Health Care Management Science
https://doi.org/10.1007/s10729-024-09681-8

# A novel approach to forecast surgery durations using machine learning techniques

**Marco Caserta**[1] · **Antonio García Romero**[1]

Received: 8 June 2022 / Accepted: 13 June 2024


**Abstract**
This study presents a methodology for predicting the duration of surgical procedures using Machine Learning (ML). The methodology incorporates a new set of predictors emphasizing the significance of surgical team dynamics and composition, including experience, familiarity, social behavior, and gender diversity. By applying ML techniques to a comprehensive dataset of over 77,000 surgeries, we achieved a 24% improvement in the mean absolute error (MAE) over a model that mimics the current approach of the decision maker. Our results also underscore the critical role of surgeon experience and team composition dynamics in enhancing prediction accuracy. These advancements can lead to more efficient operational planning and resource allocation in hospitals, potentially reducing downtime in operating rooms and improving healthcare delivery.

**Keywords** Machine learning · Surgery duration · Feature importance · Team dynamics · Team composition

## Highlights

- The study developed novel predictors like team experience and gender diversity for surgery duration forecasts.
- We used ML on 77,000+ cases, achieving a 24% reduction in MAE with respect to a baseline model that mimics the approach currently used by the decision maker.
- Surgeon experience emerged as the most critical predictor of surgery duration.
- Engineered features capturing team and gender dynamics contributed to a 47% reduction in MAE compared with a model that does not use the newly proposed features.
- This study develops forecasting models that can help improve effective resource use.

Antonio García Romero contributed equally to this work.

✉ Marco Caserta
marco.caserta@ie.edu

Antonio García Romero
antonio.garcia@ie.edu

[1] IE Business School, IE University, Paseo de la Castellana 259E, Madrid 28046, Madrid, Spain

## 1 Introduction

The healthcare sector is known for its rapid pace of innovation in new diagnostic methods and advanced treatments; however, it often lags in improving management procedures. As a result, healthcare organizations often exhibit high levels of inefficiency. For example, Shrank et al. [1] indicated that approximately 25% of healthcare expenses in the US are wasted due to various reasons, including organizational dysfunctions. A prime example is the operating room (OR) scheduling, which is usually inefficient. Since surgical departments account for approximately 42% of hospital revenues [2] and 40% of total hospital costs [3], optimizing the use of the operating rooms will significantly impact hospitals' financial and operational performance.

This study aims to estimate the duration of surgical cases in an urban mid-size general hospital in Madrid, Spain. Accurately predicting surgical case duration is crucial for scheduling operating room procedures and improving hospital operations. However, surgical case duration is influenced by various factors, such as patients' clinical status, the surgical procedure, or the surgeons' expertise. Obtaining data for all these factors is rare, and clinical data availability is often limited due to regulations that preserve the confidentiality of sensitive information. Additionally, surgeons do not work

Published online: 10 July 2024

alone in the operating room; they perform surgeries with a team, and team interaction can either have a positive or negative impact, creating a collaborative or conflictive environment.

We use various ML techniques to forecast surgical durations, including Ridge Regression, Random Forest, Boosting, and Support Vector Machines. These models utilize standard features from the original dataset and integrate novel variables that reflect the surgical team dynamics, such as team familiarity, surgeon's workload, and conflict patterns within the operating room. To our knowledge, previous studies have overlooked the surgical team dynamic. Additionally, logistic factors like the scheduling of overlapping surgeries, the distinction between elective versus emergency cases, and the classification of surgeries into outpatient versus inpatient procedures are examined. To mitigate the lack of patients' clinical information, we use the Spanish Medical Association coding system as an ordinal variable, categorizing surgical cases into seven levels based on patient health and surgical complexity, circumventing the limitations imposed by data privacy regulations.

Our results indicate competitive accuracy in predicting surgical durations, as evidenced by a mean absolute error (MAE) of less than 16 minutes across a large dataset. This analysis, conducted in two steps-first, across the entire dataset to establish a baseline, and then within each surgical specialty to tailor the predictive models-highlights the significance of specific features in determining surgical times. This study provides insights into the critical factors affecting surgery duration by extracting the importance of features from the best-performing ML model. It advances the methodology in surgical case prediction, offering substantial benefits for hospital administration and patient care logistics.

In addition, we found that the best predictor of surgery duration is the surgeons' accumulated experience. Moreover, we identified some organizational variables, e.g., the team leader's familiarity and the level of exposure of the team members to a pool of diverse partners, that affect how to best design training approaches for team members by combining rotation with well-established teams. Lastly, the significant effect of gender diversity reinforces the idea that inclusive surgical teams can improve performance [4]. We also found substantial differences among surgical specialties. Our research has yielded valuable insights insofar as a more precise forecast of planned surgery duration can facilitate better resource allocation and scheduling, potentially minimizing idle operating rooms. While it is essential to acknowledge that forecasting improvements alone may not directly translate to enhanced efficiency, they undoubtedly play a pivotal role in informing decision-making processes within hospital operations management.

The main contribution of this study is the development and inclusion of innovative variables that encapsulate the complexities of organizational behavior and dynamics within surgical teams (i.e., team experience, familiarity, diversity, and cooperation dynamics). This methodology is based on the findings of Jones et al. [5], who emphasized the significance of these factors in understanding social behavior patterns in operating rooms. By integrating these nuanced variables, our approach enriches the predictive modeling of surgical outcomes, providing a more detailed understanding of the factors that impact team performance and efficiency in surgical settings.

This paper is organized as follows: We first present a literature review of the methodologies used to forecast surgery duration; next, in Section 3, we provide a brief overview of the data used, along with some descriptive statistics of the dataset. In the same Section, we present an extensive discussion of how we engineered new features by borrowing ideas from the field of team management. Section 4 presents the results of the ML techniques employed, along with statistical analysis, to measure the goodness of the alternative models and the impact of each feature on the error of the predictive models. Finally, Section 5 summarizes the study's most important findings and provides some concluding remarks.

## 2 Background

This section reviews existing literature on predicting operating room surgery durations. We categorized them according to their methodological approaches into two groups: those using ML techniques and those using alternative methodological frameworks. Table 1 summarizes the review of cases along three dimensions, i.e., the models used, the scope of the study, and the data used. The last row of the table presents the same analysis for the approach proposed in this paper. Concerning the scope, we indicate with "global" the definition of a global model for all the surgeries, while "per specialty" means that a model for each specialty is proposed. Concerning the data used, we identified three data sources, namely patient, procedure, and team data.

### 2.1 Machine learning methods

Much research has been produced to address case surgery duration and operating room scheduling problems. As Spence et al. [6] pointed out, a literature review suggests that ML and deep learning models are more accurate at predicting the duration of surgery; however, further research is required to determine the best way to implement this technology. We direct the interested reader to Zhu et al. [7], Rahimi et al. [8], Cardoen et al. [9], and Karakaya and Tatar [10] for comprehensive reviews on the problem. In addition, a review tailored explicitly to the use of ML approaches in the context of the organization of operating rooms is provided in

**Table 1** Summary of the extant review of studies along three dimensions

| Ref. | Model | Scope of the study | Data used |
|---|---|---|---|
| Shahabikargar et al. [12] | boosting, bagging, RF | global, per specialty | patient, procedure |
| Bartek et al. [13] | RF, boosting | per specialty, surgeon | patient, team |
| Tan et al. [14] | boosting, moving avg | two-step: surgeon rank and surgery duration | surgeon, team |
| Zhao et al. [15] | linear regression, ridge lasso, RF, NN | robot-assisted surgeries | patient, procedure |
| Huang et al. [16] | RF, boosting | global | patient, procedure |
| Soh et al. [17] | hybrid models | subpopulation-based | patient |
| Yuniartha et al. [18] | categorical model estimation algorithm | global | procedure |
| Ng et al. [19] | NN | global (with uncertainty) | patient, surgeon, procedure |
| Kayış et al. [20] | elastic-net linear regression | global | procedure, team |
| Laskin et al. [21] | expert prediction | oral and maxillofacial | procedure, surgeon |
| Luangkesorn and Eren-Dogu [22] | Markov chain and expert opinion | global | procedure, surgeon |
| Jiao et al. [23] | MDN | global | patient, procedure, anticipated duration |
| ours | ridge, RF, boosting, SVM | global, per specialty | procedure, team, social behavior, gender diversity |

The last row presents the characteristics of the approach proposed in this paper
RF: Random Forest
NN: Neural Network
SVM: Support Vector Machine
ICU: Intensive Care Units
MDN: Mixture Density Network

Bellini et al. [11], where ML methods for the forecast of case surgery duration, post-anesthesia care, and surgery cancellations are collected and discussed. These survey papers and the references therein evince that ML methods are often used in operating room scheduling and management problems.

Table 1 provides an overview of the existing approaches and highlights the main contribution of our approach. While most of the studies use a combination of patient, procedure, and team data, we enrich the model by engineering features that model the social behavior of the team, as well as gender diversity dynamics and experience and familiarity across team members. With an extensive computational study, we show that these features improve the ML models' forecasting performance.

Focusing only on elective surgeries, Shahabikargar et al. [12] forecast the duration of around 60,000 cases using boosting, bagging, and random forest. Results are provided per specialty as well as over the entire dataset. The authors note that more significant prediction errors are observed for some specialties, e.g., gynecology and cardio-thoracic surgeries. Interestingly, they measure the effect of filtering out surgical episodes with more than one procedure since their duration is more complex, positively affecting the quality of the produced forecasts.

Using random forest and gradient boosting models to predict surgery duration of over 47,000 surgeries from a hospital, Barket et al. [13] improved the current state-of-the-art at the hospital, based on the use of average values produced using the electronic medical records (EMR) data. The study uses patient-specific, procedure-specific, and personnel-specific variables. A characteristic of this study is that it estimates two sets of models. First, a group of models for services – estimating one model for each service or specialty – and second, a set of surgeon-specific models (one predictive model per surgeon.)

Tan et al. [14] proposed a two-step model: The first level model estimates the rank of the first surgeon while the second level model uses this predicted rank, along with other features, to predict case duration. Attention is given to capturing some features of the surgical team by coding, e.g., the team size and the presence of a student surgeon or a consultant, among other aspects, with the idea of modeling the experience and specific composition of the team. The results of a gradient boosting model are compared with those of a moving average approach. Statistical analysis shows that the results produced by the model that includes team information are significantly better than those of the other models tested in this study.

Based on around 1,000 cases from a hospital, Tuwatananurak et al. [24] present the results of a proprietary model. This study uses several features of the patients, providers, facility, procedure, and prior events in the ML model. Results over the entire data set were divided by specialty with an average improvement of around 7 minutes over the base case obtained using averages from the EMR.

Specifically focused on robot-assisted surgeries, Zhao et al. [15] use 28 features to predict the duration, spanning from patient characteristics to procedure characteristics, and a pool of ML methods (linear regression, ridge, lasso, random forest, boosted regression tree, and neural network) is employed. The boosted regression tree produces the best results, overall and per category, over 424 cases.

A recent study by Huang et al. [16] used more than 170,000 cases to develop and evaluate various ML methods, such as random forest and extreme gradient boosting, to predict the duration of surgeries. The authors compare the performance of these methods against two baseline models, namely, the surgery average and the surgeon average completion times. The best model was the extreme gradient boosting, and the authors provided feature importance statistics for this model. To account for the fact that surgeries carried out by the same surgeon within one week and one day are not independent, a "rolling horizon" time window was used. This means that when predicting the duration of a given surgery, the performance of the same surgeon within these time windows should be considered. The number of surgeries and the working time within these time windows were included as features in the ML models.

Jiao et al. [23] propose an ML model that uses both structured information (e.g., operative variables) and unstructured text (e.g., text description of procedure names and surgical diagnosis) to predict a probability distribution of surgery duration. The model concatenates tree-based methods with a neural network, i.e., a mixture density network. The authors use a mixture of Gaussian distribution in the network output to estimate the variability of the surgery durations. They tested the model on a dataset of over 50 thousand records, and, for the best model, they reported an MAE of 18.1 minutes, while a simpler Random Forest achieves an MAE of 19.6 minutes. A feature importance analysis shows that the most important predictors of the model are the estimated scheduled duration of the surgery (the anticipated duration) and the procedure's name. In contrast, the operational features have a limited contribution to the model's accuracy.

### 2.2 Other approaches

In the current literature, we noticed fewer studies that use methods other than ML to estimate surgery durations. Specifically, we identified three main approaches: (i) statistical-based techniques, (ii) expert judgment, and (iii) hybrid models that combine both elements.

Regarding statistical methods, Soh et al. [17] present a hybrid approach in which a prediction framework based on selecting a pool of alternative models is employed to tackle different segments of the surgical population. The proposed method has been tested on data from a New Zealand hospital, with promising results. Yuniartha et al. [18] propose using an estimation algorithm to forecast the surgery duration, incorporating a limited number of features capturing only the surgical procedure parameters. They claim that using surgical procedure parameters exclusively performs better than including patient data. Ng et al. [19] develop a neural network approach to forecast the surgery duration and the uncertainty associated with each prediction. Measuring the uncertainty for each prediction improves the accuracy of the operating room scheduling, thus reducing over- and under-booking times. The validity of the proposed approach is showcased using data from a large US hospital. Finally, Kayl et al. [20] used a statistical model incorporating operational, temporal, and staff-related variables to adjust the estimates generated by the SchDur method to achieve higher accuracy.

Laskin et al. [21] used an approach based on experts' opinions. They evaluated the accuracy of surgery duration based on oral and maxillofacial surgeons' views. They found that the surgeons correctly estimated 26% of cases. Their results improved the accuracy of the estimates by reducing significant errors.

Finally, some other studies combine experts' opinions with statistical methods. For example, Luangkerson and Eren-Dogu [22] combined expert judgment, expert classification of procedures by complexity category, and historical data in a Markov Chain Monte Carlo model.

## 3 Data and methodology

This section presents the data and the methods used in this study. We first explain how we obtained and processed the data from the hospital (Section 3.1.1) and how we enriched it with new features based on the literature findings (Section 3.1.2.) We then describe the ML methods used in this study (Section 3.2.)

### 3.1 Data

#### 3.1.1 Data provided by the hospital

The dataset comprises 57,132 cases from 2016-2018 and an additional set of 20,350 cases from 2019. We used the cases from 2016-2018 to create new features based on historical

data. (In the sequel, we refer to this set of observations as "historical data.") Therefore, the training-testing dataset comprises the 20,350 observations of 2019 (the "dataset" in the sequel.) Fig. 1 shows the dataset surgery times distribution. We observe that the distribution is heavily right-skewed, suggesting that a log transformation might be beneficial when regression-type approaches are used.

Table 2 lists the variables included in the dataset and their summary statistics.

- OMC[1] Code: The type of surgical procedure. This code is used by all hospitals throughout the National Health Care System in Spain. This study does not consider the unplanned OMC Codes only available after the operation.
- OMC Score: An ordinal score with seven levels representing the patient's health status and the surgery complexity. It resembles the ASA (American Society of Anaesthesiologists) Physical Status Classification System (www.asahq.org/standards-and-guidelines/asa-physical-status-classification-system, Jan 3, 2023).
- OR: The operating room identifier.
- Type: A binary variable that indicates whether the surgery is ambulatory.
- Emergency: A binary variable that indicates whether an operation is scheduled.
- Surgical Pathology: A binary variable that indicates whether patients' tissues were removed during surgery.
- Surgeon, Anesthesiologist, Assistant Surgeon, Instrumentalist, Circulating Nurse, and Assistant are individual identifiers of these roles in the Surgical Team.

Using historical data, we computed a preliminary set of features, i.e.:

- `average_OMC`: The average surgery time for each "OMC Code" in the historical dataset. We computed 711 different average times.
- `average_surgeon`: The average surgery time for each "Surgeon". Using historical data, we computed the average for each of the 209 surgeons in the dataset.

In addition, from the date and timestamp of each surgery in the dataset, we extracted the day of the week and the shift of each case. More precisely, the variable `shift` is coded into `M` (morning) if the scheduled time for the surgery is between 8 am and 3 pm and `A` (afternoon) when the scheduled time is between 3 pm and 10 pm. We selected 12,468 cases in the morning and 7,882 cases in the afternoon. Concerning the day of the week, the operating theatres function on a 7-day plan, with around 4000 surgeries on weekdays, Monday to Friday, and a much lower workload on Saturday (846 cases) and Sunday (75 cases.)

[1] OMC stands for Spanish Medical Association.

#### 3.1.2 Additional features from the management literature

In line with the findings of Jones et al. [5] about social behavior in the OR, we collected information about the gender of the team members. In addition, we defined the variable `same_gender` as follows: We set this variable to 1 if the surgeon and the assistant surgeon have the same gender and 0 otherwise. With this variable, we wanted to capture the *relation* between the principal surgeon and assistant surgeon along the dimension *of cooperation-conflict*. As mentioned in Jones et al. [5], in the operating room (OR), clinical roles and gender composition influenced behavior patterns. Conflicts were more common in interactions between individuals of different hierarchical levels, typically initiated by those higher in rank. Increased cooperation was noted with a higher proportion of female staff. Most notably, when the attending surgeon's gender differed from most OR staff, cooperation significantly increased, highlighting the intricate relationship between professional hierarchy and gender dynamics in healthcare teamwork (Jones et al. [5]). We, therefore, included this variable to operationalize the effect of cooperation vs. collaboration on the case duration.

In addition to gender-related features, we engineered several features to code potential effects of the team composition and characteristics. In the sequel, let us indicate with $\tau$ the set of team members (i.e., surgeon, anesthesiologist, assistant surgeon, etc.) and $T$ the set of all the teams in the dataset. We then computed the following variables:

- $|\tau|$: Team size, i.e., number of team members associated with a surgery. From column `count` of Table 2, we notice that, while each case requires a surgeon, the other team members are not necessarily present. Therefore, the team's composition is associated with the surgery's type and complexity.
- $|\tau_f|$: The number of female members in the team. Together with the variable $|\tau|$, this variable captures the gender diversity within the team.
- $\delta_{gender}$: We operationalize the diversity of gender of the team members using the Blau index [25]. When talking about diversity, at least three primary constructs have been defined, namely *separation*, *variety*, and *disparity*. The Blau index primarily measures the range of a dimension within a group. This is useful when working with a categorical variable like gender and when we want to understand the distribution of what is distinctively known by members of each group [26].

Finally, we captured team dynamics by estimating team familiarity and diversity of experiences via partner exposure.

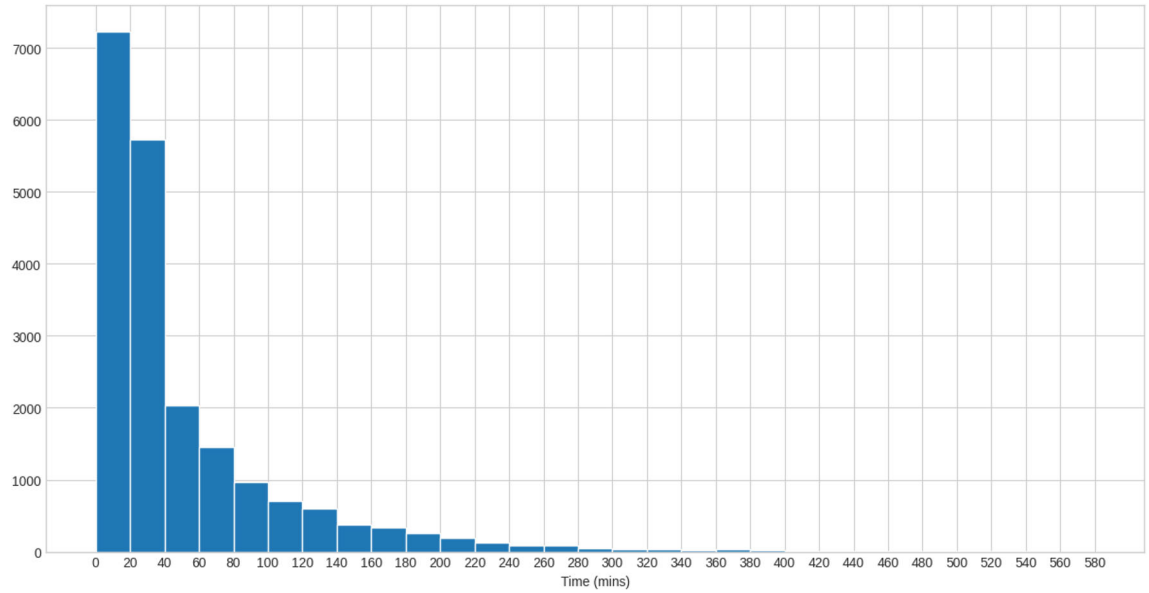



**Fig. 1** Distribution of surgery times in the dataset

Borrowing ideas from the consolidated theory of team familiarity (e.g., [27] and [28]), we account for the fact that there exist both theoretical findings and empirical evidence that changes in team membership can enhance creativity and problem-solving. (Aksin et al. [28] and references therein.) People are likely to benefit from prior partner exposure since they reap benefits by observing different ways to perform tasks and choosing the best among them. Such a broad spectrum of experiences increases the amount and novelty of the knowledge generated. This finding seems even more relevant in the context of non-standardized processes.

**Table 2** Summary statistics of the categorical variables in the dataset

| | count | unique | freq |
|---|---|---|---|
| Specialty | 20350 | 21 | 4367 |
| OMC Code | 20233 | 711 | 2114 |
| OMC Score | 19687 | 11 | 6408 |
| OR | 20350 | 16 | 4378 |
| Type | 20350 | 2 | 12936 |
| Emergency | 20350 | 2 | 17260 |
| Surgical Pathology | 20350 | 2 | 20137 |
| Surgeon | 20350 | 209 | 1551 |
| Anesthesiologist | 17716 | 77 | 1454 |
| Assistant Surgeon | 1819 | 51 | 266 |
| Instrumentalist | 2386 | 33 | 202 |
| Circulating Nurse | 10204 | 50 | 942 |
| Assistant | 12923 | 44 | 1173 |

The column "count" indicates the number of observations with a value, the column "unique" provides the number of unique values, i.e., classes for that variable, and the column "freq" gives the number of observation of the most frequent group

Let us indicate with $\mathcal{P}_i$ the set of partners of worker $i$ in the historical data, with $i \in \tau$. In addition, let $n_{ij}$ be the number of times workers $i$ and $j$ have worked together. Thus, the overall experience of worker $i$ can be computed as:

$$exp_i = \sum_{j \in \mathcal{P}_i} n_{ij}$$

We thus quantify the depth of the exposure of a worker using the Herfindahl-Hirschman Index (HHI) as follows:

$$HHI_i = \sum_{j \in \mathcal{P}_i} \left( \frac{n_{ij}}{exp_i} \right)^2$$

This measure captures the dispersion of the experience across all workers. For a worker $i$ with little prior exposure to other partners (e.g., a worker exposed to a single partner), the HHI takes the value 1, while for a worker with maximal exposure (e.g., with prior experience equally distributed across all the partners), the HHI takes the value $\frac{1}{|\mathcal{P}_i|}$. Consequently, smaller values of HHI imply greater exposure to different partners. We refer the interested reader to [28] for an example of usage of the HHI in the context of partner exposure and team familiarity.

Thus, the overall prior exposure of the members of a team $\tau \in T$ is:

$$HHI_\tau = \sum_{i \in \tau} \frac{HHI_i}{|\tau|}$$

Similarly, in line with the team productivity literature [29, 30], we created a set of familiarity scores to quantify the

familiarity of each member of the team with the others and the familiarity of the team leader (the surgeon) with the other members of the team. More specifically, using historical data, we computed the team and surgeon familiarity scores $f_\tau$ and $f_s$, respectively, as follows:

$$f_\tau = \frac{\sum_{i\in\tau}\sum_{j\in\tau} n_{ij}}{\frac{|\tau|(|\tau|-1)}{2}}$$

$$f_s = \frac{\sum_{j\in\tau} n_{sj}}{|\tau|-1}$$

where we set $n_{ii} = 0$ for all $i$, and we indicate with $s \in \tau$ the surgeon of team $\tau$. Note that, due to the surgeon's relevant role in the team dynamics in the OR, we explicitly compute a familiarity score for each surgeon. In addition, given the relevance of the dynamics between the principal surgeon and assistant surgeon, we compute a surgeon-assistant familiarity score $f_{sa}$, with $s$ indicating the surgeon and $a$ indicating the assistant surgeon of a team $\tau \in T$, as the number of times the surgeon and the assistant have worked together, i.e., $f_{sa} = n_{sa}$. To control for the cumulative experience of the surgeon with a specific surgery, we define a new variable, $exp_{sg}$, which accounts for the number of times surgeon $s$ performed surgery $g$.

All the features presented in Section 3.1.2 are summarized in Table 3. We used a one-hot encoding for each categorical variable to transform it into binary attributes. A preprocessing of the dataset further reduced the number of observations available for the training phase. The preprocessing steps were as follows:

- We eliminated all the observations for which the service, the OMC Code, or the OMC Score were unknown. This reduced the dataset size to 17013 from an initial size of 20350.
- According to the hospital's policy, any surgeries with a duration of less than 10 minutes have been removed from the schedule. The hospital divides the time window of each scheduling plan into slots of 10 minutes. This means that surgeries that take 10 minutes or less are not scheduled but are instead fit into available slots. As a result, the dataset has been further reduced to 14576 observations.

## 3.2 Methodology

We built two families of predictive models: (i) global models, i.e., models in which the entire dataset is considered; (ii) specialty-specific models, i.e., a predictive model for each specialty included in the dataset.[2]

Following the standard training/validation/testing approach, we set aside 20% of the data for the final testing of the different methods, i.e., 2915 observations are finally used to test the models. The remaining 11661 are used within a framework of 5-fold cross-validation to train and validate the different predictive models. That is, using a 5-fold cross-validation framework, we fine-tune the models' hyperparameters. Next, each model is tested on the previously unseen testing set. All the results reported in this section refer to the results obtained on the testing set. In addition, when creating the training and testing sets, we respect the time series nature of the dataset insofar as we use observations from January to September for the training set and from October to December for the testing set.

**Table 3** Notation and interpretation of the team, gender, and experience features

| Variable | Interpretation |
|---|---|
| $\text{HHI}_\tau$ | partner exposure to a diversity of experience at team level |
| $f_s$ | leader (main surgeon) familiarity with the team |
| $f_\tau$ | average familiarity across team members |
| $\|\tau\|$ | team size (number of members) |
| $exp_s$ | surgeon experience |
| $\|\tau_f\|$ | number of females in the team |
| $f_{sa}$ | familiarity between main surgeon and assistant |
| $\delta_{gender}$ | gender diversity of team members (Blau index) |

Concerning the global approach, we constructed two simple models that mimic how the decision-maker estimates surgery durations. In the sequel, we refer to these two models as the "baseline" models. These models use the historical data provided by the hospital and compute the average OR time for each surgery along two dimensions:

- `avg_OMC`: The average OR time of all the surgeries with the same OMC code.
- `avg_surgeon_OMC`: The average OR time of all the surgeries with the same OMC code (i.e., surgery type) performed by a specific surgeon.

We then built two models, using the `avg_OMC` and the `avg_surgeon_OMC` values to forecast the duration of surgeries in the dataset. The MAE produced by the models and the summary statistics of these two estimators are presented in Table 4. We use the performance of the two baseline models as a reference because any ML method must produce better results than those of these models.

Next, we proceeded with the training of the following models using a 5-fold cross-validation approach. We use the Python implementation of each method provided in the `sklearn` library. The parameter tuning for each method has been carried out using a grid search approach. (The detailed list of parameter values is provided in the repository – see

[2] The most common specialties at this hospital are General Surgery, Plastic Surgery, Eye Surgery, Trauma Surgery, and Urological Surgery.

**Table 4** Statistics and MAE of the baseline models on the testing data, measured in minutes

| | mean | std | MAE (Training) | MAE (Testing) |
|---|---|---|---|---|
| `avg_OMC` | 68.67 | 107.08 | 37.32 | 37.22 |
| `avg_surgeon_OMC` | 57.33 | 61.51 | 21.69 | 20.80 |

Section 4 for details.) The selection of the methods reported here is based on a preliminary screening. We initially tested various methods such as linear regression, Lasso regression, and neural networks. We then discarded any methods with performance metrics (MAE, RMSE, $R^2$, and PDw15) worse than those of at least one other method. After this screening, we selected the following methods:

- Ridge Regression (`RR`): A linear least square model with $L_2$ regularization (regularization parameter $\alpha = 0.5$.) [31]
- Random Forest Regressor (`RF`): A decision tree-based method with bootstrapping and subsampling [32].
- Light Gradient Boosting (`LightGBM`): A modification of the Gradient Boosting Regressor, specially designed to deal with large datasets [33].
- Support Vector Regressor (`SVM`): A regressor with kernel function and regularization [34].
- Extreme Gradient Boosting (`XGBoost`): A Gradient Boosting-based method [35].

We ran all the experiments on a Linux machine on Google Cloud, using a Virtual Machine of type `e2-standard-4` with four vCPUs and 16GB of memory. The code implemented, trained, and fine-tuned the different models is available at http://github.com/marcocaserta/OR. The data is available upon request to the authors and contingent on the approval of the hospital Research Committee.

## 4 Results

We begin by analyzing the correlation among the engineered features in Table 3. Figure 2 presents the correlation values for the engineered features and some summary statistics for each of these variables. These variables have been re-scaled when used in a predictive model. The figure gives the magnitude of the correlation between the engineered features. Ideally, we aim to minimize the correlation among these features to better capture the dependent variable's variability. We observe that the new features have moderate to low correlation except for the two familiarity variables, i.e., team leader familiarity $f_s$ and team familiarity $f_t$. The table at the bottom of the figure presents the correlation between the dependent variable `or_time` and this pool of independent variables. The correlation between the dependent and all the other independent variables, i.e., the variables of Table 2, is mild and, therefore, for brevity, is not presented here.

### 4.1 Machine learning models

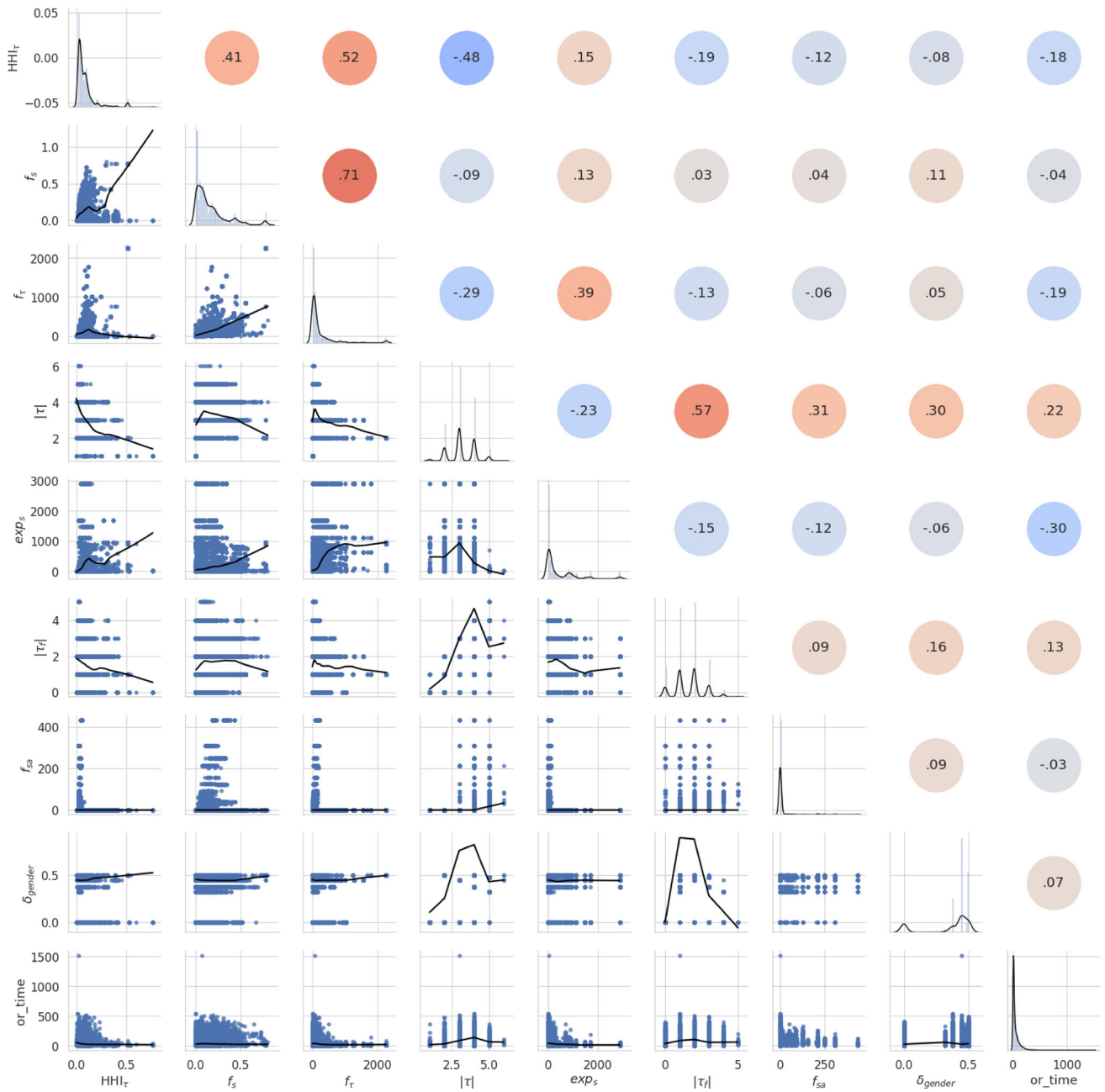
Table 5 compares the results from five methods used to forecast surgery duration and two baseline models. The first column lists the method names, while the second to fourth columns display MAE, Root Mean Squared Error (RMSE), and the $R^2$ value. The last column (PDw15) shows the percentage of forecasted cases with a deviation, in absolute value, of less than 15 minutes from the actual surgery duration.

Table 5 shows that all the methods produce better results than the baseline models along any of the four quality measures. In addition, we observe that `XGBoost`, `LightGBM`, and `RF` attain the best results, with `XGBoost` reaching the lowest MAE. In addition, we observe that `XGBoost` can forecast 73% of the testing cases with an absolute error of no more than 15 minutes. In line with the literature [15, 24], and the expert opinion at the hospital, a prediction is considered acceptable when the deviation from the actual duration is less than 15 minutes. This improves the results of the baseline model (`avg_surgeon_OMC`) by 22%.

To assess the statistical significance of performance disparities among the methods, we employed bootstrapping combined with 5-fold cross-validation across the entire dataset. Our objective is to ascertain the presence of statistically significant variances in MAE among the five methodologies outlined in this study, excluding the two baseline models due to their inferior metrics relative to the ML approaches.

According to Demsar [36], We conducted a statistical analysis to determine if the machine learning methods yield different forecasts and if we can establish a ranking of the various techniques. With this goal in mind, we performed the following two-step analysis (the detailed statistical comparison is presented in the Appendix): (i) With the Kruskal-Wallis test [37], we show that statistically significant differences across the ML methods arise; (ii) using the Nemenyi test [38], we rank the methods and we show that `XGboost` and `RF` are significantly better than the other methods, in terms of statistical metrics performance.

After establishing that `XGBoost` is significantly different from the other method, except for `RF`, in the subsequent phase of the analysis, we use `XGBoost`, i.e., the method with the




| | $HHI_\tau$ | $f_s$ | $f_\tau$ | $\lvert\tau\rvert$ | $exp_s$ | $\lvert\tau_f\rvert$ | $f_{sa}$ | $\delta_{gender}$ |
|---|---|---|---|---|---|---|---|---|
| mean | 0.09 | 223.02 | 210.28 | 3.21 | 403.08 | 1.62 | 10.78 | 0.36 |
| std | 0.1 | 417.37 | 398.55 | 0.85 | 637.86 | 0.97 | 50.28 | 0.19 |

**Fig. 2** Correlation matrix and summary statistics for the newly computed features

**Table 5** Testing results of multiple methods on the testing dataset

| Method | MAE | RMSE | $R^2$ | PDw15 |
|---|---|---|---|---|
| avg_surgeon_OMC | 20.80 | 36.34 | 0.66 | 0.60 |
| avg_OMC | 37.22 | 99.02 | 0.44 | 0.56 |
| RR | 17.51 | 32.72 | 0.77 | 0.67 |
| RF | 16.27 | 31.65 | 0.79 | 0.71 |
| LightGBM | 16.31 | 31.46 | 0.81 | 0.71 |
| SVM | 18.14 | 36.79 | 0.77 | 0.68 |
| XGBoost | 15.89 | 30.29 | 0.81 | 0.73 |

Reported results correspond to fixing the testing dataset at the end of the time window. PDw15 is measured as the percentage of cases with a deviation from the actual duration of fewer than 15 minutes

lowest MAE, to build predictive models for each surgical specialty.

We first want to evaluate the importance of the most salient features used in the XGBoost model. Figure 3 reports the top 15 features and their impact on the MAE. For the feature importance computation, we use the sklearn implementation of the permutation importance method, as presented in Breiman [32]. The importance of a feature is measured as the drop in the score (the MAE, in this case) when the values of that feature are randomly permuted. This procedure is understood to break down the relationship between the target variable and the feature; therefore, the drop in the score can be used to estimate the importance of the feature in the overall model.

The figure indicates that the most crucial factors are derived from historical data, specifically the average time taken by a particular surgeon and the average time taken for a specific OMC code (i.e.:type of surgery) across all surgeons. This implies that both the surgeon performing the procedure and the type of surgery are influential in determining the duration of the case. This aligns with the results presented in Table 5, where we report that using the surgeon's average time to predict the surgery duration allows us to predict case duration with no more than a 15-minute error around 60% of the times.

However, other features must be considered to achieve the PDw15 rate 73% obtained by the XGBoost model. Figure 3 highlights that some of the engineered features presented in Section 3 seem essential. Observe, e.g., the team ($f_\tau$) and leader ($f_s$) familiarity and the level of exposure of the team members to a pool of diverse partners ($\text{HHI}_\tau$). All these variables, which are team-specific variables, impact the quality of the final prediction. In addition, we observe some gender-related effects, captured by whether the surgeon and the assistant surgeon are of the same gender (same_gender), as well as effects due to the gender diversity within the team ($\delta_{gender}$). Finally, to a minor extent, logistics-specific information (the operating theater, the day of the week) affects the final prediction.

To measure the effect of the engineered variables on the model's predictive power, we computed the MAE and the PDw15 of the model after removing each engineered variable one at a time. In addition, we computed the forecasting error associated with a model in which the top six engineered variables are removed. Table 6 provides the results of the analysis. In line with the feature importance plot of Figure 3, we observe that these variables reduce the MAE of the forecasting model, with an improvement of 47%, from 29.84 to 15.89.

In summary, procedure-specific and team-specific factors emerge as significant predictors of surgery duration, with logistical considerations exerting a minor influence on the model's efficacy. The absence of patient-specific data in the dataset, which was not captured in this analysis, could potentially enhance the algorithm's predictive performance if included.

The results described above provide evidence for management improvement. For example, the relevance of the team leader's familiarity and the level of exposure of the team members to a pool of diverse partners may have clear implications on how to best design training approaches for surgical team members by combining rotation with well-established teams. Regarding the significant effect of gender diversity, our findings reinforce the idea that inclusive surgical teams can improve their performance [4]. For example, consider the partial dependence plots of Fig. 4.[3] Figure 4-(a) illustrates the direction of the relation between the surgery duration and the gender diversity score. We observe that, beyond a threshold value, the higher the gender diversity, the lower the expected surgery duration. The relevance of the operating room variable OR on case duration may be due to the hospital we analyze having two dedicated ORs for eye Surgery cases and another one for robotic surgery equipped with a Da Vinci(R) surgical system. Besides, the variable emergency positively correlates to case duration due to the severity and complexity of these surgeries, as shown in Fig. 4-(b). Finally, the partial dependence plot of Fig. 4-(c) shows that the surgical team size positively relates to case duration. This is because team size and case complexity correlate positively (correlation $= 0.2$, $p-\text{value} = 0.023$.)

The importance of the variable "Specialty" leads to comparing the features' importance for each main surgical field. For this reason, the last step of the empirical analysis focuses on splitting the data and training one model per surgical spe-

[3] Partial dependence plots show the dependence between the target response and a set of input features of interest, marginalizing over the values of all other input features (the 'complement' features). Intuitively, we can interpret the partial dependence as the expected target response as a function of the input feature of interest. Therefore, in a partial dependence plot, the x-axis maps the domain of the input feature, while the y-axis measures the marginal effect of the feature onto the response variable [39].

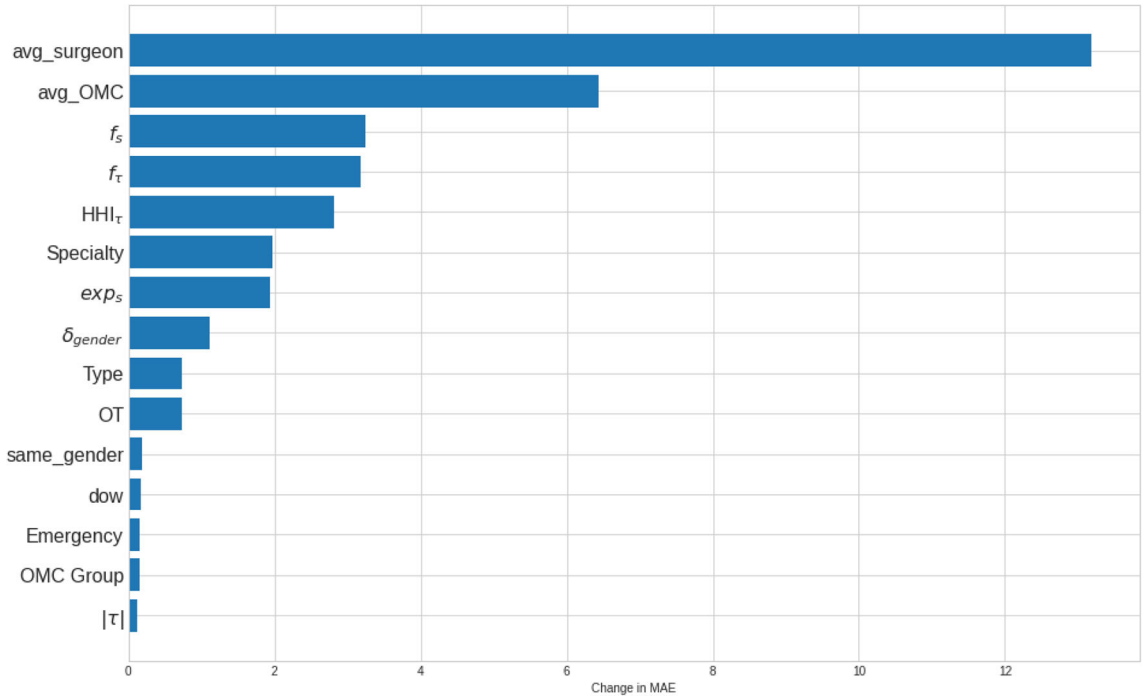
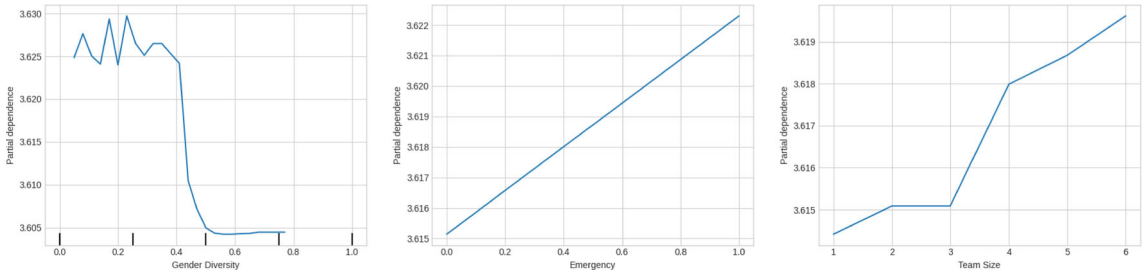



**Fig. 3** Features importance of the `XGBoost` model

**Table 6** MAE and PDw15 of different models: The first column reports the score of the full model; columns two to seven give the metrics value of the model obtained after removing the corresponding variable; the last column gives the metrics of the model after removing the six engineered variables

| | Removed variable | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | full model | $f_\tau$ | $f_s$ | $HHI_\tau$ | $exp_s$ | same_gender | $\delta_{gender}$ | all |
| MAE | 15.89 | 18.93 | 19.01 | 18.53 | 17.98 | 16.52 | 17.02 | 29.84 |
| PDw15 | 0.73 | 0.63 | 0.62 | 0.63 | 0.65 | 0.71 | 0.69 | 0.61 |




(a) (b) (c)

**Fig. 4** Partial dependence plots for the gender diversity ($\delta_{gender}$), emergency, and team size ($|\tau|$) variables

**Table 7** Summary of the performance of `XGBoost` on each specialty and comparison with the global model

| Specialty | Size | $R^2$ | MAE | RMSE | MAE Global |
|---|---|---|---|---|---|
| General Surgery | 1575 | 0.65 | 19.63 | 32.64 | 21.43 |
| Plastic Surgery | 3587 | 0.59 | 32.99 | 51.42 | 34.12 |
| Eye Surgery | 3616 | 0.66 | 8.36 | 17.4 | 11.40 |
| Trauma Surgery | 1164 | 0.51 | 19.33 | 29.87 | 19.31 |
| Urological Surgery | 1205 | 0.71 | 18.4 | 27.47 | 22.64 |

cialty. During this phase, we use the `XGBoost` model and replicate the experiment, splitting the data into training and testing and using a k-fold cross-validation approach. However, out of all the specialties in the dataset, we limit the study to the five specialties with at least one thousand observations (i.e., enough to carry out a meaningful training-validation-testing split.)

Table 7 presents the results of the `XGBoost` model for each specialty and compares the specialty-specific models with the global model on each subset. The first column reports the surgical specialty, the second column gives the sample size (we allocated 20% to testing), and columns three to five provide the results in terms of $R^2$, MAE, and Root Mean Squared Error (RMSE) on the testing dataset. Finally, the last column gives the MAE obtained by the global model when used on the data of each specialty. We first observe that each specialty-specific model performs better than the global model in terms of MAE. In addition, we notice a large variability of performance across specialties. For example, the MAE spans from 8.36 for Eye Surgery to 32.99 for Plastic Surgery. Such disparity can be explained by the type of surgical cases performed in each specialty. While in Eye Surgery, many cases are short (i.e., refractive surgery), Plastic Surgery cases require more time and typically comprise several surgeries.

Finally, Fig. 5 concludes the analysis with the feature importance plots. For each specialty, we identify the salient features of the specific model. In line with the findings of the general model, the most important features are the historical averages, both for the surgeon and the OMC code. In addition, team-related characteristics significantly affect the model's overall performance in each specialty. Especially prominent are the leader and team familiarity and the partner diversity exposure. Finally, we notice that gender-related features, such as gender diversity and the gender of some of the team members, also have a non-negligible effect on the final performance.

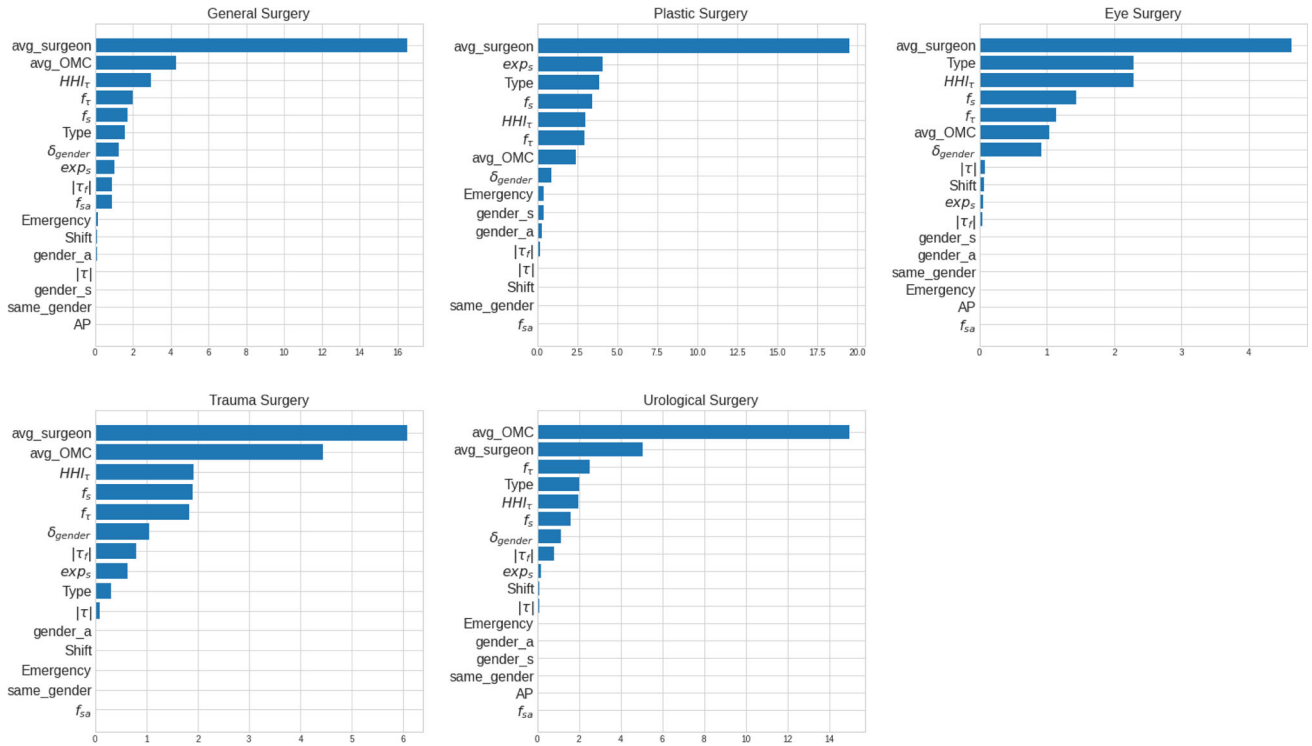



**Fig. 5** Feature importance per specialty. All the measures are provided with respect to the change in the MAE

We notice that, between the average time of the surgeon and the average time of the OMC code, the former is always more important than the latter, with a single exception. The specialty *Urological Surgery* presents an inversion of salience in that the `averarage_OMC` has a higher impact on the MAE than the `average_surgeon`. After interviewing the surgery management team at the hospital, we discerned two explanations for these results. First, the number of robotic surgeries in Urology tends to be generally higher than the number of robotic surgeries in other services. Thus, procedures in urology tend to be more "standardized" than in different areas. Consequently, the average duration of a specific surgery is a more important predictor than the average of that surgery for a particular surgeon. The same is not valid in all the other specialties, where using historical data from a specific surgeon seems to be of primary importance. Second, most Urologists in this hospital are full-time surgeons, while the other specialties are served mainly by external surgeons who work at other institutions. Consequently, there are fewer surgeons in Urology than in different specialties; hence, the variability of `average_surgeon` is smaller than in other specialties.

We also noticed that eye surgery significantly differs from the other surgical specialties. The second most relevant feature is the type of surgery (i.e., inpatient vs. outpatient cases), probably due to the differences in duration between refractive eye surgery and other types (i.e., retina, eye muscle, or glaucoma). Prior studies have also found structural duration differences among surgical specialties [12]. Some recommendations can be drawn from these results, such as implementing dedicated, i.e., specialty-specific predictive models and scheduling algorithms.

## 5 Conclusions and discussion

This paper addresses the challenge of forecasting surgery duration at a general mid-size urban hospital. The broader aim of the study is to develop reliable forecasts of surgery times, which will, in turn, be provided as input for a robust optimization of operating room scheduling formulation.

We performed feature engineering to encapsulate team performance and dynamics, creating variables for team familiarity, surgeon-team familiarity, gender diversity, and experience variety based on around 60,000 historical observations. We also incorporated cooperation measures influenced by gender distribution within team roles. Testing various ML models, we improved surgery duration forecasts, comparing our results against baseline models that emulate the approach currently used by decision-makers. In addition, we evaluated the impact of the newly engineered features on the model accuracy. Our findings highlight that including these features improves forecast accuracy, reducing the MAE by 47% compared with a model that does not use those features. In addition, by analyzing the importance of the feature, we contribute to a more effective surgery team assignment.

The study also revealed that the surgical specialty significantly influences surgical duration disparities, as observed by previous studies. For example, Urology is the only specialty for which the average surgery duration is the most relevant predictor, surpassing the surgeon's average. In Eye Surgery, the distinction between inpatient and outpatient procedures emerges as a critical factor. These results underscore the necessity for specialty-specific modeling.

Our study makes several contributions to the field of surgery duration forecasting. First, we demonstrate the value of incorporating team-related factors, such as surgeon experience and team familiarity, into predictive models. This finding highlights the importance of considering individual surgeon characteristics and broader team dynamics in surgical settings. Second, we introduce a novel set of predictors that capture organizational and behavioral features, such as surgical team cooperation dynamics, team member roles, and gender. These predictors significantly improve the accuracy of our ML models, leading to a 24% reduction in MAE over the baseline models, which mimic the approach currently used by the decision maker. Third, our study has important implications for hospital efficiency and patient outcomes. Hospitals can better allocate resources, reduce wait times, and improve patient satisfaction by accurately forecasting surgery case duration. Our findings suggest that investing in team training and professional development programs may lead to more efficient and effective surgical teams.

Our approach was able to estimate surgery duration accurately, but it has some limitations. First, the dataset was limited to a single hospital, which may limit the generalizability of our findings to other hospitals with different patient populations and surgical practices. Second, we did not have access to patient-level information, such as electronic health records (EHRs). This information could have provided additional insights into the factors influencing surgery duration, such as patient comorbidities, surgical complexity, and the surgeon's experience. To mitigate the lack of this critical information, we used an ordinal variable comprising the patient's health status and the complexity of the surgical intervention. Third, we did not collect data on other important factors, such as the affiliation and demographics of the surgical team members. For example, surgeons and anesthesiologists who work in multiple hospitals may have different surgical practices than full-time employees of a single hospital. Additionally, clinicians' age, training, and research activity may influence surgery duration. These limitations should be considered when interpreting the results of our study.

Future studies should consider gathering data from multiple hospitals, obtaining patient-level information, and collecting data on other essential factors that influence the duration of surgeries. Additionally, there is potential to explore the performance of mixture models and kernel-based methods further in this context. Furthermore, to achieve the broader objective of this research, we plan to investigate how predictive models can be integrated into a robust optimization approach to resolve the operating room scheduling problem.

## Appendix

### Statistical comparison of alternative ML models

We present a detailed statistical analysis to determine whether the tested ML models produce statistically different results. In line with the recommendations of Demsar [36], we carried out a statistical analysis aimed at asserting (i) whether the ML methods produce different forecasts and (ii) whether we can define a ranking of the different methods. With this goal in mind, we performed the following two-step analysis:

1. We first ran a Kruskal-Wallis test [37], to spot differences across methods. This test is typically used when more than two independent groups are compared. The test's null hypothesis is that the results of those algorithms are indistinguishable, *i.e.*, they all provide similar MAE values, and therefore, the ranking should be randomly distributed.
   Table 8 presents the results of the Kruskal-Wallis tests on the five methods. In the table, the first column provides the size of each sample (10 runs per method); the second column provides the degrees of freedom of the test; column three reports the $\chi^2$ statistic; and, finally, the last column provides the p-value. The null hypothesis is rejected when the p-value is below $\alpha = 0.05$. Therefore, we observe enough evidence to claim that the different methods produce significantly different results in terms of MAE.
2. Since we rejected the null hypothesis of the Kruskal-Wallis test, we ran a post hoc analysis. As suggested in Demsar [36], we use the Nemenyi test [38] to compare each method with the others, thus detecting for which methods there exists a statistically significant difference in terms of ranking. The performance of the two groups is significantly different if the corresponding average ranks differ by at least a "critical difference," computed as:

$$CD = q_\alpha \sqrt{\frac{k\,(k+1)}{6N}}$$

where the critical value $q_\alpha$ is obtained from [36]. Specifically, $k = 5$ provides the number of configurations used, and $N$ is the total number of observations in each class, $N = 10$. From [36], we obtain that, with a level of confidence of 95%, $q_\alpha = 2.728$ and, therefore, $CD = 1.92$. The results of the two algorithms are thus significantly different if their corresponding average ranks differ by at least the critical distance CD. (See [36] for more details.) From Fig. 6, we observe that `XGBoost` is significantly different from `Ridge`, `LightGBM`, and `SVM`. On the other hand, `XGBoost` and `RF` are not significantly different at an $\alpha$ level of 5%.

**Table 8** Results of the Kruskal-Wallis test on the MAE of the different algorithms

| $N$ | df | $\chi^2$ | p-value |
|---|---|---|---|
| 10 | 4 | 45.66 | <0.001 |

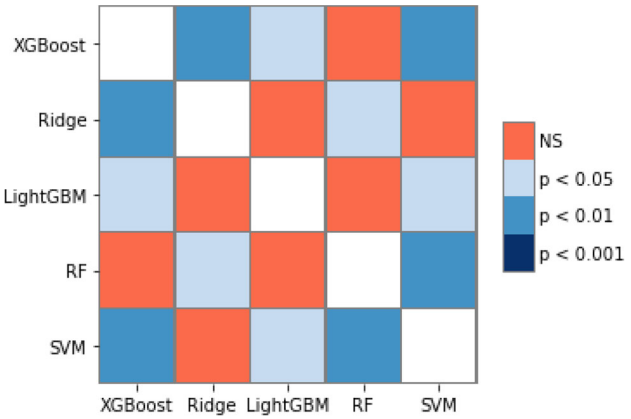



**Fig. 6** Nemenyi matrix of significant results. The color code indicates the $p$-value of the corresponding two-tail Nemenyi test

**Author Contributions** The authors have contributed equally to this work.

**Funding** The authors did not receive support from any organization for the submitted work.

**Code availability** The authors make the code available in a public repository, as indicated in the body of this manuscript.

## Declarations

**Conflicts of interest** The authors have no competing interests to declare relevant to this article's content.

**Ethics approval** We certify that ethics approval was not needed for this study under our institution's policies.

**Consent to participate** NA.

**Consent for publication** NA.

## References


1. Shrank WH, Rogstad TL, Parekh N (2019) Waste in the us health care system estimated costs and potential for savings. JAMA 322(15):1501–1509
2. Gillespie BM, Chaboyer W, Fairweather N (2012) Factors that influence the expected length of operation: results of a prospective study. BMJ Qual Satis 21:2–12
3. Bovim T, Christiansen M, Gullhav A, Range M, Hellemo L (2020) Stochastic master surgery scheduling. Eur J Oper Res 285(2):695–711
4. Minehart R, Foldy E, Long J, Weller J (2020) Challenging gender stereotypes and advancing inclusive leadership in the operating theatre. Br J Anaesth 124(3):148–154
5. Jones L, Jennings B, Higgins M, de Waal F (2018) Ethological observations of social behavior in the operating room. Proc Natl Acad Sci USA 115(29):7575–7580
6. Spence C, Shah OA, Cebula A, Tucker K, Sochart D, Kader D, V., A. (2023) Machine learning models to predict surgical case duration compared to current industry standards: scoping review. BJS Open 7(6)
7. Zhu SW, Fan WJ, Yang SL, Pei J, Pardalos PM (2019) Operating room planning and surgical case scheduling: a review of literature. J Comb Optim 37(3):757–805
8. Rahimi I, Gandomi AH (2020) A comprehensive review and analysis of operating room and surgery scheduling. Arch Comput Method Eng
9. Cardoen B, Demeulemeester E, Belien J (2010) Operating room planning and scheduling: A literature review. Eur J Oper Res 201(3):921–932
10. Karakaya Z, Tatar B (2023) Technological trend analysis for surgical operation duration estimation. The Eurasia Proceedings of Science Technology Engineering and Mathematics
11. Bellini V, Guzzon M, Bigliardi B, Mordonini M, Filippelli S, Bignami E (2019) Artificial intelligence: A new tool in operating room management role of machine learning models in operating room optimization. J Med Syst 44(1)
12. Shahabikargar Z, Khanna S, Sattar A, Lind J (2017) Improved prediction of procedure duration for elective surgery. Integ Connec Care 239:133–138
13. Bartek M, Saxena R, Solomon S, Fong C, Behara L, Venigandla R, Velagapudi K, Lang J, Nair B (2019) Improving operating room efficiency: Machine learning approach to predict case-time duration. J Am Coll Surg 229(4):346–354
14. Tan K, Francis N, Ang B, Gan J, Lam S (2019) Data-driven surgical duration prediction model for surgery scheduling: A case-study for a practice-feasible model in a public hospital. In: 15th IEEE International Conference on Automation Science and Engineering (IEEE CASE). IEEE International Conference on Automation Science and Engineering. pp 275–280
15. Zhao BQ, Waterman RS, Urman RD, Gabriel RA (2019) A machine learning approach to predicting case duration for robot-assisted surgery. J Med Sys 43(2)
16. Huang C, Lai J, Cho D, Yu J (2020) A machine learning study to improve surgical case duration prediction. medRxiv. https://doi.org/10.1101/2020.06.10.20127910
17. Soh KW, Walker C, O'Sullivan M, Wallace J (2020) An evaluation of the hybrid model for predicting surgery duration. J Med Syst 44(2):42
18. Yuniartha D, Masruroh N, Herliansyah M (2021) An evaluation of a simple model for predicting surgery duration using a set of surgical procedure parameters. Infor Med Unlocked 25:100633
19. Ng N, Gabriel R, McAuley J, Elkan C, Lipton Z (2017) Predicting surgery duration with neural heteroscedastic regression. In: Machine Learning in Health Care
20. Kayış E, Khaniyev T, Suermondt J, Sylvester K (2015) A robust estimation model for surgery durations with temporal, operational, and surgery team effects. Health Care Manag Sci 18(3):222–233
21. Laskin D, Abubaker O, Strauss R (2013) Accuracy of predicting the duration of a surgical operation. J Oral Maxillofac Surg 71(2):446–447
22. Luangkesorn KL, Eren-Dogu ZF (2016) Markov chain monte carlo methods for estimating surgery duration. J Stat Comput Simul 86(2):262–278
23. Jiao Y, Sharma A, Ben Abdallah A, Maddox T, Kannampallil T (2001) Probabilistic forecasting of surgical case duration using machine learning: Model development and validation. J Am Med Inform Assoc 27(12):1885–1893
24. Tuwatananurak J, Zadeh S, Xu X, Vacanti J, Fulton W, Ehrenfeld J, Urman R (2019) Machine learning can improve estimation of surgical case duration: A pilot study. J Med Sys 43(3)
25. Blau P (1977) Inequality and heterogeneity: A primitive theory of social structure. Free Press, New York
26. Harrison D, Klein K (2007) What's the difference? Diversity constructs as separation, variety, or disparity in organizations. Acad Manag Rev 32(4):1199–1228
27. Huckman R, Staats BR (2011) Fluid tasks and fluid teams: The impact of diversity in experience and team familiarity on team performance. Manuf Serv Oper Manag 13(3):310–328
28. Akşin Z, Deo S, Jónasson J, Ramdas K (2021) Learning from many: Partner exposure and team familiarity in fluid teams. Manage Sci 67(2):854–874
29. Huckman R, Staats BR, Upton DM (2009) Team familiarity, role experience, and performance: Evidence from indian software services. Manage Sci 55(1):85–100
30. Staats B (2012) Unpacking team familiarity: The effects of geographic location and hierarchical role. Prod Oper Manag 21(3):619–635
31. Hoerl A, Kennard R (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1):55–67
32. Breiman L (2001) Random forests. Mach Learn 1(45):5–32
33. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T (2017) LightGBM: A highly efficient gradient boosting decision tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds.) Advances in Neural Information Processing Systems, vol. 30. pp 3149–3157
34. Chang C, Lin C (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(27):1–27
35. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, NY, USA, pp 785–794
36. Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
37. Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47(260):583–621
38. Nemenyi P (1963) Distribution-free multiple comparisons. PhD thesis, Princeton University
39. Friedman J (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232