Data-Driven Prediction of Ventilation Duration in ICU Patients

Harold Haugen
School of Data Science
The University of Virginia
waa4bq@virginia.edu

Alec Pixton
School of Data Science
The University of Virginia
etk3pu@virginia.edu

Clarissa Benitez
School of Data Science
The University of Virginia
ycv3fh@virginia.edu

Emmanuel Leonce
School of Data Science
The University of Virginia
fyb7sx@virginia.edu

DS5003 Healthcare Data Science - Team 2

Target Audience

The target audience for this project proposal may include the Chief Operating Officer (COO), Director of Operations, Patient Flow Coordinator, and Nursing Supervisors. In addition, clinicians such as pulmonologists and respiratory therapists, who play a role in determining whether alternatives to invasive ventilation can be used are also important stakeholders. Leaders whose responsibilities involve hospital-wide resource allocation and logistics, particularly in areas such as bed capacity, medical equipment, and facility usage, as well as real-time monitoring of bed availability and staffing alignment during shifts, may find this proposal especially relevant and valuable.

Executive Summary

This report explores the use of machine learning to predict the length of time an ICU patient will require mechanical ventilation. Accurate predictions can assist ICU departments in allocating resources more efficiently, ultimately improving patient care.

A wide range of features were used which centered around the recorded data from the first day of a patient's stay in the ICU, which includes demographics, vital signs, and laboratory results. These were used to train several classification models: logistic regression, random forest, XGBoost, and a neural network. Model performance was evaluated using metrics including accuracy, precision, and recall.

Given that our primary performance metric was **recall** for the positive class, defined as cases where ventilation duration exceeded one day, the Random Forest classifier achieved the highest recall at 0.819, indicating strong performance in identifying patients likely to require extended ventilation. XGBoost, another ensemble-based tree model, also performed well across multiple metrics, including ROC-AUC, specificity for the negative class, and overall accuracy. Feature importance analysis using SHAP values highlighted several key laboratory and vital sign features recorded at ICU admission, most notably *partial oxygen pressure (PO_2)*, *mean airway pressure (MAP)*, and *heart rate*, as significant predictors of ventilation duration.

These results suggest that machine learning models can effectively classify patients into short-term or long-term ventilation categories. With further refinement and validation, such models could be integrated into clinical workflows to support ICU triage and resource planning.

The current model has limitations, including a relatively small sample size and the absence of time-series data. Future work should prioritize external validation, the incorporation of longitudinal data, and evaluation in real-world clinical settings.

I. Introduction

Ventilation is a primary hospital ICU procedure when patients are dealing with a compromised respiratory function. As noted by the American Association for the Surgery of Trauma, "Mechanical ventilation is one of the most common interventions implemented in the intensive care unit. More than half of the patients in the ICU are ventilated the first 24 hours after ICU admission; comprised of individuals who have acute respiratory failure, compromised lung function, difficulty in breathing, or failure to protect their airway." [2]. Consequently, when regional and global emergencies occur such as the case in 2020 with COVID-19, ICU's were faced with a supply challenge where the supply of ventilators for ICU patients could not keep pace with demand. Nature published a study in 2025 to assess the true demand for intubation-ventilator use and found that from "4,055,462 hospital admissions with a diagnosis of coronavirus disease 2019 (COVID-19) from April 2020 to December 2021", "a total of 489,390 (12.1%) patients experienced endotracheal intubation and mechanical ventilation, with the highest peak in August 2021 (48,735), followed by January 2021 (47,100) and December 2021 (43,835) [3]. The article highlighted that estimates of ventilator supply ranged from "60,000 to 160,000 in the United States in the early pandemic period" [3] and that hospitals most likely experienced supply shortages during periods of peak demand.

II. Project Proposal and Motivation

The purpose of this project is to apply statistical machine learning approaches to predict the estimated duration of mechanical ventilation required by ICU patients. To achieve this, we will utilize a combination of ICU data including procedure events, laboratory results, admission records, and charted observations, to train and evaluate several predictive models. Our selected models will be used to predict either ventilation classifications for duration or continuous time on ventilation, based on early clinical indicators. By providing these predictions, we hope this will enable both clinicians and hospital operations staff to better identify patients with more severe respiratory conditions who may require prolonged ventilation and additional resources. These insights may also inform decisions regarding the use of alternative approaches, such as non-invasive respiratory support for patients with less severe cases, thereby reducing the dependence on higher demand ventilation systems.

From our exploratory data analysis (EDA), we observed a skewed but wide distribution in patient ventilation time, ranging from as short as a few hours, to 1 day (1,440 minutes), to over 10 days (15,000 minutes) including outliers extending beyond 17 days (over 25,000 minutes). See Fig. 1.

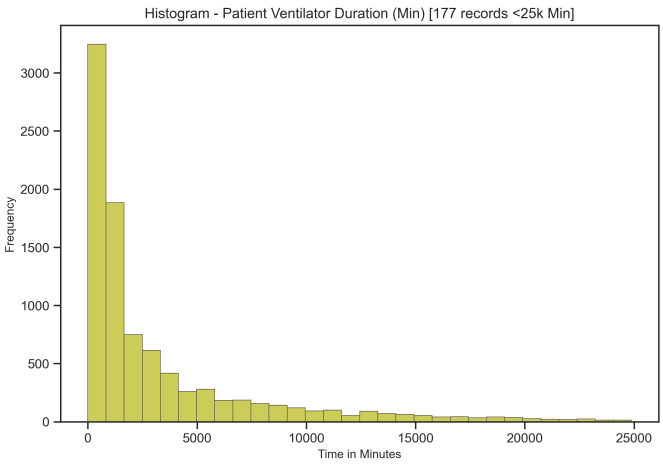

Fig. 1: Histogram of Ventilation Duration [$< 25k$ minutes]

III. Methods

A. Data Collection and Processing

To address our objective, we acquired hospital and patient-level data to identify individuals who underwent intubation and ventilation procedures. We gathered relevant biometric, diagnostic, treatment, laboratory, and charted clinical data, which we used to identify target ventilation events and perform exploratory analysis on potential independent variables.

1) Data Sources: Our team utilized the Mimic-III Clinical Database which contains data from the intensive care unit (ICU) of Beth Israel Deaconess Medical Center between 2001 to 2012. This database is comprised of data that has been de-identified in accordance with HIPPA using data cleaning and date shifting. Dates were offset at random intervals but maintained timing, so events appear from 2100 and 2200. For patients over 89, birth dates were adjusted to protect privacy. The database contains structured and linked tables that cover patient demographics, vital signs, lab results, procedures, medications and mortality outcomes. We selected this database for its public access, large and diverse population and provides detailed information over time making it well-suited for this research. [1].

2) Model-Ready Data Structuring: Developing a centralized data model was key to aligning patient and medical information with the ventilation events representing our Y outcome (ventilation duration). To build the final model-ready dataset, we performed the following:

- 2*a* Examined the `D_Items` table and its *Category* and *Label* fields to identify ItemIDs for intubation and ventilation events. We found that the `"2-Ventilation"` category, specifically the `"Invasive Ventilation"` and `"Non-invasive Ventilation"` labels, mapped directly to relevant records in the `ProcedureEvents` table. Using these IDs as pimrary keys, we extracted all ventilation records, generating a source table with approximately ∼12,000 entries.
- 2*b* Performing data cleaning using assumptions 3a and 3b below, we refined the new ventilation table to only include one unique ventilation event for each `SUBJECT_ID`, `HADM_ID` and `ICU_ID` combination. These identifiers represent the patient, hospital admission, and ICU stay, respectively. This resulted in a table including ˜9.3k records.

Table	Rows	Columns	Overview
PROCEDURE EVENTS_MV	258,066	25	Patient procedures such as mechanical ventilation.
ADMISSIONS	58,976	19	Admissions data including timing, demographic info and source of admission.
PATIENTS	46,520	8	Defines each unique patient (defines the SUBJECT_ID).
LABEVENTS	27.8m	9	Contains laboratory results for patients in hospital / out patient clinics.
CHARTEVENTS	330.7m	15	Contains all charted events or observations including vitals and lab results for patients.
D_ITEMS	12,487	10	Definitions of ICU database items by ITEMID, spanning Procedure and Chart Events.
D_LABITEMS	753	6	Definitions for all ITEMID associated with lab measurements in MIMIC-III.
ICUSTAYS	61,532	12	Defines each ICUSTAY_ID in the database.

TABLE I: MIMIC-III Data Sets

2c Using SUBJECT_ID and HADM_ID pairs as a new primary key, Patient, Admission, Lab and Chart events were pulled from their original tables to build a unique ventilation data subset for eventual consolidation. The Lab and Chart data sets were pivoted so that the independent features were transitioned into a wide columnar format representing one column per ItemId record.

2d Sub-tables from 2c were merged using Subject_ID and HADM_ID, producing a final table of 8,392 ventilation events with 73 features.

3) Data and Model Assumptions: To ensure a consistent dataset structure, we defined certain assumptions for the target variable (Y) and the identification and extraction of relevant independent variables.

3a For identified ventilation events, since each patient may have multiple hospital admissions (HADM_ID) and potentially multiple ICU stays per admission, we defined a unique ventilation event as a combination of SUBJECT_ID and HADM_ID. For each HADM_ID with multiple ventilation events, we retained the event that (a) occurred on the first hospital day and (b) had the longest duration. This approach addressed cases where events shared identical start times but had varying durations, some of which were implausibly short. We also excluded SUBJECT_ID/HADM_ID pairs with more than five events, resulting in the removal of 72 records.

3b Given our objective to predict ventilation duration, specifically distinguishing between events from onset to one day, then events lasting over one day, we isolated the first occurrence of key medical events from the laboratory and chart records for each HADM_ID. Since patients often had multiple medical entries for certain ID types recorded at varying time intervals, this method ensured consistency and also reduced the number of missing values in the initial dataset.

3c For Laboratory and Chart Events, we selected Item_ID entries where the tables contained data for at least 8,200 and 8,450 patient respectively, from the 8,516 total unique patients. These thresholds were chosen to reduce the number of null values and to minimize the number of features.

4) Feature Engineering and Transformation: Certain features and temporary variables were created during wrangling to enhance model performance.

4a Intermediate variables such as Start_Day and HADM_Count were created from ventilation records. Start_Day grouped sameday ventilation events to identify the longest duration on the first day of admission. A group-by function was also used to count (e.g., HADM_Count) the number of ventilation events per HADM_ID for data cleaning.

4b Resp_Diag_Label is a feature reduction variable that condensed 4,568 unique Diagnosis entries into a binary indicator of respiratory-related diagnoses (1 = yes, 0 = no).

4c Age_Admission To account for MIMIC-III's date-shifting for patients over 89, we assigned an age of 91.4 at admission for Subject_ID/HADM pairs with calculated ages above 89.

4d Transformation included scaling numeric and one-hot encoding categorical variables, resulting in a model ready table of size 8,392 X 246.

5) Exploratory Data Analysis (EDA): We began with exploratory data analysis (EDA) of the MIMIC-III dataset to identify tables and variables relevant to our objective. After refining our proposal, we analyzed admission and patient-level data to determine whether ventilation duration was best modeled as a regression or classification task, and whether features showed meaningful variation and/or correlation with the target.

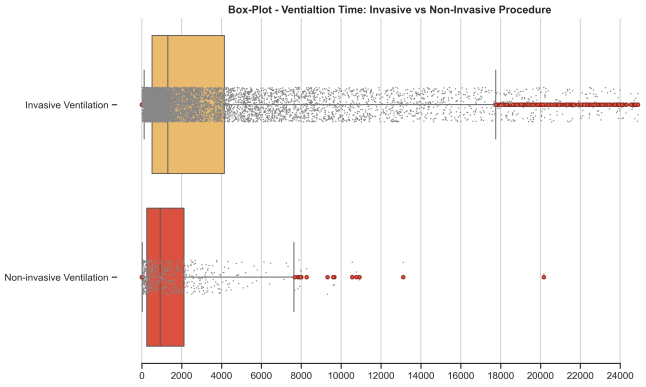

Fig. 2: Box-Plot - Ventilation Duration by Invasive Label

We observed differences in ventilation durations between Invasive and Non-Invasive labels. We were initially concerned that this may introduce target leakage to the model, given Non-Invasive cases appeared shorter and more condensed. However, further analysis indicated that a) Non-Invasive cases represented only 5.4% of the data and b) the 60th percentile duration (1,460 min) was similar to the Invasive medium level (1486 min). To further determine impact though, we modeled both the full and Invasive only subsets.

Secondly, we examined several Lab and Chart events to identify linear relationships that could support our prediction hypothesis. As shown in Fig. 3, we analyzed PO_2 values, measuring the partial pressure of oxygen in arterial blood, an indicator of how well oxygen moves from the lungs into the bloodstream, across key quartiles of ventilation duration, which revealed notable shifts as duration increased.

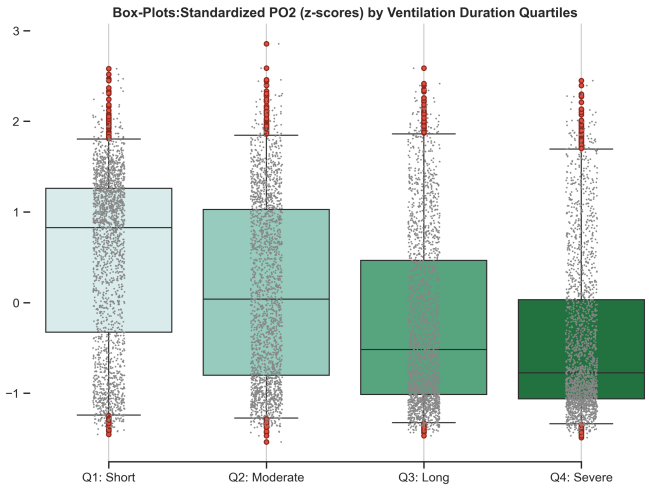

Fig. 3: Box-Plot - PO2 Standardized Results across Ventilation Duration Quartiles

B. Approach

We considered several factors when selecting the modeling approach:

- A binary classification model was selected to predict ventilation duration, using a 1-day cutoff as the threshold. This split provided a balanced dataset (approximately 50/50), which is favorable for machine learning model performance and reduces the risk of bias caused by class imbalance. Although a 3-day cutoff was considered, it resulted in a more imbalanced distribution. The binary setup simplifies the classification problem while retaining clinical relevance and statistical reliability.
- To evaluate model performance, we used two datasets: the full dataset and an Invasive subset. Initially, we considered splitting the data into two subsets, one for invasive and one for non-invasive ventilation, to compare model performance across groups. However, the non-invasive subset consisted almost entirely of patients with ventilation durations under one day, representing less severe cases. Including this group would introduce a strong label imbalance and potentially skew the model, making it less effective at predicting the longer-duration cases that are of primary interest. As a result, we focused our modeling on the full dataset and the invasive-only subset to better assess performance in predicting more severe, longer-duration outcomes. Fig.2.

C. Model Methodology

Our team tested several statistical models to determine which method provided the best predictive performance for estimating ventilation classification, including logistic regression, neural networks, random forest, and XGBoost. Most models were evaluated using an 80/20 train/test split and assessed based on key performance metrics such as ROC-AUC, accuracy, recall, sensitivity, and precision.

1) XGBoost: XGBoost is a gradient boosting algorithm based on an ensemble of decision trees to optimize predictive accuracy. Each tree corrects errors from the prior ones, allowing the model to capture complex patterns in the data. Unlike Random Forests that utilizes many independent trees, XGBoost optimizes through boosting, regularization, and weighted learning that may allow for capturing complex patterns and yielding higher accuracy on structured data.

We conducted a manual grid search to tune XGBoost hyperparameters for optimal AUC, Recall, etc. Parameters included number of trees `n_estimators` (800, 1000), `max_depth` (10, 13, 16), `learning_rate` (0.1, 0.01), and `gamma` (0.0, 0.5). We also tuned L1 LAsso (`reg_alpha`: 0, 0.5, 1.0) and L2 (`reg_lambda`: 0, 0.5) regularization to assess generalization.

2) Random Forests: A Random Forest model was selected for its ability to handle high-dimensional data, mitigate overfitting, and provide feature importance scores. The model was first trained using Scikit-learn's default parameters to establish baseline performance, assessed via accuracy, precision, recall, and F1-score. Feature importance scores identified the most influential predictors. To evaluate if reducing features could enhance performance, subsets of top features (in increments of 5) were used to retrain the model. Accuracy peaked around the top 55 features for both datasets, though these sets likely differed and were trained separately thereafter. Hyperparameter tuning was then conducted using RandomizedSearchCV over `n_estimators`, `max_depth`, `min_samples_split`, `min_samples_leaf`, and `max_features`, with cross-validation to find optimal parameters. This improved accuracy while keeping computational costs manageable.

3) Logistic Regression: The Mlxtend library was used to create logistic regression models using sequential feature selection. L1, L2, and Elasticnet regularization were all attempted with no noticeable improvement to the models. The basic logistic regression with all variables performed just as well as using feature selection with comparable performance to other types of models showing that easy to make classic models can still be practical.

4) Neural Networks: The predictive model was a fully connected, feedforward neural network designed for a binary classification task. The primary goal was to classify a patient's mechanical ventilation duration into one of two categories based on clinical data from the first 24 hours of their ICU stay.

The network architecture consisted of an input layer followed by two hidden Dense layers containing 128 and 64 neurons, respectively. The Rectified Linear Unit (ReLU) was used as the activation function for these hidden layers to introduce non-linearity. To mitigate overfitting, a Dropout layer with a rate of 0.2 was placed after the second hidden layer. The final output layer was a Dense layer with a softmax activation function, which produced a probability distribution across the two target classes. The model was compiled using the Adam optimizer and the categorical crossentropy loss function, which is standard for multiclass classification tasks. Training was monitored for validation loss, and an Early Stopping callback was implemented to halt the process and restore the best-performing weights, further preventing overfitting.

IV. Results

Our primary objective was to develop a model that accurately predicts which patients will require ventilation for more than one day. We prioritized recall as the key evaluation metric, emphasizing the model's ability to identify the majority of patients needing prolonged support (e.g., our positive class).

As noted in the EDA section above, non-invasive ventilation events were typically shorter than invasive ones. To account for this, we evaluated models on both the full dataset and an invasive-only subset, where performance differences were minimal. As shown in Table 1, recall and accuracy remained relatively stable across XGBoost, Logistic Regression, and Random Forest; however, recall declined more noticeably in the latter. Ultimately, Random Forest trained on the full dataset achieved the highest recall at 0.819.

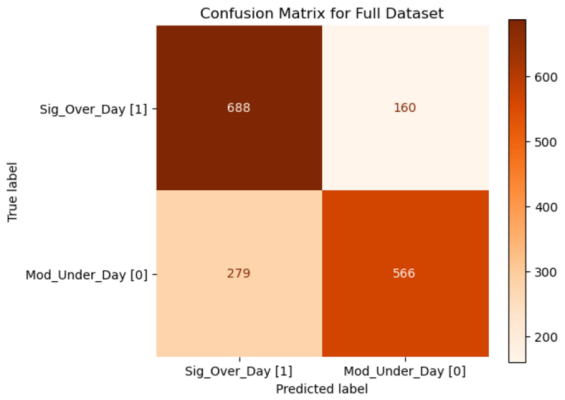

Fig. 4: Confusion Matrix Results

A. Model Performance

To kick off our modeling efforts, we conducted a series of preliminary experiments using our initial model scenarios. The results from this exploratory phase provided insights into trends across key metrics. Table II summarizes the outcomes from these baseline runs, serving as a foundation for subsequent optimization and feature tuning.

Model (Full / Invasive)	ROC-AUC	Recall	Specificity	Precision	Accuracy
XGBoost/ Full	0.823	0.811	0.692	0.726	0.752
XGBoost/ Inv.	0.821	0.803	0.708	0.738	0.756
Random Forest/ Full	0.805	0.819	0.652	0.697	0.735
Random Forest/ Inv.	0.811	0.785	0.682	0.710	0.733
Logistic Reg./ Full	0.792	0.727	0.729	0.726	0.724
Logistic Reg./ Inv.	0.792	0.725	0.723	0.726	0.725
Neural Network/ Full	0.796	0.754	0.703	0.713	0.728
Neural Network/ Inv.	0.547	0.351	0.700	0.464	0.552

TABLE II: Model Results: Full and Invasive-Only Data Sets

B. Threshold Adjustment

Threshold variation is a modeling technique that can influence performance metrics to better align predictions with medical objectives. As noted in Fig. 5, increasing recall through threshold adjustment (e.g., lowering from .5 to .4) enables the model to more effectively identify the positive class, patients likely to require ventilation for more than one day. This supports a more conservative prediction strategy aligned with ICU risk mitigation, where failure to detect high-risk patients could have critical consequences.

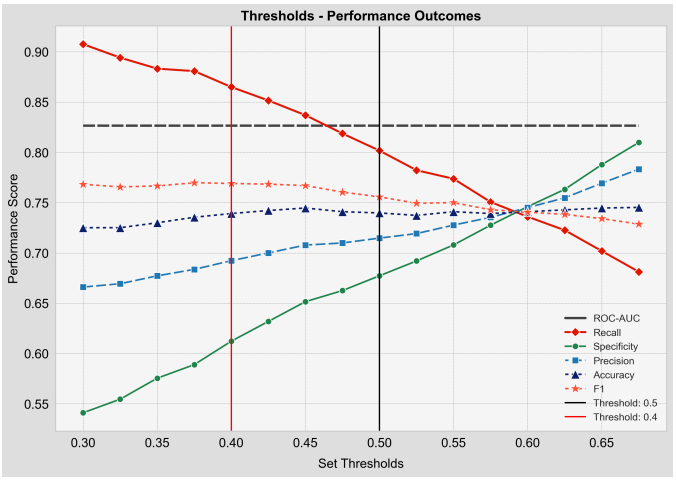

Fig. 5: Model Performance Variation by Decision Threshold [XGBoost]

C. Feature Selection

Across our modeling scenarios, we applied several feature selection techniques to identify the most informative variables and reduce dimensionality.

- **GVIF**: We used linear regression with Generalized Variance Inflation Factor (GVIF), suitable for models with categorical variables, to assess multicollinearity. 14 features with GVIF values greater than 4 were identified and later removed.

- **SHAP Values**: SHAP values, derived from Shapley values in cooperative game theory, estimate the contribution of each feature to a model's prediction [4]. We generated a ranked list based on the mean absolute SHAP values and found that several engineered features were highly informative, including `Age_Admission` (ranked 28th), `Resp_Diag_Label` (32nd), and `ICUSTAY_NUMBER` (66th). SHAP visualizations further illustrated how raw feature values influenced predictions for the positive class (e.g., `Sig_Over_Day` $>$ 1 day) [5]; see Fig. 6.
- **Sequential Feature Selection**: For the logistic regression model, we found that reduced models showed no significant improvement compared to using all available variables. This suggests a simple classic model can be practical, even with a high number of features.

Takeaways from Feature Selection: Our analysis, performed using cross validation, indicated that reducing the number of input features neither significantly improved nor diminished model performance. As illustrated in Fig. 7, performance remained relatively low with only the top 5 to 10 ranked features, then steadily increased between 10 and 30. However, gains leveled off beyond 50 features, suggesting diminishing returns as the feature count approached the maximum of 244. This pattern implies that while a core subset of features drives most of the model's predictive capability, expanding beyond this set contributes marginally and may introduce unnecessary complexity. These findings support the practicality of model deployment, as effective inference can be achieved with a streamlined subset of input variables.

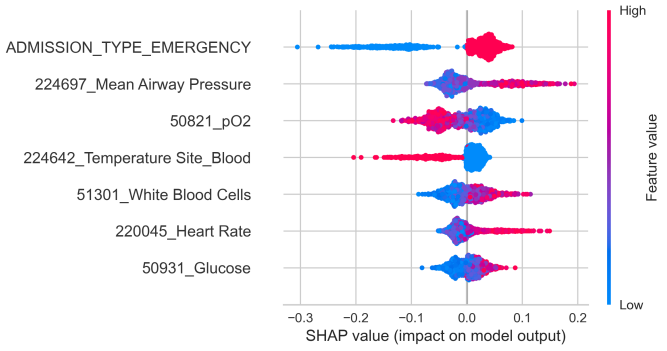

Fig. 6: SHAP Values Assessment: Top 7 Features / XGBoost

V. Conclusion and Implications

By prioritizing recall, we found that models are highly effective at identifying patients needing extended ventilation, which is crucial for hospital resource management. Our feature selection analysis also revealed that while a baseline number of features is necessary, adding more variables yields diminishing returns, suggesting that key physical characteristics are already well-represented and may allow for easier deployment in a real-world scenario.

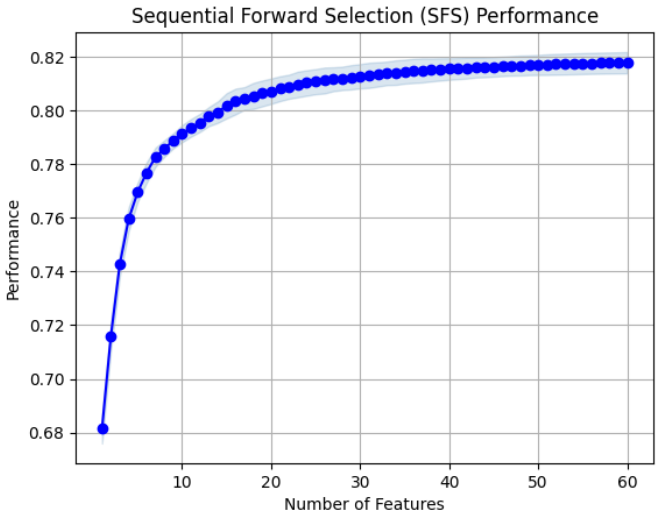

Fig. 7: Sequential Feature Selection

A. Limitations

Our study has several limitations. Our analysis was restricted to a subset of the MIMIC-III database, meaning potentially valuable information was not included. The de-identified nature of the dataset also limits the generalizability of our findings, as the patient population and clinical practices may not be representative of other healthcare systems. Finally, the retrospective design of the study means we could not account for all potential confounding factors or clinical decisions that might have influenced ventilation duration. Our models are predictive within the context of the available data but cannot fully explain the complex causal relationships in patient care.

B. Future Development

We recommend a more thorough evaluation of the full MIMIC-III database, including advanced feature engineering guided by consultation with clinicians and administrators, which could lead to the development of more effective features. Future models could also be designed to assess the efficacy of specific interventions. Additionally, further investigation into threshold adjustment is needed to maximize recall and enhance the model's utility in a real-world hospital setting.

References

[1] MIMIC-III Clinical Database; https://physionet.org/content/mimiciii/1.4/

[2] Mechanical Ventilation in the Intensive Care Unit; https://www.aast.org/resources-detail/mechanical-ventilation-in-intensive-care-unit

[3] Nationwide assessment of COVID-19 ventilator and non-invasive respiratory support burden during the early pandemic in the United States; https://www.nature.com/articles/s41598-025-99863-3

[4] Calculating XGBoost Feature Importance; https://medium.com/@emilykmarsh/xgboost-feature-importance-233ee27c33a4

[5] Basic SHAP Interaction Value Example in XGBoost; https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Basic%20SHAP%20Interaction%20Value%20Example%20in%20XGBoost.html