International Journal of Forecasting 39 (2023) 192–208

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Towards a real-time prediction of waiting times in emergency departments: A comparative analysis of machine learning techniques

Elisabetta Benevento [*], Davide Aloini, Nunzia Squicciarini

Department of Energy, Systems, Territory and Construction Engineering, University of Pisa, Largo Lucio Lazzarino,
1, 56122 Pisa, Italy

ARTICLE INFO

Keywords:
Predictive analytics
Waiting time prediction
Machine learning
Healthcare management
Emergency department

ABSTRACT

Emergency Departments (EDs) can better manage activities and resources and anticipate overcrowding through accurate estimations of waiting times. However, the complex nature of EDs imposes a challenge on waiting time prediction. In this paper, we test various machine learning techniques, using predictive analytics, applied to two large datasets from real EDs. We evaluate the predictive ability of Lasso, Random Forest, Support Vector Regression, Artificial Neural Network, and the Ensemble Method, using different error metrics and computational times. To improve the prediction accuracy, new queue-based variables, that capture the current state of the ED, are defined as additional predictors. The results show that the Ensemble Method is the most effective at predicting waiting times. In terms of both accuracy and computational efficiency, Random Forest is a reasonable trade-off. The results have significant practical implications for EDs and hospitals, suggesting that a real-time performance monitoring system that supports operational decision-making is possible.

© 2021 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

1. Introduction

Emergency Department (ED) overcrowding is a long-standing issue affecting all healthcare systems (Ahalt, Argon, Ziya, Strickler, & Mehrotra, 2016; Hoot & Aronsky, 2008). This is mainly caused by the imbalance between supply and demand for emergency services, which is becoming overwhelming due to the ageing population and the widespread implementation of restrictive cost containment policies (ACEP, 2016; Arkun et al., 2010).

The consequences of ED overcrowding include increased waiting times, delayed treatment, ambulance diversion, and financial losses (Elalouf & Wachtel, 2016; Hoot & Aronsky, 2008). In particular, prolonged waiting times reduce the quality of care and increase the likelihood of adverse outcomes for patients with serious illnesses. Patient satisfaction is also affected, and more patients leave before being visited by a physician (Arkun et al., 2010; Hobbs, Kunzman, Tandberg, & Sklar, 2000). Thus, waiting time is a key metric for measuring ED efficiency (Khalifa & Khalid, 2015).

From the hospitals' perspective, accurate estimates of waiting times can help EDs better manage activities and resources and improve patient and ambulance routings, so they can anticipate or react to potentially critical situations (Ang, Kwasnick, Bayati, Plambeck, & Aratow, 2016; Senderovich et al., 2016). In addition, the timely communication of waiting times can significantly affect patients' experiences and satisfaction (Soremekun, Takayesu, & Bohan, 2011; van der Vaart, Vastag, & Wijngaard, 2011). Hospitals are therefore becoming increasingly aware of the potential value of predictive modelling when facing overcrowding, to improve patient care and operational

[*] Corresponding author.
E-mail addresses: elisabetta.benevento@ing.unipi.it (E. Benevento), davide.aloini@unipi.it (D. Aloini), n.squicciarini@studenti.unipi.it (N. Squicciarini).

https://doi.org/10.1016/j.ijforecast.2021.10.006
0169-2070/© 2021 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

efficiency. Traditional predictive methods (e.g., rolling averages, queuing theory, linear regression, etc.) produce results that are often inaccurate and that can potentially mislead both hospitals and patients (Ang et al., 2016; Bontempi, Taieb, & l Le Borgne, 2013). Thus, prediction accuracy remains a significant challenge due to the complexity of ED processes (Rebuge & Ferreira, 2012) and the stochastic nature of the ED system, which includes patient arrivals, types of treatments, and diagnostic tests (Ang et al., 2016; Lin, Patrick, & Labeau, 2013).

The widespread adoption of Hospital Information Systems (HISs), together with recent Information Technology (IT) developments, including cloud platforms and wearable sensors, enable hospitals to obtain an ever-growing volume of diverse patient and process-related data (Galetsi & Katsaliaki, 2019; Koufi, Malamateniou, & Vassilacopoulos, 2015). This trend has led to innovative and advanced data-driven techniques, such as predictive analytics, being applied to the prediction of key performance measures, e.g., waiting times, lengths of stay, hospital admissions, etc. Araz, Olson, and Ramirez-Nafarrate (2019), Golmohammadi (2016), Islam, Hasan, Wang, Germack, and Noor-E-Alam (2018).

In this study, we test various machine learning techniques for forecasting waiting times in EDs, by applying predictive analytics and using real data from two Italian hospitals in the form of basic triage ED data that provides patient characteristics. To improve the prediction accuracy, we devise new queue-based predictors that capture the current state of the ED. The aim is to identify the learning technique that provides the most accurate and real-time estimations of waiting times for patients arriving to the ED. We evaluate the accuracy of Lasso, Random Forest, Support Vector Regression, Artificial Neural Network, and the Ensemble Method using two forecasting error measures: the mean squared error and the mean absolute error. We also assess the efficiency of each technique in terms of computational time. Finally, we propose a simulated experiment to show a potential real-time application of the predictive models in the EDs. The proposed forecasting system simulates the process of predicting the waiting time in real time for each incoming patient to the ED, using information about the current state of the system and the new patient's characteristics.

Timely and accurate waiting time estimations can help hospital managers monitor process performance in real time and manage ED resources more effectively, based on expected patient waiting times. This can also increase patient satisfaction and reduce the number of patients who leave before being seen by a physician.

The paper is structured as follows: in Section 2 we review the literature; we describe the methodology (i.e., the data, learning techniques, and evaluation methods used in this study) in Section 3; Section 4 presents the results; Section 5 describes the simulated application of the developed predictive models; and Section 6 summarises the study and the findings and identifies future directions.

2. Related work

In this section, we provide a comprehensive review of the works related to waiting time prediction in EDs. We assess all identified studies published in journals since 2007. A detailed summary is provided in Table 1.

Linear regression is traditionally the most widely used approach for predicting waiting times in EDs (Ang et al., 2016; Gul & Celik, 2018; Kuo et al., 2020). For example, Asaro, Lewis, and Boxerman (2007) develop multivariate linear regression models with the waiting time, length of stay, and boarding time as dependent variables, by using patient- and system-level data. Similarly, Hemaya and Locker (2011) propose a linear regression model to predict waiting times, using patient arrivals in the ED and temporal variables. However, linear regression is sensitive to outliers and assumes only linear relationships between predictors and the dependent variables (Asaro et al., 2007; Kuo et al., 2020; Tu, 1996). Although non-linear relationships can be captured through the transformation of variables and by introducing interaction effects and higher-order terms into the model structure (Kuo et al., 2020), linear regression does not appear to be suitable for complex and dynamic contexts such as EDs.

Scholars have increasingly focused on using predictive analytics to forecast ED waiting times, due to its superior predictive performance (Islam et al., 2018; Koh, Tan, et al., 2011). Predictive analytics and data mining encompass various learning techniques aimed at extracting hidden and potentially useful information and patterns from data in large databases, and at making predictions from such data (Friedman, Hastie, & Tibshirani, 2001; Golmohammadi, 2016; Roquette, Nagano, Marujo, & Maiorano, 2020, chap. 1). In the study of Ding et al. (2010), a quantile regression model is developed that is able to forecast waiting, treatment, and boarding times for patients across all triage levels. The authors extracted triage information, patient demographics, and clinical characteristics. Similarly, Sun, Teow, Heng, Ooi, and Tay (2012) report a quantile regression model for predicting waiting times based on triage information. However, their model does not consider patient characteristics (e.g., age and mode of arrival). This represents a limitation, because overlooking particular variables can reduce model performance. Gonçalves et al. (2018) develop a classification model to predict ED waiting times by using random forests. The waiting time is separated into categories without providing a point estimation. A set of predictors including patient characteristics and temporal variables was exploited. In the study of Ang et al. (2016), a Q-lasso model is developed that combines statistical learning and fluid model estimators to forecast ED waiting times for different patient acuity groups by using historical data from four hospitals. Finally, Kuo et al. (2020) compare basic linear regression with four learning techniques. However, the small size of the dataset (one month) affects the reliability of the results. In addition, the authors mainly focus on demonstrating the superiority of learning techniques over the traditional approach in terms of prediction accuracy.

In summary, the literature presents several applications of traditional and more advanced learning approaches for the prediction of ED waiting times. However, most of these works generally propose a single advanced technique for conducting the forecasting task by mainly using triage time, patient characteristics, and temporal variables

Table 1
Summary of papers based on waiting time prediction in EDs.

Reference	Technique(s)	NHS	Data			Performance measure(s)	
			Period	Size	Predictors	Err. metric	Computational time
Asaro et al. (2007)	Multivariate linear regression	N.A.	26 months	166k	Patient characteristics, ED arrivals, admission rates, bed utilization	R^2	N.A.
Ding et al. (2010)	Quantile regression	N.A.	1 year	50k per dataset (4 EDs)	Patient and clinical characteristics, temporal variables, ED occupancy rates	Difference between observed and predicted waiting times	N.A.
Hemaya and Locker (2011)	Linear regression	UK	8 months	43k	Mean waiting time of the last three and five patients, temporal variables	Mean absolute error	N.A.
Sun et al. (2012)	Quantile regression	Singapore	1 month	13k	Triage code, triage time, queue for consultation, flow rates at triage	Difference between the median predicted and actual waiting times	N.A.
Ang et al. (2016)	Q-Lasso, rolling average, fluid models, quantile regression	California (USA), New York (USA)	4 datasets from 12 to 18 months	60k, 90k, 120k, 70k	Queue for consultation, queue for departure, staff-based variables, best rolling average, temporal variables	Mean squared error	N.A.
Gonçalves et al. (2018)	Random Forest (classification type)	Portugal	5 years	670k	Triage time, discharge time, patient and temporal variables	F1 score	N.A.
Kuo et al. (2020)	Linear regression, stepwise multiple linear regression, Artificial Neural Network, support vector machines, gradient boosting machines	Hong Kong	1 month	13k	Triage code, arrival time, staff-based variables, queue for triage, queue for consultation, queue for departure	Mean squared error, root mean squared error	N.A.

as predictors. Few authors compare their results with basic techniques, such as a rolling average and linear regression. However, as the performance of the various learning techniques is affected by various sources of uncertainties, it is not possible to assess *a priori* if a single methodological choice performs best (Kuhn & Johnson, 2013; Wolpert, 1996), thus leading to sub-optimal solutions. Furthermore, fair comparisons among different studies cannot be done without further considerations. The studies generally analyse various datasets and the results depend on several factors, such as national health systems, ED architecture, the temporal window, the seasonality of the data, and the set of predictors. The use of various error metrics can lead to invalid results being obtained by simple comparisons between learning techniques. Table 1 shows that there is no single recognised error metric among the studies. None report the computational time of their approaches, which can determine the selection of the most appropriate technique used in a real-time forecasting tool. Thus, a discussion of the results is required, to evaluate the applicability of these learning techniques in real applications and to enable the identification of the most suitable solution.

The aim of this study is to contribute to the literature by presenting a critical and structured comparative analysis of five learning methods—including advanced linear regression techniques, non-linear regression techniques, and their combinations—for predicting ED waiting times, using two large datasets of two real EDs. Each learning technique is evaluated in terms of providing accurate estimations of waiting times in a reasonable computational time and to guide ED managers in the choice of the most suitable estimate. In addition, we aim to improve the predictive power of each learning technique through new queue-based predictors that capture the current state of the ED, as enabled by process mining (see van der Aalst (2016) for more details) as suggested by Benevento, Aloini, Squicciarini, Dulmin, and Mininno (2019). Finally, this is the first attempt to use a simulated experiment to test the predictive models' potential forecasting ability regarding the waiting time of each incoming patient in an ED in real time.

3. Materials and methods

3.1. Data

We used two real datasets from two EDs in medium-sized public hospitals, located in areas of average population in northern-central Italy. We refer to these as Hospital 1 and Hospital 2. The datasets from the two hospitals vary in timespan and number of observations. That of Hospital 1 (D1) covers January to August and consists of 40,504 ED presentations, or observations, and more than 453,000 events; the dataset from Hospital 2 (D2) covers January to December and includes 60,230 observations and more than 553,000 events. Like the datasets used in other studies, D1 and D2 contain information regarding patients (triage code and age) and their related ED episodes, from the registration and triage process to ED departure or hospital admission.

Table 2
Summary statistics of the two datasets. The waiting time of code-red patients is zero as they need immediate treatment for preservation of life.

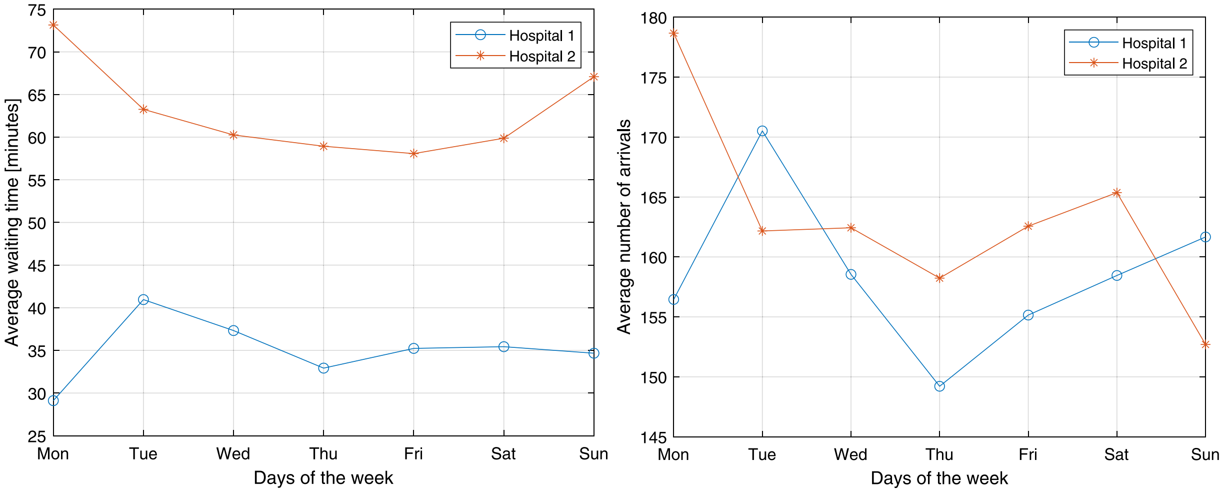
	Hospital 1	Hospital 2
Variables	% Patients	
Triage code		
Red	2	3
Yellow	30	24
Green	53	40
Blue	14	31
White	1	2
Mode of arrival		
Autonomous	70	65
By ambulance	30	35
Variables	Average (Std. Dev.)	
Age [years]	46 (33)	47 (30)
Waiting time [min]		
All patients	35 (41)	65 (102)
Yellow code	21 (19)	65 (68)
Green code	39 (46)	79 (116)
Blue code	53 (65)	50 (103)
White code	58 (72)	63 (109)

The raw data extracted from the hospital information systems included unwanted observations, missing data points, double entries, and outliers. We therefore first refined the two datasets to improve the data quality. The data cleaning procedure is described in the Supplementary Material. The refined datasets D1 and D2 consist of 38,081 and 57,887 observations, respectively.

We analysed the refined datasets in detail to gain useful insights about the variables and the relationships between them. This enabled the appropriate selection of candidate predictors. The summary statistics of the variables from these observations are presented in Table 2.

Of the 38,081 ED observations in sample D1, 85% are urgent, very urgent, and immediate patients. In D2, the percentage of such patients is lower (67%), while the percentage of blue codes is twice as high (31%). As shown in Table 2, the average waiting time in D2 is higher than that of D1. In addition, the standard deviation for D2 is greater, indicating that there is much more variability in D2 than in D1. The average patient age is similar in both datasets (46 years for D1 and 47 years for D2). Young patients experience shorter waiting times as they are immediately treated by the paediatrician, while the other age groups show similar waiting times. More details are given in the Supplementary Material. Almost 30% of patients arrive at the ED of Hospital 1 by ambulance, and the remainder arrive independently. A similar value is recorded for Hospital 2.

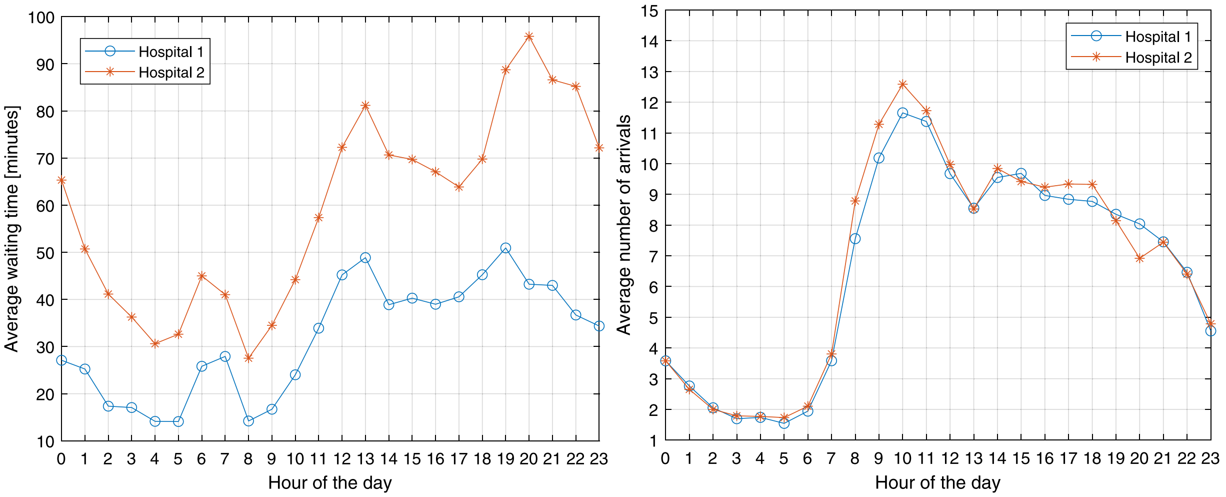
Figs. 1 and 2 compare weekly and daily variations of patient arrival rates and waiting times at the EDs of the two hospitals. For Hospital 1, the average waiting time and the number of arrivals have similar variation patterns, and are directly proportional. For example, as shown in Fig. 1, the average waiting time is high on Tuesday, when the ED experiences many arrivals, and decreases over the weekend. The significant increase in patients visiting the ED at the beginning of the week is likely due to the increase in the number who do not first



(a) Average waiting time per day in minutes

(b) Average number of arrivals per day

Fig. 1. Comparison between the average waiting time and the average number of arrivals during the week for Hospital 1 and Hospital 2.



(a) Average waiting time per hour in minutes

(b) Average number of arrivals per hour

Fig. 2. Comparison between the average waiting time and the average number of arrivals per hour.

consult their primary care physician and are then referred to the ED in times of urgency. For Hospital 2, the average waiting time is always high during the week, and peaks on Monday and Sunday. The number of arrivals per day is high on Monday, with a slight decrease on Wednesday and Sunday.

As Fig. 2 shows, for Hospital 1 on any particular day both the waiting time and the number of patients visiting the ED increase gradually from 8 a.m. to 12 a.m. In addition, there is a decrease in waiting time at the beginning of each shift (at 8 a.m. and 2 p.m.). Similar behaviour is also observed in Hospital 2.

We first conducted an in-depth identification of the variables and a collection process, and then considered a set of 26 predictors, as listed in Table 3, to effectively capture the ED state and accurately predict the waiting time of patients. We considered three aspects when selecting the contributing variables: previous research on waiting times, data availability, and interviews with ED staff. The same predictor approach was used for both D1 and D2.

We included patient characteristics (age, arrival mode, and acuity level) that may potentially influence the variances in forecasting waiting times (Arkun et al., 2010; Ding et al., 2010). To account for daily variations, we investigated a set of variables for the time of day along with the day of the week (Ang et al., 2016; Arkun et al., 2010; Ding et al., 2010; Sun et al., 2012). Following Ang et al. (2016), we incorporated the best rolling average of the waiting times of the last 15 patients who presented to the ED as an additional candidate predictor. Alongside the variables commonly used in the healthcare literature (Ang et al., 2016; Araz et al., 2019; Arkun et al., 2010; Sun et al., 2012), we also explored the use of new-arrival and queue-based predictors, which report the current state

Table 3
Characterisation of predictors.

Category	Name	Type
Patient-related	Triage code	Categorical
	Age	Numerical
	Arrival mode	Binary
Temporal	Day of the week	Categorical
	Hour of the day	Categorical
Arrivals-based	Estimated average number of arrivals - red codes	Numerical
	Estimated average number of arrivals - yellow codes	Numerical
	Estimated average number of arrivals - green codes	Numerical
	Estimated average number of arrivals - blue codes	Numerical
Queue-based	Patients who are triaged but not yet treated – red codes	Numerical
	Patients who are triaged but not yet treated – yellow codes	Numerical
	Patients who are triaged but not yet treated – green codes	Numerical
	Patients who are triaged but not yet treated – blue codes	Numerical
	Patients who are triaged but not yet treated – white codes	Numerical
	Patients who are treated but not yet discharged – red codes	Numerical
	Patients who are treated but not yet discharged – yellow codes	Numerical
	Patients who are treated but not yet discharged – green codes	Numerical
	Patients who are treated but not yet discharged – blue codes	Numerical
	Patients who are treated but not yet discharged – white codes	Numerical
	Patients between triage and exit	Numerical
	Patients who are treated but waiting to receive the laboratory report	Numerical
	Patients who are treated but waiting to receive the radiological report	Numerical
	Patients who are treated but waiting for the orthopaedic consultation	Numerical
	Patients who are treated but waiting for the paediatric consultation	Numerical
	Patients who are transferred to the short observation unit	Numerical
Others	Best rolling average	Numerical

of the ED, to improve prediction accuracy, as suggested by Benevento et al. (2019).

Arrivals-based variables capture the influence of new ED arrivals on the waiting time of patients within the ED. The higher the priority of new arrivals, the more significant their impact on the waiting time. Thus, we developed four predictors corresponding to red, yellow, green, and blue codes. White codes were omitted because they denote the lowest priority, and thus they do not affect the waiting time of patients already queuing. To compute arrival-based predictors for each observation, we measured the average number of arrivals in the previous months that occurred on the same day of the week and at the same hour. Our estimate of the predictive power of the arrivals-based variables is reported in the Supplementary Material.

Queue-based variables determine the crowding level of the ED system. They measure the queue of patients waiting at various stages, from registration to disposition (e.g., the queue for lab tests or for transfer to the Short Observation Unit). To derive internal queue-based predictors, we followed the process mining approach (van der Aalst, 2016), as suggested by Benevento et al. (2019).

3.2. Methods

The dependent variable is the waiting time between triage and the first visit for each incoming patient to the ED. To predict the ED waiting time, we tested several learning techniques, both linear and nonlinear, by leveraging the predictors defined in Section 3.1.

We compared Lasso, Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Network (ANNs), and the Ensemble Method (EM), to evaluate their accuracy in terms of providing information useful for operational decisions in the ED (Araz et al., 2019). These learning techniques are representative of the main families of algorithms used in machine learning applications (Delen, Cogdell, & Kasap, 2012; Hastie, Tibshirani, & Friedman, 2009; Koh et al., 2011). They are used because they have an advantage in handling large datasets, flexibly explore data connections, identify significant predictors, and prevent data over-fitting.

The accuracy of the learning techniques was evaluated and compared using the mean squared error (MSE) and mean absolute error (MAE), which are the most widely used error measurements (Hyndman & Koehler, 2006; Shcherbakov et al., 2013). The MSE and MAE are particularly useful when comparing different predictive models applied to the same set of data (Hyndman & Koehler, 2006; Shcherbakov et al., 2013). The MSE is the most popular measure because of its theoretical significance in statistical modelling. However, it is significantly influenced by outliers. The MAE can mitigate this problem, as it considers only the absolute value of the error vector components (Hyndman & Koehler, 2006).

In addition to model accuracy, the efficiency levels of Lasso, RF, SVR, ANN, and EM were evaluated in terms of computational time.

To test the impact of the queue-based and arrivals-based variables described in the previous section on the accuracy of waiting time prediction in EDs, we defined three sets of candidate predictors (see Table 3 in Section 3):

(a) *Basic predictors (B)*. This includes only the traditional predictors, that is, patient-related variables,

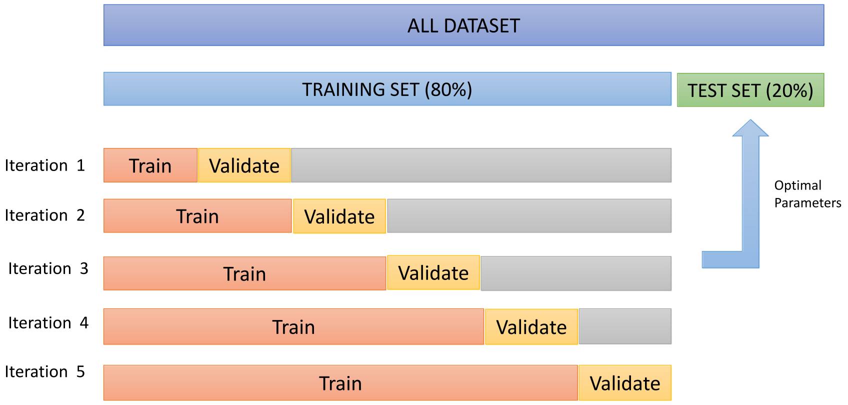


Fig. 3. Overview of the procedure for model selection and hyperparameter tuning.

temporal variables, and the best rolling average variable.

(b) *Basic (B) + arrivals-based (A) predictors*. This includes the predictors in set B plus arrivals-based predictors.

(c) *Basic (B) + arrivals-based (A) + queue-based (Q) predictors*. This includes all predictors.

The accuracy of the Lasso, RF, SVR, ANN, and EM models were then evaluated for all three sets using the MSE and MAE. The evaluation was conducted for D1 and D2.

Finally, to implement and test the learning techniques for waiting time prediction, we split both D1 and D2 into two subsets: 80% of patient observations represented the *training set*, which was used to tune the hyperparameters of each technique; and the remaining 20% represented the *test set*, which was used to evaluate the accuracy of each technique. The hyperparameters of the Lasso, RF, SVR, ANN, and EM techniques were tuned by using the time series cross-validation technique. The training set was initially split into six equal parts, to allow for five iterations. At the *i*th iteration, the training subset consists of *i* of these parts, while the validation set is represented by the *(i+1)*th part. As each technique comes with various hyperparameters and may perform differently by varying them within a discrete set, we conducted time series cross-validations for all hyperparameter values and finally chose the technique that gave the lowest cross-validation average MSE among the five iterations. An overview of the procedure for model selection and hyperparameter tuning is given in Fig. 3.

We then implemented the selected learning techniques by using MATLAB® internal routines, as described in the following sections.

3.2.1. Lasso

Lasso is a regularised linear regression method that can be used to select significant parameters among the predictor variables (Kuhn & Johnson, 2013; Tibshirani, 1996). Lasso has the advantage of reducing the number of variables within the model. Thus, if a dataset has many features, Lasso can identify and extract the most important (Tibshirani, 1996). However, if a group of predictors is highly correlated, Lasso tends to randomly choose one of the multi-collinear predictors, regardless of the context (Zou & Hastie, 2005).

Table 4
Best Lasso model for D1 and D2.

Dataset	Best λ	MSE [min^2]	MAE [min]
D1	0.34	1428	26.2
D2	0.91	4883	53.1

Lasso estimates the regression coefficients vector $\boldsymbol{\beta}$ as the minimiser of the Lasso objective function (see Eq. (1)$_3$), composed of the sum of a loss function (RSS, defined in Eq. (1)$_2$) and a penalty term (Ogutu & Piepho, 2014):

$$\begin{aligned} \mathbf{y} &= \beta_0\mathbf{e} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathrm{RSS} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \mathrm{Lasso}(\boldsymbol{\beta}) &= \arg\min_{\boldsymbol{\beta}} \left[\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|\right] \end{aligned} \quad (1)$$

In Eq. (1), $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the predictor array (X_{ij} is the *j*th predictor for the *i*th patient) being p the number of predictors and n the number of observations, $\mathbf{y} \in \mathbb{R}^n$ is the waiting time vector (each component is relative to a patient), $\boldsymbol{\beta} \in \mathbb{R}^p$ is the regression coefficients vector, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the error terms vector, $\mathbf{e} \in \mathbb{R}^n$ is a vector of ones. As customary, the vector $\mathbf{y}$ has been centered. The penalty multiplier λ prevents over-fitting the prediction function by forcing Lasso to use only a subset of the candidate predictors, i.e., those with the largest signal for a given λ, and to set the $\boldsymbol{\beta}$ coefficients of the others to zero (Ang et al., 2016; Tibshirani, 1996).

In our case, $p = 26$ and $n = 30464$ for D1, while $n = 46309$ for D2. To find the optimal value of λ, we varied it in the range of [0.01, 100], sampled in 100 logarithmically spaced values. We then retrained the Lasso model with the best λ. The best-performing Lasso models for D1 and D2 are summarised in Table 4. For more details on Lasso hyperparameter tuning, see the Supplementary Material.

3.2.2. Random Forest

RF is a learning technique that generates and combines binary decision trees, while also aggregating the results (Breiman, 2001; Hastie et al., 2009, chap. 15). Decision trees are constructed by using a bootstrap sample of the training data and randomly choosing a subset of predictors at each node.

The algorithm for RF regression is summarised in Algorithm 1 (Dudek, 2015; Hastie et al., 2009, chap. 15), where K is the number of trees, P is the number of input predictors at each split, p is the total number of predictors, and m is the minimum node size.

Algorithm 1 RF algorithm.

```
for $i = 1, \cdots, K$ do
    Draw a bootstrap sample $B$ of size $S$ from the training data.
    while node size $\neq m$ do                    ▷ Grow an RF tree $T_i$
        for every node do
            Randomly select $P$ predictors out of $p$.
            Pick the best predictor/split-point among the $P$.
            Split the node into two daughter nodes.
        end for
    end while
end for
return $\{T_i\}_{i=1,\cdots,K}$
```

To make a prediction at a new point $\mathbf{x}$, Eq. (2) is used:

$$f(\mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K} T_i(\mathbf{x}) \tag{2}$$

In our case, $p = 26$, and we initially fixed the number of trees K (Probst & Boulesteix, 2017) as 400. We determined m and P by varying them in the ranges of [2, 10] and [6, 20], respectively. The best combination of (m, P) that minimises the MSE is $(7, 9)$ for D1 and D2. We then retrained the RF model with the best (m, P). The best RF model performances are listed in Table 5. For more details on RF hyperparameter tuning, see the Supplementary Material.

3.2.3. Support Vector Regression

SVR belongs to the family of generalised linear models, which are aimed at achieving a prediction decision based on a linear combination of features derived from variables (Araz et al., 2019; Delen et al., 2012).

The general idea of SVR is to construct an ε-tube. Errors are ignored if they are inside the tube, and penalised if they are outside of the tube (Carrasco, López, & Maldonado, 2019; Drucker, Burges, Kaufman, Smola, & Vapnik, 1996).

The SVR can be formalised by the following optimisation problem:

$$\begin{aligned} &\min_{\boldsymbol{\beta},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \frac{1}{2}\|\boldsymbol{\beta}\|_{L^2}^2 + C\mathbf{e}^T(\boldsymbol{\xi}+\boldsymbol{\xi}^*) \qquad \text{subject to} \\ &\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - b\mathbf{e} \le \varepsilon\mathbf{e} + \boldsymbol{\xi}, \\ &-\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + b\mathbf{e} \le \varepsilon\mathbf{e} + \boldsymbol{\xi}^*, \\ &\boldsymbol{\xi},\ \boldsymbol{\xi}^* \ge \mathbf{0}, \end{aligned} \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the predictor array, n is the number of observations, p is the number of predictors, $\mathbf{y} \in \mathbb{R}^n$ is the waiting time vector, $\boldsymbol{\xi} \in \mathbb{R}^n$ and $\boldsymbol{\xi}^* \in \mathbb{R}^n$ are slack variables vectors, $\mathbf{e} \in \mathbb{R}^n$ is a vector of ones, b is the intercept of the regression, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the regression coefficient vector. The positive box-constraint parameter C to be tuned affects the tolerance of points lying outside the error margin ε. That is, it determines the tradeoff between the flatness of the regression function and the maximum tolerated number of deviations larger than ε (Gunn et al., 1998; Smola & Schölkopf, 2004). It can be shown that problem (3) can be formulated in dual terms and can be extended to the nonlinear case as follows (Carrasco et al., 2019; Smola & Schölkopf, 2004):

Table 5
Best RF model for D1 and D2.

Dataset	Best m	Best P	MSE [min^2]	MAE [min]
D1	7	9	1105	21.2
D2	7	9	4073	46.2

$$\begin{aligned} &\min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \left\{ \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\mathbf{K}(\mathbf{X},\mathbf{X}^T)(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) \right. \\ &\quad \left. + \varepsilon\mathbf{e}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*) - \mathbf{y}^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) \right\} \qquad \text{subject to} \\ &\mathbf{e}^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) = 0, \\ &\mathbf{0} \le \boldsymbol{\alpha}, \boldsymbol{\alpha}^* \le C\mathbf{e}, \end{aligned} \tag{4}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\boldsymbol{\alpha}^* \in \mathbb{R}^n$ are the Lagrange multiplier vectors associated with the first two (vector) constraints appearing in Eq. (3), $\mathbf{K}(\mathbf{X},\mathbf{X}^T) \in \mathbb{R}^{n\times n}$ is the Gram matrix whose components $(\mathbf{K})_{ij} = k(\mathbf{X}_i,\mathbf{X}_j)$ $(\mathbf{X}_i, \mathbf{X}_j \in \mathbb{R}^p)$ are called kernel functions and are defined as the scalar product $< \Phi(\mathbf{X}_i), \Phi(\mathbf{X}_j) >$, where $\Phi : \mathbf{X} \to \mathcal{F}$ maps the training samples into some higher-dimensional feature space $\mathcal{F}$. According to problem (4), the prediction is given by

$$\mathbf{y}(\mathbf{X}) = \mathbf{K}(\mathbf{X},\mathbf{X}^T)(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + b\mathbf{e}. \tag{5}$$

It is noteworthy that, in the non-linear case, problem (4) corresponds to finding the flattest function in the feature space, not in the input space. Moreover, note that in the non-linear case, the kernel functions are introduced to map the data into the feature space $\mathcal{F}$ to make linear regression possible. In the linear case, the components $(\mathbf{K})_{ij}$ are given simply by $< \mathbf{X}_i, \mathbf{X}_j >$, and problem (4) is the dual of (3) (Carrasco et al., 2019; Gunn et al., 1998; Smola & Schölkopf, 2004). In this work, the Gaussian kernel function was used:

$$k(\mathbf{X}_i,\mathbf{X}_j) = \exp\left(-\frac{\|\mathbf{X}_i-\mathbf{X}_j\|_{L^2}^2}{2\sigma^2}\right), \tag{6}$$

where $\mathbf{X}_i, \mathbf{X}_j \in \mathbb{R}^p$ represent two training samples and $\sigma > 0$ is the kernel shape parameter.

The hyperparameter of SVR to be tuned is the box-constraint parameter C. In our case, we varied C in the range of [5, 400]. The best value of C for minimising the MSE between the predicted and actual waiting time is 75 for D1 and D2. We then retrained the SVR model with the best C. The best SVR model performance is given in Table 6. For more details on SVR hyperparameter tuning, see the Supplementary Material.

Table 6
Best SVR model for D1 and D2.

Dataset	Best C	MSE [min^2]	MAE [min]
D1	75	1224	21.4
D2	75	4617	46.6

Table 7
Best ANN model for D1 and D2.

Dataset	NHL	Best NNPL	MSE [min^2]	MAE [min]
D1	1	4	1198	22.7
D2	1	4	4307	48.3

3.2.4. Artificial Neural Network

ANNs aim to simulate the human brain when collecting and processing data for the purpose of learning (Golmohammadi, 2016; Golmohammadi & Radnia, 2016).

An ANN consists of interconnected nodes that mimic the actual neurons of a biological brain. The nodes are connected by signals, which are the weighted sums of the inputs. Each of the connections, or edges, transmits a signal from one neuron to another. Artificial neurons and edges have a weight that is self-determined as learning proceeds. The training procedure describes how the values of the weights are determined, allowing the network to accurately grasp the behaviour of the system to be described. Typically, artificial neurons are aggregated into layers. Different layers can perform different types of transformations on their inputs (Gardner & Dorling, 1998).

Many types of ANNs have been applied, depending on how the neurons are connected. The most used is the multilayer feed-forward neural network (Gardner & Dorling, 1998). The architecture of a multilayer network is variable, but in general it consists of several layers of neurons. The output of a node is scaled by the connecting weight and fed forward to be an input for the nodes in the next layer. The number of neurons in each layer depends on the specific problem. An ANN requires determining the number of hidden layers (NHLs), the number of nodes per layer (NNPL), and the type of training algorithm (TA).

In our case, we built a feed-forward network with sigmoidal activation functions, trained with a back-propagation TA, called *Trainbr*. We varied the NNPL among $\{1, 2, 3, 4, 5, 10, 20, 50\}$, and we used one NHL, as suggested by Blackard and Dean (1999), Hornik, Stinchcombe, and White (1989). We then retrained the ANN model with the best NNPL. The performances of the best ANN model are listed in Table 7. For more details on ANN hyperparameter tuning, see the Supplementary Material.

3.2.5. Ensemble Method

EM represents an ensemble of all of the previous methods. The main idea is to combine and assign weights to several learning techniques, to obtain a new predictive technique that can outperform any single method it is composed of (Polikar, 2012; Rokach, 2009; Valentini & Masulli, 2002). Although different methods of combining single forecasts can be devised, we simply computed a weighted sum of the outputs of Lasso, RF, SVR, and ANN.

Table 8
Best EM model for D1 and D2.

Dataset	Best β	MSE [min^2]	MAE [min]
D1	[0, 0.74, 0.10, 0.16]	1100	20.8
D2	[0, 0.63, 0.03, 0.34]	4059	46

The procedure for EM training and tuning is shown in Fig. 4.

The initial training set, consisting of the 80% of the data, was split into training and validation sets. Each learning technique was trained and tuned on the training set, following the time series cross-validation procedure. The optimal weights were determined based on the validation set. The distance from the actual waiting times for EM was evaluated in the test set. The best value of the weight vector $\boldsymbol{\beta}$ is the result of the minimisation problem of Eq. (7):

$$\min_{\beta} \mathcal{E}\left(\sum_{i=1}^{4} \beta_i \mathbf{y}_{pred-i}, \mathbf{y}_{act}\right) \quad \text{subject to} \qquad \sum_{i=1}^{4} \beta_i = 1, \tag{7}$$

where $\mathbf{y}_{pred-i}$ is the predicted waiting time vector for the ith technique (Lasso, RF, SVR, or ANN), $\mathbf{y}_{act}$ is the actual waiting time vector and $\mathcal{E}$ is the MSE error measure. The best performances of EM are listed in Table 8.

4. Results

Tables 9 and 10 summarise the performance of the different methods in terms of the MSE and MAE with D1 and D2, respectively. All methods were compared with the ordinary least squares (OLS) method, commonly named linear regression, which is traditionally a widely used approach for forecasting waiting times (Ang et al., 2016; Kuo et al., 2020).

We can infer that EM outperforms other techniques, as it achieves the lowest error for each metric. It improves accuracy by between 28% and 23% for D1, depending on the error metric relative to OLS. The same applies to D2, with a decrease in the MSE and MAE of 17% and 14%, respectively. OLS obtains the largest error within all error metrics with both datasets when compared to the stand-alone techniques. This is followed by Lasso, which achieves a very similar level of performance, as it yields the lowest error reduction (up to 6% with D1 and up to 0.6% with D2). For the MAE, which does not penalise outliers, RF and SVR perform similarly with both datasets. However, with the MSE, RF achieves the best results, with an increase in accuracy of up to 27% in D1 and up to 17% in D2.

We note a significant worsening in the performance of the predictive models with D2. This is mainly due to the greater variability in the waiting times of D2, as confirmed by the standard deviation in Table 2.

All of the results we obtained are statistically significant, as described in Appendix.

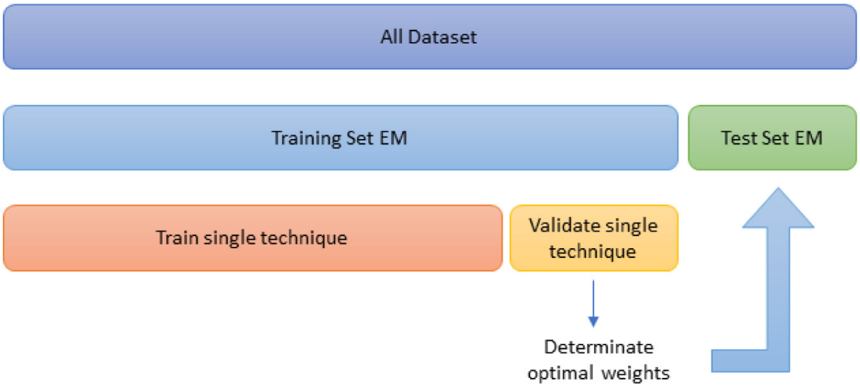


Fig. 4. Overview of the procedure for EM training and tuning.

Table 9
Model comparison with D1.

	OLS	Lasso	%Change[a]	RF	%Change[a]	SVR	%Change[a]	ANN	%Change[a]	EM	%Change[a]
MSE [min^2]	1520	1428	−6	1105	−27	1224	−19	1198	−21	1100	−28
MAE [min]	27	26.2	−3	21.2	−22	21.4	−21	22.7	−16	20.8	−23

[a]The variation is measured with respect to OLS.

Table 10
Model comparison with D2.

	OLS	Lasso	%Change[a]	RF	%Change[a]	SVR	%Change[a]	ANN	%Change[a]	EM	%Change[a]
MSE [min^2]	4895	4883	−0.3	4073	−17	4617	−6	4307	−12	4059	−17
MAE [min]	53.4	53.1	−0.6	46.2	−14	46.6	−13	48.3	−10	46	−14

[a]The variation is measured with respect to OLS.

Figs. 5 and 6 show the quantiles for the prediction and absolute prediction errors for D1 and D2, respectively. The behaviour of the prediction error and absolute prediction error curves is almost the same for both datasets. This confirms that Lasso achieves the poorest performance among the proposed techniques. EM, RF, and SVR have the same trend, and they achieve generally better performance. In more detail, Figs. 5(a) and 6(a) show that the waiting time is underestimated for about 40% of patients in the test set for all the forecasting methods. Of these, RF and EM are the best for underestimations (they underestimate less). Conversely, Lasso and ANN tend to overestimate the prediction when compared to EM or SVR. Fig. 5(b) shows that, up to the 80th percentile, the absolute prediction error is kept within 30 minutes for all techniques, except for Lasso, with D1. This is a promising result in a complex and dynamic context like an ED. However, Fig. 6(b) shows that the 30-minute threshold is respected only up to 50th percentile, when using D2. This result confirms that the greater variability within D2 can negatively affect the performance of the predictive models.

In terms of computational efficiency, Table 11 summarises the computational time of all techniques with D1 and D2. Simulations were run on an Intel i7-8750H CPU @2.20 GHz notebook.

Table 11
Average computational time.

		OLS	Lasso	RF	SVR	ANN	EM
D1	CPU time [s]	0.3	0.8	37	100	60	198
D2	CPU time [s]	0.5	2	60	137	86	285

As expected, increasing the size of the dataset leads to an incremental increase in the computational times. However, the behaviour of the techniques remains similar. EM has the highest computational time with D1 and D2, as it is derived from the combination of four techniques. For the standalone techniques, despite their prediction accuracy, it is evident that SVR suffers from a high training time, which directly influences the computational time of EM. Lasso achieves the lowest running time, and is comparable to that of simple traditional methods like OLS. RF also obtains an acceptable computational time value, which remains under one minute with both datasets, without penalising its predictive performance. The low computational time of RF is partially influenced by the low number of trees. The computational time may increase with a larger number of trees. This result suggests that a predictive model with acceptable accuracy can be developed in a feasible amount of time, allowing for applications in real contexts.

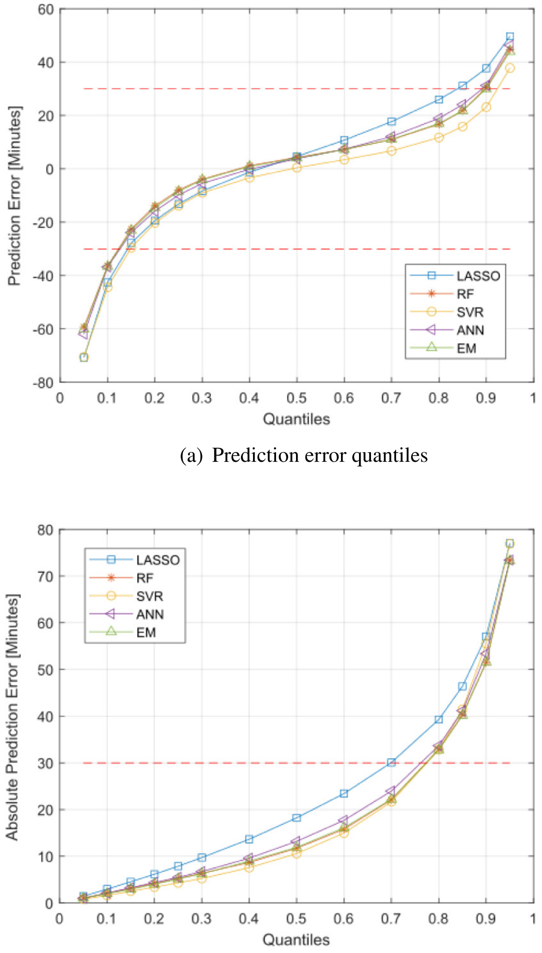
Table 12 summarises the main results of the evaluation of the impact of queue-based and arrivals-based variables on the prediction accuracy.

Enriching the set of predictors with arrivals-based and queue-based variables increases the accuracy of waiting time predictions with both D1 and D2. RF and EM bene-

Table 12
MSE and MAE obtained by Lasso, RF, SVR, ANN, and EM with each set of predictors.

	Predictors	Lasso				RF				SVR				ANN				EM			
		MSE	MAE	%Change[a]		MSE	MAE	%Change[a]		MSE	MAE	%Change[a]		MSE	MAE	%Change[a]		MSE	MAE	%Change[a]	
D1	B	1583	26.8	–	–	1320	22.6	–	–	1353	21.7	–	–	1296	23.1	–	–	1288	22.2	–	–
	B+A	1498	26.5	−5	−2	1216	22.2	−8	−2	1280	21.5	−6	−1	1228	22.9	−5	−1	1207	21.9	−6	−1
	B+A+Q	1428	26.2	−10	−3	1105	21.2	−16	−7	1224	21.4	−9	−2	1198	22.7	−8	−2	1100	20.8	−15	−6
D2	B	5459	55.2	–	–	4942	49.5	–	–	5322	48.1	–	–	4934	51	–	–	4823	49.3	–	–
	B+A	4950	53.7	−9	−3	4308	47.7	−13	−4	4762	46.9	−11	−3	4469	49.3	−10	−4	4285	47.9	−11	−3
	B+A+Q	4883	53.1	−11	−4	4073	46.2	−18	−7	4617	46.6	−13	−3	4307	48.3	−13	−5	4059	46	−16	−7

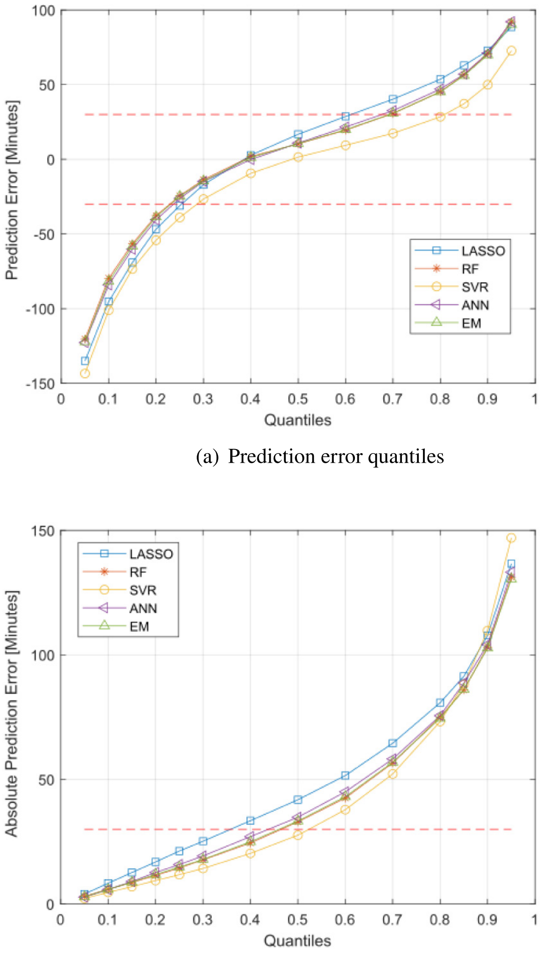
[a] The variation is measured with respect to B.



(a) Prediction error quantiles

(b) Absolute prediction error quantiles

Fig. 5. Comparison of prediction and absolute prediction errors, in minutes, for Lasso, RF, SVR, ANN, and EM with D1.



(a) Prediction error quantiles

(b) Absolute prediction error quantiles

Fig. 6. Comparison of prediction and absolute prediction errors, in minutes, for Lasso, RF, SVR, ANN, and EM with D2.

fited most from the introduction of such new predictors, with a decrease in the MSE and MAE ranging between 15% and 18% and between 6% and 7%, respectively. There was a greater reduction in the MSE (up to 18%) compared to the MAE (up to 7%) for all predictive models.

We then drill down into the results to explore the impact of each category of predictors on the error reduction rate. Arrivals-based predictors (B+A models) lead to a significant decrease in the MSE (up to 13%) compared to traditional predictors (B models) with D2, while a minor decrease (up to 8%) is observed with D1. Including queue-based variables (B+A+Q models) considerably improves the prediction accuracy with both D1 and D2 (up to 16% and 18%, respectively) compared to using only traditional predictors (B models). Thus, these new predictors that capture the current ED state seem to significantly influence waiting times in the ED.

Our results are consistent with other studies (Benevento et al., 2019; Kuo et al., 2020) that demonstrate the predictive power of queue-based and arrivals-based variables by using both linear and non-linear learning techniques.

Table 13
Comparison between the rolling average and EM.

		Rolling average	EM	%Change
D1	MSE [min^2]	2058	1100	−47
	MAE [min]	30.9	20.8	−33
D2	MSE [min^2]	6270	4059	−35
	MAE [min]	59.7	46	−23

Finally, Table 13 shows the predictive performance of the best-performing EM and the commonly applied rolling average, which is widely used in hospitals and EDs as an order-zero estimation technique (Dong, Yom-Tov, & Yom-Tov, 2018), with both datasets. The rolling average was measured over a time span of three hours, similar to Ang et al. (2016). As expected, the more sophisticated EM technique clearly outperforms the rolling average, with an increase in accuracy of between 23% and 47%, depending on the considered metric and on the dataset.

To summarise, the EM technique is the best choice if the accuracy of the results is the priority, because it combines good performing methods through proper weight-

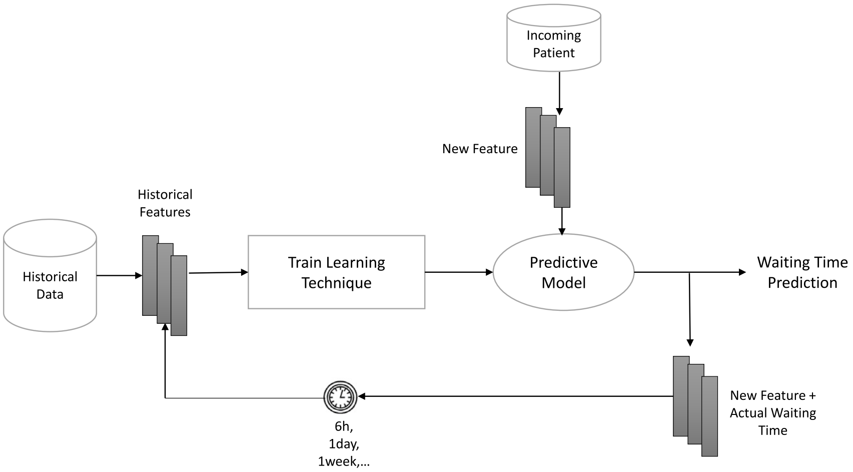


Fig. 7. Scheme of the real-time forecasting system.

ing. However, RF reasonably balances computational cost and the accuracy of the predictions.

5. Simulated experiment of a real-time application of the proposed predictive model in EDs

We developed a forecasting system based on the proposed machine learning technique, with the aim of providing real-time estimations of waiting times for each incoming patient to the ED. The system works as follows (see Fig. 7). Historical data are used to train the learning technique. Then, the data of the incoming patient are recorded and used to predict that patient's waiting time. These new data are then added to the historical dataset and used to retrain the model for new forecasts. Retraining is performed after a certain time interval (every day, every hour, etc.).

To test the real-time implementation of the proposed forecasting system, we conducted the following simulated experiment:

1. We selected D1 as the reference dataset and RF as the learning technique of the model.
2. The first six months of D1 were used to train the predictive model.
3. The trained model was used to make real-time predictions of the waiting times for all incoming patients, starting from the first day after the six months. For each patient, the set of predictors was built, thus capturing the current state of the ED.
4. We retrained the predictive model by using the new patient data with an expanded-window approach. We reassessed the model both at the end of each day and every six hours. The hyperparameters were tuned at the end of the day in both cases. Indeed, adding a few observations to a large dataset does not warrant more frequent re-optimisation of the model hyperparameters.
5. We assessed the accuracy of real-time predictions for each time-window as the absolute value of the difference between the predicted waiting times and the actual times (from D1). The accuracy of the error was evaluated against the two acceptance thresholds of 15 min and 30 min, as defined in agreement with the ED managers.

For the sake of brevity, we report point estimates of the forecasting error per patient only for the first two days. For the remaining period, we show the trend of the daily average error of the single predictions.

Figs. 8(a) and 8(b) show the error trends on days 1 and 2, respectively, with daily retraining of the model. The mean error per day is 19 minutes on day 1, and 22 minutes on day 2.

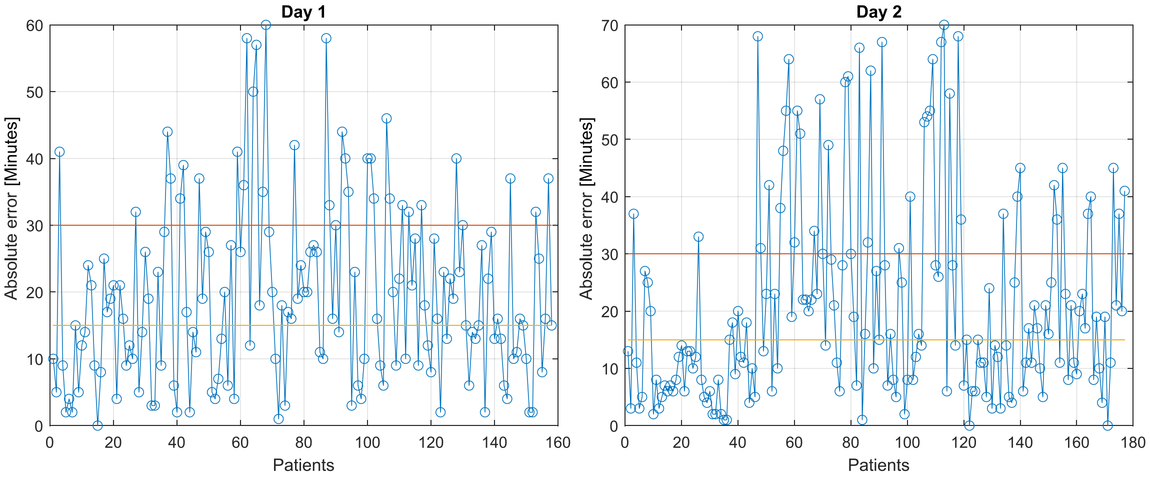
As expected, the stochastic nature of the ED system leads to greater error variability, with peaks during rush hours. However, the error is under 30 min in 80% of cases on day 1, and in 76% of cases during day 2. A slightly lower percentage of cases demonstrate errors of less than 15 min, i.e., 45% on day 1 and 49% on day 2.

Figs. 9(a) and 9(b) show the error trends on days 1 and 2, respectively, with one retraining of the model every six hours.

The behaviour of the errors is quite similar. The mean error per day is lower: 16 min for day 1 and 20 min for day 2. In addition, the percentage of cases under the thresholds is higher than with daily training. For day 1, 87% of cases are under 30 min and 61% are under 15 min, while for day 2 76% of cases are under 30 min and 50% are under 15 min.

Fig. 10 shows the daily mean absolute error for the last two months of the dataset with daily and six-hour training. In both cases, the trend is flat, with some peaks on certain days. However, these peaks do not exceed 30 min. In more detail, as shown in Figs. 10(a) and 10(b), the average error over the two months is about 22.4 min with daily training, while it is about 21.1 min with six-hour training. Furthermore, the percentage of cases below the mean value is higher with more frequent training (about 58%).

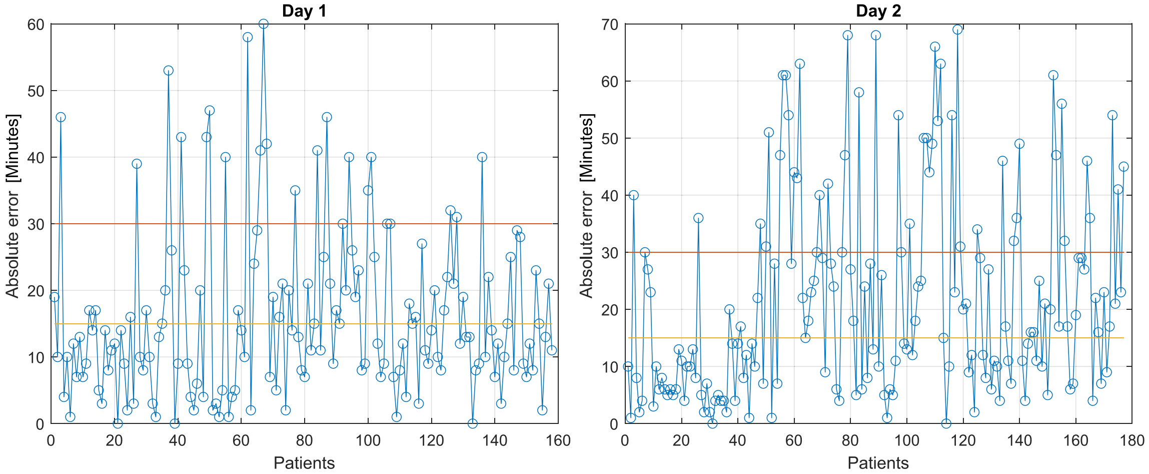
The experiment suggests three main conclusions. First, real-time estimations of waiting times with an error of 30 min can be obtained in about 80% of cases. Second, the predictive model should be retrained as frequently



(a) Day 1

(b) Day 2

Fig. 8. Absolute error per day, in minutes (daily training).



(a) Day 1

(b) Day 2

Fig. 9. Absolute error per day, in minutes (six-hour training).

as possible, e.g., every six hours. Finally, by using RF within the forecasting system, the computational time is relatively low. This evidence supports our choice of RF for keeping the computational time at a minimum. These are promising results for a complex scenario like the ED, suggesting that a forecasting system with satisfactory accuracy can be developed.

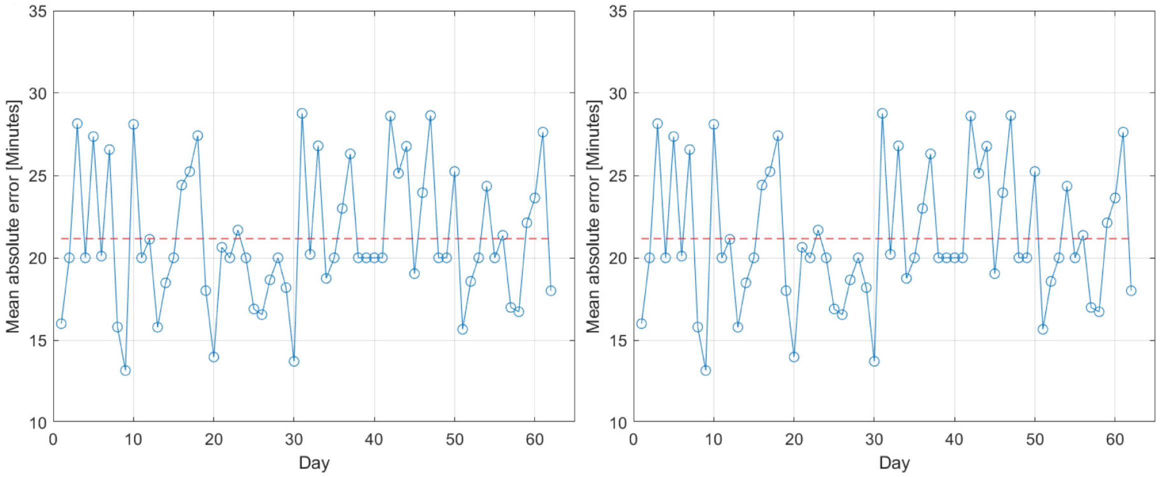
6. Discussion and conclusions

This paper answers the call for accurate waiting time predictions that can help to solve operational decision-making problems in EDs (Ang et al., 2016; Araz et al., 2019; Dong et al., 2018).

We tested several learning techniques aimed at accurately forecasting waiting times in almost real time, based on data extracted from the HISs of two Italian EDs. The study contributes to the current literature on predictive analytics in EDs, as an extensive and structured comparative analysis that establishes the actual value of different learning techniques in complex and dynamic environments is lacking. We demonstrated the accuracy and suitability of predictive models in a real application context.

The models developed were able to provide accurate waiting time estimations, due to new queue-based predictors that effectively capture the current ED state. According to the literature, enriching the set of predictors with variables that describe the crowding level of activities within the ED system can improve the predictive power of the learning techniques (Benevento et al., 2019; Kuo et al., 2020).

The experiments revealed that EM was the most effective at predicting waiting times, significantly outperforming all other techniques in terms of the MSE and MAE. This result supports other research in the field suggesting that ensemble techniques are much more accurate than the individual base learners they consist of (Dietterich, 2000; Gigoni et al., 2017; Valentini & Masulli, 2002). Lasso



(a) Daily training

(b) Six-hour training

Fig. 10. Absolute error per day, in minutes.

achieved the worst results in terms of prediction accuracy. This was expected, as linear regression techniques, like OLS, Lasso, etc., have disadvantages when estimating unknown non-linear relationships (Asaro et al., 2007), which are typical of dynamic and complex systems like EDs. In terms of computational efficiency, the training times of EM, SVR, ANN, and RF were higher than that of Lasso. However, it is noteworthy that such times, particularly for ANN and RF, do not prohibit real-life applications in EDs. Moreover, this study demonstrated that the forecasting power of EM, SVR, and ANN was significantly superior to Lasso. When considering both prediction accuracy and computational efficiency, RF appears to be a reasonable tradeoff. We also found that, as expected, the higher the data variability, the lower the predictive power of the models.

In general, the results suggest that a forecasting system based on the proposed predictive models can provide waiting time estimations with an error of about 30 min in 80% of cases. This is a valuable result, as it demonstrates the potential applicability and usefulness of the developed predictive models in a real and complex context like the ED.

The results from our study have significant practical implications for EDs and hospitals. They provide new perspectives for EDs and hospital managers through a forecasting tool based on a real-time monitoring system, which enables them to be constantly informed about the expected state of the ED (for example, the volume of patients and the expected waiting time). This information is essential for EDs, to help them make more proactive responses if long waiting times can be reduced. Timely and updated predictions at the triage stage can also enable hospitals to better identify patients who join or abandon the queue to receive treatment. Thus, hospital managers can assess in advance the potential extent of ED overcrowding and react accordingly (Derlet, 2002). The implementation of the developed predictive models within a decision support tool can help hospital decision makers more effectively manage ED activities and resources, based on the expected waiting time, and adjust their strategies aimed at reducing long delays. This can prevent overcrowding or enable them to react dynamically to possibly critical occurrences or conditions (e.g., improving the routing of patients who arrive by rescue vehicles in the ED network).

The study also provided indications in terms of patient communication. Accurate information about waiting times can help manage the expectations of patients waiting for treatment. Accordingly, waiting time announcements can positively affect patient behaviour by increasing their tolerance for waiting and reducing their levels of stress and anxiety (Jouini, Akşin, & Dallery, 2011). Prolonged perceived waiting times and the level of communication are strongly correlated with patient perceptions regarding service quality. Thus, communicating information about the waiting time can increase patient satisfaction and reduce the likelihood of patients leaving without being seen by a physician (Soremekun et al., 2011). Greater benefits can also be achieved if multiple nearby EDs provide and share real-time waiting time estimates for prospective patients. This can help distribute and balance the load among EDs, and thus reduce the overall waiting time.

A potential limitation of this study is the size of the dataset. A one-year dataset is reasonable but may lead to less generalizability. Using a larger dataset can be more valuable, as seasonal variations in waiting times can be considered, during both model training and testing, and to obtain more reliable results. In addition, the effectiveness of the developed predictive models depends on the availability of reliable data. However, due to the multiple data sources and high degree of data granularity, data quality issues such as missing values and outliers may be encountered in real-life datasets, and such issues are difficult to handle.

In the future, we aim to use larger datasets and explore external sources of data (e.g., road traffic data, temperature, weather forecasts, and epidemiological data) to identify potential exogenous predictors that could improve

Table 14
Statistical significance of the accuracies of the predictive models with D1 and D2.

A better than B	p-value	
	D1	D2
EM better than RF	9×10^{-5}	2×10^{-6}
EM better than ANN	8×10^{-16}	2×10^{-13}
EM better than SVR	4×10^{-17}	2×10^{-7}
EM better than Lasso	1×10^{-15}	1×10^{-16}
RF better than ANN	0.002	0.003
RF better than SVR	7×10^{-10}	0.001
RF better than Lasso	3×10^{-16}	3×10^{-15}
ANN better than SVR	7×10^{-8}	2×10^{-7}
ANN better than Lasso	2×10^{-16}	7×10^{-16}
SVR better than Lasso	1×10^{-15}	6×10^{-16}

the accuracy of the predictive models. Our future work will also involve developing more sophisticated arrivals-based predictors by using time series forecasting methods, which can consider the stochastic behaviour of patient arrivals. Finally, we plan to conduct prediction experiments with additional key measures (e.g., hospital admissions and lengths of stay), and the traceability of patient position and routes within the ED, by means of wearable sensors or other environmental sensors. Implementing a more effective and complete decision support tool that uses constantly updated information about the ED's current situation would be beneficial to support operational decision making.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Statistical test to evaluate the significance of the results

To evaluate the reliability and the significance of the results, we conducted a Wilcoxon signed-rank test on the errors from the test set (Flores, 1989).

In more detail, the p-value was computed for the null hypothesis that the models A and B have equal performance, by considering a significance level α equal to 0.05. We used the Bonferroni correction method to consider multiple tests. Thus, we corrected α to α/number of tests. If the p-value is smaller than the corrected α, we reject the null hypothesis and accept that there is a significant difference between the two models.

The results are reported in Table 14. All comparisons reject the null hypothesis, and, thus, the results we obtained are statistically relevant. EM and RF have significantly higher predictive power compared to ANN, SVR, and Lasso. Conversely, Lasso achieves statistically lower accuracy than the other techniques.

References

ACEP (2016). Emergency department crowding: High-impact solutions. URL https://www.acep.org/globalassets/sites/acep/media/crowding/empc_crowding-ip_092016.pdf.

Ahalt, V., Argon, N. T., Ziya, S., Strickler, J., & Mehrotra, A. (2016). Comparison of emergency department crowding scores: a discrete-event simulation approach. *Health Care Management Science*, *21*(1), 144–155. http://dx.doi.org/10.1007/s10729-016-9385-z.

Ang, E., Kwasnick, S., Bayati, M., Plambeck, E. L., & Aratow, M. (2016). Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management*, *18*(1), 141–156. http://dx.doi.org/10.1287/msom.2015.0560.

Araz, O. M., Olson, D., & Ramirez-Nafarrate, A. (2019). Predictive analytics for hospital admissions from the emergency department using triage information. *International Journal of Production Economics*, *208*, 199–207. http://dx.doi.org/10.1016/j.ijpe.2018.11.024.

Arkun, A., Briggs, W., Patel, S., Dattilo, P., Bove, J., & Birkhahn, R. (2010). Emergency department crowding: Factors influencing flow. *The Western Journal of Emergency Medicine*, *11*, 10–15.

Asaro, P. V., Lewis, L. M., & Boxerman, S. B. (2007). The impact of input and output factors on emergency department throughput. *Academic Emergency Medicine*, *14*(3), 235–242. http://dx.doi.org/10.1197/j.aem.2006.10.104.

Benevento, E., Aloini, D., Squicciarini, N., Dulmin, R., & Mininno, V. (2019). Queue-based features for dynamic waiting time prediction in emergency department. *Measuring Business Excellence*, *23*(4), 458–471. http://dx.doi.org/10.1108/mbe-12-2018-0108.

Blackard, J. A., & Dean, D. J. (1999). Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, *24*(3), 131–151. http://dx.doi.org/10.1016/s0168-1699(99)00046-0.

Bontempi, G., Taieb, S. B., & l Le Borgne, Y.-A. (2013). Machine learning strategies for time series forecasting. In *Business Intelligence* (pp. 62–77). Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-642-36318-4_3.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Carrasco, M., López, J., & Maldonado, S. (2019). Epsilon-nonparallel support vector regression. *Applied Intelligence*, *49*(12), 4223–4236. http://dx.doi.org/10.1007/s10489-019-01498-1.

Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting*, *28*(2), 543–552. http://dx.doi.org/10.1016/j.ijforecast.2011.05.002.

Derlet, R. W. (2002). Overcrowding in emergency departments: Increased demand and decreased capacity. *Annals of Emergency Medicine*, *39*(4), 430–432. http://dx.doi.org/10.1067/mem.2002.122707.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139–157. http://dx.doi.org/10.1023/a:1007607513941.

Ding, R., McCarthy, M. L., Desmond, J. S., Lee, J. S., Aronsky, D., & Zeger, S. L. (2010). Characterizing waiting room time, treatment time, and boarding time in the emergency department using quantile regression. *Academic Emergency Medicine*, *17*(8), 813–823. http://dx.doi.org/10.1111/j.1553-2712.2010.00812.x.

Dong, J., Yom-Tov, E., & Yom-Tov, G. B. (2018). The impact of delay announcements on hospital network coordination and waiting times. *Management Science*, http://dx.doi.org/10.1287/mnsc.2018.3048.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155–161.

Dudek, G. (2015). Short-term load forecasting using random forests. In *Advances in Intelligent Systems and Computing* (pp. 821–828). Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-11310-4_71.

Elalouf, A., & Wachtel, G. (2016). An alternative scheduling approach for improving emergency department performance. *International Journal of Production Economics*, *178*, 65–71. http://dx.doi.org/10.1016/j.ijpe.2016.05.002.

Flores, B. E. (1989). The utilization of the wilcoxon test to compare forecasting methods: A note. *International Journal of Forecasting*, *5*(4), 529–535. http://dx.doi.org/10.1016/0169-2070(89)90008-3.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *vol. 1*, *The Elements of Statistical Learning*. (10), Springer series in statistics New York.

Galetsi, P., & Katsaliaki, K. (2019). A review of the literature on big data analytics in healthcare. *Journal of the Operational Research Society*, 1–19. http://dx.doi.org/10.1080/01605682.2019.1630328.

Gardner, M., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Enviroment*, *32*(14–15), 2627–2636. http://dx.doi.org/10.1016/s1352-2310(97)00447-0.

Gigoni, L., Betti, A., Crisostomi, E., Franco, A., Tucci, M., Bizzarri, F., et al. (2017). Day-ahead hourly forecasting of power generation from photovoltaic plants. *IEEE Transactions on Sustainable Energy*, *9*(2), 831–842.

Golmohammadi, D. (2016). Predicting hospital admissions to reduce emergency department boarding. *International Journal of Production Economics*, *182*, 535–544. http://dx.doi.org/10.1016/j.ijpe.2016.09.020.

Golmohammadi, D., & Radnia, N. (2016). Prediction modeling and pattern recognition for patient readmission. *International Journal of Production Economics*, *171*, 151–161. http://dx.doi.org/10.1016/j.ijpe.2015.09.027.

Gonçalves, F., Pereira, R., ao Ferreira, J., Vasconcelos, J. B., Melo, F., & Velez, I. (2018). Predictive analysis in healthcare: Emergency wait time prediction. In *Advances in Intelligent Systems and Computing* (pp. 138–145). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-01746-0_16.

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

Gunn, S. R., et al. (1998). Support vector machines for classification and regression. *ISIS Technical Report*, *14*(1), 5–16.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hemaya, S. A. K., & Locker, T. E. (2011). How accurate are predicted waiting times, determined upon a patient's arrival in the emergency department? *Emergency Medicine Journal*, *29*(4), 316–318. http://dx.doi.org/10.1136/emj.2010.106534.

Hobbs, D., Kunzman, S. C., Tandberg, D., & Sklar, D. (2000). Hospital factors associated with emergency center patients leaving without being seen. *The American Journal of Emergency Medicine*, *18*(7), 767–772. http://dx.doi.org/10.1053/ajem.2000.18075.

Hoot, N. R., & Aronsky, D. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, *52*(2), 126–136.e1. http://dx.doi.org/10.1016/j.annemergmed.2008.03.014.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. http://dx.doi.org/10.1016/0893-6080(89)90020-8.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Islam, M., Hasan, M., Wang, X., Germack, H., & Noor-E-Alam, M. (2018). A systematic review on healthcare analytics: Application and theoretical perspective of data mining. *Healthcare*, *6*(2), 54. http://dx.doi.org/10.3390/healthcare6020054.

Jouini, O., Akşin, Z., & Dallery, Y. (2011). Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management*, *13*(4), 534–548. http://dx.doi.org/10.1287/msom.1110.0339.

Khalifa, M., & Khalid, P. (2015). Developing strategic health care key performance indicators: A case study on a tertiary care hospital. *Procedia Computer Science*, *63*, 459–466. http://dx.doi.org/10.1016/j.procs.2015.08.368.

Koh, H. C., Tan, G., et al. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management*, *19*(2), 65.

Koufi, V., Malamateniou, F., & Vassilacopoulos, G. (2015). A big data-driven model for the optimization of healthcare processes. *Studies in Health Technology and Informatics*, *210*, 697–701.

Kuhn, M., & Johnson, K. (2013). A short tour of the predictive modeling process. In *Applied Predictive Modeling* (pp. 19–26). Springer New York, http://dx.doi.org/10.1007/978-1-4614-6849-3_2.

Kuo, Y.-H., Chan, N. B., Leung, J. M., Meng, H., So, A. M.-C., Tsoi, K. K., et al. (2020). An integrated approach of machine learning and systems thinking for waiting time prediction in an emergency department. *International Journal of Medical Informatics*, *139*, Article 104143. http://dx.doi.org/10.1016/j.ijmedinf.2020.104143.

Lin, D., Patrick, J., & Labeau, F. (2013). Estimating the waiting time of multi-priority emergency patients with downstream blocking. *Health Care Management Science*, *17*(1), 88–99. http://dx.doi.org/10.1007/s10729-013-9241-3.

Ogutu, J. O., & Piepho, H.-P. (2014). Regularized group regression methods for genomic prediction: Bridge, MCP, SCAD, group bridge, group lasso, sparse group lasso, group MCP and group SCAD. *BMC Proceedings*, *8*(S5), http://dx.doi.org/10.1186/1753-6561-8-s5-s7.

Polikar, R. (2012). Ensemble learning. In *Ensemble Machine Learning* (pp. 1–34). Springer.

Probst, P., & Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, *18*(1), 6673–6690.

Rebuge, A., & Ferreira, D. R. (2012). Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, *37*(2), 99–116. http://dx.doi.org/10.1016/j.is.2011.01.003.

Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1–2), 1–39. http://dx.doi.org/10.1007/s10462-009-9124-7.

Roquette, B. P., Nagano, H., Marujo, E. C., & Maiorano, A. C. (2020). Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. *Neural Networks*, *126*, 170–177. http://dx.doi.org/10.1016/j.neunet.2020.03.012.

Senderovich, A., Leemans, S. J. J., Harel, S., Gal, A., Mandelbaum, A., & van der Aalst, W. M. P. (2016). Discovering queues from event logs with varying levels of information. In *Business Process Management Workshops* (pp. 154–166). Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-42887-1_13.

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. (2013). A survey of forecast error measures. *World Applied Sciences Journal*, *24*(24), 171–176.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. http://dx.doi.org/10.1023/b:stco.0000035301.49549.88.

Soremekun, O. A., Takayesu, J. K., & Bohan, S. J. (2011). Framework for analyzing wait times and other factors that impact patient satisfaction in the emergency department. *The Journal of Emergency Medicine*, *41*(6), 686–692. http://dx.doi.org/10.1016/j.jemermed.2011.01.018.

Sun, Y., Teow, K. L., Heng, B. H., Ooi, C. K., & Tay, S. Y. (2012). Real-time prediction of waiting time in the emergency department, using quantile regression. *Annals of Emergency Medicine*, *60*(3), 299–308. http://dx.doi.org/10.1016/j.annemergmed.2012.03.011.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288. http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225–1231. http://dx.doi.org/10.1016/s0895-4356(96)00002-9.

van der Vaart, T., Vastag, G., & Wijngaard, J. (2011). Facets of operational performance in an emergency room (ER). *International Journal of Production Economics*, *133*(1), 201–211. http://dx.doi.org/10.1016/j.ijpe.2010.04.023.

Valentini, G., & Masulli, F. (2002). Ensembles of learning machines. In *Neural Nets* (pp. 3–20). Springer Berlin Heidelberg, http://dx.doi.org/10.1007/3-540-45808-5_1.

van der Aalst, W. (2016). *Process Mining*. Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-662-49851-4.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*(7), 1341–1390. http://dx.doi.org/10.1162/neco.1996.8.7.1341.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*(2), 301–320. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.