JRMC Journal of Regional Medical Campuses

Machine Learning to Improve Resident Scheduling: Harnessing Artificial Intelligence to Enhance Resident Wellness

Aazad Abbas, MD, HBSc; Jay Toor, MD, MBA; Denis Margalik, MASc, BMSc; Jin Tong Du, MD; Anne Versteeg, MD, PhD, MHSc; Nicholas J Yee, MD, MSc; Joel A. Finkelstein, MD, MSc, FRCSC; Jihad Abouali, MD, BSc, FRCSC; Markku T. Nousiainen, MD, MSc, FRCSC; Hans J Kreder, MD, MPH, FRCSC-SB; Jeremy A Hall, MD, FRCSC, Med; Cari Whyne, PhD, FIOR; Jeremie Larouche, MD, MSc, FRCSC

DOI: https://doi.org/10.24926/jrmc.v8i1.6332

Journal of Regional Medical Campuses, Vol. *8*, Issue 1 (2025)

z.umn.edu/JRMC

Published by University of Minnesota Libraries Publishing

LIBRARIES PUBLISHING

Machine Learning to Improve Resident Scheduling: Harnessing Artificial Intelligence to Enhance Resident Wellness

Aazad Abbas, MD, HBSc; Jay Toor, MD, MBA; Denis Margalik, MASc, BMSc; Jin Tong Du, MD; Anne Versteeg, MD, PhD, MHSc; Nicholas J Yee, MD, MSc; Joel A. Finkelstein, MD, MSc, FRCSC; Jihad Abouali, MD, BSc, FRCSC; Markku T. Nousiainen, MD, MSc, FRCSC; Hans J Kreder, MD, MPH, FRCSC-SB; Jeremy A Hall, MD, FRCSC, Med; Cari Whyne, PhD, FIOR; Jeremie Larouche, MD, MSc, FRCSC

Abstract

Introduction: Excessive resident duty hours (RDH) is a recognized issue with implications for physician well-being and patient safety. A significant component of the RDH concern is on-call duty. While other industries have adopted machine learning models (MLMs) to optimize scheduling and employee well-being, medicine has lagged. This study aimed to investigate the use of MLMs to predict demand on orthopaedic residents to optimize scheduling.

Denis Margalik, MASc, BMSc Role: Manuscript preparation Email: denis.margalik@mail.utoronto.ca ORCID: https://orcid.org/0009-0005-2825-6272 1 King's College Circle, Toronto, ON, M5S 1A8, Canada Phone: 1-416-823-4230
Aazad Abbas, MD, HBSc.Role: Primary investigator. Ideation. Data collection. Data analysis. Manuscript preparation. Email: aazad.abbas@mail.utoronto.ca ORCID: https://orcid.org/0000-0001-7414-1701 1 King's College Circle, Toronto, ON, M5S 1A8, Canada Phone: 1-613-407-6991
Jay Toor, MD, MBA Role: Co-investigator. Ideation. Data collection. Data analysis. Manuscript preparation. Email: jay.toor@mail.utoronto.ca ORCID: https://orcid.org/0000-0002-9443-455X 750 Bannatyne Ave, Winnipeg, MB R3E 0W2, Canada Phone: 1-416-918-9519
Jin Tong Du, MD Role: Co-investigator. Ideation. Manuscript preparation. Email: jintong.du@mail.utoronto.ca ORCID: https://orcid.org/0000-0002-9443-455X 1 King's College Circle, Toronto, ON, M5S 1A8, Canada Phone: 1-647-270-5411
Anne Versteeg, MD, PhD, MHSc Role: Co-investigator. Ideation. Manuscript preparation. Email: anne.versteeg@mail.utoronto.ca ORCID: https://orcid.org/0000-0003-3251-9694 149 College Street Room 508-A, Toronto, ON, M5T 1P5, Canada Phone: 1-437-989-5517
Nicholas J Yee, MD, MSc Role: Co-investigator. Ideation. Manuscript preparation. Email: njsyee@gmail.com ORCID: 0000-0002-0819-6381 149 College Street Room 508-A, Toronto, ON, M5T 1P5, Canada Phone: 1-437-989-5517
Joel A. Finkelstein, MD, MSc, FRCSC Role: Co-investigator. Ideation. Manuscript preparation. Email: joel.finkelstein@sunnybrook.ca 2075 Bayview Avenue, Toronto, ON, M4N 3M5, Canada Tel: 1-416-480-6774
Jihad Abouali, MD, BSc, FRCSC. Role: Co-investigator. Ideation. Manuscript preparation. Email: jihad.abouali@gmail.com ORCID: https://orcid.org/0000-0002-6988-8429 825 Coxwell Ave Toronto, ON M4C 5T2 Phone: 1-416-937-5860
Markku T. Nousiainen, MD, MSc, FRCSC Role: Co-investigator. Ideation. Manuscript preparation. Email: markku.nousiainen@sunnybrook.ca ORCID: 0000-0002-6750-7768 43 Wellesley St. East, Room 621, Toronto, ON M4Y 1H1, Canada Tel: 1-416-967-8639
Hans J Kreder, MD, MPH, FRCSC-SB Role: Co-investigator. Ideation. Manuscript preparation. Email: hans.kreder@sunnybrook.ca. ORCID: 0000-0001-9406-5232 2075 Bayview Avenue, MG-365, Toronto, ON, M4N 3M5, Canada Tel: 1-416-480-6100 ext. 6816
Jeremy A Hall, MD, FRCSC, Med Role: Co-investigator. Ideation. Manuscript preparation. Email: Jeremy.Hall@unityhealth.to ORCID: 36 Queen St E, Toronto, ON, M5B 1W8, Canada. Tel: 416-864-5580
Cari Whyne, PhD, FIOR Role: Co-investigator. Ideation. Project supervision. Manuscript preparation. Email: cari.whyne@sunnybrook.ca ORCID: 0000-0002-6822-8314 Sunnybrook Research Institute, Orthopaedics Biomechanics Laboratory, 2075 Bayview Avenue – S620, Toronto, ON, M4N 3M5, Canada Phone: 1-416-480-6100.
Jeremie Larouche, MD, MSc, FRCSC Role: Principal investigator. Ideation. Manuscript preparation. Email: Jeremie.Larouche@sunnybrook.ca ORCID: https://orcid.org/0000-0001-9636-2273 Sunnybrook Health Sciences Centre, 2075 Bayview Avenue – MG 375, Toronto, ON, M4N 3M5, Canada. Phone: 1-416-480-6775

Corresponding author: Denis Margalik, MASc, BMSc Role: Manuscript preparation Email: denis.margalik@mail.utoronto.ca ORCID: https://orcid.org/0009-0005-2825-6272 1 King's College Circle, Toronto, ON, M5S 1A8, Canada Phone: 1-416-823-4230

Methods: Daily surgical handover emails over an eight-year (2012-2019) period at a level I trauma centre were used to model demand on residents. Various MLMs were trained to predict the workload, with their results compared to the current approach. Quality of models was determined by using the area under the receiver operator curve (AUC) and accuracy. The top ten most important variables were extracted from the most successful model.
Results: The reduction in orthopaedic resident shifts possible per annum was 24.7%. The most successful model during testing was the neural network (AUC: 0.81, accuracy: 73.7%). All models were better than the current approach (AUC: 0.50, accuracy: 50.1%). Key variables used by the neural network model were (descending order): spine call duty (y/n), year, weekday/weekend, month, and day of the week.
Conclusion: This was the first study using MLMs to predict demand for orthopaedic residents at a major academic institution. All MLMs were more successful than the current scheduling approach. Future work should look to incorporate predictive models with optimization strategies, matching scheduling with demand to improve resident well-being and patient care.

Keywords: resident; medical education; orthopaedics; machine learning; scheduling

Introduction

Excessive resident duty hours (RDH) is a recognized issue within modern postgraduate medical education.[1-3] Research has shown that excessive RDH is linked to worse patient outcomes, decreased medical trainee satisfaction, and increased rates of physician burnout.[4-8] As such, various institutions worldwide have implemented work hour restrictions to improve patient safety and resident well-being.[4, 9-12] However, this has not yet resulted in evidence demonstrating improved patient care nor resident wellness, and instead has negatively impacted resident education through decreased learning opportunities, limited resident supervision and decreased learning opportunities.[8, 13, 14]

A major component of the RDH concern is on-call duty. Call coverage is traditionally scheduled according to a fixed schedule rather than based on the anticipated amount of work required (i.e., demand), leading to over-scheduling call coverage to prevent a service gap. Research in operations management has attempted to solve the resident scheduling problem (RSP) through various models, heuristics, and approaches.[15-21] However, a key drawback of these models is that they solve the RSP as a scheduling optimization problem without considering the varying demand placed on residents throughout the year. These models have failed to be implemented in resident scheduling practice despite in-depth analysis, mostly due to their complexity and lack of practical applicability.

While considerable research has been put forth to address the problem of reducing resident call workload and solving the RSP, there is a paucity in the current literature looking at predicting and forecasting the demand on surgical residents. If one can accurately predict the demand on residents at any given time, one will be able to schedule residents according to the demand appropriately. Accordingly, machine learning models (MLMs) have been extensively used in other industries to predict demand and prevent such issues as a service supply-demand mismatch.[22-25] Now the use of MLMs are slowly gaining traction in healthcare. For instance, predictive models have been used to forecast the demand on emergency departments to improve nursing staff scheduling.[26, 27] Additionally, there are many studies using MLMs to optimize hospital management and patient scheduling. [28, 29, 30] However, there has not been a large focus on fully addressing the unique aspects of resident scheduling, such as training requirements and educational opportunities. We believe that predictive MLMs may be used to accurately forecast the workload demand placed on residents during on-call hours. We hypothesize that external factors such as seasonality and day of the week may predict the demand on residents. Accordingly, the aim of this study was to: 1) identify the optimization potential available in the current orthopaedic resident schedule, 2) identify key variables involved in determining the demand on residents, and 3) develop MLMs to predict the demand on residents.

Methods

This single-centre quality improvement study was conducted in the Division of Orthopaedic Surgery at Sunnybrook Health Sciences Centre, a major academic hospital and level 1 trauma centre in Toronto, Canada. As a quality improvement initiative, this retrospective study was exempt from requiring Research Ethics Board consideration.

Determining Resident Demand

The main contributors to the demand of work on the residents during each call shift consist of operating room (OR) hours, traumas, admissions, consults and rounding on patients. As such, the demand on residents can be modelled mathematically by determining the number of hours spent doing the following critical tasks: 1) operating, 2) seeing traumas, 3) admitting patients, 4) seeing consults, and 5) rounding on patients. The mathematical definition of the demand may be found in Appendix I, Equation I.

Timed Study

To determine the average time required to complete each critical task, a detailed time study of 100 observations of each task was completed over one year (i.e., the critical tasks were observed over 100 call shifts), with mean and standard deviation (SD) determined for each task.

Data Extraction

To determine the number of critical tasks conducted during each shift, handover emails were collated over an eight-year period (January 1st, 2012 - December 31st, 2019). The following data was extracted using a natural language processing algorithm built using Python programming language: 1) date, 2) number of ORs completed, 3) number of traumas, 4) number of admissions, 5) number of consults, 6) spine call duty (orthopaedics/neurosurgery residents), and 7) hip fracture case presence (yes/no).[31] Note, residents from orthopaedics and neurosurgery share spine call at our institution, thus, spine call duty is an important factor to consider in the requirement for orthopaedic on-call residents. In addition, whether a hip fracture procedure was completed or not is an important consideration at our institution as they need to be completed within a scheduled time per government policy.[32]

Shift Optimization

Classically, two residents (one junior and one senior) are placed on each shift at our institution, with 730 resident shifts scheduled per year. Reduction capacity was defined as the number of unnecessary second-call resident shifts, which occurred when the daily demand was less than 20 hours. A 20-hour cut-off was chosen through a focus group of twenty-three residents whereby they were asked how many hours of work in a 24-hour shift they felt competent doing alone before requiring a second resident on call. If the demand was less than 20 hours, then one resident would be sufficient to handle the workload, and if the demand was greater than 20 hours, then a second resident would be required. Accordingly, the mathematical definition of shift optimization available is described in Appendix I, Equation 2.

Prediction of Demand

MLMs were deployed to predict whether the demand on residents exceeds the cut-off of 20 hours (positive prediction) or falls below 20 hours (negative prediction). The dataset was randomized and split into a training/testing ratio of 80/20, with training and validation of MLMs conducted on the training set and testing of MLMs on the testing set.

Features used for the MLMs were 1) season, 2) year, 3) day of the week, 4) month, 5) weekday, 6) spine call duty, and 7) hip fracture case. Refer to Table 1 for a detailed description of these variables. The correlation between features was examined by determining the Pearson correlation coefficient between each feature, and results were visualized through a network plot.

Table 1. Characterization of variables available to predict the demand.

Variable Name	Variable Type	Example Values
Season	Categorical	Summer, Winter, Fall, Spring
Year	Continuous	2012, 2013, 2014
Day of the week	Categorical	Monday, Tuesday, Wednesday
Month	Categorical	January, February, March
Weekday	Binary	Yes, No
Spine call duty	Categorical	Orthopaedics, Neurosurgery
Hip fracture case	Binary	Yes, No

DOI: https://doi.org/10.24926/jrmc.v8i1.6332

MLMs utilized were logistic regression,[33] random forest,[34] k-nearest neighbour,[35] earth (multivariate adaptive regression splines),[36] naive bayes,[37] adaboost classification trees,[38] partial least squares,[39] and neural networks.[40] Models were compared to the current approach of resident scheduling, i.e., two residents for each call shift. Outcome metrics for the models was area under the receiver operator curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The model with the largest AUC was considered optimal. Statistical significance was determined by comparing the AUC metrics using analysis of variance (ANOVA) with Tukey's test post-hoc sampling. The significance level was set at p <.05. The most important features for the model with the highest AUC on the test set were determined. All analysis was completed using RStudio (RStudio Inc; Version 1.2.5042),[41] running R software (R Foundation, Version 3.6.3).[42] Models were constructed using the Caret package in the R programming language.[43]

Results

Resident Demand

Details of the average amount of time and number of units of each task measured in the timed study are reported in Table 2. The number of shifts according to demand on residents across all eight years in hours is shown in Figure 1. The mean demand was 23 hours, with 50.6% of the shifts over 20 hours.

Table 2. Mean and standard deviation (SD) of tasks required of residents during each shift.

Resident Task	Number of units*	Mean Time (hours)
Operating	1.79 (1.64)	4.01 (2.03)
Triaging trauma	4.13 (2.70)	1.09 (0.19)
Seeing consult	2.57 (1.99)	1.14 (0.36)
Admitting patient	1.13 (1.36)	1.02 (0.33)
Rounding†	1	1.59 (0.14)

*Units are the average number of each task per call shift.
†Rounding was not recorded in emails and so one unit of rounding was done per shift.

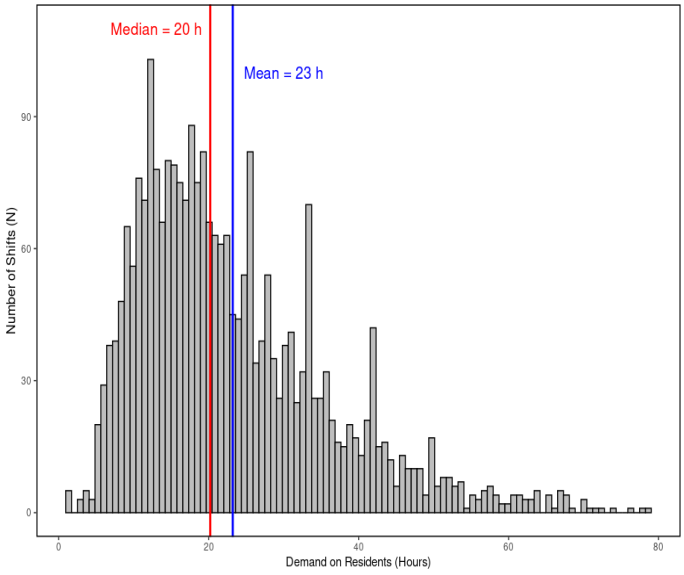


Figure 1. Histogram of demand on residents in hours. Blue vertical line indicates mean (23 hours) and red vertical indicates median (20 hours).

Shift Optimization

A detailed description of the demand and optimization potential on an annual average is provided in Table 3.

The total number of shifts that may be removed per annum was 180 out of 730 shifts (24.7% of all shifts). The number of shifts (percentage) that may be reduced per annum by season was 37 (20.3%) in summer, 41 (22.6%) in fall, 50 (27.5%) in winter, and 52 (28.5%) in spring. Regarding days of the week, Wednesdays and Thursdays had the maximum amount of shift optimization possible, with an average of 32 (31.2%) and 34 (32.9%) shifts to be removed per annum. Conversely, Saturdays and Sundays had the maximum demand and thus least optimization possible, with 15 (14.2%) and 16 (15.1%) shift reductions per annum. The months of the year with maximum optimization possible were March and January, with 20 (32.4%) and 19 (31.3%) shift reductions per annum. Conversely, the months with the least reduction possible were August and October, with 12 (19.6%) and 12 (19.8%) shift reductions per annum.

While orthopaedic residents were on spine call, 79 (17.7%) of their shifts could be reduced per annum. While orthopaedic residents were not on spine call, 102 (35.7%) of their shifts could be reduced per annum. Similarly, the percentage of days with extra

call shifts when orthopaedic residents were or were not on spine calls were 35.3% and 71.4%, respectively. When there was a hip fracture, 62 (24.5%) of their shifts could be reduced per annum. While there was no hip fracture case, 121 (24.8%) of their shifts could be reduced per annum. Similarly, the percentage of days with extra call shifts when there was a hip fracture case versus not were 49.0% and 49.7%, respectively.

Table 3. The reduction in shifts characterized on an annual average across the years studied (2012-2019).

Season	Demand* (hours)	Days with extra call shifts (%)	Reduction in shifts possible (%)	Historic number of shifts (N)*	Number of optimized shifts (N)	Number of extra shifts (N)
Season						
Summer	25.5 (7.0)	40.5	20.3	183	146	37
Fall	24.4 (6.9)	45.2	22.6	182	141	41
Winter	21.8 (6.3)	54.9	27.5	183	133	50
Spring	21.2 (5.9)	56.9	28.5	182	130	52
Day of the week						
Monday	22.7 (6.3)	51.2	25.6	104	77	27
Tuesday	22.1 (6.0)	50.0	25.0	104	78	26
Wednesday	19.5 (5.1)	62.4	31.2	104	72	32
Thursday	18.9 (4.9)	65.8	32.9	104	70	34
Friday	20.8 (4.9)	55.8	27.9	104	75	29
Saturday	30.8 (9.3)	28.3	14.2	104	89	15
Sunday	28.5 (9.4)	30.3	15.1	104	88	16
Month						
January	21.4 (6.1)	62.5	31.3	62	43	19
February	25.9 (7.1)	59.0	29.5	56	39	17
March	24.5 (7.2)	64.7	32.4	62	42	20
April	20.0 (5.5)	55.6	27.8	60	43	17
May	20.6 (6.1)	49.5	24.8	62	47	15
June	25.3 (6.9)	41.5	20.8	60	48	12
July	25.4 (7.1)	40.9	20.5	62	49	13
August	19.7 (5.5)	39.2	19.6	62	50	12
September	22.8 (6.3)	48.2	24.1	60	46	14
October	24.2 (6.9)	39.6	19.8	62	50	12
November	25.3 (7.4)	48.3	24.2	60	45	15
December	23.7 (6.2)	43.7	21.9	62	48	14
Spine call coverage						
Orthopaedics	27.4 (8.3)	35.3	17.7	444	365	79
Neurosurgery	16.8 (3.9)	71.4	35.7	286	184	102

*Demand displayed in mean (standard deviation).

Machine Learning Demand Prediction

Figure 2 displays a network plot of the correlation between all features used to train the models. Figure 3 shows the ROC curves for all models during training and testing. A display of the confusion matrices for all the models tested can be found in Appendix II.

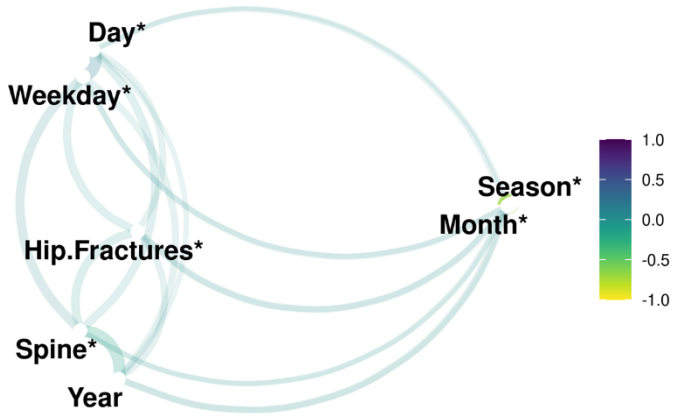


Figure 2. Network plot representing the correlation of features in the training set. Colour represents direction according to the scale on the right-hand side. Line thickness and proximity of features represents strength of correlation. *Represent categorical features.

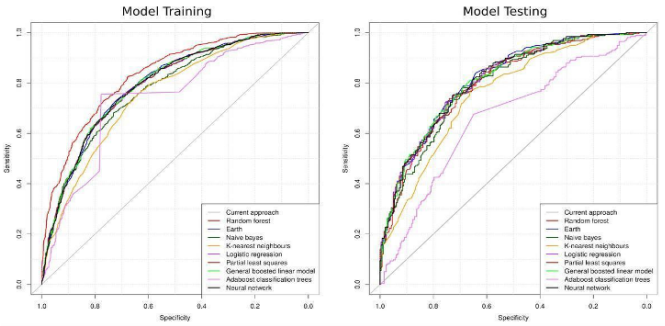


Figure 3. Receiver operator curve (ROC) plots for the models during training and testing. Models with larger areas under the ROC represent better models. The best models during training were the general boosted linear model, logistic regression, and partial least squares (AUC 0.78) and the best models during testing were the neural network, general boosted linear model, partial least squares, logistic regression, and earth (AUC 0.81). The current approach (always 2 residents) is a straight diagonal line as there is always a 50/50 chance it will be either above or below the cut-off across all sensitivity and specificity ranges.

Model Results

The models that generated the highest AUC during training were the neural network, general boosted

linear model, logistic regression, and partial least squares, all with an AUC of 0.78 (SD 0.03) (Table 4). The worst model was the adaboost classification trees, with an AUC of 0.60 (SD 0.04) and an accuracy of 62.8% (SD 3.50). All models proved better than the current approach, with an AUC of 0.50 (SD 0.01) and an accuracy of 50.1% (SD 0.0). The multivariate adaptive regression splines (earth) model had the highest accuracy during training at 71.7% (SD 3.13). According to specificity, positive predictive value, and negative predictive value, all models proved better than the current approach. Sensitivity of the current approach was better than the MLMs (Table 4).

Table 4. The mean and standard deviation (SD) of the resampled accuracy, area under the receiver operator curve (AUC), sensitivity, specificity, positive predictive value (PPN) and negative predictive value (NPV) during training and cross validation.

Model	Accuracy (%)	AUC	Sensitivity	Specificity	PPV	NPV
Current approach	50.1 (0.0)	0.50 (0.01)	1 (0)	0 (0)	0.51 (0.01)	NA*
Random forest	69.7 (2.92)	0.76 (0.03)	0.73 (0.05)	0.66 (0.06)	0.69 (0.03)	0.71 (0.03)
Earth	71.7 (3.13)	0.77 (0.03)	0.75 (0.05)	0.69 (0.06)	0.71 (0.04)	0.73 (0.04)
Naive bayes	69.2 (3.35)	0.76 (0.04)	0.68 (0.05)	0.70 (0.04)	0.70 (0.03)	0.68 (0.04)
K-nearest neighbours	68.1 (3.62)	0.73 (0.04)	0.64 (0.05)	0.73 (0.05)	0.70 (0.04)	0.66 (0.03)
Logistic regression	70.9 (3.08)	0.78 (0.03)	0.74 (0.05)	0.68 (0.05)	0.70 (0.03)	0.72 (0.04)
Partial least squares	70.8 (3.30)	0.78 (0.03)	0.74 (0.05)	0.68 (0.05)	0.70 (0.03)	0.72 (0.04)
General boosted linear model	71.1 (3.03)	0.78 (0.03)	0.75 (0.05)	0.67 (0.05)	0.70 (0.03)	0.73 (0.04)
Adaboost classification trees	62.8 (3.50)	0.60 (0.04)	0.62 (0.04)	0.64 (0.05)	0.64 (0.04)	0.62 (0.04)
Neural network	71.0 (2.94)	0.78 (0.03)	0.71 (0.04)	0.71 (0.04)	0.72 (0.03)	0.71 (0.03)

AUC: area under the receiver operator curve, NA: not applicable, NPV: negative predictive value, PPV: positive predictive value.
*Not applicable as the current approach makes no negative predictions i.e., it always predicts that the demand is above the cut-off as it schedules two residents for each call shift.

During testing, the models that generated the highest AUC were the neural network, general boosted linear model, partial least squares, logistic regression, and earth, all with an AUC of 0.81 (Table 5). The worst MLM was the adaboost classification trees, with an AUC of 0.66 and an accuracy of 66.8%. All models proved better than the current approach, with an AUC of 0.50 and an accuracy of 50.1%. The neural network model had the highest accuracy of 73.7%. All models proved better than the current approach according to accuracy, AUC, specificity, positive predictive value, and negative predictive value (Table 5). Sensitivity of the current approach was better than the MLMs (Table 5).

Table 5. The accuracy, area under the receiver operator curve (AUC), sensitivity, specificity, positive predictive value (PPN) and negative predictive value (NPV) during testing.

Model	Accuracy (%)	AUC	Sensitivity	Specificity	PPV	NPV
Current approach	50.1	0.50	1	0	0.501	NA*
Random forest	71.0	0.80	0.75	0.67	0.70	0.72
Earth	72.6	0.81	0.78	0.67	0.72	0.75
Naive bayes	72.2	0.79	0.71	0.73	0.73	0.71
K-nearest neighbours	71.0	0.76	0.66	0.76	0.74	0.69
Logistic regression	72.6	0.81	0.76	0.69	0.72	0.74
Partial least squares	71.4	0.81	0.76	0.67	0.70	0.73
General boosted linear model	72.8	0.81	0.77	0.68	0.71	0.75
Adaboost classification trees	66.8	0.66	0.69	0.65	0.67	0.67
Neural network	73.7	0.81	0.73	0.75	0.75	0.73

AUC: area under the receiver operator curve, NA: not applicable, NPV: negative predictive value, PPV: positive predictive value.
*Not applicable as the current approach makes no negative calls, i.e. it always predicts that the demand is above the cut-off.

ANOVA analysis comparing the AUC metrics of different models during training and testing revealed a significant difference between models ($p < .0001$). Post-hoc sampling revealed that all models significantly improved upon the current scheduling approach ($p < .0001$). Post-hoc sampling proved the logistic regression and neural network to have no significant difference in AUC metrics across validation folds, with a difference of 0 (95% confidence interval (CI): -4.3×10^{-16}, 4.3×10^{-16}), $p < .0001$. The general boosted linear model was the most improved compared to the current approach, with a difference of -0.811 (95% CI: -0.812, -0.810), $p < .0001$.

Feature Importance

Feature importance for the top ten features of the neural network model is displayed in Figure 5. Features considered significant by the model in order

of importance include spine call duty, year, weekday/weekend, month, and day of the week.

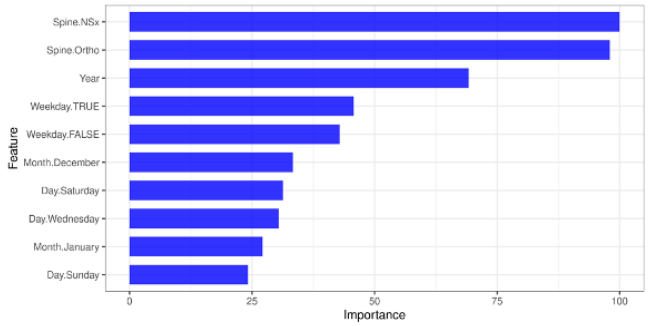


Figure 4. Bar plot representing the top ten most important features used by the neural network model.

Discussion

This study demonstrated a novel approach to addressing excessive resident duty hours using machine learning, with results showing significant potential in scheduling optimization of orthopaedic resident on-call shifts. We identified that although the mean resident demand for an on-call shift was 23 hours, it was not uncommon for residents to report working beyond the required 24 hours. These extended shifts were most likely due to the requirement of completing handover, outstanding tasks, and other substantial clinical responsibilities, but regardless demonstrate the need for a more adaptive scheduling approach that can align resident shift duration with predicted demand.

With our models, the annual optimization potential was 24.7%, which is nearly a quarter of all shifts. This was a surprising realization, especially since only two residents were scheduled per shift. Compared to the current scheduling approach, such a high optimization potential demonstrates a massive potential for scheduling optimization at our institution. Furthermore, it highlights the need for institutions to harness the mathematical techniques that have been used extensively in the operations research literature at a practical and local level to improve the scheduling of their residents.[15-21]

Notably, as the number of residents on-call increases, so too does the possible RDH savings using MLM. These findings are timely in the context of increasing interest from governing bodies and medical institutions worldwide in reducing resident workload. In current efforts to address RDH, there is often a trade-off of service coverage which needs to be filled using physician extenders. However, our study offers an alternative approach that avoids these trade-offs by optimizing existing resident schedules based on predicted demand. This strategy not only reduces unnecessary shifts but also preserves the integrity of resident education and patient care.[1-3]

With respect to the predictive models themselves, MLMs with high predictive accuracy were generated, demonstrating the possibility of forecasting the demand on residents based on external factors, allowing for subsequent schedule optimization. The MLM that proved to be the most effective in predicting resident demand was the neural network. This was not a surprise, as MLMs built on neural network architecture can learn non-linear relationships that may otherwise not be apparent to other MLMs.[40] Neural networks are particularly adept at capturing complex interactions between variables, such as temporal patterns and workload fluctuations, which are critical in the context of resident scheduling. Despite this, other MLMs such as the general boosted linear model, partial least squares, logistic regression, and earth models demonstrated similar accuracy and AUCs compared to the neural network. This comparable performance across different MLMs suggests that the choice of model may depend on specific institutional needs, such as interpretability versus predictive power. This may be due to the small size of the dataset, as neural networks tend to perform even better against shallow MLMs on larger datasets.[42] Overall, the quality of all the MLMs generated was superior at predicting demand compared to the current static scheduling approach, as demonstrated by the confusion matrices presented in Appendix II.[45] An interesting observation was that the sensitivity of the current approach was better than all MLMs during training and testing. However, this is inherent as the current approach always assumes two residents per shift, suggesting a large potential for overscheduling. The reduced sensitivity of the MLMs reflects a more balanced approach to scheduling, where the goal is not merely to avoid under scheduling but to optimize the allocation of residents based on actual demand, thereby reducing unnecessary workload without compromising patient care.

A finding highlighted by both the shift optimization and MLMs was the strong dependence on time and on-call shift coverage. It was clear that certain seasons, days of the week, and months of the year had much higher shift reduction possible, while many of the top features used by the neural network were temporal factors. This demonstrates that not only is there a temporal variation of the demand on residents, but the demand is highly amenable to prediction via MLMs. The ability to predict these variations allows for proactive scheduling adjustments, which can significantly improve both resident well-being and operational efficiency. Being amenable to prediction is crucial as it allows future work to construct live forecasting models that will ensure even more accurate demand predictions. However, this study has its limitations. Firstly, this work was conducted at a single trauma hospital in Toronto, although it represents an academic site typical of medical education training institutions. While the specific optimization potential and algorithms may not necessarily generalize to other non-trauma sites, the approach of 1) identifying optimization potential and 2) modelling the demand on residents is generalizable. These same techniques may be applied at other institutions using their historic data to optimize the scheduling and predict the on-call demand of their residents. Secondly, factors such as air and ground ambulance duty, weather, and statutory holidays were not incorporated into the models. These factors may further improve the accuracy of these models, which may be addressed in future works. Thirdly, regarding our mathematical model to define resident demand, we acknowledge that task complexity along with variability in task duration could influence resident demand. Where complex cases may require additional time and resources, affecting the overall demand on residents. Including measures of task complexity in future machine learning models (MLMs) could enhance predictive accuracy by allowing the model to differentiate between tasks of varying complexity within each category. Moreover, refining how we account for variability by incorporating standard deviation or other spread measures might help the models handle fluctuations in workload more effectively. Fourthly, specific details regarding the surgical case (e.g., multi-trauma vs minor fracture) and patient factors were not incorporated. Previous work has shown that these factors may effectively predict case costing, length of stay, and duration of surgery. [46-49] Accordingly, incorporating these factors would help improve the accuracy of the models in this study. Fifthly, this study does not address the intangible educational value of call shifts for junior residents as they are mentored by their senior resident counterparts. This limits the real-world application of these models as they would need to consider educational opportunities for juniors during downtime on a call shift. However, major academic centres often have a fellow scheduled on the same call shift, which has the potential to provide this mentorship and guidance to junior residents. Furthermore, additional responsibilities (which are often the work of junior residents, e.g., nighttime ward calls) were not considered, further limiting the application of these models in their current form. Sixthly, this study does not address the impact on patient outcomes and institutional costs of possibly under-scheduling residents during an unanticipated increase in workload. This is best suited in future work whereby these models are implemented in hospitals with their real-world impact analyzed through prospective clinical trials. Lastly, the optimization potential identified may not always be realizable, as the junior resident performs critical tasks such as seeing consults and admitting patients while the senior resident operates and sees traumas. This discrepancy in task responsibility based on resident category could not be incorporated within our MLM. In addition, many of these tasks may require both residents to be on site as they may occur simultaneously, limiting the ability to optimize the workload of residents. These practical constraints highlight the need for a flexible approach to scheduling optimization, one that can adapt to the dynamic and often unpredictable nature of clinical work. Future work should aim to incorporate the previously mentioned constraints to maximize the realizable scheduling optimization possible. Additionally, future studies could consider exploring the use of framing on call duties in terms of Entrustable Professional Activities (EPAs). With EPAs being considered as observable units of work that break down the complex competencies needed in a specialty, tasks such as patient admissions, trauma evaluations, consults, and procedural responsibilities could each be aligned with specific EPAs. [50] This

approach may help identify not only the intensity and variability in workload across shifts but also the critical learning experiences available to residents on call. Integrating EPAs into call scheduling could further enhance scheduling optimization by identifying shifts that present valuable learning opportunities, especially for junior residents, and by ensuring that duty hours prioritize tasks aligned with their learning needs.

In conclusion, this work has demonstrated a large potential for reducing orthopaedic resident on-call shift scheduling, while MLMs predicting resident on-call demand with a high level of accuracy have been generated and tested. These findings are timely in the context of increasing interest from governing bodies and medical institutions worldwide in reducing resident workload. This could prove advantageous to residents as the reduction in work hours would allow for more daytime clinical and surgical responsibility while being more focused and less sleep-deprived, creating an improved learning environment and patient care. As healthcare continues to evolve with increasing reliance on data-driven decision-making, our approach offers a promising avenue for integrating advanced predictive analytics into resident scheduling practices. Future work should aim to incorporate additional extraneous factors (i.e., air and ground ambulance scheduling, weather, and statutory holidays) to improve predictive power and aim to evaluate the efficacy of these predictive and scheduling algorithms in practice, including their effect on physician well-being and patient outcomes. Next steps would also include implementing our MLM-predicted changes into resident scheduling. This approach would allow us to systematically assess the real-world impact of our MLM-driven scheduling adjustments on the various outcomes focused on this study, and directly observe the reliable impact MLM-scheduling can provide onto resident wellness, workload balance, and patient care.

References

1. Fletcher KE, Davis SQ, Underwood W, Mangrulkar RS, McMahon LF, Saint S. Systematic review: effects of resident work hours on patient safety. *Ann Intern Med*. 2004;141(11):851-7. doi: 10.7326/0003-4819-141-11-200412070-00009.
2. Fletcher KE, Underwood W, Davis SQ, Mangrulkar RS, McMahon LF, Saint S. Effects of work hour reduction on residents' lives: a systematic review. *JAMA*. 2005;294(9):1088-100. doi: 10.1001/jama.294.9.1088
3. Fletcher KE, Reed DA, Arora VM. Patient safety, resident education and resident well-being following implementation of the 2003 ACGME duty hour rules. *J Gen Intern Med*. 2011;26(8):907-19. doi: 10.1007/s11606-011-1657-1.
4. Friedman WA. Resident Duty Hours in American Neurosurgery. *Neurosurgery.* 2004;54(4):925-33. doi: 10.1227/01.neu.0000115153.30283.f5.
5. Britt LD, Sachdeva AK, Healy GB, Whalen TV, Blair PG. Resident duty hours in surgery for ensuring patient safety, providing optimum resident education and training, and promoting resident well-being: A response from the American College of Surgeons to the Report of the Institute of Medicine, “Resident Duty Hours: Enhancing Sleep, Supervision, and Safety”. *Surgery*. 2009;146(3):398-409. doi: 10.1016/j.surg.2009.07.002.
6. Hameed TK, Masuadi E, Al Asmary NA, Al-Anzi F, Al Dubayee MS. A study of resident duty hours and burnout in a sample of Saudi residents. *BMC Med Educ*. 2018;18(1):1-6. doi: 10.1186/s12909-018-1300-5
7. Ulmer C, Wolman DM, Johns MME. Resident Duty Hours: Enhancing Sleep, Supervision, and Safety. Washington, D.C.: *The National Academies Press.* 2009.
8. Bolster L, Rourke L. The Effect of Restricting Residents' Duty Hours on Patient Safety, Resident Well-Being, and Resident Education: An Updated Systematic Review. *J Grad Med Educ.* 2015;7(3):349-63. doi: 10.4300/JGME-D-14-00612.1.

9. Choma NN, Vasilevskis EE, Sponsler KC, Hathaway J, Kripalani S. Effect of the ACGME 16-hour rule on efficiency and quality of care: duty hours 2.0. *JAMA Intern Med.* 2013;173(9):819-21. doi: 10.1001/jamainternmed.2013.3014.
10. Theobald CN, Stover DG, Choma NN, Hathaway J, Green JK, Peterson NB, et al. The effect of reducing maximum shift lengths to 16 hours on internal medicine interns' educational opportunities. *Acad Med*. 2013;88(4):512-8. doi: 10.1097/ACM.0b013e318285800f.
11. Philibert I, Nasca T, Brigham T, Shapiro J. Duty-hour limits and patient care and resident outcomes: can high-quality studies offer insight into complex relationships? *Annu Rev Med.* 2013;64:467-83. doi: 10.1146/annurev-med-120711-135717.
12. Auger KA, Landrigan CP, Gonzalez del Rey, Javier A., Sieplinga KR, Sucharew HJ, Simmons JM. Better rested, but more stressed? Evidence of the effects of resident work hour restrictions. *Acad Pediatr.* 2012;12(4):335-43. doi: 10.1016/j.acap.2012.02.006.
13. Temple J. Resident duty hours around the globe: where are we now? *BMC Med Educ*. 2014;14(1):1-5. doi: 10.1186/1472-6920-14-S1-S8
14. Reed DA, Fletcher KE, Arora VM. Systematic review: association of shift length, protected sleep time, and night float with patient care, residents' health, and education. *Ann Intern Med.* 2010;153(12):829-42. doi: 10.1059/0003-4819-153-12-201012210-00010
15. Topaloglu S. A shift scheduling model for employees with different seniority levels and an application in healthcare. *Eur J Oper Res.* 2009;198(3):943-57. doi: 10.1016/j.ejor.2008.10.032
16. Day TE, Napoli JT, Kuo PC. Scheduling the Resident 80-Hour Work Week: An Operations Research Algorithm. *Curr Surg.* 2006;63(2):136-41. doi: 10.1016/j.cursur.2005.12.001.
17. Topaloglu S. A multi-objective programming model for scheduling emergency medicine residents. *Comput Ind Eng.* 2006; 51(3):375-88. doi: 10.1016/j.cie.2006.08.003
18. Franz LS, Miller JL. Scheduling Medical Residents to Rotations: Solving the Large-Scale Multiperiod Staff Assignment Problem. *Operations Research.* 1993;41(2):269-79. doi: https://doi.org/10.1287/opre.41.2.269.
19. Sherali HD, Ramahi MH, Saifee QJ. Hospital resident scheduling problem. *Prod Plan Control.* 2010;13(2):220-33. doi: 10.1080/09537280110069667
20. Ozkarahan I. A scheduling model for hospital residents. *J Med Syst.* 1994;18(5):251-65. doi: 10.1007/BF00996605.
21. Topaloglu S, Ozkarahan I. A constraint programming-based solution approach for medical resident scheduling problems. *Comput Oper Res.* 2011;38(1):246-55. doi: 10.1016/j.cor.2010.04.018
22. Song H, Liu H. Predicting Tourist Demand Using Big Data. *Analytics in Smart Tourism Design.* 2017;13-29. doi: 10.1007/978-3-319-44263-1_2
23. Moreira-Matias L, Gama J, Ferreira M, Mendes-Moreira J, Damas L. Predicting taxi-passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*. 2013;14(3):1393-402. doi: 10.1109/TITS.2013.2262376.
24. Rodger JA. A fuzzy nearest neighbor neural network statistical model for predicting demand for natural gas and energy cost savings in public buildings. *Expert Syst Appl.* 2014;41(4):1813-29. doi: 10.1016/j.eswa.2013.08.080
25. Souza GC. Supply chain analytics. *Bus Horiz.* 2014; 57(5):595-605. doi: 10.1016/j.bushor.2014.06.004
26. Storme Ramsey K. *Using Predictive and Descriptive Models to Improve Nurse Staff Using Predictive and Descriptive Models to Improve Nurse Staff Planning and Scheduling Planning and Scheduling [dissertation]*. University of Tennessee; 2014.
27. Jones SS, Thomas A, Evans RS, Welch SJ, Haug PJ, Snow GL. Forecasting daily patient volumes in the emergency department. *Acad Emerg Med.* 2008;15(2):159-70. doi: 10.1111/j.1553-2712.2007.00032.x.
28. Valenzuela-Nunez C, Latorre-Nunez G, Troncoso F. Smart medical appointment

scheduling: optimization, machine learning and overbooking to enhance resource utilization. *IEEE Access.* 2024;12:7551-7562. doi: 10.1109/ACCESS.2024.3349953.
29. Masroor F, Gopalakrishnan A, Goveas N. Machine Learning-Driven Patient Scheduling in Healthcare: A Fairness-Centric Approach for Optimized Resource Allocation. *IEEE Wireless Communications and Networking Conference (WCNC)*. 2024:01-06, doi: 10.1109/WCNC57260.2024.10571017
30. Shi Y, Mahdian S, Blanchet J, Glynn P, Shin AY, Scheinker D. Surgical scheduling via optimization and machine learning with long-tailed data. *Health Care Manag Sci.* 2023;26:692–718. doi: 10.1007/s10729-023-09649-0
31. Pilgrim M, Willison S. Dive Into Python 3. 2nd ed. *Springer*; 2009.
32. Hip fracture - health quality ontario. https://www.hqontario.ca/portals/0/documents/evidence/quality-standards/qs-hip-fracture-clinical-guide-en.pdf
33. Hilbe JM. Logistic Regression Models. 1st ed. *Chapman & Hall*; 2009.
34. Breiman L. Random Forests. *Machine Learning.* 2001;45(1):5-32. doi: 10.1023/A:1010933404324
35. Cunningham P, Cunningham P, Delany SJ. k-Nearest Neighbour Classifiers. *ACM Comput Surv.* 2007;54(6):1-25. doi: 10.1145/3459665
36. Friedman JH. Multivariate Adaptive Regression Splines. *Ann. Statist.* 1991;19(1):1-67. doi: 10.1214/aos/1176347963.
37. Murphy KP. *Machine Learning: A Probabilistic Perspective*. Cambridge, Massachusetts: Massachusetts Institute of Technology; 2012.
38. Kingsford C, Salzberg SL. What are decision trees? *Nat Biotechnol*. 2008;26(9):1011-3. doi: 10.1038/nbt0908-1011
39. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics Intellig Lab Syst.* 2001;58(2):109-30. doi: 10.1016/S0169-7439(01)00155-1
40. Kubat M. Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7. *The Knowledge Engineering Review.* 1999;13(4):409-12. doi: 10.1017/S0269888998214044.
41. RStudio Team. RStudio: Integrated Development for R. 2020.
42. Karkkainen T. Extreme minimal learning machine: ridge regression with distance based basis. *Neurocomputing.* 2019;342:33-48. doi: 10.1016/j.neucom.2018.12.078
43. Core Team R. R: A Language and Environment for Statistical Computing. 2021.
44. Kuhn M. Building predictive models in R using the caret package. *J Stat Soft*. 2008;28(1):1-26. doi: 10.18637/jss.v028.i05
45. Kulkarni A, Chong D, Batarseh FA. 5 - Foundations of data imbalance and solutions for a data democracy. In: Batarseh FA, Yang R, editors. Data Democracy. Academic Press; 2020. p.83-106.
46. Abbas A, Mosseri J, Lex JR, Toor J, Ravi B, Khalil EB, et al. Machine learning using preoperative patient factors can predict duration of surgery and length of stay for total knee arthroplasty. *Int J Med Inform*. 2022;158:104670. doi: 10.1016/j.ijmedinf.2021.104670.
47. Stepaniak PS, Heij C, Vries GD. Modeling and prediction of surgical procedure times. *Statistica Neerlandica.* 2010;64(1):1-18. doi: 10.1111/j.1467-9574.2009.00440.x
48. Navarro SM, Wang EY, Haeberle HS, Mont MA, Krebs VE, Patterson BM, et al. Machine Learning and Primary Total Knee Arthroplasty: Patient Forecasting for a Patient-Specific Payment Model. *J Arthroplasty.* 2018;33(12):3617-23. doi: 10.1016/j.arth.2018.08.028.
49. Ramkumar PN, Karnuta JM, Navarro SM, Haeberle HS, Iorio R, Mont MA, et al. Preoperative Prediction of Value Metrics and a Patient-Specific Payment Model for Primary Total Hip Arthroplasty: Development and Validation of a Deep Learning Model. *J Arthroplasty*. 2019;34(10):2228-2234. doi: 10.1016/j.arth.2019.04.055.
50. Shorey S, Lau TC, Lau ST, Ang E. Entrustable professional activities in health care education: a scoping review. *Medical Education.* 2019;53:766–777. doi: 10.1111/medu.13879.

Appendices

Appendix I. Mathematical definition of demand and shift optimization.

Equation 1. Definition of the demand.

Suppose we want to model the demand $D\ (d)$ on a particular day(d) according to the various critical tasks (CTs) that are expected of orthopaedic residents, where $D \in R, D\ \geq 0$. Then, the demand on a particular day(d)may be defined as:

$$D\ (d)\ =\ O(d) \cdot \underline{t_O} + T(d) \cdot \underline{t_T} + A(d) \cdot \underline{t_A} + C(d) \cdot \underline{t_C}\ +\ R$$

, where $O(d)$ is the number of ORs complete per day, $T(d)$ is the number of traumas seen per day, $A(d)$ is the number of admissions per day, $C(d)$ is the number of consults seen per day, and R is the baseline amount of time required to round per day. Similarly, $\underline{t_O}$ is the average amount of time it takes to complete an OR, $\underline{t_T}$ is the average amount of time it takes to complete a trauma, $\underline{t_A}$ is the average amount of time it takes to complete an admission, and $\underline{t_C}$ is the average amount of time it takes to complete a consult.

Equation 2. Definition of optimization potential possible.

Suppose N_{OS} is the total number of shifts of the optimized schedule on an annual basis, where $n_{OS}(d)$ is the number of residents in the optimized schedule on a particular day d. Suppose C is the cut-off on the demand of the residents, where $C \in R, C\ \geq 0$. Then, the number of residents on a particular day can be modeled by:

$$n_{OS}(d)\ =\ \begin{cases} 1, if\ D(d) \leq C \\ 2, if\ D(d) > C \end{cases}$$

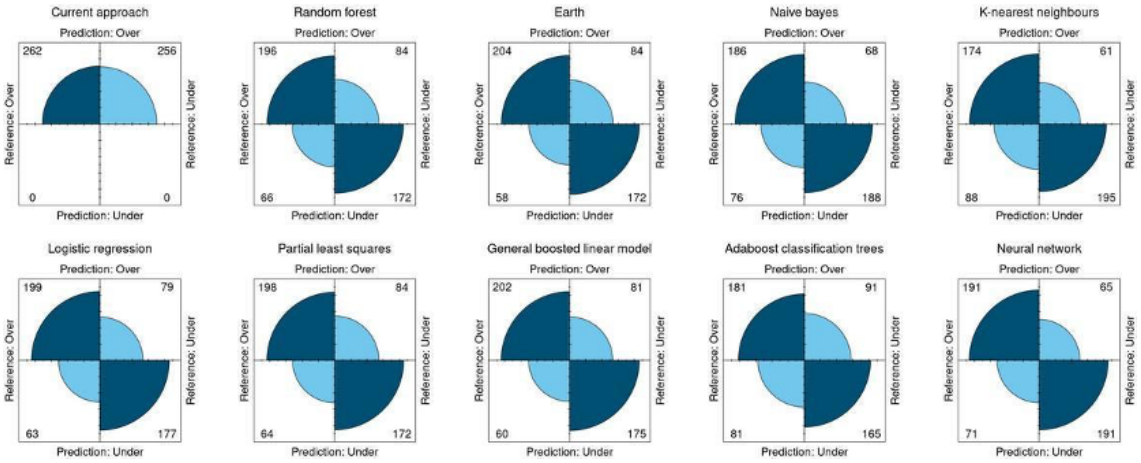
The total number of shifts of the optimized schedule may be defined by:

$$N_{OS} = \sum_{d=1}^{D} n_{OS}(d)$$

Accordingly, the annual reduction possible may be defined as:

$$Reduction\ Possible\ = \frac{N_{HS} - N_{OS}}{N_{HS}} \cdot 100\%$$

, where N_{HS} is the historic number of residents scheduled, and N_{OS} is the optimized number of residents scheduled.



Appendix II. Confusion matrices for all models on the testing set.. Under means the demand is under the 20-hour cut-off, while over is over the 20-hour cut-off. Each of these plots may be interpreted like a 2-by-2 epidemiologic table, with true positives and true negatives on the top left and bottom right corners and false positives and false negatives in the top right and bottom left corners.