Machine Learning to Predict Surgery Duration: Towards Implementing AI and Digital Twin for Effective Scheduling

Madhu Sudan Sapkota
Knowledge Transfer Partnership Programme
University of Essex
Colchester, UK
ms23436@essex.ac.uk

Faiyaz Doctor
School of Comp. Science & Electr. Engineering
University of Essex
Colchester, UK
fdocto@essex.ac.uk

Hugo Herrera
Dept. of Business Informat. & Analytics
ESNEFT
Colchester, UK
hugo.herrera@esneft.nhs.uk

Michael Kampouridis
Sch. of Comp. Science & Electr. Eng.
University of Essex
Colchester, UK
mkampo@essex.ac.uk

Xinan Yang
Sch. of Maths., Stat. & Actuarial Science
University of Essex
Colchester, UK
xyangk@essex.ac.uk

***Abstract*—The COVID-19 pandemic has exacerbated the backlog of Referral-to-Treatment (RTT) patients awaiting surgery, presenting a significant challenge within the UK's National Health Service (NHS). As waiting times continue to rise, optimising surgical theatre planning becomes crucial for effective patient care.**

Traditionally, scheduling relies on plan-makers' subjective estimates or historical averages, leading to inefficiencies such as surgery cancellations or underutilisation of resources. Machine learning (M/L)-based predictive algorithms offer a promising solution by leveraging data-driven models to forecast surgical times more reliably. However, their application in NHS hospital settings remains limited. This also causes restriction for the broader adoption of Digital Twin (DT) technology and Artificial Intelligence (AI) within the healthcare area, despite their significant potential in today's era.

This study explores the implementation of multiple M/L algorithms for surgical time estimation for Trauma and Orthopaedics related procedures in an NHS Trust hospital. Results indicate that Neural Networks, along with ElasticNet, Gradient Boosting, and Bayesian Ridge models, demonstrate robust performance. Additionally, expansion of modelling to procedure-specific is adopted where models are built separately for each surgical procedure. The consideration of procedure-specific modelling is promising.

The study contributes insights into the integration of M/L algorithms into healthcare digital resources, paving the way for enhanced surgical planning strategies. Future research will focus on integrating the predictive models into a comprehensive framework (referred to as a Digital Twin) for simulation and optimisation-driven automated decision-making.

***Index Terms*—Surgical Time Prediction, Theatre Planning, Machine Learning, Artificial Intelligence, Health-Care Digital Twin**

This work is an outcome of a Knowledge Transfer Partnership (KTP) between the University of Essex and East Suffolk & North Essex NHS Foundation Trust (ESNEFT), with funding provided by ESNEFT.

I. Introduction

The backlog of Referral-to-Treatment (RTT) waiting patients requiring surgery has been exacerbated by the COVID-19 pandemic [1], and continues to present a formidable challenge. In the UK, regardless of the National Health Service (NHS), being an efficient health-care system, the number of waiting patients requiring some sort of surgery and waiting more than a year continues to increase [2]. With such increasing cumulative waiting times for patients requiring surgery, the importance of surgical-theatres and their effective planning becomes even more pronounced.

The availability of surgical theatre's and their planning, which play a crucial role in hospitals for patient care and well-being, are often constrained, since they demand a higher percentage of the total allocated hospital costs [3]. Any enhancement in the utilisation of the theatres and related resources directly affects patients waiting time, especially in the context of elective surgeries. Consequently, cost-effective surgical planning becomes imperative to ensure judicious utilisation of available and/or affordable theatres' associated resources, aiming at reducing the RTT list [4].

It is undoubtedly accepted that the overall effectiveness of theatre planning/ Scheduling mostly relies upon the estimation of surgical duration for the planned surgical procedure(s) [5]. Accurate schedules hinge on reliable estimations of how long each surgery will take. However, scheduling in most hospitals relies on estimations from surgeon and/or averages of historical procedure-time duration [6]. This dependencies have limited accuracy, since they don't fully consider broader range of features and variables that could affect the surgical procedure time. Thus, currently used methods of planning have higher chances of leading to the major issues: surgery

cancellations due to overrun time in previous procedures or under-utilisation of theatres when procedures performed in less time than was allocated [7]. Under-utilisation means wasted resources causing revenue-loss, while, cancellations leads to patient's disappointment towards the hospital setting. Therefore, the solution lies in improved surgery duration estimation. With more accurate forecasting of procedure time, hospitals can create optimal operating room schedules. This translates to several benefits: efficient resource utilisation, increased surgical capacity and optimised case arrangement.

In the evolving landscape of healthcare, the role of machine learning (M/L) based predictive algorithms has been receiving significant attention for estimating surgical times [5, 8, 9]. Time-predictive models built using these algorithms have the potential to enhance decision-making by providing accurate estimates of procedure times during surgical schedule planning. However, the use of M/L algorithms and their derived tools for procedure time estimation is not sufficiently exploited and remains very limited within NHS-related hospital settings in the UK.

This limitation, in turn, restricts the broader adoption of Digital Twin (DT) technology and Artificial Intelligence (AI), which hold significant promise for revolutionising healthcare [10]. DTs can create highly accurate virtual replicas of physical systems, allowing for advanced simulations and real-time optimisations. AI, when integrated with DTs, can provide personalised treatment plans, predict patient outcomes, and optimise resource allocation. These technologies are particularly effective in optimising surgical scheduling and enhancing personalised medicine. Effective exploitation of DT and AI in healthcare heavily relies on M/L algorithms and data-assisted analysis and prediction for activities involving human intervention [11].

The rest of the paper is structured as follows: Section II discusses approaches for estimating surgical durations in hospital settings and examines trends of predictive modelling in surgical time estimation. Section III outlines the steps this research-work follows for the task such as data collection, processing, and feature selection. Section IV details the model-building process, including the algorithms and approaches used or proposed. Section V presents empirical results from different model types, compares them and provides a detailed empirical analysis of the results. This includes evaluation of the performance of different models to identify the most reliable approaches for surgical time estimation. Section VI concludes the paper and outlines future research directions.

II. Background

Studies have been conducted, within the realm of predicting or forecasting the surgical duration, leading to some significant level of developments in this area. Distribution fitting model can be considered as first stream attempt in the estimation of procedure time and, applied, to a certain extent for predicting time, based on fitted distributions [12, 13].

The subsequent advancement in predictive modelling involves statistical modelling, wherein factors influencing surgical duration(s) are assigned with relative importance to create a simpler linear regression model. Eijkemans et al. [5] developed a multi-variable linear regression model considering specific estimators, including surgeon's time estimate, surgery-specific features, surgeon team features, and patient-level features. They examined the effects of these features by adding them individually or in groups to the base model, which solely comprises procedure-related information as an input feature. The study aimed to identify the impact of these additions on the overall accuracy (R^2) of the model. The findings suggest that team characteristics play a crucial role in influencing the duration of surgical procedures, with surgeon's estimates of team characteristics being the most influential factor. In contrast, patient characteristics have a limited impact on procedure duration. However, using surgeon's time estimates as input features may hinder practical replicability in this context.

Kayis et al. [14] attempted to harvest fittings model's predictions by adding features to the base fitting model which is the mean value of the last 5 particular procedures' duration. The Last 5 particular procedures' duration is considered such a way that the particular surgeon has performed at least 5 surgeries of the particular type in the last year, and otherwise the estimates are generated from all the same type of procedures performed by any surgeon. They trained an Elastic Net regularised linear regression with added other features considering the Last 5 estimate as the core input feature. The added features as temporal features were time of day, day of week, etc.), and further operational features (Theatre room assignment, etc.). Though, having mean of previous procedure-related duration as input features maximises practicality in contrast to the modelling with surgeon's estimate as input features [5], the Last 5 mean feature creates ambiguity for features' importance analysis. This is because most or all of the features already affect the procedure time for the last 5 selected procedures.

ShahabiKargar et al. [15] on other side, has avoided both of the above case i.e., primarily estimated procedure time as input features, and also focused on practically available pre-operative surgical data. An exhaustive review of available information sources was made to select the potential predictors together with discussions with clinical experts and hospital administrators.

In the last decade, M/L techniques beyond linear regression, such as Decision-tree-based, Deep-learning-based, Bayesian-approaches or hybrid models, have been implemented for theatre time forecasting [8, 16, 17, 18]. This is primarily driven by their effectiveness in handling outliers and missing data, as well as their capability to model non-linear relationships. Master et al. [8] conducted a study where they trained multiple decision-tree-based M/L models, including random forest regression, gradient-boosted regression trees, and hybrid combinations. The goal was to achieve more robust predictions compared to linear regression models. In the study of Sahadev, Lovegrove, Kunz [18], both RF and xGboost regression model performed with no significance difference in performance results for the given orthopaedic-related elective surgery dataset from 'East Kent Hospitals University NHS

Foundation Trust'. Likewise, Bartek et al. [19] evaluated two decision-tree based models Random-forest (RF), and xGBoost, with xGboost outperforming RF model predicting with the accuracy of 50% considering $\pm 10\%$ flexibility for the duration requirements to be acceptable.

Regarding the data and domain of interest in this study-area, there have been different trends observed in the research area. Some researchers tend to incorporate multiple specialties into a single model [14, 15], while others deal separating them [17] or focus on a particular specialty [18, 20]. Similarly, many research studies has filtered the top-performing procedures from one or more specialties [8, 18], to establish the role of machine learning (M/L) in forecasting. Having the model trained on only selected procedures, however, has limitation in obtaining insight covering broader range and also not enough for practical application.

Other than above, next prevailing trend involves modelling for selected procedure(s) only that is/are both frequent and crucial for most hospitals [9, 21, 22]. Being procedure-specific, this approach enables a robust analysis of the model avoiding generalisation among the procedures. Upon achieving a significant performance level, it facilitates planning for these specific procedures with reduced risks, thereby maximising the chances of practical replicability. For example, [9] developed machine learning models to predict the duration of Total Knee Arthroplasty (TKA). In their study, the neural network-based model outperformed the tree-based model. However, in reality, the planning team for any specialty require to plan for a broader range of procedures, which is a limitation for such single procedure specific related time-estimation. This limitation can be mitigated by expanding the procedure-specific modelling concept for multiple procedure categories.

In summary, a range of statistical to machine learning techniques have been implemented in predicting surgery time, where most of the research efforts outperform existing hospital's manual estimation methods. However, these models still holds considerable predictive errors, which is why they lack confidence in practical applications. This can be attributed to factors such as the lack of benchmarks for practical pre-surgery viable features affecting procedure time, the limited balanced availability of surgery related data from individual hospital settings required to train a complex model, limited implementation and analysis of algorithms due to restricted access to inter-hospital/trust-related data, and the inherent complexity of surgical procedures. Also, the question on effectiveness of building single (holistic) or multiple specialty/procedure specific models remains unanswered: should all procedure or specialty data be combined for model-building to increase training data volume with generalisability on variables' contribution on procedure-time, or should they be separated which will decrease training data volume for each model but allows for the exploration of procedure/specialty-specific contributions of variables?

In this study, we have focused on a broader range of machine learning (M/L) algorithms and different approaches, leveraging real-world data from a NHS trust. The contributions of this study, are articulated as follows:

- We investigate a diverse array of machine learning (M/L) algorithms that are or can be utilised to predict surgical duration, aiming for a comprehensive evaluation of their performance.
- The study highlights the ranges of surgical-procedures, offering insights into the variability and complexity inherent in surgical scheduling.
- Our use of real-world data from an NHS Trust enhances the study's contribution by providing a realistic and context-rich dataset.
- Acknowledging the importance of feature selection, we follow a thorough process of feature selection focusing on pre-surgical variables. The rigorous approach ensures a proper balance of input features with the data available for model training without compromising predictive accuracy.
- By implementing procedure-specific modelling, we assess its impact on predictive accuracy, offering insights into its effectiveness compared to generalised models.
- The findings aim to establish benchmarks for procedure-time estimation within specific surgical contexts (specialty or procedure-wise), paving the way for future research into multi-procedural analyses across various specialties.

III. Methodology: Data Collection and Modelling Pipeline

Figure 1 outlines the overall model-building related steps which are later discussed in this and later sections. The step includes: data collection and pre-processing, feature selection, model(s) selection, model(s) training and performance testing of the models.

A. Data Collection and pre-processing

This study explores the data associated to elective surgical procedure at Colchester Hospital under East Suffolk and North Essex NHS Foundation Trust (ESNEFT), UK. The focus of this study in this stage is the specialty of "Trauma and Orthopaedics (T&O)", therefore, data was retrieved being limited to the specialty. Limiting to a specific specialty is motivated by the goal of assessing whether building multiple models for each procedure and/or sub-specialty would be more effective than having a single holistic model for all specialties (or procedures). Likewise, the goal is to predict times for elective procedures, which is why emergency surgeries are not considered.

The dataset covers details of planned and performed elective surgical procedures between Jan 1st, 2019, and May 30th, 2023 (4 years). Records with missing, inconsistent, or duplicate data are then removed. Additionally, procedures performed less than 5 times during the timespan or unassigned to procedure codes, are excluded. Outliers are then identified and removed on a procedure-wise basis since the overall procedure time range from 10 to 300 minutes, and generalisation of outliers could still leave corrupt data un-noticed. For example, data

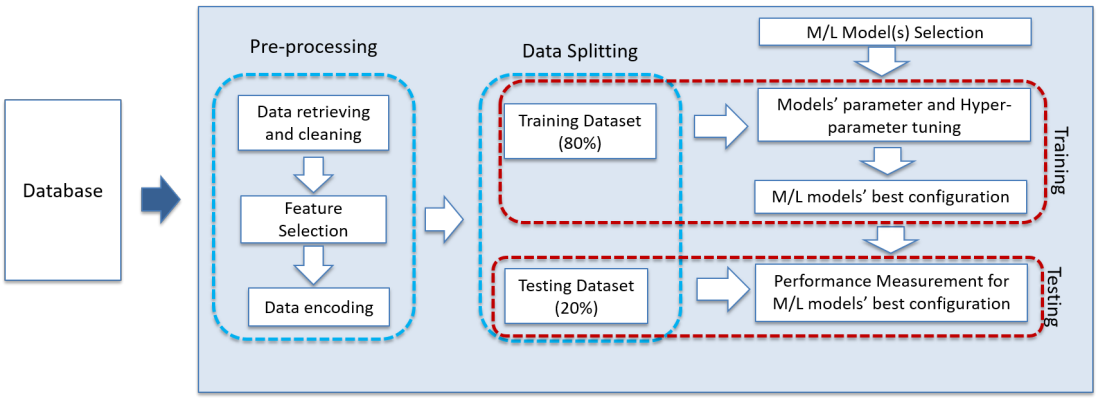


Fig. 1. Steps showing the data collection, model building and validation process

may contain noise due to mismatches between procedure codes and durations. For the overall dataset, the accepted maximum z-score $\left(\frac{|\text{value}-\text{mean}|}{\text{std}}\right)$ is set to 2.75, while for each procedure category data, it is set to 1.75. Data points not meeting this z-score criterion are considered outliers, which may result from data corruption or other untraceable causes leading to such extreme patterns. The z-score values were chosen after a pre-analysis of the data, considering representative procedure-specific data, where outliers (not-possible, rare exceptional) were identified through both visual and analytical methods.

Approximately 24% of the cases from the original dataset, were excluded based on the criterion outlined in Figure 2 . With this elimination and filtering, the dataset count retrieved

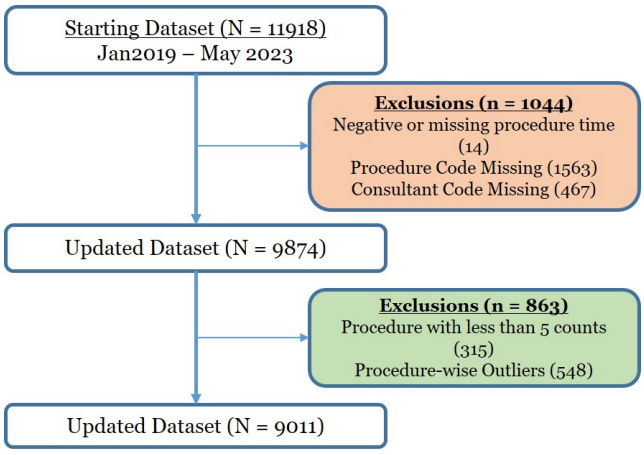


Fig. 2. Dataset cleaning and filtering steps with given data count for those steps: Exclusions of unfit data and filtering on remaining in order to have better model's performance

for the development and evaluation of the predictive model is 9011, encompassing 163 procedural categories.

Categorical features data are then one-hot-encoded, using the tool available under pandas and sklearn packages. One-hot encoding is a technique used to convert categorical variables into a form that can be provided to machine learning algorithms. Each category value is converted into a new binary column, where an entry has a value of 1 if the original categorical variable had that category, and 0 otherwise. Missing categorical features within the feature other than 'Procedure Code' and 'Consultant Code' are termed as 'Unknown' and kept.

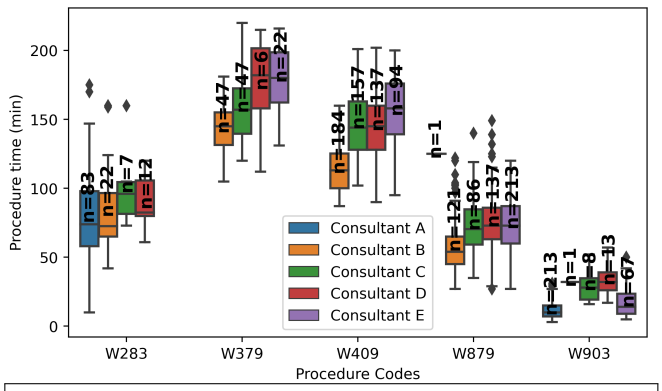
From the final dataset, 80% will be used for training the model(s) while 20% will be utilised for model's performance testing.

B. Target Variable and Selecting Input Features

The duration of a surgical procedure can be understood as either the time during which surgical tasks are performed or the entirety of the time that a patient occupies the operating theatre (i.e., from entering to leaving). In this study, this duration is equivalent for the time-range during which the patient is under assessment by the theatre staff, starting from pre-anesthesia and ending when the patient leaves or is taken out of the theatre. This time duration, which will serve as the target variable for prediction, is commonly referred to as '*H4 Minutes*' at ESNEFT.

For input features selection, a list of potential predictors (features) of procedure time were identified initially by reviewing relevant literatures. Beyond this, the features initially considered for selection were also motivated by pre-analysis of the data, suggestions from consultants, or analytics involved in the area. For example, based on related research and pre-analysis, two new features were considered in this study: *Theatre Area*, which categorises theatres based on location and facilities, and *Covid-flag*, which represents the pandemic period. The pre-analysis includes procedure timing (*H4 Minutes*) comparison across potential features associated data. For instance, a pre-analysis involving data categorisation based on procedure type and consultant code, as illustrated in Figure 3, demonstrated higher variability in actual procedure times between the procedure categories. Then, again within each procedure type, variability can be observed among the consultants involved in the surgery. This analysis hinted the procedure time is very much influenced by the procedure type itself and the consultant involved.

The similar type of variability can be observed with the current approach of planning at ESNEFT, depicted in the first



W283: Removal of internal fixation from bone nec
W379: Total prosthetic replacement of hip joint using cement unspecified.
W409: Total prosthetic replacement of knee joint using cement unspecified.
W879: Diagnostic endoscopic examination of knee joint unspecified
W903: Injection of therapeutic substance into joint

Fig. 3. Box-plot showing procedure times (in min) categorically for the selected procedure categories (related codes and description presented in the box) and also further segmented by consultants' code within each category (n representing total counts after data-categorisation)

sub-figure of Figure 7 (discrete nature of scatter-plot in vertical y-axis, between and also among procedures). This variation in planning for the same procedure arises from differences in the individual consultants' planning approaches. Suggestions about incorporating past experiences of consultants (or surgeons) and also of the theatre team as input variables to improve the model's performance can be found in research-study [23, 24] . However, due to the limited available data, this study will only consider the consultant's name (anonymous code) and anaesthetists' count for inclusion.

Furthermore, in order to capture additional attributing features, experimental based refining was performed. This process involved employing the forward method of feature selection [25], where features (variables) are progressively included with greedy search (best performance) strategies. This experiment-based analysis was carried out using two distinct model types—elastic regression and decision-tree-based models. The objective was to ensure the capture of significant features while avoiding redundancy. Clinical case related features (except the procedure type) are avoided with the focus to operational and temporal data that could be typically available in advance while scheduling. This is to maximise the practical replicability of the approach, i.e., utilising the final product (model) on planning theatres in advance.

The following are the final features selected after refining the listed possible variables (grouped based upon their characteristics):

- **Patient characteristics:** *Age-group, Gender, Obesity*
- **Operation characteristics:** *Procedure type (Code), anaesthetist expected?, Number of Procedures*
- **Team and Theatre characteristics:** *Consultant Code, Anaesthetist count, Theatre Area*
- **Temporal Characteristics:** *Day of the week, Covid-flag*

The selected features mostly contain categorical string data, while a few, such as *Number of Procedures* and *Anaesthetists Count*, have numerical continuous integer values. However, due to the limited range of values for these latter two features, they were also treated as categorical for consistency. The following steps were taken to one-hot encode the categorical features into a format suitable for machine learning models:

- **Patient characteristics:** *Age group* was encoded into multiple binary features representing different age ranges, such as *Age group_5to15*, *Age group_15to30*, and so on up to *Age group_71to80*. Gender was encoded into binary features *Gender_Male* and *Gender_Female*, indicating the patient's gender. Obesity status was represented by *Obesity_Yes* and *Obesity_No*, indicating whether the patient is obese or not.
- **Operation characteristics:** For considered 163 different procedure codes, each were encoded as a binary feature such as *Procedure Code_W409*, *Procedure Code_W903*, *Procedure Code_W879* and so on. The feature indicating if an anaesthetist was expected was encoded into *Anaesthetist Expected?_Yes* and *Anaesthetist Expected?_No*. The number of procedures was categorised into discrete binary features like *N_procedures_1*, *N_procedures_2* and *N_procedures_3*.
- **Team and Theatre characteristics:** Each of consultants (total total count 24) were encoded with anonymisation into binary features such as *Consultant_A*, *Consultant_B*, and so on. The *Anaesthetist Count* data were again encoded into categories such as *Anaesthetist count_1* and *Anaesthetist count_2*. Theatre areas were one-hot encoded into binary features like *Theatre Area_1* and *Theatre Area_2*.
- **Temporal characteristics:** Each day of the week is one-hot encoded into binary features such as *Day_Monday*, *Day_Tuesday* and so on. The Covid-flag is encoded into *Covid Flag_pre-covid*, *Covid Flag_covid* and *Covid Flag_post-covid*, indicating whether the surgery took place before, during or after the Covid pandemic.

IV. Models Development

This section details the process of models' development including the selection of multiple algorithms.

A. Models and Algorithms Selection

Multiple Machine Learning (M/L) algorithms were considered for the model development including: linear regression, decision-tree-based, Bayesian-based, and Deep learning (Neural-Network). This selection encompasses the diverse range of algorithms and tools commonly utilised within this research area (as discussed in Section 2), and allows us to conduct a thorough examination of which techniques are most suitable for the specific problem (theatre procedure time estimation).

Elastic Net [26] which combines the L_1 and L_2 penalties to minimise both the sum of absolute values and the sum of squares for the metrics was utilised for linear regression.

Two different decision-tree-based algorithms were considered following the trend set by decision-tree-based regression models: Standard Decision-tree, which are simple yet powerful models used for both regression and classification tasks, and Gradient Boosted Trees, ensemble methods that sequentially combine multiple weak learners to build a strong predictive model [27]. Additionally, Bayesian Ridge regression, a linear regression model that introduces regularisation, is considered. Bayesian algorithms are adopted in many areas due to their ability to overcome over-fitting by placing a prior distribution over the model parameters, making them particularly useful for handling multi-collinearity and noisy data [28]. The algorithm considered from the deep learning model perspective was a simple Feedforward Neural Network (FNN) architecture, as FNNs are well-suited for regression tasks where the goal is to predict continuous values based on input features. It consists of multiple layers: one input layer, one or more hidden intermediate layers, and an output layer.

Python-based scikit-learn (sklearn) [29] and PyTorch [30] packages were chosen as the primary modeling tools. Within scikit-learn, modules ElasticNetRegressor (from *linear_model*), DecisionTree (from *tree*), GradientBoosting (from *ensemble*), and BayesianRidge (from *linear_model*) were utilised to construct above discussed models. Similarly, the PyTorch nn module was employed for the implementation of the FNN model.

Within each type of model, two distinct approaches were undertaken. The first approach considers Procedure Code feature together with remaining selected input features (Section III-B) comprehensively to create a holistic model. In the second approach, a modular approach to modelling is adopted, focusing on categorised procedure related data so that Procedure Code feature get removed.

B. Holistic and Procedure Specific Categorical Modelling

The holistic model encompasses all data from T&O Specialty under the categorical feature '*Procedure Code*', providing a comprehensive overview of the specialty's data. This comprehensive approach generalises the effect of other features, regardless of procedure type, in procedure-time prediction which has both pros and cons. It can be beneficial for procedures with less data, as it helps establish the effect of features (procedure code as well as other) on procedure timing; however, this generalisation may obscure the unique impact of the other features among different procedures.

While, the all-procedures inclusive holistic model(s) considers all data under the categorical feature '*Procedure Code*' collectively, the categorical modelling takes a more nuanced approach. It involves separating unique procedures and constructing models individually for each procedure, given a threshold of minimum data count. Since, procedure-specific modelling is another less explored topic of research as discussed in Section II, this will further allows for a more granular understanding of the predictive power of the procedure specific-models. Moreover, it allows us to study the unique impact of selected input variables on the procedure duration for each specific procedure, thus suggesting higher potential of the approach in procedure time prediction considering the varying nature of each procedure. However, there will be data limitations to implement the approach for all the procedures.

For the eligibility of procedure-specific categorical modeling, we set a threshold of a minimum of 200 occurrences in the available data. This means that only procedures that have been performed at least 200 times within the specified time range and have at least 200 data points available after pre-processing (Section III-A) would be considered. Using this criterion, 11 procedures were deemed eligible for the categorical modeling approach (This includes procedures presented in Figure 3).

To conduct a comparative analysis with the overall (holistic) model, we will evaluate the performance of the overall model on the selected procedures using the specified performance metrics. This assessment aims to provide insights into how the categorical models for the filtered procedures perform in comparison to the holistic model.

C. Model Training along with Hyper-parameter Tuning

Single designated or multiple algorithms were applied on training each model type, in order to capture the underlying relationships between input features and the target variable. The goal of model training would be minimising the discrepancy between predicted and actual target values by iterative adjustment of prediction related parameters. For the situation where multiple choices of algorithms or model's setting as training options be available, the options were passed as hyper-parameters. For each model, along with their relevant set of chosen hyper-parameters, training the model(s) with the provided training data involves the following steps:

- The Elastic Net model was trained by adjusting the coefficients for input features, balancing between $L1$ and $L2$ regularisation.
- The Decision Tree model adjusted its feature splits (significance) to minimise prediction error.
- The Gradient Boosting model refined input features' significance through boosting stages.
- The Bayesian Ridge model was trained by adjusting the coefficients for input features, incorporating priors on the weights and regularising them based on Bayesian inference.
- The Neural Network in PyTorch used predefined learning rate, number of epochs, hidden layer dimensions, optimiser, and batch size. During this training of this model, node-related weights in the multi-layer forward neural network were adjusted iteratively through back-propagation to minimise the loss function.

Hyper-parameter tuning was crucial for identifying the best options that suit the data. For instance, in Gradient Boosting, hyper-parameters such as learning rate, tree depth, and the number of boosting rounds played pivotal roles in shaping the model's performance. Subsequently, the hyper-parameter tuning procedure was conducted to enhance model performance and prediction accuracy.

- For the Elastic Net model, hyper-parameters such as *alpha* (regularisation strength) and *l1_ratio* (balance between $L1$ and $L2$ regularisation) were varied.
- In the Gradient Boosting model, hyper-parameters including the *number of boosting stages*, *learning rate*, and *maximum tree depth* were adjusted.
- For the Decision Tree model, hyper-parameters such as *maximum depth*, *minimum samples split*, and *minimum samples per leaf* were fine-tuned.
- The Bayesian Ridge model's hyper-parameters included *max_iter*, *alpha_1* and *alpha_2* (priors on the precision of the weights), and *lambda_1* and *lambda_2* (priors on the precision of the noise).
- For the Neural Network, hyper-parameters included *learning rates*, *number of epochs*, *hidden layer dimensions*, *optimizer classes*, and *batch sizes*.

Grid search with cross-validation was employed to systematically explore different hyper-parameter combinations for each model, identifying the settings that provided the best predictive performance, i.e., to have minimum possible discrepancy between predicted and actual procedure times (*H4 Minutes*).

The process of model training was followed for both holistic and categorical modelling scenarios. The trained models with best hyper-parameters combination for each model-type were then assessed for their performance using the testing dataset.

V. Models' Performance Results and Analysis

Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (R2) were utilised as three different metrics to assess the performance of each model. These metrics are commonly employed in regression tasks due to their ability to capture different aspects of predictive accuracy and model fit. Therefore, the metrics are selected to ensure a comprehensive assessment of model variability and effectiveness. Additionally, Consultants' estimations considered as the base model are also presented for significance comparison of the models. The overall evaluation was conducted using the testing dataset, which comprises 20% of the total data after pre-processing (Figure 1).

A. Performance of Different Algorithms and related Models

TABLE I
Holistic (built for all procedures) Models' Overall (average) Performance on testing dataset encompassing all 163 procedure codes related data

	RMSE	**MAPE**	**R2**
Base model	31.63	0.57	0.7
ElasticNet	22.10	0.25	0.85
GBoost	22.71	0.24	0.85
Decision Tree	27.86	0.28	0.77
Bayesian Ridge	22.11	0.25	0.85
NN Model	21.46	0.22	0.86

Table I comprehensively presents the performance of all considered holistic models with 3 different performance metrics mentioned before. Specifically saying, the table includes the overall performance for each of the holistic model type, and performances are measured against all procedures considered in building the model(s). Figure 4 illustrates comparison between predicted and true values via. the scatter plots for each Holistic model type. The plots facilitate analysis by complementing the limitations of tabular data, which only offer metrics-based comparisons of model performance. Tabular results alone fail to provide sufficient insights into the distribution range of procedure time between predicted and real-world values, as well as details regarding errors. The scatter sub-plots (Figure 4), together with predictive-error (residual) distribution related sub-plots (Figure 5), assist in identifying the potential biases in the predictive nature of the model. The residual analysis allows for interpretation of the systematic patterns or trends in prediction errors, such as distribution of normality, skewness, etc.

In terms of performance, all of the models surpassed the Base-Model significantly. Among them, NN model showcase superior predictive accuracy in overall, obtained with lower Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and higher R-squared (R2) values (Table I). ElasticNet, Gradient Boosting and Bayesian Ridge also demonstrate competitive performance, displaying relatively low RMSE and high R2 values across all procedures. Decision-Tree (DT) remains as least performing from all metrics perspective.

To validate the differences between the models' output and their performance, we implemented a statistical-test, named Friedman test, started setting a null hypothesis (H_0). The H_0 posited that there are no differences in the central tendencies (median ranks) among the models output per case, implying that all models perform equivalently, lets say by absolute Percentage Error per case. The Friedman test yielded a test statistic of 57.9776 and a p-value of 7.7141e-12, which gave p-value significantly lower than the common significance level of 0.05. Given this p-value, we reject the null hypothesis, indicating strong evidence that one model's performance is significantly different from the others. With results suggesting that the models do not perform identically we ranked their performance. We calculated the ranks based on the absolute percentage error for all cases from from the test dataset i.e., across all of the testing data point. The average ranks ((Table II) also indicated that NeuralNet has the best performance among the models tested, followed by Gradient Boosting and ElasticNet Regression. This rank-based analysis, where Neural Network outperforms all other algorithms is again statistically validated at the 5% significance level. It can be observed by the p-values of the Bonferroni post-hoc test (Table II), where all of these p-values are below 5%.

Evaluation of the residual distribution related to the Holistic Models (Figure 5), showed the density plot tends to be biased towards the negative in most cases. This indicated greater predicted time compared to the true duration and can be

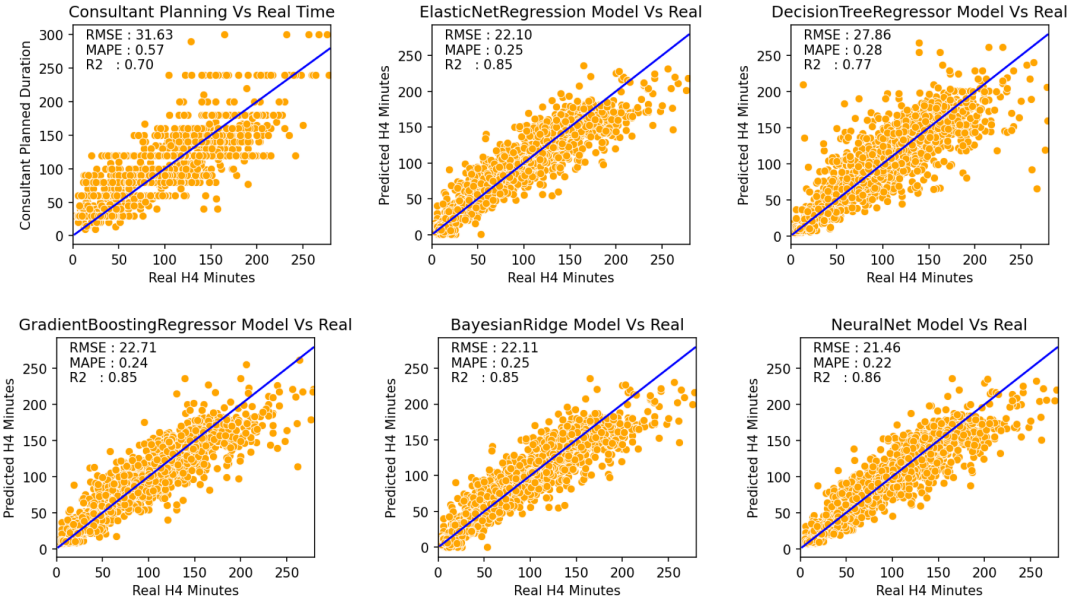


Fig. 4. Models' Overall Prediction Performance on the Testing Dataset: Scatter Plot Analysis (Prediction vs True H4 minutes) for All Models, including plot for existing Consultant Planning data

TABLE II
MODELS AVERAGE RANKS BASED ON ABSOLUTE PERCENTAGE ERROR FOR CASES IN TEST DATASET, ALONG WITH THE BON-FERRONI'S POST-HOC (P_{Bonf}) TEST (PAIRWISE COMPARISON WITH NEURALNET RELATED DATA FOR SIGNIFICANT DIFFERENCES).

Model	**Average Rank**	P_{Bonf}
NeuralNet	3.114809	-
GradientBoostingRegressor	3.354964	2.852E-05
RegressionPipeline&ElNet	3.364947	1.546E-05
BayesianRidge	3.396007	1.743E-06
DecisionTreeRegressor	3.564060	1.028E-13
Base Model	4.186911	2.422E-67

disregarded if the bias is not excessively high. In fact, a slight bias towards the negative side might be preferable, as it could help minimise risks during planning, compared to being biased towards the positive side. Bayesian Ridge and NN models exhibit higher peaks than others, with Decision-Tree showing a lower peak. While Gradient Boosting Regressor and NN models have lesser variance than the others, the Decision Tree Regressor and ElasticNet Regression models also show notable improvements, with better-balanced error distributions and reduced skewness. The Gradient Boosting Regressor and Bayesian Ridge models, while showing reduced variability, still exhibit heavy tails, indicating frequent extreme errors. Overall, the machine learning models provide more reliable and accurate predictions, with the Neural Network Model leading in performance.

B. SHAP Analysis

As a part of analysis of models, SHAP analysis was conducted to reveal critical insights into the influence of features on surgical time. The feature importance findings obtained from SHAP analysis across all model types were almost identical. Here, we have discussed one of one representative Holistic model (gradient boosting).

Among the categorical features, operational factors were observed to have a higher influence on predicting 'H4 Minutes'. In the SHAP analysis plot (Figure 6) with only top significant shown, red points indicate high feature values (1 in one-hot encoded case), while blue points represent low (0 in one-hot encoded case) feature values. Likewise, the higher negative or positive value signifies their impact on the surgical time estimation, in negative or positive way. The colour gradient indicating the higher feature values (red) for the variables *Anaesthetist Expected?_No* and *Theatre Area_2* are both associated with negative SHAP values. Notably, the absence of an expected anaesthetist (i.e., when binary 1 for *Anaesthetist Expected?_No*) significantly impacts the model's predictions, leading to a decrease in the predicted surgical time. Following closely behind is a factor related to 'Team and Theatre Characteristics', i.e., *Theatre Area*, which categorises theatres based on location and facilities. With this, the feature *Theatre Area_2* is associated with shorter surgical times.

The pre-considered crucial feature i.e., *Procedure Code* (Section II, Section III-B) demonstrated distinct impacts, for certain procedures like *W879*, *W903*, *W379, W399*, and also others notably influencing the time prediction. The analysis also suggests the involvement of different consultants varying impacts on surgical time, with certain consultants (e.g., B, J) being linked to reducing durations and *Consultant L* linked to longer durations. Additionally, 'Temporal Factors' such as the *Covid flag* indicate that *pre-Covid* conditions generally have a neutral to slightly negative impact on surgical time, whereas

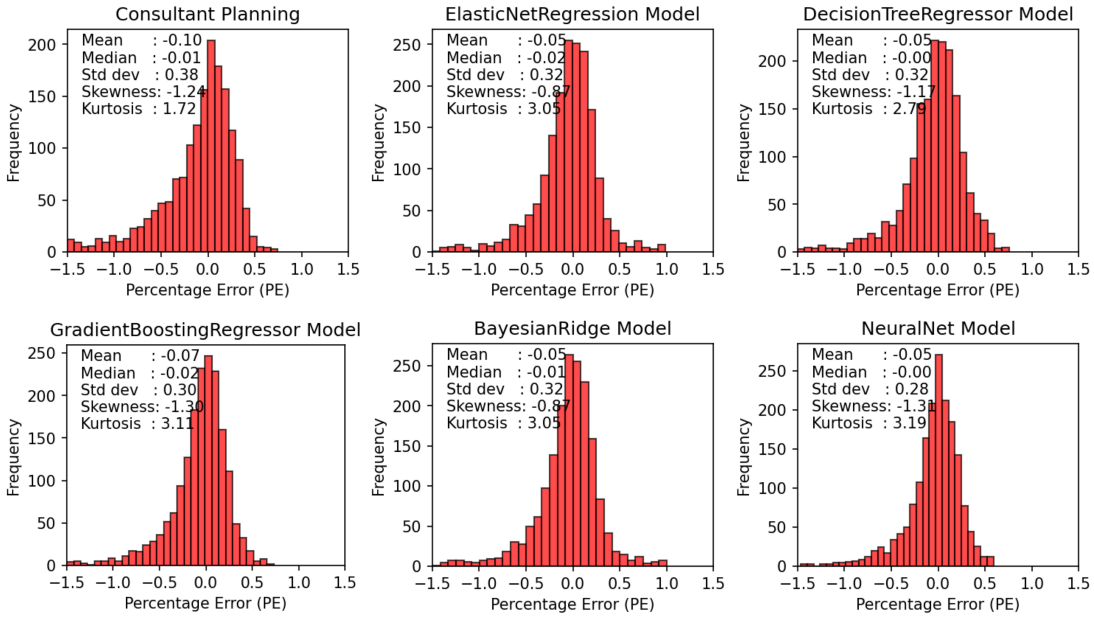


Fig. 5. Residual (Percentage Errors) distribution for Models and data related to Figure 4

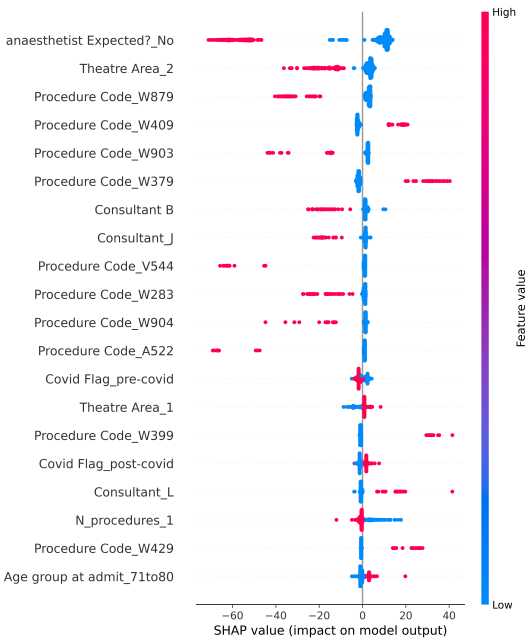


Fig. 6. A representative plot of SHAP (SHapley Additive exPlanations) values (related to GradientBoosting model) to represent the impact of some top influencing input features to the model

post-Covid scenarios might slightly increase the duration. Lastly, patients' demographic characteristics, such as *gender* and *age*, exhibit less significance. However, the *age group of 71-80* demonstrates some influence, suggesting that older age groups tend to experience marginally longer surgical times.

C. Categorical Modelling: Results and Analysis

In the Holistic-modelling approach, the procedure-related variable was found to be a significant feature to variability in the model (Figure 6). This justify the approach for development of categorical models for procedures that occur more frequently, aiming to improve prediction accuracy.

To study the impact of categorical modelling (Section IV-B), we initially identified the performance of the Holistic model discussed earlier on the selected 11 procedures (Table III). The performance trend of the Holistic Models remained largely consistent with the results for all 163 procedure cases (Table I), showing some metric-based improvement when evaluated against the 11 most frequently performed procedures (Table III). However, a notable exception was the Decision-Tree (DT) Holistic model, which performed significantly better for the 11 selected procedures. This suggests that the data distribution influences the DT model, making it more effective and potentially biased towards procedures with higher counts. It is also noteworthy that consultant's planning (also base model) for these higher frequency 11 procedures is better compared to the situations with all procedures included.

Then, we also recorded the combined performance of the 11 procedure-specific models related to same 11 procedures, each of which is tailored to a particular procedure. This approach is repeated for all 5 different types of predictive model being considered.

Prediction vs. real procedure duration related illustrations (scatter-plot) are provided for categorical models, i.e., procedure specific-models in Figure 7. Each of these sub-plots illustrate real and predicted procedure times by each of categorical-models for each procedure.

The impact of categorical modelling is notably advanta-

TABLE III
PERFORMANCE OF VARIOUS MODELS (INCLUDING RELATIVE IMPROVEMENT TO BASE MODEL) ON 11 SELECTED PROCEDURES FOR TWO DIFFERENT MODELLING TYPES: (I) HOLISTIC MODEL AND (II) CATEGORICAL MODEL

		RMSE		MAPE		R2	
		Value	± % base model	Value	± % base model	Value	± % base model
Base model	-	30.62	-	0.71	-	0.72	-
ElasticNet	Holistic	20.72	0.32	0.27	0.62	0.87	0.21
	Categorical	19.82	0.35	0.23	0.68	0.88	0.22
GBoost	Holistic	20.12	0.34	0.25	0.65	0.88	0.22
	Categorical	19.89	0.35	0.23	0.68	0.88	0.22
Decision Tree	Holistic	22.99	0.25	0.26	0.63	0.84	0.17
	Categorical	21.71	0.29	0.24	0.66	0.86	0.19
Bayesian Ridge	Holistic	20.73	0.32	0.27	0.62	0.87	0.21
	Categorical	19.76	0.35	0.23	0.68	0.88	0.22
NN Model	Holistic	19.87	0.35	0.23	0.68	0.88	0.24
	Categorical	18.90	0.38	0.22	0.69	0.89	0.24

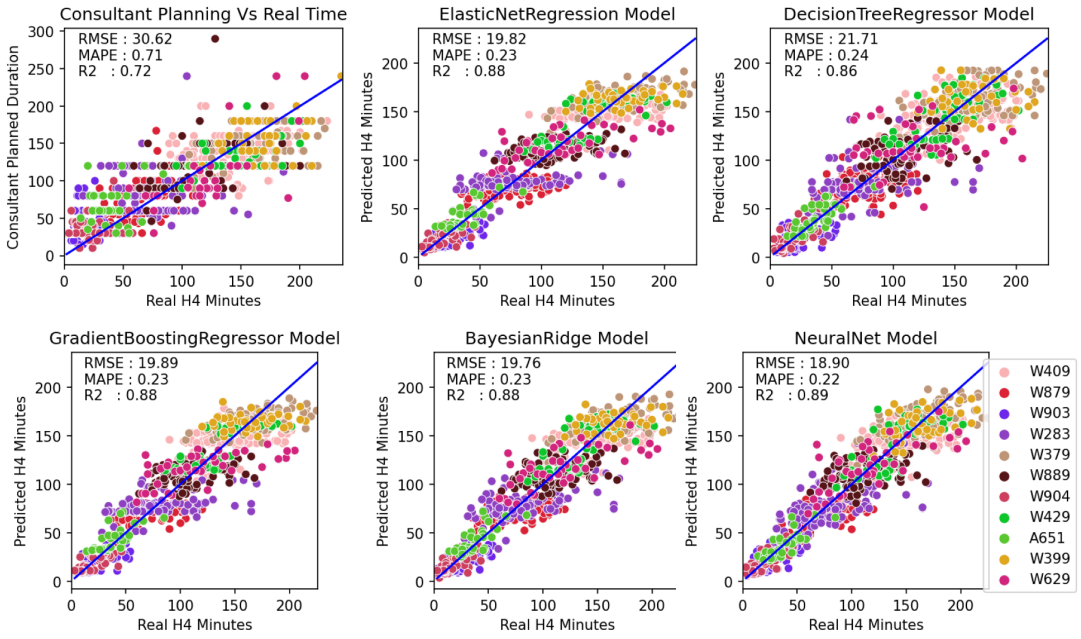


Fig. 7. Prediction Performance of Categorical Models on the Testing Dataset for Various Model Types: Models are individually constructed for each procedure within each model type, and the performance results are then aggregated

geous across most considered model types, demonstrating superior average performance compared to the holistic approach (Table III). To substantiate these findings, a Kolmogorov-Smirnov test was employed to evaluate performance differences between categorical and holistic modelling approaches across various model types (as in Section V-A). The test made to check the discrepancies between the approach based upon the prediction-values indicate statistically significant disparities ($p < 0.05$), for the ElasticNet and Bayesian Ridge models. Conversely, no statistically significant differences ($p > 0.05$) were found for DecisionTreeRegressor, GradientBoostingRegressor, and NeuralNet, suggesting comparable performance between the two approaches for these models.

Furthermore, a frequency-based percentage analysis was performed based on absolute errors at each testing data point (Table IV). It showed that the categorical modelling approach in overall performs better than the holistic approach across various model types but also revealed varying preferences

TABLE IV
PERCENTAGE WHERE CATEGORICAL APPROACH PERFORMS BETTER MODEL-WISE

Model Type	Overall Percentage favouring Categorical Approach
ElasticNet	54.66%
DecisionTreeRegressor	44.27%
GBoostingRegressor	51.75%
BayesianRidge	54.17%
NeuralNet	48.93%

across models. The analysis pointed DecisionTreeRegressor leans towards the holistic approach, while BayesianRidge and ElasticNet show a preference for the categorical approach. GradientBoostingRegressor and NeuralNet demonstrate balanced performance between the two approaches, underscoring the nuanced performance dynamics across different model types.

In summary, linear regression models (ElasticNet and Bayesian Ridge) show enhanced accuracy and explanatory power, emphasising the significance of procedure-specific modelling whenever abundant data are available. The deep-learning and decision-tree based categorical models though not differ much from holistic in performance does not perform less than the holistic. These mixed findings suggest implementing optimal model selection strategies based on specific modeling approaches in practical applications. Moreover, the findings also suggest more experimentation with data expansion for the concept of categorical modelling. It need not be limited to procedures-specific, but surgeon-specific models also can be explored further in improving prediction accuracy [8, 19].

The performance results of considered M/L models were presented in this section, together with comparative analysis both among themselves and in relation to the base model. In summary, the Neural Network models, appears best choices for predictive modelling, offering superior performance in capturing complex relationships within the data, but requires more data for performance-improvement with categorical approach. The ElasticNet, Gradient-Boosting and Bayesian Ridge provide reliable alternatives. In overall, the outcome with consideration of procedure-specific modelling is promising.

VI. CONCLUSION AND FURTHER WORK

In this study, we developed and analysed multiple M/L predictive models for the estimation of surgical duration, with their intended utilisation in theatre(s) scheduling. For this, we used data from four years of elective surgeries conducted in an NHS Trust hospital within the specialty of 'Trauma and Orthopaedics'. We focused on patient and hospital setting related features that could be known at the time of theatre planning as input variables to the model, so that the realism of the model be enhanced. Regarding performance of each of the models, our analysis suggested that Neural Network models among utilised in this study could be the optimal choice for predictive modeling, while ElasticNet, Gradient-Boosting, and Bayesian Ridge models also offer reliable alternatives.

Other unique approach we employed is the procedure-specific categorical modelling, where models were built separately for each procedure using the related procedural data. Procedure-specific modelling for the procedures with sufficient data counts showed promising results. While most authors point out that models with limited coverage of procedures and/or sub-specialties are a limitation [19, 31] , the result of this study suggests creating multiple models to cover the broader range of procedures and sub-specialties instead of broader range of data incorporation for a single model. This approach, however, still has the limitation that we will require the holistic level model for the procedures which are not performed frequently enough enabling the building of the unique model. Having procedure-specific models allows for more insightful procedure-based performance analysis, thus adaptation of the model(s) would be easier and effective when identified procedure-time attributing additional features (procedures wise).

While the resulting models and their predictive capabilities still require broader validation and enhancement before being relied upon for theatre planning, they can offer partial assistance. The models have already demonstrated their capability in estimating procedure time with less inaccuracy than the current approach, thus aiding in reducing over-utilisation or under-utilisation of theatres. This, in turn, facilitates effective theatre planning to reduce the Referral to Treatment Time (RTT) of waiting patients, whether by optimising existing resources or adding to them. Furthermore, the exploitation of M/L algorithms in predictive analysis allows for the broader adoption of Artificial Intelligence (AI) and Digital Twin technology across the hospital and theatres.

For future work, there is still scope for model improvement, particularly in the category of 'Team and Theatre Characteristics', by incorporating information about staff (surgeons, anaesthetists, and nurses) as input variables into the models. Incorporating team dynamics into the models not only improves predictive accuracy but also facilitates the creation of a Digital Twin of the hospital and its operational system.

Furthermore, integrating predictive machine learning models into a comprehensive framework (referred to as a Digital Twin) for simulation and optimisation-driven decision-making is increasingly vital in healthcare. Additionally, experimentation using the datasets from other Trusts and/or Hospitals will help identify and improve the generalisability of the obtained results.

REFERENCES

[1] A. Jain, B. K. Dai T, C. Myers. "Covid-19 created an elective surgery backlog: how can hospitals get back on track". In: *Harvard Business Review* 10.1 (2020).

[2] L. Farrow et al. "Impact of COVID-19 on opioid use in those awaiting hip and knee arthroplasty: a retrospective cohort study". In: *BMJ Quality & Safety* 32.8 (2023), pp. 479–484.

[3] C. P. Childers, M. Maggard-Gibbons. "Understanding costs of care in the operating room". In: *JAMA surgery* 153.4 (2018), e176233–e176233.

[4] A. Ala, F. Chen. "Appointment scheduling problem in complexity systems of the healthcare services: A comprehensive review". In: *Journal of Healthcare Engineering* 2022 (2022).

[5] M. J. Eijkemans et al. "Predicting the unpredictable: a new prediction model for operating room times using individual characteristics and the surgeon's estimate". In: *The Journal of the American Society of Anesthesiologists* 112.1 (2010), pp. 41–49.

[6] J. H. May et al. "The surgical scheduling problem: Current research and future opportunities". In: *Production and Operations Management* 20.3 (2011), pp. 392–405.

[7] M. Koushan, L. C. Wood, R. Greatbanks. "Evaluating factors associated with the cancellation and delay of elective surgical procedures: a systematic review". In: *International Journal for Quality in Health Care* 33.2 (2021), mzab092.

[8] N. Master et al. "Improving predictions of pediatric surgical durations with supervised learning". In: *International Journal of Data Science and Analytics* 4 (2017), pp. 35–52.

[9] A. Abbas et al. "Machine learning using preoperative patient factors can predict duration of surgery and length of stay for total knee arthroplasty". In: *International journal of medical informatics* 158 (2022), p. 104670.

[10] A. Vallée. "Digital twin for healthcare systems". In: *Frontiers in Digital Health* 5 (2023), p. 1253050.

[11] D. M. Botín-Sanabria et al. "Digital twin technology challenges and applications: A comprehensive review". In: *Remote Sensing* 14.6 (2022), p. 1335.

[12] D. Strum, J. May, L. Vargas. "Surgical procedure times are well modeled by the lognormal distribution". In: *Anesthesia & Analgesia* 86.2S (1998), 47S.

[13] W. E. Spangler et al. "Estimating procedure times for surgeries by determining location parameters for the lognormal model". In: *Health care management science* 7 (2004), pp. 97–104.

[14] E. Kayis et al. "Improving prediction of surgery duration using operational and temporal factors". In: *AMIA Annual Symposium Proceedings*. Vol. 2012. American Medical Informatics Association. 2012, p. 456.

[15] Z. ShahabiKargar et al. "Predicting procedure duration to improve scheduling of elective surgery". In: *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings 13*. Springer. 2014, pp. 998–1009.

[16] F. Dexter et al. "Value of a scheduled duration quantified in terms of equivalent numbers of historical cases". In: *Anesthesia & Analgesia* 117.1 (2013), pp. 205–210.

[17] Z. ShahabiKargar et al. "Improved Prediction of Procedure Duration for Elective Surgery." In: *HIC*. 2017, pp. 133–138.

[18] D. Sahadev, T. Lovegrove, H. Kunz. "A Machine Learning Solution to Predict Elective Orthopedic Surgery Case Duration." In: *Studies in Health Technology and Informatics* 295 (2022), pp. 559–561.

[19] M. A. Bartek et al. "Improving operating room efficiency: machine learning approach to predict case-time duration". In: *Journal of the American College of Surgeons* 229.4 (2019), pp. 346–354.

[20] N. Curtis et al. "Artificial neural network individualised prediction of time to colorectal cancer surgery". In: *Gastroenterology Research and Practice* 2019 (2019).

[21] B. M. Bradley et al. "The effect of obesity and increasing age on operative time and length of stay in primary hip and knee arthroplasty". In: *The Journal of arthroplasty* 29.10 (2014), pp. 1906–1910.

[22] C. Han et al. "To predict the length of hospital stay after total knee arthroplasty in an orthopedic center in China: the use of machine learning algorithms". In: *Frontiers in surgery* 8 (2021), p. 606038.

[23] E. Kayış et al. "A robust estimation model for surgery durations with temporal, operational, and surgery team effects". In: *Health care management science* 18 (2015), pp. 222–233.

[24] O. Martinez et al. "Machine learning for surgical time prediction". In: *Computer Methods and Programs in Biomedicine* 208 (2021), p. 106220.

[25] I. Guyon, A. Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[26] H. Zou, T. Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.

[27] A. C. Müller, S. Guido. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.

[28] D. J. MacKay. "Bayesian interpolation". In: *Neural computation* 4.3 (1992), pp. 415–447.

[29] F. Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[30] A. Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[31] J. Lai et al. "Improving and Interpreting Surgical Case Duration Prediction with Machine Learning Methodology". In: *medRxiv* (2020), pp. 2020–06.