International Journal of Forecasting 40 (2024) 1410–1420

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting emergency department occupancy with advanced machine learning models and multivariable input☆

Jalmari Tuominen [a,*], Eetu Pulkkinen [a], Jaakko Peltonen [b], Juho Kanniainen [b], Niku Oksala [a,c], Ari Palomäki [a,d], Antti Roine [a]

[a] Faculty of Medicine and Health Technology, Tampere University, Finland
[b] Faculty of Information Technology and Communication Sciences, Tampere University, Finland
[c] Centre for Vascular Surgery and Interventional Radiology, Tampere University Hospital, Finland
[d] Kanta-Häme Central Hospital, Hämeenlinna, Finland

ARTICLE INFO

Article history:
Dataset link: https://github.com/tuomijal/ed-ml-multivar

Keywords:
Emergency department
Crowding
Overcrowding
Forecasting
Multivariable analysis
Occupancy

ABSTRACT

Emergency department (ED) crowding is a significant threat to patient safety and it has been repeatedly associated with increased mortality. Forecasting future service demand has the potential to improve patient outcomes. Despite active research on the subject, proposed forecasting models have become outdated, due to the quick influx of advanced machine learning models and because the amount of multivariable input data has been limited. In this study, we document the performance of a set of advanced machine learning models in forecasting ED occupancy 24 h ahead. We use electronic health record data from a large, combined ED with an extensive set of explanatory variables, including the availability of beds in catchment area hospitals, traffic data from local observation stations, weather variables, and more. We show that DeepAR, N-BEATS, TFT, and LightGBM all outperform traditional benchmarks, with up to 15% improvement. The inclusion of the explanatory variables enhances the performance of TFT and DeepAR but fails to significantly improve the performance of LightGBM. To the best of our knowledge, this is the first study to extensively document the superiority of machine learning over statistical benchmarks in the context of ED forecasting.

1. Introduction

Emergency department (ED) crowding is a well-known threat to patient safety (Boyle, Beniuk, Higginson, & Atkinson, 2012), and the documented adverse effects range from a decrease in the work satisfaction of ED staff (Eriksson, Gellerstedt, Hillerås, & Craftman, 2018) to increased length of stay (McCarthy et al., 2009) and increased mortality (Berg et al., 2019; Guttmann, Schull, Vermeulen, & Stukel, 2011; Jo et al., 2014; Richardson, 2006). In contrast to outpatient clinics[1] or elective surgery, EDs are unable to adjust the inflow of patients and are thus exposed to both a stochastic incidence of diseases and changes in patients' care-seeking behavior. Moreover, EDs are usually unable to freely adjust the outflow of patients, since they depend on other healthcare facilities to organize follow-up care when necessary. The only component that an ED can independently adjust is throughput, which is mostly affected by the quantity (Bucheli & Martina, 2004) and quality (Trotzky et al., 2021) of staff. These restrictions lead to repeated crowding which has developed into a global public health crisis (Pearce, Marchand,

☆ The results presented in the paper were reproduced by the Editor-in-Chief on the 18th of December 2023.

* Corresponding author.
E-mail address: jalmari.tuominen@tuni.fi (J. Tuominen).

[1] An outpatient clinic is a healthcare facility where patients receive treatment without being admitted to the hospital

https://doi.org/10.1016/j.ijforecast.2023.12.002

Shannon, Ganshorn, & Lang, 2023). Sufficiently accurate forecasts of future service demand would enable proactive administrative decisions aiming to alleviate or even prevent crowding, and has the potential to improve patient outcomes. This rationale has motivated an increasing amount of ED forecasting articles (Gul & Celik, 2018), but for some reason, readily available commercial solutions have not emerged. We believe this to be due, at least in part, to increasingly outdated forecasting methods and the lack of relevant multivariable input.

First, a significant amount of ED forecasting literature has focused on the autoregressive integrated moving average (ARIMA) or its variants (Gul & Celik, 2020). The tendency to favor an ARIMA model over advanced models is understandable, as it has up to very recent years repeatedly defended its place over more complex solutions (Cheng et al., 2021; Whitt & Zhang, 2019; Zhou, Zhao, Wu, Cheng, & Huang, 2018) and has been considered a pinnacle of the statistical approach in time series forecasting in general (Oreshkin, Carpov, Chapados, & Bengio, 2019). This has been changing rapidly. In 2020, a statistical and deep learning (DL) hybrid proposed by Smyl (2020) outperformed statistical benchmark models in the renowned M4 time series forecasting competition for the first time in the history of the competition (Makridakis, Spiliotis, & Assimakopoulos, 2020). Following this result, several time-series-specific DL architectures have been introduced, such as the temporal fusion transformer (TFT) by Lim, Arık, Loeff, and Pfister (2021), neural expansion analysis for interpretable time series forecasting (N-BEATS) by Oreshkin et al. (2019), and DeepAR by Salinas, Flunkert, Gasthaus, and Januschowski (2020). In the latest M5 competition in 2022 (Makridakis, Spiliotis, & Assimakopoulos, 2022), all of the best-performing models were pure ML implementations, and out of the five best-performing solutions, four utilized a multivariable LightGBM model (Ke et al., 2017). In addition to LightGBM, DeepAR and N-BEATS were highlighted for showing forecasting potential. However, these models have not been tested using ED data.

Second, the amount and quality of used input data have been limited. This is an important deficit because of the highly interdependent nature of the ED. As suggested by Asplin et al. (2003), ED crowding is a sum of three operational components: input (number of arrivals), throughput (length of stay, mainly affected by staffing resources), and output (mainly affected by the availability of follow-up care beds). A disturbance in one of these components alone can lead to crowding, but the most severe situations are observed when two or more of them are detrimentally affected. This is in line with findings of M5, in which one of the seven key implications of the competition was the importance of exogenous variables (Makridakis et al., 2022). Regardless, ED forecasting input data have repeatedly consisted of simple calendar and weather variables, mounting up to 29 input vectors in total (Holleman, Bowling, & Gathy, 1996; Jiang, Chin, & Tsui, 2018; Whitt & Zhang, 2019), which leaves a significant proportion of Asplin's model unaccounted for.

Third, the studies that have suggested the utility of novel input variables, such as website visits (Ekström et al., 2015), road traffic flow (Rauch, Hübner, Denter, & Babitsch, 2019), or the emergency department severity index (Cheng et al., 2021), have done so by utilizing only one of them at a time. This runs the risk of an overoptimistic evaluation of variable importance, due to the inevitable multicollinearity between them. We thus believe that a data-centric approach with a high number of input variables has the potential to increase predictive accuracy, provide a better understanding of the factors underlying crowding, and even inform policies among local healthcare providers. This is a continuation of our previous work, in which we used simulated annealing and floating search to perform feature selection in order to enhance accuracy when using conventional statistical models (Tuominen, Lomio, & Palomäki, 2022).

To conclude, our contributions are as follows: (1) we investigate the performance of state-of-the-art ML models in predicting ED occupancy using data spanning over two years in a large, combined ED; (2) we use the largest-to-date collection of explanatory variables, containing not only weather and calendar variables but also the availability of hospital beds, traffic information, local public events, website visits, and more, and (3) we analyze the proportional importance of these variables when used in conjunction with one another. We show that ML models outperform statistical benchmarks with ED data, reproduce the well-established superiority of LightGBM over other ML models, and show that while the explanatory variables enhance the performance of TFT and DeepAR, they do not significantly improve the performance of LightGBM.

2. Materials and methods

2.1. Datasets and data splitting

Tampere University Hospital is an academic hospital located in Tampere, Finland. It serves a population of 535,000 in the Pirkanmaa Hospital District and, as a tertiary hospital, an additional population of 365,700, providing level 1 trauma center capabilities. The hospital ED, *Acuta*, is a combined ED with a total capacity of 111–118 patients, with 70 beds (and an additional seven beds in reserve) and 41 seats for walk-in patients. Approximately 100,000 patients are treated annually. For this study, all registered ED visits were obtained from a hospital database created during the sample period from January 1, 2017, to June 19, 2019. All remote consultations and certifications of death without prior medical interventions, as well as hourly duplicates, were excluded.

Data splitting for hyperparameter optimization. To optimize hyperparameters, we divided the dataset into training and validation sets. To account for yearly seasonal patterns, we ensured that the training set covered a 12-month period, capturing the complete spectrum of seasonal variations. The training set spanned January 1, 2017, to December 31, 2017, comprising 8760 data points, while the validation set extended from January 1, 2018, to June 19, 2018, containing 4080 data points. During hyperparameter optimization, the models were trained on the training set, and their performance was assessed using

the validation set. The optimal hyperparameters were determined based on the validation set's performance. All models underwent hyperparameter optimization via the tree-structured Parzen estimator (TPE) method (Bergstra, Bardenet, Bengio, & Kégl, 2011), using the optimization framework by Akiba, Sano, Yanase, Ohta, and Koyama (2019). Details on the number of tested hyperparameter combinations and search spaces can be found in Appendix A.

Re-training protocol and testing. Due to the temporal nature of data, it is important to train the final models on the recent data available. For that reason, after optimizing the model's hyperparameters, we adopted the data-splitting principle for subsequent re-training.

We evaluated the models using data from June 20, 2018, to June 19, 2019, a span of 365 days. To facilitate periodic re-training, the data were divided into 13 folds. The first 12 folds each consisted of 720 data points, equivalent to 24 h multiplied by 30 days, roughly approximating a month. The final fold consisted of a residual of 120 h. The models were re-trained at the start of each month using the preceding 12,840 data points. For instance, the training duration for the initial fold ranged from January 1, 2017, to June 19, 2018. This period, which totals 12,840 data points (8760 + 4080), matched the dataset used for hyperparameter optimization. The re-training window was then rolled forward for evaluations on the subsequent folds.

Each prediction spanned a horizon of 24 h. Every day, a 24 h forecast was generated at 00:00 based on the models re-trained at the start of the respective month (fold). Given that the testing phase encompasses 365 days across all folds, this produces matrices of dimensions 365×24 for each model. For DL models, the data were normalized to fit within the range [0, 1] before analysis.

2.1.1. Explanatory variables

For the purposes of this study, we collected 167 explanatory variables from multiple data sources with the goal of covering as much of the three components of Asplin's model as possible. These variables as summarized in Table 1 and briefly introduced below. All covariates are divided into two categories: past covariates (P) and future covariates (F). P features refer to variables that are not known in the future (e.g. hospital bed capacity), in contrast to F features, which are always known both in the future and in the past (e.g. hour of the day).

Hospital beds. The temporal availability of hospital beds in 24 individual hospitals or healthcare centers in the catchment area was included as provided by the patient logistics system Uoma® by Unitary Healthcare Ltd. in hourly resolution. The impact of the availability of beds on ED service demand is two-fold. First, a low availability of beds leads to a prolonged length of stay, since patients remain in the ED after initial treatment while they wait for an available follow-up care bed. This kind of access block leads to the accumulation of patients in the ED, and both clinical and empirical evidence has shown that this effect is a significant contributor to overcrowding (Morley, Unwin, Peterson, Stankovich, & Kinsman, 2018). Second, a low availability of beds sometimes forces primary healthcare physicians to refer patients to an ED merely to organize the bed that the patient requires, which again contributes to occupancy. Bed capacity statistics are visualized in Fig. 1, and the locations of the facilities, along with their distance from the study hospital, are provided in Appendix C.

Traffic. Hourly traffic data were obtained from an open database maintained by Fintraffic Ltd., a company operating under the ownership and direction of the Finnish Ministry of Transport and Communications (*Fintraffic: Digitraffic open data from Finnish roads*, 2021). Data from all 33 bidirectional observation stations in the Pirkanmaa Region were included, resulting in 66 traffic feature vectors, each containing the number of cars that passed the observation station each hour. The acquisition of traffic variables was motivated by the work by Rauch et al. (2019), which suggested that traffic variables might increase predictive accuracy when used as an input in an ARIMAX model. The locations of the observation stations, along with their distance from the study hospital, are provided in Appendix C.

Weather. Ten historical weather variables were collected from the nearest observation station located in Härmälä, Tampere, 600 meters from the city center, using open data provided by the Finnish Meteorological Institute (*Finnish Meteorological Institute Open Weather Data*, 2020). The inclusion of weather variables was inspired by the work by Whitt and Zhang (2019). We assumed that weather can be forecasted with satisfying accuracy one day in advance and, for this reason, used next-day weather variables as future covariates.

Public events. City of Tampere officials provided us with an exhaustive historical event calendar, containing all public events ranging from small to large gatherings that were organized during the sample period in the Tampere area. Using these data, we created a time series containing the number of public events organized each day in the sample period. Hypothetically, an increased number of citizens engaging in festivities—often with increased substance consumption—might be associated with ED service demand.

Website visits. Data on website visits to two hospital domains were provided by the hospital IT department. Data were available on two domains: tays.fi (Domain 1, D1) and tays.fi/acuta (Domain 2, D2), the former being the hospital's homepage and the latter the homepage of the hospital's ED. D1 visit data were available in hourly resolution, whereas D2 data were only available in daily resolution. Using D1 visits, we also summed up visits between 18:00 and 00:00 in an identical manner to that proposed by Ekström et al. (2015) (Domain 1_{EV}). In addition, we included a stationary version of this variable by dividing evening visits by earlier visits during the day (Domain 1_{ER}). The number of Google searches for the search term *Acuta* were also extracted from Google Trends (*Google Trends*, 2020).

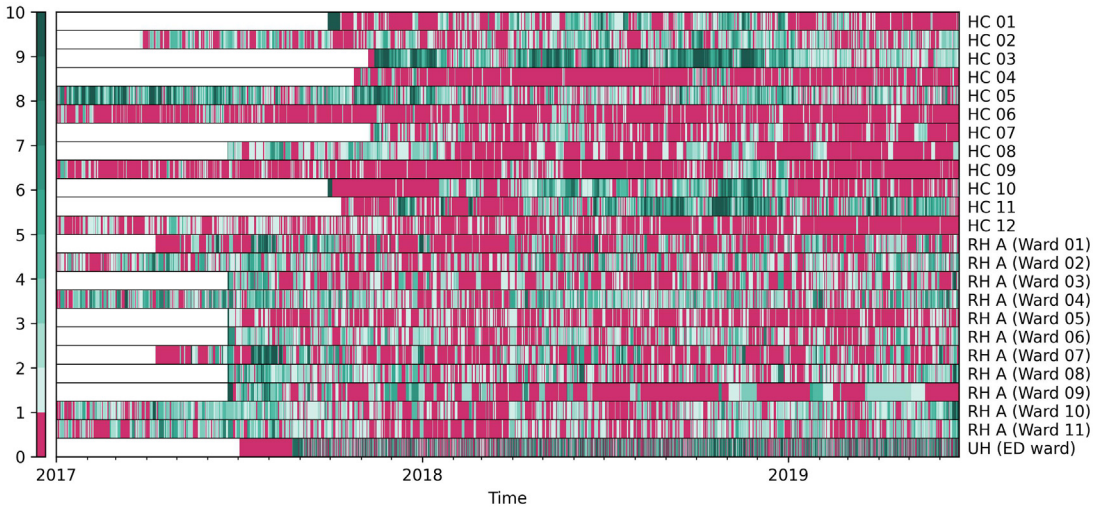


Fig. 1. Hospital bed capacity statistics. HC = health center, RH = regional hospital, UH = university hospital. Red color indicates that no beds are available, and number of available beds is shown with hues of green.

Calendar variables. Weekdays and months were included as categorical variables. Timestamps of national holidays were provided by the University Almanac Office (*University of Helsinki Almanac Office*, 2020), and each of them was included as a binary vector. Inspired Whitt and Zhang (2019), we included so-called holiday lags, which encode whether the three previous or three following days were holidays. We also included the number of previous consecutive holidays, encoding how many consecutive holidays preceded the day of interest. A binary encoding of a day's status as a working or non-working day was also included.

Technical analysis. In addition to the exogenous explanatory features described above, we engineered 30 features using the endogenous signal of the target variables. These variables range from a set of moving averages and mathematical moments to econometric indicators, and they are introduced in detail in Appendix B.

2.1.2. Feature sets

Using a high number of input features from multiple data sources poses a significant challenge if the predictive model is to be implemented in a real-life clinical setting, which increases the cost of building and maintaining the system as well as its fragility. For this reason, we tested the models with two sets of inputs: one containing all variables listed above (feature set $A, n = 167$), and one containing nothing but the target variable history (feature set U) as an input. Each multivariable model was tested with both A and U inputs, and they are distinguished from one another with naming convention of M_F, in which M stands for the model name and F for the feature set. For example, LGBM_A refers to a LightGBM model trained and tested with all available data.

2.1.3. Target variable

In this study we focus on predicting absolute non-stratified hourly ED occupancy. This includes both bed-occupying and walk-in patients in all treatment spaces of the ED. Occupancy was selected as the target variable because it is affected by all three components of Asplin's model, since input, throughput, and output all contribute to total occupancy—in contrast to arrivals, which is by definition affected primarily by input.

Table 1
Explanatory variable list. P = past covariate, i.e. a value that is not known into the future at the prediction time; F = future covariate, i.e. a value that is known both in the past and in the future. Some variables are provided in the appendices for brevity.

Feature group	Name	Number
Hospital beds	See Appendix C	$P_{1–33}$
Calendar variables	Holiday name	F_{34}
	Holiday lags	$F_{35–41}$
	Hour	F_{42}
	Working day	F_{43}
	Month	F_{44}
	Preceding holidays	F_{45}
	Weekday	F_{46}
Public events	All events	F_{47}
TA indicators	See Appendix B	$P_{48–85}$
Traffic	See Appendix C	$P_{86–151}$
	Google Trends	P_{152}
Weather	Air pressure	F_{153}
	Air temperature	F_{154}
	Cloud count	F_{155}
	Day air temperature max	F_{156}
	Day air temperature min	F_{159}
	Dew point temperature	F_{158}
	Rain intensity	F_{159}
	Relative humidity	F_{160}
	Slipperiness	F_{161}
	Snow depth	F_{162}
	Visibility	F_{163}
Website visits	Domain 1	P_{164}
	Domain 1_{EV}	P_{165}
	Domain 1_{ER}	P_{166}
	Domain 2	P_{167}
	Total	A_{167}

Table 2
Hourly seasonality of the absolute occupancy.

Hours	Mean	Std	Min	25%	50%	75%	Max
00:00–03:00	24.70	10.00	2	17	24	31	75
04:00–07:00	16.27	6.22	2	12	16	20	57
08:00–11:00	31.56	10.97	5	23	31	39	82
12:00–15:00	59.14	13.12	22	50	59	68	117
16:00–19:00	62.74	14.86	22	52	62	73	124
20:00–23:00	46.58	13.85	11	37	46	56	99

2.1.4. Performance metrics

Continuous metrics. We provide three continuous error metrics: the mean absolute error (MAE) and root mean squared error (RMSE) for point forecasts, and the mean scaled interval score (MSIS) for prediction intervals. The MAE is calculated as follows:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|, \tag{1}$$

where y_t is the ground truth, and $\hat{y}$ is the prediction. The MAE was used to calculate the proportional difference between the models and relevant benchmark, and for statistical tests.

Point forecasts were also evaluated using the RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{2}$$

All the models investigated in this study are probabilistic in nature and were configured to produce 95% prediction intervals in addition to point forecasts. We thus also quantify the performance of these prediction intervals using the MSIS, as proposed by Gneiting and Raftery (2007):

$$\text{MSIS} = \frac{\Sigma_{t=n+1}^{n+h}(u_t - l_t) + \frac{2}{\alpha}\mathbb{1}\{y_t < l_t\} + \frac{2}{\alpha}(y_t - l_t)\mathbb{1}\{y_t > u_t\}}{h \times \frac{1}{n-m}\Sigma_{t=m+1}^{n}|y_t - y_{t-m}|} \tag{3}$$

where u_t and l_t are the upper and lower bounds, respectively, x_t is the ground truth, and α is the significance level which, was set to 0.05 based on the levels of the generated prediction intervals.

Statistical significance. Statistical significance testing of hourly absolute error rates between ARIMA and the other models was performed using Kruskal–Wallis one-way analysis for variance with Dunn's post hoc test and Holm's correction for multiple pairwise comparisons. This aimed to provide a similar approach to that for the M3 competition (Koning, Franses, Hibon, & Stekler, 2005). The significance level was specified as $p < 0.05$. Statistical tests were performed using software by Virtanen et al. (2020) and Terpilowski (2019).

2.1.5. Feature importance analysis

Studying how the model selects and weights the features the predictions are based upon can provide insight into the factors affecting occupancy and the reasons underlying good or bad forecasting performance. In this study, this was performed using Shapley additive explanations (SHAP), as proposed by Lundberg and Lee (2017). SHAP assigns a unique importance value to each feature in a prediction by quantifying its contribution to the prediction outcome. SHAP values are based on cooperative game theory principles, calculating the average marginal contribution of each feature across different coalitions of features and providing a unified and interpretable explanation for individual predictions. For brevity, we limit our attention to the importance statistics of the best-performing model.

2.2. Models

Model definition, training, and backtesting were handled using software by Herzen et al. (2022), which provided a unified interface to the underlying models and their implementation, unless otherwise stated. We document the performance of four forecasting models:

- **TFT** (the temporal fusion transformer) is a deep learning model designed for interpretable time series forecasting, combining recurrent layers, multivariable attention mechanisms, and static covariate encoders to capture complex temporal patterns and interdependencies (Lim et al., 2021).
- **N-BEATS** (neural basis expansion analysis) is a deep learning architecture that decomposes the past values of a time series using a set of basis expansion blocks, eliminating the need for prior knowledge of the underlying temporal patterns (Oreshkin et al., 2019).
- **DeepAR** is a probabilistic forecasting model utilizing an autoregressive recurrent network structure, typically trained on large collections of related time series, to produce point and probabilistic forecasts (Salinas et al., 2020).
- **LightGBM** is a gradient boosting framework that employs a histogram-based algorithm, optimized for speed and efficiency, while handling large datasets and supporting both classification and regression tasks (Ke et al., 2017).

Benchmark models. Four models were used for benchmarking purposes: seasonal naïve (SN); the autoregressive integrated moving average (ARIMA); and two ETS models, namely Holt–Winter's seasonal damped method (HWDM) and Holt–Winter's additive method (HWAM). A 168-hour sliding window was used for all models. ARIMA parameters were defined with AutoARIMA as initially described by Hyndman and Khandakar (2008), using the stepwise approach and Python implementation by Garza, Canseco, Chall, and Olivares (2022). A priori known hourly seasonality was provided to the AutoARIMA model as a parameter.

3. Results

Descriptive statistics. The inclusion criteria resulted in a sample of 210,019 individual visits that occurred during the 21,600-hour sample window. The hourly seasonality of the absolute occupancy demonstrated a sinusoidal

Table 3
Continuous performance of the tested models. FS = feature set, MAE = mean absolute error, RMSE = mean squared error, MSIS = mean scaled interval score. 95% confidence intervals in parenthesis.

Model	FS	MAE	Delta (%)	p	RMSE	MSIS
SN	U	9.53 (9.39–9.67)	−22	<.001	12.58 (12.40–12.76)	–
HWAM	U	9.44 (9.30–9.55)	−21	<.001	12.20 (12.03–12.39)	56
HWDM	U	9.20 (9.06–9.34)	−18	<.001	11.91 (11.76–12.08)	55
DeepAR	U	8.56 (8.43–8.68)	−10	<.001	11.50 (11.34–11.68)	56
ARIMA	U	7.78 (7.67–7.89)	0	–	10.50 (10.33–10.64)	65
TFT	U	7.18 (7.06–7.30)	8	<.001	9.58 (9.42–9.72)	43
DeepAR	A	7.10 (6.98–7.20)	9	<.001	9.39 (9.25–9.54)	62
N-BEATS	U	6.98 (6.89–7.08)	10	<.001	9.33 (9.20–9.47)	45
TFT	A	6.86 (6.74–6.96)	12	<.001	9.04 (8.92–9.16)	48
LightGBM	U	6.66 (6.54–6.75)	14	<.001	8.95 (8.80–9.07)	55
LightGBM	A	6.63 (6.52–6.73)	15	<.001	8.77 (8.64–8.90)	54

shape (see Table 2), with the lowest median occupancy of 16 between 04:00 and 07:00 and the highest median occupancy of 62 between 16:00 and 19:00 The minimum, median, and maximum occupancies were 2, 38, and 124, respectively.

Missing data. There was a significant amount of missing data in the case of available hospital beds, as can be seen in Fig. 1. In total, data were missing for 77,636 h (11%) out of the total 518,400 h for all facilities combined. The amount varied significantly between facilities, from 0–7486 h (0%–35%), due to the gradual introduction of the Uoma® software to each facility. All missing data were imputed with the mean of the other hospitals at a given time.

3.1. Model performance

Aggregated performance. Continuous performance results are provided in Table 3. Kruskall–Wallis showed statistically significant differences between the models with $p = 0.0$. LightGBM$_A$ was the best-performing model, with MAE and RMSE values of 6.63 and 8.77, respectively, yielding a 15% improvement over ARIMA ($p < 0.001$). LightGBM$_U$ was the second-best model, with an MAE of 6.66 and a proportional improvement of 14% over ARIMA ($p < 0.001$). TFT$_A$ outperformed ARIMA, with an MAE of 6.86 ($p < 0.001$) yielding a 12% proportional improvement. N-BEATS was the fourth-best model, with an MAE of 6.98 ($p < 0.001$) and a 10% proportional improvement, followed by DeepAR$_A$ with 9% ($p < 0.001$) and finally TFT$_U$ with 8% ($p < 0.001$). DeepAR$_U$ was 10% worse than ARIMA ($p < 0.001$). TFT$_U$, N-BEATS, and TFT$_U$ had the lowest MSISs of 43, 45, and 48, respectively, compared to ARIMA's 65.

Horizontal performance. The hourly accuracy of each model stratified by the forecasting horizon is provided in Fig. 2. The figure shows that both the absolute errors and the differences between the errors are greatest in the afternoon, which follows the sinusoidal shape of the variance of the target variable (see Table 2). LightGBM$_U$ was consistently the best univariable model regardless of the forecast horizon. Multivariable models performed in a very similar manner to one another, except for the very short forecast horizons, during which LightGBM$_A$ outperformed the others.

Monthly performance. The performance of the univariable models over different months of the test set is provided in Fig. 3. TFT, LightGBM, and N-BEATS outperformed the benchmarks consistently, whereas DeepAR was several times bested by statistical models. Overall, the errors were higher in December and October—compared to April or January, for example.

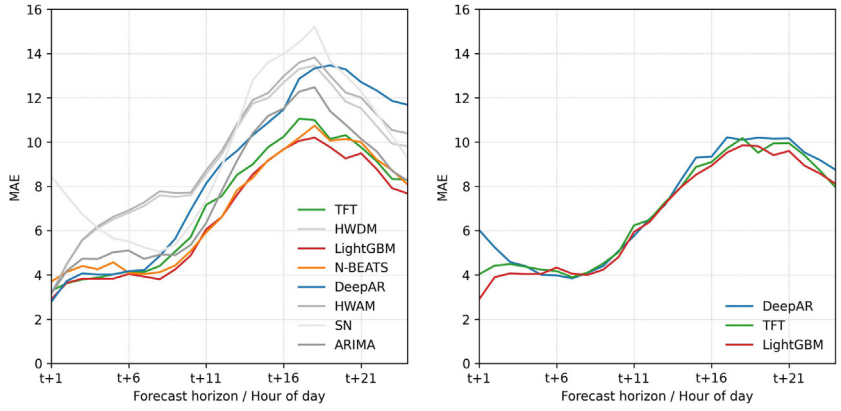
3.2. Feature importance analysis

The proportional absolute mean SHAP values for the 20 most important features for LightGBM$_A$ model are visualized in Fig. 4 separately for horizons $t+1$ and $t+24$. For $t + 1$, the target variable itself at lag $t - 1$ was the most important variable, followed by the CMO and RSI indicators and 16 traffic variables. For $t + 24$ predictions, 18 traffic variables were included in the top 20, along with website visit statistics to domain 1 and the AO indicator.

4. Discussion

In this study, we establish four main findings regarding the use of ML and multivariable input data for forecasting aggregated ED occupancy. We discuss each of these separately below.

ML models outperform statistical models. All of the tested ML models significantly outperformed the ARIMA benchmark with 8%–15% proportional improvements. In the case of N-BEATS, Oreshkin et al. (2019) initially reported a 10% improvement over statistical models, which aligns with the 10% improvement reported here. For DeepAR, Salinas et al. (2020) reported an average 15% improvement over innovation-state space and autoregressive models, which is higher than the 8% reported here. This might be due to our relatively small dataset, since DeepAR is designed to work at scale. TFT outperformed the other DL models, which aligns with both Lim et al. (2021) and Elsayed, Thyssens, Rashed, Jomaa, and Schmidt-Thieme (2021). However, it was outperformed by LightGBM, which contradicts Elsayed et al. (2021), who documented TFT to be superior to gradient boosting regression trees such



(a) Univariable models (b) Multivariable models

Fig. 2. Horizontal error, as measured by the mean absolute error (MAE). The errors follow the sinusoidal shape of the standard deviation of the target variable (see Table 2), which peaks at 16:00–19:00 and then decreases towards the end of the day.

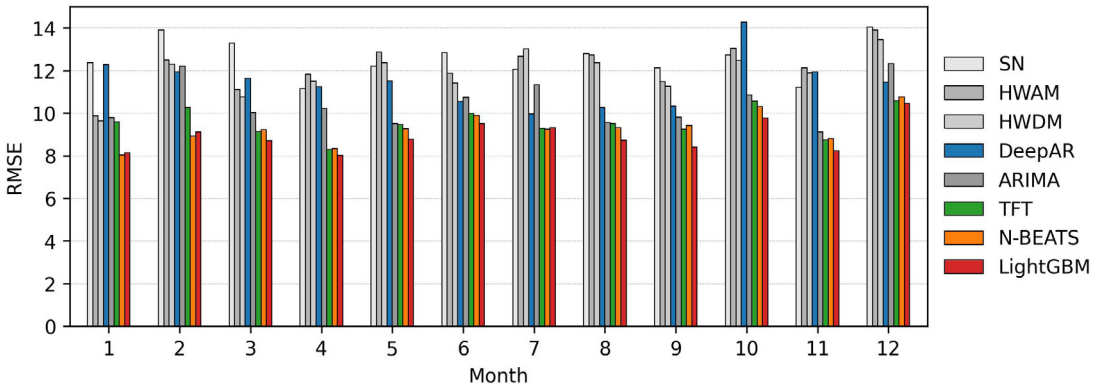
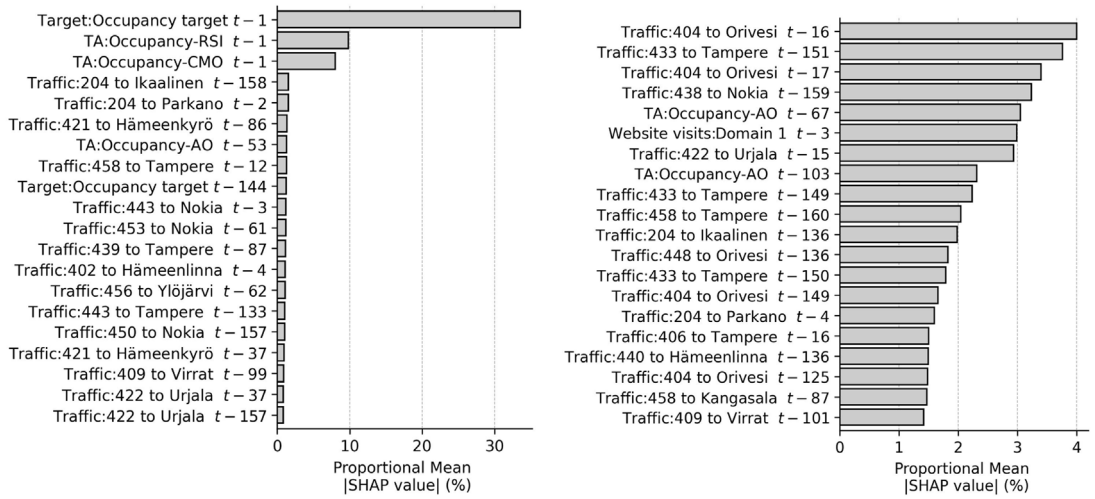


Fig. 3. Performance of the univariable models over different months of the test set.



(a) $t + 1$ importance (b) $t + 24$ importance

Fig. 4. Feature importance SHAP statistics for the 20 most important features used with the LightGBM$_A$ model.

as LightGBM. Regardless, the breakthrough of ML in time series applications seems to apply to ED data as well.

LightGBM outperforms DL models. LightGBM was the best-performing model using both univariable data and multivariable data. In fact, the univariable LightGBM$_U$ outperformed multivariable DL models, which highlights the performance of the model and undermines the value of multivariable data in this problem setting, as discussed in more detail below. The performance of the model is in line with that of the M5 competition (Makridakis et al., 2022) and serves to show that ED statistics do not differ in terms of predictability from other time series of interest. The superior performance of tree-based methods over deep learning algorithms in processing tabular data is a recognized phenomenon. This advantage can be linked to two factors: their proficiency in managing uninformative features, and their ability to identify and leverage irregular patterns within the target function. Additionally, tree-based methods can exploit the specific structure of the data as it is, even with data whose descriptors are not invariant to transformations, such as rotation, translation, and scaling (Grinsztajn, Oyallon, & Varoquaux, 2022).

Regardless, LightGBM is a good candidate for a model that is not only excellent in performance but also efficient in terms of required computation, since the univariable LightGBM took only 18 min to execute on a CPU, which is 86%–96% less than the DL models on a GPU (see Appendix D).

Exogenous multivariable data are of limited importance. Exogenous variables improved the performance of both TFT and DeepAR by 4% and 17%, respectively. However, for the top-performing model, LightGBM, there was negligible difference between its multivariable and univariable versions. This indicates that richer models do not necessarily outperform when applied to out-of-sample data. The observed limited importance of exogenous variables, along with LightGBM's superiority, aligns with Occam's razor. This principle, also called the law of parsimony, advocates for the use of models and procedures that contain all that is necessary for modeling but nothing more (Hawkins, 2004).

This finding has two practical implications. First, showing that one or more exogenous variables improve the performance of a model does not mean that a better univariable model does not exist. It is thus important to prioritize finding high-performing univariable models before extensively collecting all potential exogenous data. Second, from a practical standpoint, it is fortunate that multivariable data do not significantly enhance the overall performance in our specific context, since implementing a univariable forecasting model is considerably simpler than implementing a multivariable one, particularly when the latter demands continual data collection from multiple sources.

Lack of prominent covariates. The intuitive association between follow-up care capacity and increased occupancy is strong, and we expected improved performance from including these variables. However, this was not observed, perhaps due to a non-trivial amount of missing data in hospital bed variables, as described in Section 2.1. The traffic monitoring nodes and their associated time lags appear to have been chosen somewhat arbitrarily. Moreover, the absolute SHAP values of the traffic variables increase in relation to the forecast horizon, as seen in Fig. 4. We attribute this to multicollinearity between the calendar and traffic variables. The traffic variables not only capture the target variable's weekly seasonality but also provide a marginal yet beneficial signal. This marginal advantage contrasts with Rauch et al. (2019), who suggested that incorporating these variables could enhance forecasting accuracy by 10%–20%.

We believe that the most important underlying reason for the unexpectedly low value of exogenous covariates is the unstratified sample used in this study. The ED under consideration operates as a Nordic combined facility, serving a highly heterogeneous population that includes medical, surgical, neurological, and psychiatric patients. These patients are further categorized into walk-in versus bed-occupying patients, as well as discharged versus admitted patients. In this study, we combined all patient categories into a single aggregated occupancy metric. This is important because, for example, only 39% of the presented patients were admitted after the initial assessment, meaning that follow-up care capacity had no direct impact on the remaining 61% of the patients. It is also possible that there are associations between the explanatory variables and some of the subgroups, but these associations are contradicted by other opposite associations between other subgroups. This heterogeneity has to be accounted for in future work.

4.1. Limitations

This study is limited by its retrospective single-center setup, and further work is required to investigate the applicability of our approach to other facilities, preferably in a prospective setting by integrating a production prototype to hospital information infrastructure. In the training set, we observed missing data on hospital bed variables, which might obscure their importance. This study focused on an unstratified sample, meaning that pooled occupancy statistics of all treatment rooms were used as the target variable. This is a limitation because occupancies of different treatment spaces might have very different interactions with included explanatory variables. We retrained models each month to limit the computational cost in the backtesting phase (and even with this restriction, it took 34 h of computation to produce the TFT results; see Appendix D). It is likely that more frequent retraining (e.g. weekly) would enhance the performance even more.

5. Conclusions

In this study, we set out to investigate the applicability of advanced ML models and intuitively relevant multivariable data in forecasting ED occupancy. Our results suggest that, in an ED forecasting context, (1) ML outperforms conventional statistical models, as has been demonstrated with other datasets; (2) LightGBM outperforms other DL

methods, which is in line with previous work; (3) extensive multivariable input data do not significantly improve model performance when forecasting unstratified occupancy statistics, and (4) a clear association of occupancy and any of the used covariates was not observed.

We identify several directions for follow-up studies. Stratifying the total visit statistics by functional subunits of the ED (walk-in clinic, medical, surgical, etc.) would likely produce different feature importance statistics, potentially improve model performance, and certainly be of operational value. We also note that classification methods should be investigated in forecasting binary crowded states. Finally, the performance of the models has to be evaluated in a prospective setting.

CRediT authorship contribution statement

Jalmari Tuominen: Study design, Data collection, Data analysis, Manuscript preparation. **Eetu Pulkkinen:** Data analysis, Manuscript preparation. **Jaakko Peltonen:** Study design, Technical supervision. **Juho Kanniainen:** Study design, Technical supervision. **Niku Oksala:** Study design, Data collection, Medical supervision. **Ari Palomäki:** Study design, Medical supervision. **Antti Roine:** Study design, Medical supervision, Manuscript preparation.

Declaration of competing interest

NO is a shareholder in Unitary Healthcare Ltd., which has developed a patient logistics system currently used in the study emergency department. JT, AR, and EP are shareholders in Aika Analytics Ltd., a company specializing in time series forecasting.

Data and code availability

Data and code used to produce (and reproduce) the results are openly available at https://github.com/tuomijal/ed-ml-multivar.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used GPT-4 by OpenAI (USA) in order to enhance the language and readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

We acknowledge Unitary Healthcare Ltd. for providing the dataset on available hospital beds, the City of Tampere for providing timestamps for public events, and Tampere University Hospital's information management for providing website visit statistics. We also acknowledge CSC - IT Center for Science for providing computational resources, and specifically D.Sc. Mats Sjöberg for technical support. All authors have read and approved the final manuscript.

Funding

The study was funded by the Finnish Ministry of Health and Social Welfare via the Medical Research Fund of Kanta-Häme Central Hospital; the Finnish Medical Foundation, Finland; the Competitive State Research Financing of the Expert Responsibility Area of Tampere University Hospital, Pirkanmaa Hospital District, Grant 9X040 and Academy of Finland Grant 310617; Hauho Oma Savings Bank Foundation; and Renko Oma Savings Bank Foundation.

Ethics approval and consent to participate

Since our study was retrospective in nature, ethics committee approval was not required. An institutional approval was acquired prior to data collection, with the following specifications:

- Name: Potilaslogistiikan häiriötekijöiden tunnistaminen ja mallintaminen
- Number: PSHP/R19565
- Date: 16 June 2019

Consent for publication

Not applicable

Appendix A–D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2023.12.002.

References

Akiba, Takuya, Sano, Shotaro, Yanase, Toshihiko, Ohta, Takeru, & Koyama, Masanori (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2623–2631).

Asplin, Brent R., Magid, David J., Rhodes, Karin V., Solberg, Leif I., Lurie, Nicole, & Camargo, Carlos A. (2003). A conceptual model of emergency department crowding. *Annals of Emergency Medicine*, *42*(2), 173–180. http://dx.doi.org/10.1067/mem.2003.302.

Berg, Lena M., Ehrenberg, Anna, Florin, Jan, Östergren, Jan, Discacciati, Andrea, & Göransson, Katarina E. (2019). Associations between crowding and ten-day mortality among patients allocated lower triage acuity levels without need of acute hospital care on departure from the emergency department. *Annals of Emergency Medicine*, *74*(3), 345–356. http://dx.doi.org/10.1016/j.annemergmed.2019.04.012.

Bergstra, James, Bardenet, Rémi, Bengio, Yoshua, & Kégl, Balázs (2011). Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems 24: 25th annual conference on neural information processing systems 2011* NIPS 2011, (pp. 1–9).

Boyle, Adrian, Beniuk, Kathleen, Higginson, Ian, & Atkinson, Paul (2012). Emergency department crowding: Time for interventions and policy evaluations. *Emergency Medicine International*, *2012*, 1–8. http://dx.doi.org/10.1155/2012/838610.

Bucheli, Bruno, & Martina, Benedict (2004). Reduced length of stay in medical emergency department patients: A prospective controlled study on emergency physician staffing. *European Journal of Emergency Medicine*, *11*(1), 29–34. http://dx.doi.org/10.1097/00063110-200402000-00006.

Cheng, Qian, Tanik, Nilay, Scott, Christopher, Liu, Yufeng, Platts-mills, Timothy F, & Ziya, Serhan (2021). American journal of emergency medicine forecasting emergency department hourly occupancy using time series analysis. *American Journal of Emergency Medicine*, *48*, 177–182. http://dx.doi.org/10.1016/j.ajem.2021.04.075.

Ekström, Andreas, Ed, M., Kurland, Lisa, Farrokhnia, Nasim, Castrén, Maaret, & Nordberg, Martin (2015). Forecasting emergency department visits using internet data. *Annals of Emergency Medicine*, *65*(4), 436–442.e1. http://dx.doi.org/10.1016/j.annemergmed.2014.10.008.

Elsayed, Shereen, Thyssens, Daniela, Rashed, Ahmed, Jomaa, Hadi Samer, & Schmidt-Thieme, Lars (2021). Do we really need deep learning models for time series forecasting?

Eriksson, Julia, Gellerstedt, Linda, Hillerås, Pernilla, & Craftman, Åsa G. G. (2018). Registered nurses' perceptions of safe care in overcrowded emergency departments. *Journal of Clinical Nursing*, *27*(5–6), e1061–e1067. http://dx.doi.org/10.1111/jocn.14143.

Finnish Meteorological Institute Open Weather Data. (2020). https://www.ilmatieteenlaitos.fi/avoin-data. (Accessed: 02 February 2020).

Fintraffic: Digitraffic open data from Finnish roads. (2021). https://www.digitraffic.fi/en/road-traffic/. (Accessed: 10 May 2021).

Garza, Federico, Canseco, Max Mergenthaler, Chall, Cristian, & Olivares, Kin G. (2022). StatsForecast: Lightning fast forecasting with statistical and econometric models. URL https://github.com/Nixtla/statsforecast.

Gneiting, Tilmann, & Raftery, Adrian E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. http://dx.doi.org/10.1198/016214506000001437.

Google Trends. (2020). https://www.google.com/trends. (Accessed: 07 June 2020).

Grinsztajn, Léo, Oyallon, Edouard, & Varoquaux, Gaël (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, *35*.

Gul, Muhammet, & Celik, Erkan (2018). An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *Health Systems*, *00*(00), 1–22. http://dx.doi.org/10.1080/20476965.2018.1547348.

Gul, Muhammet, & Celik, Erkan (2020). An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *Health Systems*, *9*(4), 263–284.

Guttmann, Astrid, Schull, Michael J., Vermeulen, Marian J., & Stukel, Therese A. (2011). Association between waiting times and short term mortality and hospital admission after departure from emergency department: Population based cohort study from Ontario, Canada. *Bmj*, *342*(7809), http://dx.doi.org/10.1136/bmj.d2983.

Hawkins, Douglas M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, *44*(1), 1–12.

Herzen, Julien, Lässig, Francesco, Piazzetta, Samuele Giuliano, Neuer, Thomas, Tafti, Léo, Raille, Guillaume, et al. (2022). Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research*, *23*(124), 1–6, URL http://jmlr.org/papers/v23/21-1177.html.

Holleman, Donald R., Bowling, Renee L., & Gathy, Charlane (1996). Predicting daily visits to a walk-in clinic and emergency department using calendar and weather data. *Journal of General Internal Medicine*, *11*(4), 237–239. http://dx.doi.org/10.1007/BF02642481.

Hyndman, Rob J., & Khandakar, Yeasmin (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *26*(3), 1–22, URL https://www.jstatsoft.org/article/view/v027i03.

Jiang, Shancheng, Chin, Kwai Sang, & Tsui, Kwok L. (2018). A universal deep learning approach for modeling the flow of patients under different severities. *Computer Methods and Programs in Biomedicine*, *154*, 191–203. http://dx.doi.org/10.1016/j.cmpb.2017.11.003.

Jo, Sion, Jin, Young Ho, Lee, Jae Baek, Jeong, Taeoh, Yoon, Jaechol, & Park, Boyoung (2014). Emergency department occupancy ratio is associated with increased early mortality. *Journal of Emergency Medicine*, *46*(2), 241–249. http://dx.doi.org/10.1016/j.jemermed.2013.05.026.

Ke, Guolin, Meng, Qi, Finley, Thomas, Wang, Taifeng, Chen, Wei, Ma, Weidong, et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 3147–3155.

Koning, Alex J., Franses, Philip Hans, Hibon, Michèle, & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, *21*(3), 397–409. http://dx.doi.org/10.1016/j.ijforecast.2004.10.003.

Lim, Bryan, Arık, Sercan, Loeff, Nicolas, & Pfister, Tomas (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(4), 1748–1764.

Lundberg, Scott M., & Lee, Su In (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-December*(Section 2), 4766–4775.

Makridakis, Spyros, Spiliotis, Evangelos, & Assimakopoulos, Vassilios (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. http://dx.doi.org/10.1016/j.ijforecast.2019.04.014.

Makridakis, Spyros, Spiliotis, Evangelos, & Assimakopoulos, Vassilios (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346–1364. http://dx.doi.org/10.1016/j.ijforecast.2021.11.013.

McCarthy, Melissa L., Zeger, Scott L., Ding, Ru, Levin, Scott R., Desmond, Jeffrey S., Lee, Jennifer, et al. (2009). Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of Emergency Medicine*, *54*(4), 492–503.e4. http://dx.doi.org/10.1016/j.annemergmed.2009.03.006.

Morley, Claire, Unwin, Maria, Peterson, Gregory M., Stankovich, Jim, & Kinsman, Leigh (2018). Emergency department crowding: A systematic review of causes, consequences and solutions. In *PLoS One 13*(8), 1–42. http://dx.doi.org/10.1371/journal.pone.0203316.

Oreshkin, Boris N., Carpov, Dmitri, Chapados, Nicolas, & Bengio, Yoshua (2019). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. (pp. 1–31).

Pearce, Sabrina, Marchand, Tyara, Shannon, Tara, Ganshorn, Heather, & Lang, Eddy (2023). Emergency department crowding: an overview of reviews describing measures causes, and harms. *Internal and Emergency Medicine*, *18*(4), 1137–1158. http://dx.doi.org/10.1007/s11739-023-03239-2.

Rauch, Jens, Hübner, Ursula, Denter, Mathias, & Babitsch, Birgit (2019). Improving the prediction of emergency department crowding: A time series analysis including road traffic flow. *Studies in Health Technology and Informatics*, *260*, 57–64. http://dx.doi.org/10.3233/978-1-61499-971-3-57.

Richardson, Drew B. (2006). Increase in patient mortality at 10 days associated with emergency department overcrowding. *Medical Journal of Australia*, *184*(5), 213–216. http://dx.doi.org/10.5694/j.1326-5377.2006.tb00204.x.

Salinas, David, Flunkert, Valentin, Gasthaus, Jan, & Januschowski, Tim (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181–1191.

Smyl, Slawek (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75–85. http://dx.doi.org/10.1016/j.ijforecast.2019.03.017.

Terpilowski, Maksim (2019). Scikit-posthocs: Pairwise multiple comparison tests in python. *The Journal of Open Source Software*, *4*(36), 1169. http://dx.doi.org/10.21105/joss.01169.

Trotzky, Daniel, Tsur, Avishai M., Fordham, Daniel E., Halpern, Pinchas, Ironi, Avinoah, Ziv-Baran, Tomer, et al. (2021). Medical expertise as a critical influencing factor on the length of stay in the ED: A retrospective cohort study. *Medicine*, *100*(19), Article e25911. http://dx.doi.org/10.1097/MD.0000000000025911.

Tuominen, Jalmari, Lomio, Francesco, & Palomäki, Ari (2022). Forecasting daily emergency department arrivals using high-dimensional multivariate data : A feature selection approach. *BMC Medical*

Informatics and Decision Making, *7*, 0–37. http://dx.doi.org/10.1186/s12911-022-01878-7.

University of Helsinki Almanac Office. (2020). https://almanakka.helsinki.fi/en/. (Accessed: 20 July 2020).

Virtanen, Pauli, Gommers, Ralf, Oliphant, Travis E., Haberland, Matt, Reddy, Tyler, Cournapeau, David, et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2.

Whitt, Ward, & Zhang, Xiaopei (2019). Operations research for health care forecasting arrivals and occupancy levels in an emergency department. *Operations Research for Health Care*, *21*, 1–18. http://dx.doi.org/10.1016/j.orhc.2019.01.002.

Zhou, Lingling, Zhao, Ping, Wu, Dongdong, Cheng, Cheng, & Huang, Hao (2018). Time series model for forecasting the number of new admission inpatients. *BMC Medical Informatics and Decision Making*, *18*(1), 1–11. http://dx.doi.org/10.1186/s12911-018-0616-8.