Beyond the Clock: Improving Elective Surgical Duration Predictions with Machine Learning

Mia deLadurantaye[1], Elise Williamson[1], Michael Dertke[1], Mackenzie Craig[1], Jacob Singer[1], Nathaniel Kusic[1], Constanza Lorca[1,*], Daniel Otero-León[1,2], Robert J. Riggs[1]
Corresponding Author: wrk2pp@virginia.edu

[1] *University of Virginia, Department of Systems and Information Engineering Charlottesville, Virginia, USA*

[2] *University of Virginia, Department of Public Health Sciences Charlottesville, Virginia, USA*

***Abstract*—Surgery is a cornerstone of modern healthcare, with the operating room (OR) as one of the most resource-intensive units within a hospital. Since surgeries account for a substantial portion of hospital revenue, accurate prediction of surgery durations and efficient OR scheduling are critical to minimizing delays for patients as well as reducing overtime for staff and OR idle time. However, OR scheduling is complex due to variability including patient history, unforeseen complications, and emergency cases that can disrupt planned schedules. This complexity is compounded by the current OR scheduling approach, which remains a highly manual process driven largely by human decision-making and crude statistical averages. The goal of this project is to develop a model that improves elective surgery duration predictions by identifying patterns from historical data from the UVA Department of Surgery. Given the large number of unique surgeries, we clustered the data to gain greater insights into underlying patterns. Using this processed data, we developed XGBoost and Random Forest models that identify the most influential variables affecting surgery duration and improve time estimates for future procedures. Preliminary models show a 39.7-minute improvement in Root Mean Squared Error (RMSE) for time prediction, a 59% decrease from the current prediction method.**

***Keywords*—Machine Learning, Healthcare, Surgery Prediction, Clustering.**

I. Introduction

Operating room (OR) scheduling is a central component of hospital operations, requiring coordination between surgical staff, equipment access, and facility availability. At the University of Virginia (UVA) Health's Department of Surgery, elective procedure schedules are created about a week in advance by the scheduling department based on the requested surgeries. Procedure durations are typically estimated using a combination of surgeon-provided predictions informed by experience and historical averages for similar procedures, often categorized using Current Procedural Terminology (CPT) codes. In some cases, basic statistical models that group procedures are also applied.

Despite steps to improve planning efforts, accurately predicting case durations remains a challenge. Similar surgical procedures can exhibit notable variability due to factors such as surgeon specific operating patterns, patient characteristics, case complexity, and procedure rarity. When estimated durations differ substantially from actual times, schedules may experience overruns or underutilization, leading to increased hospital costs from staff overtime and inefficient resource use. These inefficiencies can also lead to case delays, cancellations, and reduced overall patient safety.

Beyond the estimation challenges, surgical scheduling must also account for resource constraints which include OR availability, surgical teams, and specialized equipment. Schedules must balance efficiency and flexibility while accounting for turnover times and potential disruptions. Duration uncertainty directly impacts operating room throughput, staffing decisions, and financial performance. Current estimation approaches are limited in their ability to capture different factors, particularly in highly variable or rare procedures.

Given the operational importance of accurate estimates, improving surgical duration prediction presents an important opportunity to enhance hospital efficiency. The objective of this capstone project is to use surgery data from UVA Health to develop clustering-based features and machine learning models for predicting surgical case durations. Model performance is compared against baseline estimation methods currently being used, and improvements are evaluated using standard error metrics to assess gains in predictive accuracy.

II. Related Works

Early approaches to surgical duration prediction relied on statistical methods, including regression-based models and historical averaging techniques [1], [2]. In practice, duration estimates are often based on surgeon-provided predictions or averages from prior cases, as well as procedure types defined by CPT codes. Many hospital systems rely on historical averaging approaches, such as moving averages of prior case durations, to estimate surgical time [3]. These methods are simple and practical, making them widely used in real-world

scheduling workflows [1], [3]. However, prior research has shown that surgical durations have substantial variability due to factors such as case complexity, which are not fully captured by simple averaging approaches [4], [5]. As a result, these methods often perform poorly for rare or highly variable procedures.

More recently, research has explored the use of machine learning (ML) models to improve surgical duration prediction [5], [6]. Algorithms such as Random Forest, gradient boosting methods (e.g., XGBoost), Decision Trees, k-Nearest Neighbors (KNN), and neural networks have been widely applied due to their ability to model nonlinear relationships and interactions among multiple variables [5]. These models incorporate a broader set of features, including procedural details, provider experience, and case-level attributes, enabling more flexible and accurate predictions [2], [3]. Across studies, machine learning approaches have generally demonstrated improved predictive performance compared to traditional statistical methods [4]–[6]. Prior work shows that machine learning models can achieve more accurate predictions in comparison to historical averages and baseline estimation approaches [4], [6]. Additionally, data-driven machine learning methods decrease reliance on manual estimation and reduce biases introduced by human scheduling decisions. Improvements are commonly evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), as well as goodness-of-fit measures like R^2 [4].

Beyond improvements in predictive accuracy, several studies have examined the operational impact of improved duration estimates. Accurate prediction of surgical duration is critical for optimizing operating room utilization and minimizing scheduling inefficiencies, a key focus in recent operating room scheduling research [7]. More accurate predictions have been shown to reduce staff overtime, decrease case delays, and improve resource allocation across surgical schedules [6], [8]. Additionally, systematic reviews highlight that integrating machine learning with optimization strategies can further improve resource utilization and overall scheduling efficiency [5].

In addition to predictive modeling, clustering and optimization techniques have been explored to improve surgical scheduling. Prior work demonstrates that grouping procedures with similar duration characteristics can reduce variability and improve the reliability of master surgical schedules [9]. These approaches emphasize the importance of minimizing within group variability to support more balanced operating room utilization and reduce the risk of overtime. Such findings support the use of clustering-based feature engineering to capture underlying structure in surgical data.

Though the success of machine learning methods is highlighted in surgical literature, these methods are not widely implemented in many real-world hospital settings. Many healthcare systems, including the UVA Health Department of Surgery, continue to rely on traditional estimation methods and manual scheduling processes, often supported by specialized scheduling personnel rather than fully adopting data-driven approaches [7], [8]. At UVA Health, surgical scheduling and duration estimation are primarily managed through human-centered processes, with specialized staff responsible for constructing and maintaining operating room schedules. This highlights an opportunity to integrate data-driven ML methods into existing surgical workflows to improve predictions while supporting current operational practices.

III. Methods

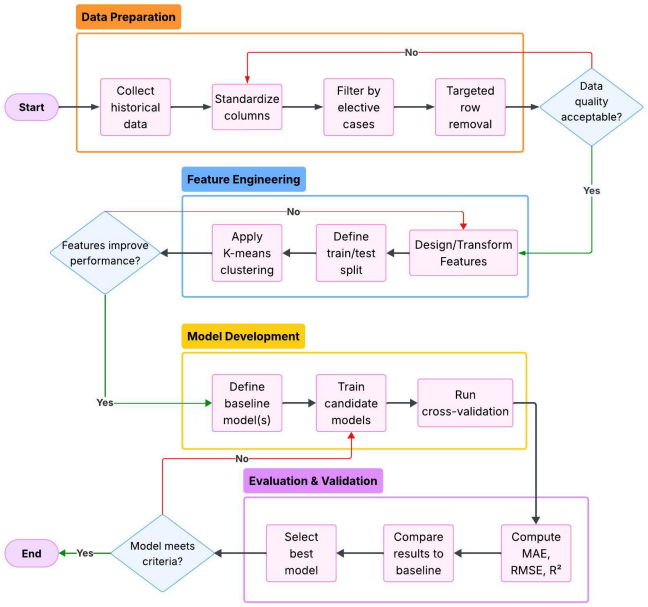

Fig. 1. Applied machine learning pipeline showing the sequential stages of data preparation, feature engineering, model development, and evaluation with iterative feedback loops

This study uses surgical data from the UVA Health Department of Surgery, spanning July 2022 to June 2025 and focusing on elective surgery procedures across all primary surgery services. Each observation represents a single surgical case. The target variable is operating room minutes utilized (OR MINS UTILIZED), which represents the total duration of time that a single surgical case occupies the operating room. This measure was chosen because it reflects both the actual surgical duration and the associated use of operating room resources.

Data preparation emphasized schema consistency and model readiness. Source columns were standardized by renaming common variants and replacing spaces in column names with underscores. Date and time fields were parsed into standard ISO 8601 datetime format when present. Cases were restricted to elective status when a case-class field was available. Rather than broad deletion for missingness, row removal was targeted to variables required for model construction: procedure text, clustering duration signal, and outcome. For categorical predictors, missing values were retained via a dedicated MISSING category before encoding.

Additional features were derived to capture temporal patterns in surgical scheduling. These include start hour and start

minute extracted from scheduled start time, as well as surgery month, day, and year derived from the scheduled surgery date. A numeric weekday variable was also created to represent the day of the week. These features allow the model to capture time based variation in operating room utilization. Outliers were retained in the dataset, as extreme values are critical since they often correspond to complex surgical cases. Preserving these observations ensures that the model captures real world variations.

Natural language processing and clustering were applied to group similar surgical cases based on procedure descriptions and observed duration patterns. The dataset included 3,290 unique procedure descriptions, which meant clustering was necessary to include them in the machine learning model. To prevent data leakage, clustering was performed exclusively on the training set, and cluster assignments for the test set were generated using the trained cluster models. Term frequency-inverse document frequency (TF-IDF) and sentence embeddings using the pretrained all-MiniLM-L6-v2 were used to convert the descriptions of surgeries into numeric values for K-means clustering. These vectors were then reduced using TruncatedSVD to improve clustering performance. The Surgery Duration column was scaled to a standard normal distribution and multiplied by a weight of 2 and added to the feature matrix for clustering. The duration weight was determined by minimizing the number of clusters with a procedure time standard deviation of more than 75 percent of the mean. Clustering was performed using K-means with 160 clusters, and the optimal K value was chosen by using a combination of visual inspection of the clusters, silhouette score, and Davies-Bouldin Index. Each case was assigned a cluster label, which was included as an additional feature in the machine learning model to capture similarities between procedures not fully reflected by the categorical variables in the dataset.

Key features include procedural variables (e.g., primary procedure, anesthesia type, and surgical specialty), scheduling information (e.g., start time, surgery date and weekday), and staff level attributes such as the provider's primary service.

To evaluate the effectiveness of the machine learning approach, we compared the models' performance to the current scheduling practices. Current schedulers use a moving average approach, which predicts surgical duration based on the historical average operating room time for prior cases with the same primary procedure. Additionally, they may use the surgeon's estimation, if given, and the scheduler's own expertise and experience. We used the variable scheduled duration, calculated as the difference between scheduled start and stop times, as a baseline performance metric that reflects the estimates currently used in operating room scheduling workflows and provides a reference point for assessing whether machine learning models offer meaningful improvements. This duration, since it represents the schedulers' expert opinion, was also used as a feature. The primary model used in this study is Extreme Gradient Boosting (XGBoost), a tree-based ensemble method well suited for structured tabular data. XGBoost was selected for its ability to model nonlinear relationships and interactions among features. The dataset was chronologically split into training (80%) and testing (20%) subsets to evaluate model generalization. An initial baseline XGBoost model was trained using default hyperparameters. Then, hyperparameter tuning was performed using RandomizedSearchCV, exploring parameters such as the number of estimators, maximum tree depth, learning rate, subsample ratio, and regularization terms. Model stability was supplemented with 3-fold cross-validated RMSE on the training partition.

When deciding on a model, we initially evaluated five diverse algorithms: Random Forest, k-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). Model performance was evaluated using multiple metrics to provide a comprehensive assessment of predictive accuracy. We defined error as the difference between the predicted duration and the actual duration for the test set. We used MAE to measure the average magnitude of prediction errors, while RMSE places greater emphasis on larger deviations. The coefficient of determination (R^2) was also used to assess the proportion of variance in surgical duration explained by the model. We prioritized RMSE for final selection as it penalizes extreme errors more heavily. In addition, a custom weighted error score was used to reflect overtime being more significant than undertime by penalizing underprediction (overtime risk) at twice the weight of overprediction. For cases under 100 minutes, a ± 15 minute on-time window was used. For cases over 100 minutes, a $\pm 15\%$ on-time window was used. These metrics were calculated for both training and testing datasets to evaluate model performance and assess potential overfitting.

IV. Results

The performance of the machine learning models was evaluated against a baseline estimation approach based on expert prediction, defined as the scheduled surgical duration used in current operating room workflows. The results were assessed using MAE, RMSE, and R^2. The numerical values are detailed in Table I.

TABLE I
PERFORMANCE COMPARISON OF SURGICAL DURATION PREDICTION MODELS

Model	CV RMSE (3-fold)	CV MAE (3-fold)	CV R^2 (3-fold)	RMSE	MAE	R^2	Score
Baseline	N/A	N/A	N/A	67.46	43.52	0.73	70.75
XGBoost	30.61	20.96	0.94	27.74	18.10	0.95	26.91
Random Forest	38.44	24.74	0.91	34.29	21.47	0.93	31.25

Overall, machine learning models demonstrated substantial improvements in predictive accuracy compared to the baseline approach. As shown in Table I, the current scheduling system ("Expert Prediction") produces a MAE of approximately 43.5 minutes. In contrast, the Random Forest model reduces MAE to 21.5 minutes, and the XGBoost model further improves performance to 18.1 minutes. This represents nearly a 58 percent reduction in prediction error compared to current

estimation practices. The same trend is observed in the weighted score metric (Table I), which incorporates overtime penalties. Expert predictions yield a score of 70.8, compared to 31.3 for Random Forest and 26.9 for XGBoost, indicating that improved prediction accuracy directly translates to better operational outcomes.

Prediction accuracy was further evaluated across varying tolerance thresholds experimentally. Machine learning models consistently outperformed expert predictions, achieving higher accuracy across all tolerance levels. These results were obtained using expert prediction as a feature, since in the hospital environment, it is available before the surgery is conducted. At a tolerance of 30 minutes, XGBoost achieves over 90 percent prediction accuracy, compared to approximately 75 percent for expert predictions. The gap between models is especially pronounced at lower tolerance levels, indicating that machine learning models are not only more accurate overall but also more precise in tighter scheduling scenarios.

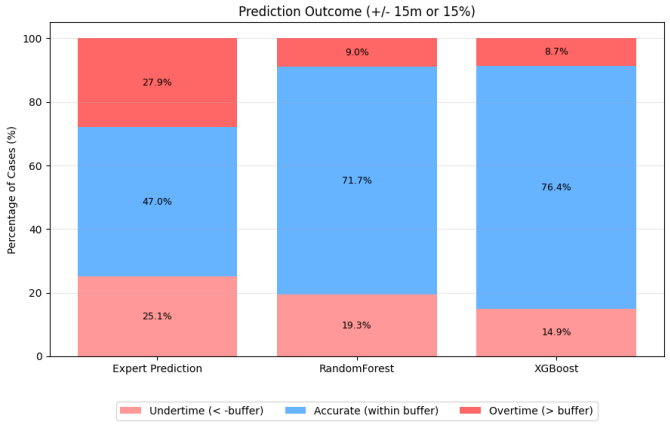

Fig. 2. Undertime, accurate, and overtime performances of ML models and baseline

The stacked bar chart in Fig. 2 shows prediction outcomes within a ±15 minute (or 15 percent) tolerance range. The expert predictions correctly estimate durations within the tolerance for 45.9 percent of cases, while Random Forest and XGBoost increase this to 64.8 percent and 67.7 percent, respectively. Machine learning models reduce both overestimation and underestimation rates, resulting in a more balanced distribution of prediction errors. This balance is important in practice, as it reduces the likelihood of both idle time and unexpected overruns.

Operational performance improvements were also observed in terms of overtime reduction. As tolerance thresholds increased, machine learning models consistently resulted in fewer cases with unplanned overtime compared to expert predictions. At lower tolerance levels, expert predictions resulted in substantially higher overtime rates, while both Random Forest and XGBoost demonstrated a mathematically significant decrease in overtime as tolerance levels relaxed.

Additional analysis was conducted to evaluate the impact of feature engineering on model performance. In this experiment, model performance was reassessed using RMSE and R^2 to better capture overall model fit and variance explained. The baseline scheduling system produced an RMSE of approximately 67.5 and an R^2 of 0.73, reflecting current expert predictions. After incorporating additional features, including procedure count and scheduling-related variables, the XGBoost model achieved an RMSE of approximately 27.7 and R^2 of 0.95, representing a substantial improvement over both the baseline system and earlier model configurations. Random Forest demonstrated similar gains, while simpler models such as Decision Trees and KNN showed less consistent improvement.

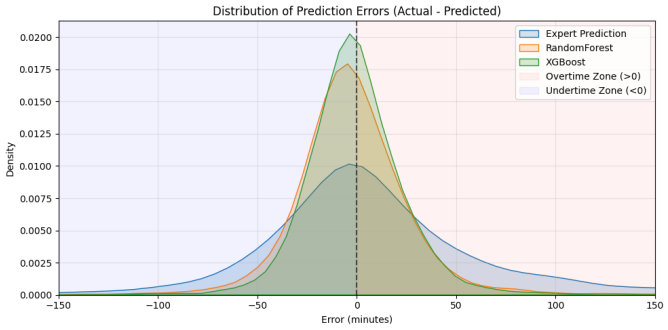

Fig. 3. Distribution of prediction errors of ML models and baseline

Error distribution analysis further supports these findings. As shown in the error distribution plots in Fig. 3, machine learning model errors are more tightly centered around zero, indicating lower variance and fewer extreme prediction errors compared to expert estimates. In contrast, expert predictions exhibit a wider spread, reflecting greater inconsistency and a higher likelihood of significant over- or underestimation.

Feature importance analysis from the trained Random Forest model (left side of Fig. 4) indicates that the most important feature for model prediction was the expert prediction, followed by the TF-IDF cluster and the MiniLM cluster. The expert prediction refers to the duration selected by the schedulers, which often relies on the moving average of the past procedures. Similarly, for the XGBoost model, the top 3 features are expert prediction, the admit base class (which identifies whether the patient was classified as inpatient or outpatient for care), and the TF-IDF cluster. The prominence of both TF-IDF and MiniLM clusters among the top features validates the dual-NLP approach used in our methodology. TF-IDF can capture exact matches in specific medical terminology, whereas the dense embeddings from MiniLM are able to correlate semantically similar procedures that may be documented with varying nomenclature. Also, the high importance of the admit base class for XGBoost highlights the differences in case complexity between inpatient procedures, which often involve more critical conditions and extensive preparations, and outpatient procedures. Procedural characteristics matter for explaining variability in surgical duration and support the inclusion of clustering-based features to capture underlying structures in the data.

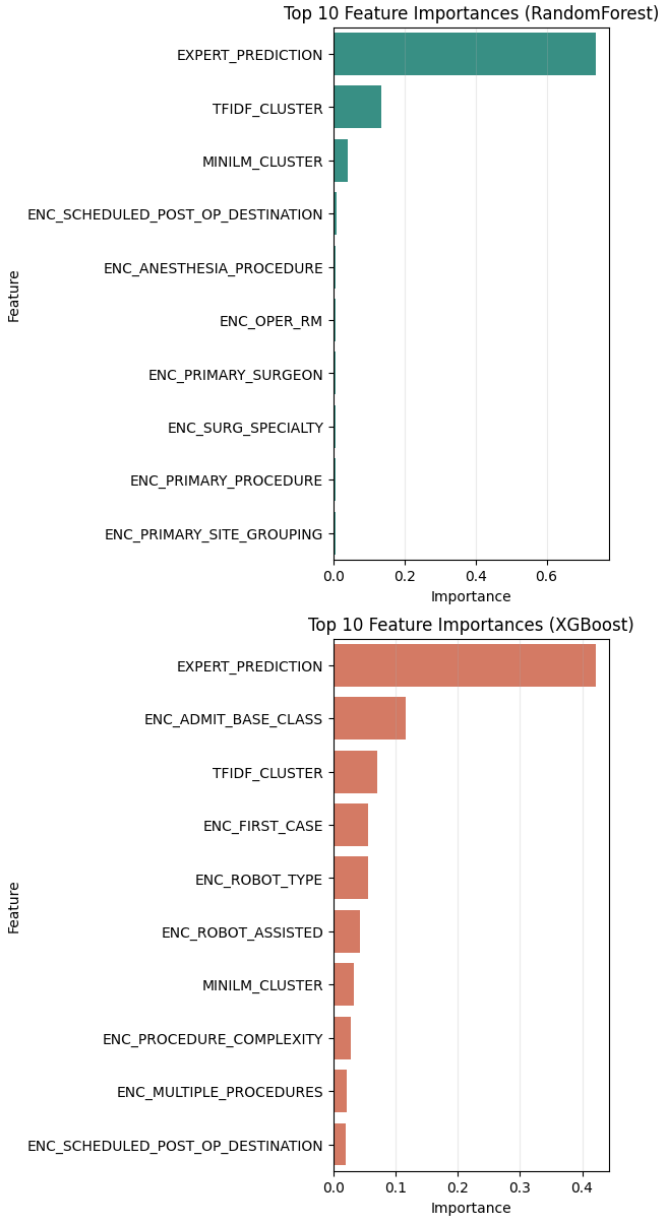

Fig. 4. Top 10 feature importance for Random Forest and XGBoost models.

TABLE II
PERFORMANCE COMPARISON BY SURGEON EXPERIENCE GROUP

Model	Group	MAE	RMSE	N
RandomForest	≥10 surgeries	20.44	32.49	7445
RandomForest	<10 surgeries	27.03	42.74	1378
XGBoost	≥10 surgeries	17.40	26.42	7445
XGBoost	<10 surgeries	21.89	34.01	1378
Expert Prediction	≥10 surgeries	41.14	63.12	7445
Expert Prediction	<10 surgeries	56.39	87.26	1378

As shown in Table II, both ML models also increase performance over the baseline for common (≥10 occurrences in the dataset) and uncommon (<10 occurrences) surgeries. The Random Forest model improves the MAE compared to the baseline scheduling methods by a little over 50% for both types of surgeries. The XGBoost model performs even better in terms of both RMSE and MAE - the XGBoost model improves both the MAE and RMSE for unique surgeries by 61%.

Overall, the ML models, especially XGBoost, demonstrate improved predictive accuracy, reduced error variance, and better operational outcomes compared to baseline estimation methods. These results suggest that data driven ML approaches can significantly enhance surgical duration prediction and support more efficient operating room scheduling.

V. DISCUSSION

This project demonstrates that it is possible to develop machine learning models that outperform the current surgical duration estimation methods used at UVA Health. Both the XGBoost and Random Forest models produced accurate and consistent predictions, showing clear improvement over the baseline scheduling approach. These findings suggest that historical procedural data contains meaningful patterns that can be leveraged to improve future prediction accuracy beyond traditional heuristic or surgeon-estimated methods.

While the difference in performance between XGBoost and Random Forest was relatively small, XGBoost generally performed better, with a few exceptions, such as overtime performance with a small accuracy window. This variation likely reflects differences in procedural complexity and variability across specialties. This indicates that model selection could be optimized depending on what metrics are desired, rather than relying on a single universal model, allowing for more tailored and effective predictions.

Despite these promising results, several limitations must be acknowledged. Due to strict compliance with the Health Insurance Portability and Accountability Act (HIPAA), the dataset was fully de-identified, removing all Protected Health Information (PHI) as well as surgeon-identifying variables. As a result, important patient-specific clinical factors such as BMI, comorbidities, and prior surgical history were excluded. The absence of these variables limits the model's ability to account for individualized physiological differences that may significantly influence surgical duration and likely contribute to residual prediction error. Additionally, these results were obtained by including expert prediction as a feature. In the hospital setting, this is known before the surgery is scheduled. Since expert prediction has the higher importance of any feature, removing it would significantly affect the performance of the model.

Additionally, the scope of this study was restricted to elective procedures, as defined by the "Case Class" variable. Urgent, emergent, and priority cases were excluded due to their inherently high variability and non-standardized workflows, reducing the dataset from 50,685 to 44,180 cases. While this

restriction allows for more stable and predictable modeling, it limits the applicability of the model in environments where emergency procedures are common. As such, the model is best suited for settings where procedures follow relatively consistent and planned workflows.

More broadly, the model is constrained by the inherent stochasticity of the hospital environment. This model is most effective when procedures are routine and uneventful. As seen in Table II, performance increases for common surgeries. However, it cannot explicitly capture unforeseen intraoperative complications, equipment malfunctions, or perioperative delays as independent variables. These irregular events are effectively "smoothed" into the dataset, which may introduce noise and reduce predictive precision, particularly for procedures prone to higher variability.

In practice, this model would be most effectively implemented as a decision-support tool for the surgical scheduling team. By integrating the model into existing scheduling systems, basic procedural information entered during case booking, alongside surgeon-provided estimates, could be used to generate improved predictions of required operating room time. In this role, the model would not replace human judgment, but rather augment it, enabling schedulers to allocate operating room blocks more efficiently and reduce the likelihood of delays or underutilization.

Looking ahead, the model has strong potential for further improvement if trained on more comprehensive datasets that include patient-specific clinical variables, provided that such use remains compliant with HIPAA regulations. With additional data and continued refinement, this approach could meaningfully improve operating room utilization, reduce scheduling inefficiencies, and lower operational costs. While this study focuses on UVA Health, the underlying methodology is broadly applicable and could be adapted to other healthcare systems seeking to optimize surgical scheduling through data-driven approaches.

VI. Conclusion

Overall, this project showed that combining clustering and machine learning methods with the scheduler's expert opinion can vastly improve surgical duration predictions. This project does not intend to replace surgical scheduling teams, but to work alongside them to improve OR scheduling and reduce delays, overtime and idle time.

References

[1] E. Kayis et al., "Improving prediction of surgery duration using operational and temporal factors," *AMIA Annu. Symp. Proc.*, vol. 2012, pp. 456–462, 2012. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3540440/

[2] O. Martinez, C. Martinez, C. A. Parra, S. Rugeles, and D. R. Suarez, "Machine learning for surgical time prediction," *Comput. Methods Programs Biomed.*, vol. 208, p. 106220, 2021, doi: 10.1016/j.cmpb.2021.106220.

[3] S. S. W. Lam et al., "Estimation of Surgery Durations Using Machine Learning Methods—A Cross-Country Multi-Site Collaborative Study," *Healthcare*, vol. 10, no. 7, p. 1191, 2022, doi: 10.3390/healthcare10071191.

[4] C. Spence et al., "Machine learning models to predict surgical case duration compared to current industry standards: scoping review," *BJS Open*, vol. 7, no. 6, p. zrad113, 2023, doi: 10.1093/bjsopen/zrad113.

[5] B. Entezari et al., "Improving resource utilization for arthroplasty care by leveraging machine learning and optimization: A systematic review," *Arthroplasty Today*, vol. 20, p. 101116, 2023, doi: 10.1016/j.artd.2023.101116.

[6] M. A. Bartek et al., "Improving operating room efficiency: Machine learning approach to predict case-time duration," *J. Am. Coll. Surg.*, vol. 229, no. 4, pp. 346–354, 2019, doi: 10.1016/j.jamcollsurg.2019.06.002.

[7] M. Al Amin et al., "Exploring the landscape of operating room scheduling: A bibliometric analysis of recent advancements and future prospects," *Biomed. Eng. Comput. Biol.*, vol. 16, pp. 1–16, 2025, doi: 10.1177/11795972241271549.

[8] K. W. Tan, F. N. H. L. Nguyen, B. Y. Ang, J. Gan, and S. W. Lam, "Data-driven surgical duration prediction model for surgery scheduling: A case-study for a practice-feasible model in a public hospital," in *2019 IEEE 15th Int. Conf. Autom. Sci. Eng. (CASE)*, 2019, pp. 275–280, doi: 10.1109/COASE.2019.8843299.

[9] A. Kressner and K. Schimmelpfeng, "Clustering surgical procedures for master surgical scheduling," Hohenheim Discussion Papers in Business, Economics and Social Sciences, No. 28-2017, 2017. [Online]. Available: https://nbn-resolving.de/urn:nbn:de:bsz:100-opus-14123