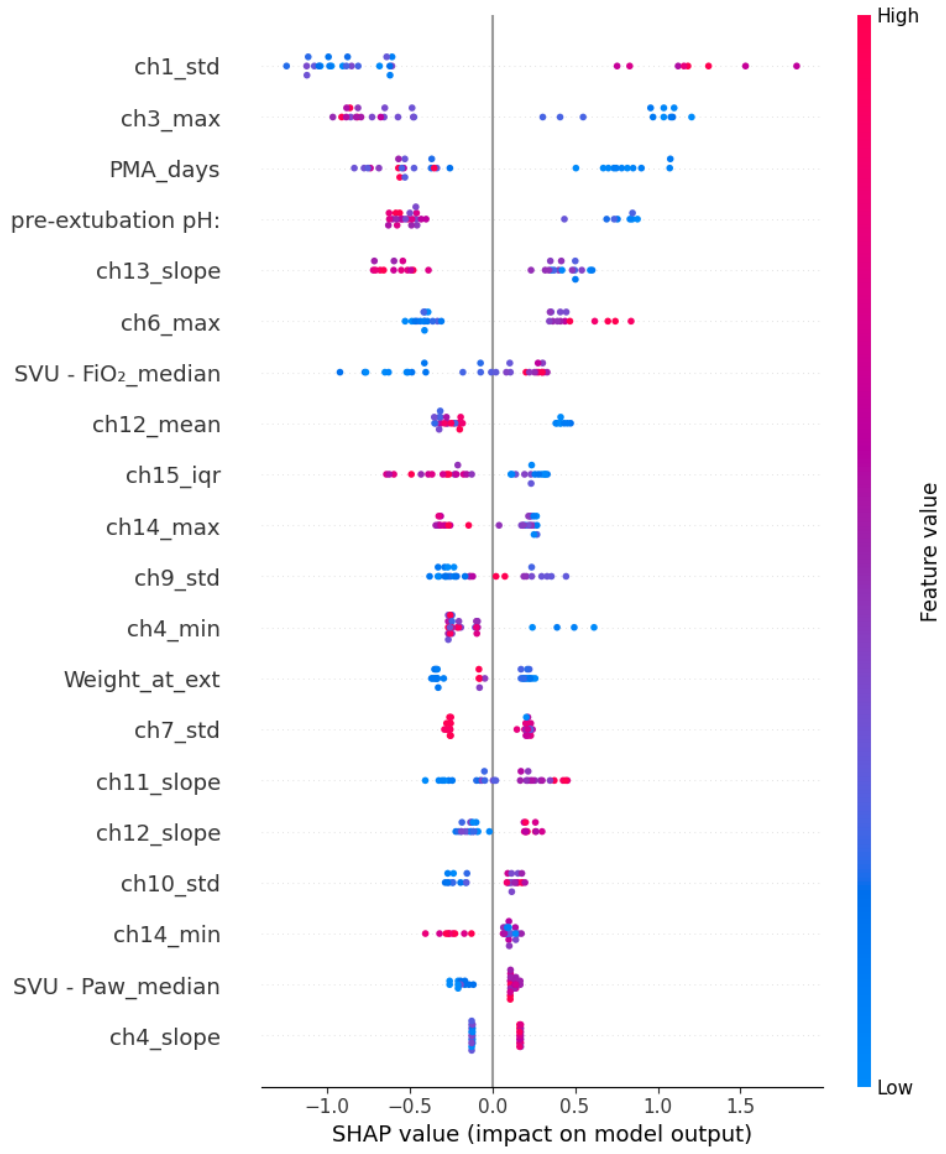


Figure 11. SHAP summary for best Phase 3 configuration.



Ideas/aims for future extramural project:

The preliminary data developed for Aim 1 has been used for a K23 NIH grant submission with the goal for expanding this study to a multicenter study to increase overall training data size and validate the algorithm at external sites.

The deep learning and hybrid ML approaches showed promise and will likely perform better with a larger database. Future work will include evaluating both the Phase 1 and Phase 3 methodologies outlined in this pipeline. The enhanced multicenter sample size that will be provided by the K23 funding will lead to a future R01 funding to test the implementation and clinical impact of a bedside clinical prediction tool. The goal would be to develop a transparent bedside decision-support workflow with clinician-guided threshold governance and periodic recalibration of the ML model.