Porto and Fogliatto
*BMC Medical Informatics and Decision Making* (2024) 24:377
https://doi.org/10.1186/s12911-024-02788-6

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

# Enhanced forecasting of emergency department patient arrivals using feature engineering approach and machine learning

Bruno Matos Porto[1,2*] and Flavio Sanson Fogliatto[1,2]

## Abstract

**Background** Emergency department (ED) overcrowding is an important problem in many countries. Accurate predictions of ED patient arrivals can help management to better allocate staff and medical resources. In this study, we investigate the use of calendar and meteorological predictors, as well as feature-engineered variables, to predict daily patient arrivals using datasets from eleven different EDs across three countries.

**Methods** Six machine learning (ML) algorithms were tested on forecasting horizons of 7 and 45 days. Three of them – Light Gradient Boosting Machine (LightGBM), Support Vector Machine with Radial Basis Function (SVM-RBF), and Neural Network Autoregression (NNAR) – were never before reported for predicting ED patient arrivals. Algorithms' hyperparameters were tuned through a grid-search with cross-validation. Prediction performance was assessed using fivefold cross-validation and four performance metrics.

**Results** The eXtreme Gradient Boosting (XGBoost) was the best-performing model on both prediction horizons, also outperforming results reported in past studies on ED arrival prediction. XGBoost and NNAR achieved the best performance in nine out of the eleven analyzed datasets, with MAPE values ranging from 5.03% to 14.1%. Feature engineering (FE) improved the performance of the ML algorithms.

**Conclusion** Accuracy in predicting ED arrivals, achieved through the FE approach, is key for managing human and material resources, as well as reducing patient waiting times and lengths of stay.

**Keywords** Emergency department, Patient arrivals, Feature engineering, Machine learning algorithms, Patient visits forecast, Time series forecasting

## Introduction

Emergency department (ED) overcrowding, a global issue [1–3], poses significant challenges in managing these environments [3–6]. It refers to an imbalance between the demand and supply of emergency services. This imbalance occurs when demand for emergency beds surpasses the current capacity of the ED, including human and material resources for patient care [7, 8]. Addressing the increasing incidence of ED overcrowding calls for interventions to minimize its impact [9, 10].

ED overcrowding impacts patient satisfaction [11–13], leading to emotional exhaustion among healthcare teams [12, 14], and extends patient stays [12]. It results from external factors such as population growth and the incidence of epidemic events, as well as internal issues such as delays in patient care and inadequate ED resources [1]. Accurate prediction of patient arrivals helps optimize resource allocation and improve care quality [15, 16].

*Correspondence:
Bruno Matos Porto
bmatosporto@gmail.com
[1] Industrial Engineering Department, Federal University of Rio Grande do Sul, Av. Osvaldo Aranha, 99, 5th floor, Porto Alegre, RS 90020-035, Brazil
[2] Industrial Engineering Department, Federal University of Rio Grande do Sul, Av. Osvaldo Aranha, 99, 5th floor, Porto Alegre 90035-190, Brazil

Several studies have focused on predicting patient arrivals in EDs, primarily using autoregressive integrated moving average (ARIMA) models [17–23]. However, while effective for data with systematic variations, these models are challenged by irregular fluctuations [16, 24]. To overcome this limitation, researchers have turned to machine learning (ML) algorithms, such as artificial neural networks (ANN) [7, 15]. Recent advancements propose hybrid approaches [20, 21], combining statistical models with ML algorithms and text mining techniques [25, 26]. Hybrid approaches have demonstrated superior predictive performance compared to individual models in forecasting ED patient arrivals [18, 27].

Despite the proven effectiveness of Feature Engineering (FE) in enhancing ML model performance, existing studies on patient arrival prediction have yet to explore its potential. FE has consistently demonstrated significant improvements in accuracy across various domains, including patient flow modeling in healthcare [28], forecasting competitions [29], and other predictive modeling fields [30]. By creating new features through domain knowledge or exploratory data analysis [31–33], FE can enhance model performance. As Kuhn and Johnson [31] note, techniques such as Principal Component Analysis, one-hot encoding, and other FE methods can substantially improve ML algorithms. This approach has been successful in prediction competitions such as Kaggle, where extracting additional features (e.g., those derived from time data) has led to improved predictive accuracy [29, 34].

The objectives of this article are threefold: (*i*) to compare the performance of six ML algorithms (namely, XGBoost—eXtreme Gradient Boosting, LightGBM—Light Gradient Boosting Machine, RF—Random Forest, SVM-RBF—Support Vector Machine with Radial Basis Function, NNAR—Neural Network Autoregression, and GLMNET—Lasso and Elastic-Net Generalized Linear Model) and identify the most accurate ones for predicting daily patient arrivals using data from eleven different EDs; (*ii*) to apply FE to create calendar-related features, which are included as predictors in the ML algorithms; and (*iii*) to compare prediction accuracy in datasets treated with FE (i.e., combining FE variables with meteorological and calendar predictors), and subsequently implement a variable selection step based on the RF technique for all types of EDs analyzed. Two reasons support the use of an FE approach in this study: (*i*) its documented ability to improve ML algorithm performance [30, 34], and (*ii*) the absence of previous applications in predicting ED patient arrivals. Meteorological and calendar variables were chosen as predictors, given their wide use in the area of predicting patient arrivals and general applicability across different contexts.

Our research contributes to the state-of-the-art in ED patient arrival prediction studies in three ways aligned with our objectives. First, we address a gap in research by applying ML algorithms—specifically LightGBM, SVM-RBF, and NNAR—to predict daily ED arrivals, which is novel in the literature (Sect. " Background" and recent studies [20, 35]).

Second, we introduce FE as a means to enhance the performance of ML algorithms in predicting ED patient arrivals. While existing literature highlights the use of ML models and hybrid approaches, there remains a gap in exploring novel predictor variables for ED arrival forecasting. Our approach addresses this gap by identifying new predictor variables that significantly improve the performance of ML models. The latest systematic review on this topic [36] identified the discovery of new predictor variables as an underexplored area, calling for future studies to investigate this further. By demonstrating that FE-generated variables, across datasets from different countries, are more influential than traditional predictors such as meteorological factors, our work not only improves prediction accuracy but also contributes to advancing this important aspect of ED arrival forecasting.

Building on the conclusions of systematic reviews by Wangon et al. [37] and Jiang et al. [38], which found that calendar variables are more influential than meteorological ones in predicting ED patient arrivals, we propose an FE approach that creates new variables based on time-series signatures (timestamps). These time-based variables have proven to be strong predictors across multiple datasets. In contrast to prior research that primarily focused on conventional meteorological or calendar variables, our approach applies FE to generate additional variables from temporal signatures. Predictive analysis across datasets from different countries demonstrates that these feature-engineered variables outperform traditional predictors in terms of predictive power.

Our analysis of FE datasets with the XGBoost algorithm yields unprecedented results in the current literature. Third, our study expands the scope by comparing ML algorithms using data from 11 EDs across three countries. Most previous studies have focused on single ED predictions [7, 16, 21, 39–41], which may limit statistical significance and generalizability. Boyle et al. [22] analyzed data from 27 EDs but did not explore ML algorithms. Additionally, few studies have employed rigorous comparison methods such as cross-validation for ED arrival prediction (e.g., [3, 7, 20]). In our study, we employ grid-search with cross-validation to optimize hyperparameters across all algorithms and evaluate performance over two distinct horizons using five-fold cross-validation. We also use a variable selection step based on RF.

## Background

The literature on patient arrival prediction in EDs has expanded in recent years, including a number of systematic literature review articles, e.g., [12, 13, 37, 38]. To avoid overlap with existing studies, specific criteria were applied in the selection of articles discussed in this section. First, only studies whose main objective is the prediction of patient arrivals in EDs were included. Second, only studies published in the past ten years were included, as the literature reviews by Wargon et al. [37] (covering the period from 1981 to 2007) and Gul and Celik [13] (covering the period from 2001 to 2017) showed that studies prior to 2012 used only traditional forecasting methods. The third criterion is to select works that used calendar and meteorological predictors or only the time series in the prediction.

Thirty-three articles met the search criteria and are summarized in Table 1. The articles were classified based on the following characteristics: EDs analyzed and database time frame, forecast object, period and horizon, predictors tested, forecasting methods applied, most frequently retained predictors, partitioning of the dataset for validation purposes, performance metrics, and main results.

Here is a summary of the studies presented in Table 1, which also includes a glossary of abbreviations and acronyms used throughout the paper. Data periodicity is typically daily (66.66% of the studies) and hourly (33.33% of the studies), with 15.15% including both hourly and daily forecasts. Most articles focus on daily arrival predictions; hourly predictions are less commonly studied, as most EDs operate on daily staffing and resource planning [4, 5, 17]. The time frame of the analyzed databases ranged from one to ten years, with 48.48% of the studies covering up to three years of observations. A 7-day forecasting horizon was most commonly adopted (51.51% of the studies), which is considered more useful operationally [15], given that most EDs rely on short-term planning schedules.

Calendar predictors most frequently tested were weekdays (60.60% of the studies), month of the year (45.45%), holidays (45.45%), school holidays (15.15%), days before or after holidays (12.12%), and time of day (12.12%). Meteorological variables most commonly tested were temperature (42.42% of the studies), precipitation (18.18%), wind speed (18.18%), and relative humidity (12.12%). Additionally, 24.24% of the studies considered only the time series to predict patient arrivals. Meteorological and calendar variables most frequently retained in the models were weekdays (82% of the studies that tested these predictors), holidays (57%), month of the year (51%), and temperature (100%). Such findings are consistent with those reported by Gul and Celik [13], who identified the most commonly used independent variables for predicting patient arrivals in EDs as time of day, weekday, month of the year, days before or after holidays, vacation days, maximum and minimum temperatures, precipitation, and wind speed. Jiang et al. [38], in their literature review on ED patient arrivals, concluded that calendar variables are predominant compared to other types of independent variables.

The models for ED arrival prediction listed in Table 1 can be classified into four groups: time series models, regression models, ML algorithms, and hybrid methods. Among the time series models, ARIMA [1, 17] and its variants Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) [18, 54], Seasonal Autoregressive Integrated Moving Average (SARIMA) [5, 21], and Seasonal Autoregressive Integrated Moving Average with external variables (SARIMAX) [3, 46] have been widely used. Naive models, Seasonal Naive (Snaive) [47, 54], Error-trend-seasonal (ETS) [20, 45], and Exponential Smoothing (ES) models [5, 46] are also reported, with emphasis on the seasonal Holt-Winters (HW) model [1, 23]. The second group includes Logistic regression-based approaches [4, 6], logistic [55, 56], and Poisson models [48], which are causal models, unlike most time series models. More recently, ML algorithms have been employed to overcome limitations of causal and time series models. For example, Multilayer Perceptron Neural Network (MLP) [3, 27], Long Short-Term Memory (LSTM) [24, 49], *k*-nearest neighbours (KNN) [20, 41], XGBoost [5, 55], RF [54, 55], Support Vector Regression (SVR) [41, 57], Deep Neural Network-based algorithms [7, 56], such as Recurrent Neural Networks (RNNs) [5, 24], and Convolutional Neural Networks (CNNs) [7, 49], have been reported in studies. In addition to these three groups, hybrid approaches have also been used [5, 20, 24, 27]. For instance, Autoregressive Integrated Moving Average with Linear Regression (ARIMA-LR) [18, 21], Autoregressive Integrated Moving Average with Artificial Neural Network (ARIMA-ANN) [4, 18], and Autoregressive Integrated Moving Average with Support Vector Regression (ARIMA-SVR) [35] offer advantages over single models as they exploit the strengths of each individual model to improve prediction accuracy. Reviews by Gul and Celik [13] and Wargon et al. [37] demonstrate the predominance of time series models in predicting ED patient arrivals. In contrast, studies listed in Table 1 demonstrate the increasing application of ML algorithms and hybrid approaches combining statistical and ML models, particularly in more recent articles.

Regarding model validation, the most frequent procedure was to split the data into training and testing sets (70% of the studies). Among these studies, only 33%

**Table 1** Overview of studies on forecasting patient arrivals in emergency departments ($n = 33$)

| Author (Year) / Country | EDs analyzed / Time frame | Forecast object, period and horizon | Predictors tested | Forecasting models | Most frequently retained variables | Partitioning of the dataset for training, test and validation | Performance metrics | Main Results |
|---|---|---|---|---|---|---|---|---|
| Chen et al. [17] (2011) / Taiwan | ED of a regional teaching hospital with intensive care beds in Kaohsiung City/ 01/2005 – 09/2009 | Number of monthly general, traumatology, and pediatric arrivals over a 9-month horizon | - | ARIMA | - | Training: 01/2005 –12/2008. Test: 01/2009–09/2009 | MAPE | General patients: ARIMA model with MAPEs in the interval (13.71%—41.61%) and average 19.59% Traumatology patients: ARIMA model with MAPEs in the interval (10.16%—19.12%) and average 12.39% Pediatric patients: ARIMA model with MAPEs in the interval (5.73%—54.24%) and average 29.08% |
| Boyle et al. [22] (2012)/ Austrália | EDs of 27 public hospitals mixed urban e regional in Queensland / 2005–2009 | Monthly arrivals over a 12-month horizon | Calendar: day of week, month of year, holidays, and days before and after holidays | ARIMA, MLR and ES | All variables in the model MLR | Training: 01/2005-12/2008. Test: 01/2009–12/2009 | MAPE | For all 27 hospitals: MAPEs in the interval (7%—25%) and average 12.30% |
| Marcilio et al. [42] (2013) / Brazil | High complexity ED in the southern region of the city of São Paulo, operating 24/7 / 01/2008–12/2010 | Daily and monthly arrivals over 7 and 30-day horizons | Calendar: day of week, holidays, and days before and after a holidayClimate: daily average temperature | GLM, GEE and SARIMA | All variables in the best performing models (GLM and GEE) | Training: 01/2008–09/2010 Test: 10/2010–12/2010. After forecasting 7 and 30 days in October, the values observed of October were incorporated into the training set and the models re-estimated to November forecast. This process was repeated in the month of December | MAPE | GLM and GEE with MAPEs of 4.5–9.9% over 7-day horizon, with and without temperature as predictor. GLM and GEE with MAPEs ranging of 8.7–12.8% over 30-day horizon, with and without temperature as predictor |

**Table 1** (continued)

| Author (Year) / Country | EDs analyzed / Time frame | Forecast object, period and horizon | Predictors tested | Forecasting models | Most frequently retained variables | Partitioning of the dataset for training, test and validation | Performance metrics | Main Results |
|---|---|---|---|---|---|---|---|---|
| Menke et al. [43] (2014) / Country not reported | ED of a tertiary hospital, location not disclosed / 02/2007–12/2009 | Daily arrivals. Horizon not reported | Calendar: day of week and special events<br>Climate: weather and air quality | MLP with back propagation | All variables to estimate the MLP algorithm | There was no division between training and testing | $R^2$ | The adjusted $R^2$ is 0.957. 95% of the time, the MLP estimated an error within 20 arrivals |
| Kadri et al. [44] (2014) / France | Paediatric Emergency Department of the Lille Regional Hospital Centre, operating 24/7 / 01/2012–12/2012 | Total number of daily arrivals, unplanned arrivals (G2) and unexpected arrivals (G4) over 1 to 7-day horizons | - | ARMA | - | Training: 01/2012–12/2012. Test: 12/2012–12/2012 | ME | Total arrivals: ME of 3.79–9.03. Unplanned arrivals (G2): ME ranged of 5.5–9.26. Unexpected arrivals (G4): ME of 31–72, over all horizons |
| Bergs et al. [45] (2014)/ Belgium | Four EDs of a university hospital and three regional hospitals in the Flemish region / 01/2005–12/2011 | Monthly arrivals over a 12-month horizon | - | ETS | - | Training: 01/2005–12/2010<br>Test: 01/2011–12/2011 | MAPE, MASE and MAE | MAPEs ranged of 2.63–4.76% and MASE of 0.53–0.68 in forecasting the four EDs |
| Calegari et al. [46] (2016) /Brazil | ED of a public, tertiary teaching hospital in the city of Porto Alegre, serving high complexity patients, operating 24/7 / 01/2013–05/2015 | Total number of daily arrivals, urgent arrivals and very urgent arrivals over horizons of 1, 7, 14, 21 and 30 days | Calendar: month of year and day of week. Climate: minimum, maximum and average temperature, amount of rain, wind speed, relative humidity and hours of insolation | SS, HW, SARIMA and MSARIMA | All calendar and climate variables in the MSARIMA models | Training: 01/2013–03/2015<br>Test 30-days: (i) 03/2015–04/2015, (ii) 01/2015–02/2015 and (iii) 05/2015–05/2015. After forecasting of (i). The values observed of (i) were incorporated into the training set and the models re-estimated to forecast of (ii). This process was repeated on test set (iii) | MAPE | Total arrivals of all horizons: SS with MAPEs of 2.91%–11.35%. Urgent and very urgent arrivals of all horizons: SARIMA obtained MAPEs ranging of 3.98%–15.71% and 7.19%–17.23% respectively |

**Table 1** (continued)

| Author (Year) / Country | EDs analyzed / Time frame | Forecast object, period and horizon | Predictors tested | Forecasting models | Most frequently retained variables | Partitioning of the dataset for training, test and validation | Performance metrics | Main Results |
|---|---|---|---|---|---|---|---|---|
| Xu (2016) [18] / China | EDs from two hospitals in DaLian, LiaoNing Province. ED-A is large while ED-B is small / 01/2012–12/2013 | Daily arrivals over 1 and 7-day horizons | Calendar: day of week, month of year, public holidays and school holidays, day after a holiday Climate: maximum and minimum daily temperature | ARIMA, ARIMAX, ARIMA-LR, ARIMA-LR with smoothing, ARIMA- ANN and GLM | Group 1: temperature variables, group 2: holiday variables, and group 3: calendar variables. In ED-A, the three groups of variables were retained by the GLM, ARIMAX, ARIMA-LR and ARIMA-ANN models. In ED-B, group 2 was retained in all models | Training: 01/2012–06/2013 (547-days). Test: 07/2013–12/2013 (184-days) | MAPE and RMSE | ED-A and ED-B: ARIMA-LR obtained values of MAPEs ranging of 5.8%–13.1% and RMSEs of 5.37–136.4 over 1-day horizon. ED-A and ED-B: ARIMA-LR (smoothing) obtained MAPEs ranging of 6.1%–12.9% and RMSEs of 5.33–147 over 7-day horizon |
| Juang et al. (2017) [19] /Taiwan | ED of a medical center in southern Taiwan / 01/2009–12/2016 | Monthly arrivals within a 12-month horizon | - | ARIMA | - | Training: 01/2009–12/2015 Test: 01/2016–12/2016 | MAPE | ARIMA (0, 0, 1), obtained a MAPE of 8.91% |
| Hertzum (2017) [47] /Denmark | Four EDs of mid-sized hospitals in the Zealand region / 01/2012–01/2015 | Hourly arrivals over 1, 2, 4, 8 and 24-h horizons | Calendar: time of day, day of week and month of year | MLR, ARIMA and Naive | The months of year, days of week, and times of day in MLR | Training: 01/2012–12/2014 Test: 01/2015–01/2015 | MAPE, MASE and MAE | 1-h prediction in the 4 EDs: ARIMA e MLR obtiveram MAPEs variando entre 47%–58% e MASE de 0.72–0.77. 2, 4, 8, and 24-h predictions for ED2: ARIMA and MLR with MAPEs of 41% and 37% (2-h), 34% and 34% (4-h), 21% and 26% (8-h), and 11% and 9.9% (24-h) |
| Carvalho-Silva et al. [23] (2018) / Portugal | ED of a public hospital in the Minho region / 01/2012–12/2014 | Weekly and monthly arrivals over one-week, three-week, one-month and twelve-month horizons | - | ARIMA, HW, Multiplicative HW, Moving average and ES | - | Training: 01/2012–12/2013 Test: 01/2014–12/2014 and three weeks of 01/2014 | MAPE | ARIMA: MAPEs of 5.22% and 6.34%, in weekly and monthly forecasts |

**Table 1** (continued)

| Author (Year) / Country | EDs analyzed / Time frame | Forecast object, period and horizon | Predictors tested | Forecasting models | Most frequently retained variables | Partitioning of the dataset for training, test and validation | Performance metrics | Main Results |
|---|---|---|---|---|---|---|---|---|
| Yucesan (2018) [4] / Turkey | ED of a private hospital in Trabzon / 05/2015–04/2017 | Daily arrivals. Horizon not reported | Calendar: day of week, day of year and month of year | MLR, ARIMA, ANN, ES, ARIMA-ANN and ARIMA-LR | MLR and ANN models retained all calendar variables. Best performing hybrid methods did not retain variables | Training: 2016 (365-days). Didn't have the test set | MAPE | ARIMA-ANN and ARIMA-LR: MAPEs of 0.49% and 0.92% respectively |
| Asheim et al. [48] (2019) /Noruega | ED of St. Olav's University Hospital, Trondheim, operating 24/7 / 01/2010–12/2018 | Hourly arrivals over 1, 2 and 3-h horizons | - | Poisson regression | - | Training: 01/2010–12/2016 Validation: 01/2017–12/2017 Test: 1, 2 and 3-h | MAPE | 1, 2, and 3-h forecasts: MAPEs were 57%, 43% and 36%, respectively |
| Jilani et al. [15] (2019) / United Kingdom | Four EDs in the UK, locations not disclosed / 01/2011–12/2015 | Weekly and monthly arrivals over 4-week and 4-month horizons | - | Fuzzy Time Series (FTS), ARIMA and NN | - | Training: 01/2011–12/2015 (260 datapoints). Test: 4-weeks and 4-months | MAPE and RMSE | For all EDs: MAPEs ranged from 2.6%–4.7% using FTS in weekly forecasts and MAPEs from 2.01%–2.81% and RMSE from 57.30–167.89 in monthly forecasts |
| Whitt et al. [3] (2019) /Israel | ED of Rambam Hospital, Haifa, operating 24/7 / 01/2004–10/2007 | Daily arrivals over 1 to 7-day horizons | Calendar: days of week, month of year and holidays. Climate: precipitation, maximum and minimum daily temperature | MLR, SARIMA, SARIMAX and MLP | The days of week, the month of year, holidays and temperatures in MLR, SARIMAX and MLP | tenfold cross validation in all datasets for MLP | MAPE and MSE | SARIMAX: 8.4% MAPE in 1-day forecast. MSE between 193–211, in forecasts of 1 to 7-days |
| Khaldi et al. [27] (2019) /Morocco | ED in the city of Fes, operating 24/7 / 2010–2016 | Weekly arrivals over one-week week horizon | - | ANN, ARIMA EEMD-ANN and DWT-ANN | - | ARIMA used 80%/20% train/test partitions. Remaining models used training: 70%, Validation: 15%, Test: 15% | RMSE, MAE and $R^2$ | EEMD-ANN had the best performance: RMSE of 52.86 and MAE of 39.88 |

**Table 1** (continued)

| Author (Year) / Country | EDs analyzed / Time frame | Forecast object, period and horizon | Predictors tested | Forecasting models | Most frequently retained variables | Partitioning of the dataset for training, test and validation | Performance metrics | Main Results |
|---|---|---|---|---|---|---|---|---|
| Zhang et al. [35] (2019) /China | Radiology ED of a large hospital in Sichuan / 01/2013–12/2016 | Daily arrivals over a 1-day horizon | Calendar: temporal factors related to annual, quarterly, monthly and weekly periodicities and the effect of holidays | ARIMA, SVR and ARIMA-SVR | Temporal factors annual, quarterly, monthly, weekly and the effect of holidays | Training: 80% Test: 20% tenfold cross validation only in training set to determine the hyperparameters of the SVR and ARIMA-SVR methods | MAPE, RMSE, MAE and relative error (RE) | ARIMA-SVR achieved the best performance: MAPE of 7.02% and RMSE of 19.20 |
| Choudhury and Urena [1] (2020) / USA | ED of a hospital in Des Moines, Iowa / 2014–2017 | Hourly arrivals over a 30-day horizon | - | TBATS, HW, NN and ARIMA | - | Training: 01/2014–07/2017 Test: 08/2017 | RMSE and ME | ARIMA had the best performance: RMSE of 1.55 and ME of 1.00 |
| Yousefi et al. [21] (2020) /Brazil | ED of a hospital in Belo Horizonte, operating 24/7 / 01/2014–11/2016 | Daily arrivals over 1 to 7-day horizons | Calendar: soccer match events, weekends, holidays, day before and after holidays | LSTM | Soccer match events, weekends, holidays | Training: 70% Test: 30% | MAPE and $R^2$ | MAPE values from 4.89% to 6.31% (average 5.55%) and $R^2$ values from 0.86 to 0.99 (average 0.940) over all horizons |
| Harrou et al. [16] (2020) / France | Paediatric ED of the CHRU-Lille Hospital, operating 24/7 / 01/2011–11/2013 | Hourly and daily arrivals over a 1 to 4-h and 1-day horizons | - | RNN, LSTM, BiLSTM, ConvLSTM, RBM, CNN, GRU and VAE | - | Training: 70% Test: 30% | RMSE, MAE, $R^2$ and EV | VAE was the best performing model, with RMSE values from 0.41 to 2.74 and MAE values from 0.30 to 2.32 over all horizons |
| Rocha and Rodrigues [5] (2021) / Portugal | ED of a Portuguese hospital, location not disclosed / 01/2009–12/2018 | Daily and hourly arrivals over 4 to 24-h horizons | Calendar: year, month of year, day of week, time of day and holidays | Snaive, ESD, SARIMA, AR-NN, RNN, XGBoost, RNN-1L, RNN-3L, RNN-1L–XGBoost and Ensemble model | AR-NN, RNN, XGBoost, RNN-1L, RNN-3L, RNN-1L–XGBoost and Ensemble models retained all variables | Training: 01/2009–12/2016 Validation: 01/2017–12/2017 Test: 01/2018–12/2018 | RMSE, sMAPE and MAE | RNN-1L: RMSE ranging of 4.8–26 and sMAPE of 4.3%–21.3%; RNN-3L: RMSE ranging of 4.7–26.1 and sMAPE of 4.2%–23.1%; and RNN-1L–XGBoost: RMSE ranging of 4.7–28.4 and sMAPE of 4.7%–23.2% over all horizons |

**Table 1** (continued)

| Author (Year) / Country | EDs analyzed / Time frame | Forecast object, period and horizon | Predictors tested | Forecasting models | Most frequently retained variables | Partitioning of the dataset for training, test and validation | Performance metrics | Main Results |
|---|---|---|---|---|---|---|---|---|
| Vollmer et al. [20] (2021) / England | EDs at St Mary's Hospital and Charing Cross Hospital in London, both operating 24/7 / 01/2011–12/2018 | Daily arrivals over 1, 3 and 7-day horizons | Calendar: day of week, month of year, public holidays, and school holidays. Climate: precipitation, maximum and minimum temperature on previous day | ARIMA, ETS, STLM, StructTS, GLMNET, RF, GBM and KNN | ML algorithms (GLMNET, RF, GBM and KNN) retained all variables | sixfold time series cross validation in all datasets | MAPE and MAE | St Mary's hospital: MAPEs of 6.9%–8.3% using time series models and MAPEs of 6.8%–7.4% using ML algorithms. Charing Cross Hospital: MAPEs ranging of 8.5%–12% using time series models and MAPEs from 8.6%–10.1% using ML algorithms |
| Erkamp et al. [6] (2021) / Netherlands | ED of Jeroen Bosch Hospital in the Hertogenbosch, operating 24/7 / 06/2016–12/2019 | Daily arrivals over a horizon not reported | Calendar: day of week, month of year, summer vacation, school holidays and public holidays Climate: wind speed, minimum, average and maximum temperatures, radiation, pressure, visibility, cloudiness, humidity and precipitation | MLR | Calendar variables, temperature maximum, radiation, pressure, visibility and humidity | Training: 06/2016–12/2018 Test: 01/2019–12/2019 | MAPE | MAPEs of 8.71%, retaining only calendar predictors and 8.68%, retaining calendar and climate predictors |
| Sudarshan et al. (2021) [7] / Denmark | ED of a public teaching hospital in Esbjerg, operating 24/7 / 05/2015–10/2017 and 01/2019–12/2019 | Daily and weekly arrivals over 1, 3, and 7-day horizons | Calendar: time of day, day of week, day of month, month of year, day of year, holidays and school holidays. Climate: temperature, wind speed, wind direction, cloud visibility, cloud cover and dew point | RF, LSTM and CNN | Day of week, day of month, month of year, day of year, holidays, temperature, wind speed, cloud visibility, cloud cover and dew point retained by all algorithms | tenfold cross validation in all datasets | MAPE and MSE | LSTM: average MAPEs of 9.31% and 8.91%, average MSEs of 193.25 and 190.46, on the three and seven-days forecasts. CNN: average MAPEs of 9.24% and 10.69%, average MSEs of 192.84 and 232.39, on the three and seven-days forecasts |

**Table 1** (continued)

| Author (Year) / Country | EDs analyzed / Time frame | Forecast object, period and horizon | Predictors tested | Forecasting models | Most frequently retained variables | Partitioning of the dataset for training, test and validation | Performance metrics | Main Results |
|---|---|---|---|---|---|---|---|---|
| Pekel et al. [39] (2021) /Turkey | ED of a public hospital in Istanbul / 01/2011–12/2012 | Daily arrivals within 148 days horizon | Calendar: month of year, day of week, and holidays. Climate: maximum daily temperature | Bayesian ANN model, GA-ANN and PSO-ANN | ML algorithms (ANN, GA-ANN and PSO-ANN) retained all variables | tenfold cross validation only in training set | MAPE, RMSE, MAE, MSE and $R^2$ | PSO-ANN obtained the lowest values in all performance metrics evaluated, with MAPE of 6% and RMSE of 53.29 |
| Harrou et al. [24] (2022) /France | Paediatric ED of the CHRU-Lille Hospital, operating 24/7 / 01/2011–12/2012 | Daily arrivals of the following types: non-urgent, urgent, unexpected, biology, radiology, scanner and echography over a 150-day horizon | - | DBN, RBM, LSTM, GRU, CNN-GRU, CNN-LSTM, GAN-RNN, SVR and RR | - | Training: 80% (first 580-days of each time series) Test: 20% (remaining 150 days). Algorithms' hyperparameters determined through minimization of the cross-entropy error | MAPE, RMSE, $R^2$, EV and MAE | DBN, RBM and CNN-GRU showed superior performance, with mean values of MAPEs between 4.09%–7.52% and RMSE of 0.63–0.94 |
| Petsis [40] (2022) / Greece | ED of a general and public hospital in Ioannina, operating on odd days / 03/2013–12/2019 | Daily arrivals over a 1- and 2-days horizon | Calendar: day of year, month of year, day of month, day of week and week of year. Public and local holidays, school holidays and special events. Climate: daily average, minimum and maximum temperatures, amount of rain per day and average wind speed per day | XGBoost | Algorithm retained all variables | Training: 80% (first 870-days of the time series) Test: 20% (remaining 217-days). Algorithm trained through cross-validation | RMSE, MAPE and MAE | One-Day forecasting: RMSE of 22.96 and MAPE of 6.5%. Two-Day forecasting: RMSE of 23.9 and MAPE of 6.91% |

**Table 1** (continued)

| Author (Year) / Country | EDs analyzed / Time frame | Forecast object, period and horizon | Predictors tested | Forecasting models | Most frequently retained variables | Partitioning of the dataset for training, test and validation | Performance metrics | Main Results |
|---|---|---|---|---|---|---|---|---|
| Zhang et al. [41] (2022) /China | ED of a public hospital in Hefei / 11/2019–11/2020 | Daily and hourly arrivals over a 90-day horizon | Calendar: day of week, time of day, month of year, season of year and holiday Climate: Daily temperature (min, max and mean), mean wind speed, air quality level, precipitation, weather and variables related to the air quality index | MLR, KNN, SVR, Ridge, XGBoost, RF, AdaBoost, Gradient Boosting, Bagging and LSTM | All variables (except month of year) retained by all ML algorithms on hourly arrivals. All variables (except time of day, month of year) retained by all ML algorithms and MLR on daily arrivals | Training: 11/2019–08/2020 Test: 09/2020–11/2020 | RMSE, MAPE and MAE | Daily Arrivals: SVR was superior with RMSE of 26.84 and MAPE of 8.81%. Hourly Arrivals: LSTM and XGBoost were best performers with RMSEs of 4 and 4.5 and MAPEs of 49% and 44%, respectively |
| Zhao et al. [49] (2022) /Singapore | ED of the Singapore General Hospital, operating 24/7 / 01/2015–12/2019 | Daily arrivals over horizons of 546, 327, 108 and 53 days | Calendar: day of week Climate: temperature and daily relative humidity | ARIMA, Prophet, CNN, ConvLSTM, BiLSTM, DLSTM, DGRU and DRNN | ML algorithms (CNN, ConvLSTM, BiLSTM, DLSTM, DGRU and DRNN) retained all variables | Training: 70% Test: 30% for 6 months to 5-year-long time series sizes | MAPE, RMSE and MAE | For 1 and 5-year-long time series, DLSTM yielded the best results with MAPE values of 5.67% and 5.72%, and RMSEs values of 25.29 and 24.71, respectively For 6 months and 3-year-long time series, DRNN yielded the best results with MAPEs values of 5.41% e 5.70% |

**Table 1** (continued)

| Author (Year) / Country | EDs analyzed / Time frame | Forecast object, period and horizon | Predictors tested | Forecasting models | Most frequently retained variables | Partitioning of the dataset for training, test and validation | Performance metrics | Main Results |
|---|---|---|---|---|---|---|---|---|
| Gafni-Pappas et al. [50] (2023) / USA | ED of the St. Joseph Mercy Hospital in Michigan, operating 24/7 / 01/2017–12/2019 | Daily arrivals over a 365-day horizon | Calendar: day of week, week of month, month of year Climate: Daily temperature maximum, precipitation, cloud cover, relative humidity, average pressure, daily change pressure, wet bulb, solar radiation, air quality level, ozone concentration, and percent flu visits | ARIMA, Prophet, ETS, GBM, RF, and Prophet-XGBoost | The day of week, week of month, temperature, average pressure, percent flu visits, retained by all ML algorithms | Training: 01/2017–12/2018 Test: 01/2019–12/2019 fivefold time series cross validation only in training set in ML algorithms | RMSE | RF showed superior performance, with mean value of RMSE of 18.94 |
| Hu et al. [51] (2023) / USA | ED of a quaternary-care teaching hospital in New York, operating 24/7 / 01/2018–01/2021 | Hourly arrivals over 12-h shifts | Calendar: day of week, day versus night, month of year, season, holidays, Recent arrival count 1 and 7-day prior, 30-day moving average, Climate: temperatures (min, max), precipitation, snow, wind, and hot-weather indicator | LR, SARIMA, SARIMAX, XGBoost, and regression tree (RT) | All variables (except month of year, recent arrival count 1-day prior, temperature max, and hot-weather indicator) in SARIMAX model | Training: 01/2018–01/2019 Test: 02/2019–01/2020 COVID test set: 02/2020–01/2021 tenfold cross validation only in training set in ML algorithms | MAPE and RMSE | SARIMAX showed superior performance, with mean values of MAPE of 8.70% and RMSE of 14.65 for non-COVID test set |
| Reboredo et al. [52] (2023) / Spain | ED of a teaching hospital in Santiago, operating 24/7 / 01/2015–12/2020 | Daily arrivals over horizons of 1 and 730-days | Calendar: day of week, weekends, season and past mean values | Poisson regression and INGARCH model | All variables in INGARCH model | Training: 01/2015–12/2018 Test: 01/2019–12/2020 | MAE and MSE | INGARCH showed superior performance, with values of MAE of 24.52 and MSE of 979.06 in forecast of 730-days |

**Table 1** (continued)

| Author (Year) / Country | EDs analyzed / Time frame | Forecast object, period and horizon | Predictors tested | Forecasting models | Most frequently retained variables | Partitioning of the dataset for training, test and validation | Performance metrics | Main Results |
|---|---|---|---|---|---|---|---|---|
| Rostami-Tabar et al. [53] (2023) / UK | ED of a hospital, location not disclosed / 04/2014–02/2019 | Hourly arrivals over 1 to 48-h horizons | Calendar: hour of day, day of week, day of year, week of year, holidays, and 24 h lags of holidays and events sporting Climate: temperature | ADAM, Poisson regression, QR, GAM, GAMLSS, Naive, ETS, TBATS, and Prophet, GBM | All variables in ADAM, Poisson regression, QR, and GBM | Training: 04/2014–02/2018 Test: 03/2018–02/2019 | RMSE, Quantile Bias, and Pinball score | ADAM showed the best performance, with value of RMSE of 0.0896 in forecast of 48-h |

*ADAM* Augmented Dynamic Adaptive Model, *ANN* Artificial Neural Networks, *ARMA* Auto Regressive-Moving Average, *ARIMA* Autoregressive Integrated Moving Average, *ARAMI-ANN* Autoregressive Integrated Moving Average with Artificial Neural Network, *ARAMI-LR* Autoregressive Integrated Moving Average with Linear Regression, *ARAMI-SVR* Autoregressive Integrated Moving Average with Support Vector Regression, *ARIMAX* Autoregressive Integrated Moving Average with Explanatory Variable *AR-NN* Autoregressive Neural Network, *BiLSTM* Bidirectional Long Short-Term Memory, *CNN* Convolutional Neural Networks, *CNN-GRU* Gated Recurrent Unit with Convolutional Neural Networks, *CNN-LSTM* Long Short-Term Memory with Convolutional Neural Networks, *ConvLSTM* Convolutional Long Short-Term Memory, *DBN* Deep Belief Network, *DGRU* Deep Stacked Architecture with Gated Recurrent Units, *DWT* Discrete Wavelet Transform, *DNN* Deep Neural Networks, *EEMD* Ensemble Empirical Mode decomposition, *ES* Exponential Smoothing, *ESD* Exponential smoothing decomposition, *ETS* Exponential smoothing state space, *EV* Explained variance, *FTS* Fuzzy Time Series, *GA-ANN* Genetic Algorithm-based ANN, *GAMLSS* Generalised Additive Models for Location, Scale and Shape, *GAN-RNN* Generative Adversarial Network based on Recurrent Neural Networks, *GBM* Gradient Boosting Machines, *GEE* Generalized Estimating Equations, *GLM* Generalized Linear Model, *GLMNET* Generalized Linear Models via Coordinate Descent, *GAMs* Generalised additive models, *INGARCH* Integer-valued generalized autoregressive conditional heteroscedastic model, *HW* Holt-Winters, *KNN* k-Nearest Neighbours, *LSTM* Long Short-Term Memory, *DLSTM* Deep Stacked Architecture with Long Short-Term Memory, *LR* Linear Regression, *MAE* Mean Absolute Error, *MASE* Mean Absolute Scaled Error, *MAPE* Mean Absolute Percentage Error, *ME* Mean Error, *MLP* Multilayer Perceptron Neural Network, *MLR* Multiple Linear Regression, *MSARIMA* Multivariate Autoregressive Integrated Moving Average, *MSE* Mean Squared Error, *Naïve* Naive Forecast, *Snaïve* Seasonal Naive Method, *NN* Neural Network, *NNAR* Neural Network Autoregression, *PSO-ANN* Particle Swarm Optimization algorithm-based ANN, *QR* Quantile regression, $R^2$ Coefficient of Determination, *RBM* Restricted Boltzmann Machines, *RNN* Recurrent Neural Networks, *DRNN* Deep Stacked Architecture with Recurrent Neural Network, *RNN-1L* Recurrent Neural Network with One Layer, *RF* Random Forest, *RMSE* Root Mean Square Error, *RR* Ridge Regression, *SARIMA* Seasonal Autoregressive Integrated Moving Average, *SARIMAX* Seasonal Autoregressive Integrated Moving Average with External Variables, *sMAPE* symmetric Mean Absolute Percentage Error, *SS* Simple Seasonal Exponential Smoothing, *STLM* Seasonal Decomposition of Time Series by LOESS method, *STLM-ETS* Seasonal Trend Decomposition using Loess and Exponential Smoothing State Space, *StructTS* Structural Time Series Model, *SVM* Support Vector Machine, *SVM-RBF* Support Vector Machine with Radial Basis Function, *SVR* Support Vector Regression, *TBATS* Trigonometric Exponential Smoothing State Space model with Box-Cox Transformation, *VAE* Variational AutoEncoder algorithm, *VAR* Vector Autoregression Model, *XGBoost* eXtreme Gradient Boosting

reported the proportion of division between training and testing sets, with a predominance of 70%/30% and 80%/20%. Furthermore, only 24.24% of the studies in Table 1 used some form of cross-validation to assess the quality of the predictions, with five studies [35, 39, 40] conducting cross-validation only on the training set and three studies [3, 7, 20] on the complete datasets. Cross-validation provides a more robust and reliable way to measure model performance [58].

To assess the accuracy of predictions, the most frequently employed error metrics were Mean Absolute Percentage Error—MAPE (75.75% of the studies), Root Mean Square Error—RMSE (45.45%), and Mean Absolute Error—MAE (39.39%). The prevalence of MAPE and RMSE can be attributed to their scale-independence and interpretability, which makes them suitable for comparative analysis across studies [6, 42, 59]. Consistently, Gul and Celik [13] also identified MAPE, RMSE, and MAE as the top three error metrics commonly used in ED patient arrival prediction studies.

Regarding prediction performance, our review indicates that 75% of the studies show ML algorithms outperforming time series and regression models. Specifically, when comparing only ML algorithms, the LSTM and SVR demonstrate superior performance. Furthermore, all studies considering hybrid approaches report better performance compared to time series and regression models. Among studies comparing ML algorithms and hybrid approaches, 78% report hybrid approaches as having the best performance.

Our review reveals some limitations of the ED patient arrival prediction literature. First, the number of EDs analyzed is generally limited, often from geographically close locations, resulting in low generalizability of the prediction methods and lack of external validation. Second, most studies do not employ cross-validation procedures. Last, most studies are not reproducible due to closed data sources and limited or no access to computer codes used in the analyses.

## Materials and methods

### Overview of the datasets

This retrospective and multicenter study uses datasets from 11 EDs in hospitals located in Australia, the USA, and the Netherlands. The datasets were extracted from the publicly available Harvard Dataverse [60], covering the period from January 1, 2014, to December 31, 2016.

Two criteria determined our choice of EDs to be included in the analysis. The first criterion is diversity. The sample of EDs includes both public and private hospitals that provide general and specialized care, located in countries with varying climatic conditions. Including multiple datasets from different hospitals and EDs aims to enhance the generalizability of the results. Such approach is supported by several authors (e.g., [4, 18, 27, 55, 61]), who suggest the comparison of ML prediction methods across different EDs as a research opportunity. To the best of our knowledge, there are currently no available studies that evaluate ML algorithms for predicting ED arrivals using data from different countries. The second criterion is the public availability of data. Research on forecasting is deemed significant when the datasets used are publicly available, allowing for comparability with other studies and replicability [62].

The majority of EDs included in the study are from public hospitals (7 out of 11), providing general care (7 out of 11) of medium and low complexity (9 out of 11). Out of the total number of EDs, four belong to teaching hospitals. All EDs operate 24 h per day, 7 days of the week. Table A1 gives the characterization of the EDs and descriptive statistics for the analyzed period. Complete three-year data were available for all EDs. The 11 complete datasets contained 1,096 observations. The annual average number of patient arrival events over all datasets is 46,495. On average, EDs had between 36 and 268 daily arrivals from January 2014 to December 2016.

Time series signatures use the date entry to generate a set of time-based variables, namely, day of the month and year, week of the month and year, defining when each observation occurred. The theoretical and empirical justifications for using variables created through FE were as follows: (i) the time signatures can capture common seasonal and trend patterns in time series of patient arrivals in the EDs; (ii) the distinct seasonal patterns identified in the exploratory data analysis of both weekly and monthly data confirm that the FE approach for extracting temporal features was well-suited for this analysis, as nuanced seasonal patterns can be more precisely captured by specific elements of the arrival time signatures; (iii) the FE approach improves the performance of ML algorithms [29, 30, 34]; (iv) building on findings of Wangon et al. [37] and Jiang et al. [38], which demonstrated that calendar variables are more predictive than meteorological variables for forecasting ED arrivals, we adopt a FE approach that generates new variables from time-series timestamps. Specifically, day-of-the-week patterns are frequently retained as critical predictors due to their strong association with patient arrival volumes. For instance, multiple studies [16, 20] and [21] have shown that Mondays often experience higher arrival rates, underscoring the significance of such temporal variables in improving predictive accuracy. This approach not only aligns with previous research but also enhances the model's ability to capture relevant temporal patterns in ED demand;

(v) the presence of weekly cycles and annual seasonality in patient arrival time series is widely documented in the literature. Including variables created through FE based on temporal patterns allows these cyclical patterns to be captured, which is essential for more accurate predictive models. The FE predictor variables used are presented in Table 3.

To assess the impact of FE on the performance of ML prediction algorithms, the predictors of interest contained in Tables 2 and 3 were selected based on the RF algorithm, which includes distinct subsets of predictor

**Table 2** Predictors and outcome variable

| Feature type | Variable name | Description | Variable type |
|---|---|---|---|
| Calendar data | Day of the week | Day of week in which the patient arrived at the ED. The categorical variable was deployed into seven dummy variables, each indicating a day of the week | Binary |
| | Month of the year | Month of the year in which the patient arrived at the ED. The categorical variable was deployed into twelve dummy variables, each indicating a month of the year | Binary |
| Meteorological data | Minimum daily temperature | Minimum temperature (in Celsius) of the arrival day | Continuous |
| | Mean daily temperature | Mean temperature (in Celsius) of the arrival day | Continuous |
| | Maximum daily temperature | Maximum temperature (in Celsius) of the arrival day | Continuous |
| Daily patient arrivals at the ED | Arrivals | Total number of daily patient arrivals at the ED | Discrete |

**Table 3** Feature engineered predictors and their description

| Calendar feature name | Description | Calendar feature type |
|---|---|---|
| date_index.num | Time is converted into a numerical value in seconds from a fixed base date set at 2014–01-01 00:00:00 = 0, where 2014–01–02 corresponds to 86,400 s, 2014–01–03 = 172,800, and so on. The variable represents the number of seconds elapsed from 2014–01-01 to 2016–01–31 | Numeric |
| date_half | The variable indicates whether the date falls in the first or second half of the year (e.g., 2014–01-01 = 1, 2014–07-01 = 2) | Categorical [1 or 2] |
| date_quarter | It represents the quarterly component of the index. The year is divided into four quarters, each including three consecutive months. The variable indicates to which quarter of the year a specific date belongs (e.g., January 15 = 1; April 28 = 2), enabling data analysis based on quarterly patterns or trends | Categorical [1 to 4] |
| date_mday | The variable indicates the day of the month associated with a particular date (e.g., January 15 = 15) | Categorical [1 to 31] |
| date_qday | The variable represents the day of the quarter, ranging from 1 to 92 for a given date, with each quarter including from 90 to 92 days (e.g., June 30 = 91, as it is the 91st day of the second quarter; September 30 = 92, as it is the 92nd day of the third quarter) | Categorical [1 to 92] |
| date_yday | The variable represents the day of the year, ranging from 1 to 365, for a given date (e.g., March 31 is the 90th day of the year), enabling analysis and grouping of data based on annual patterns or trends | Categorical [1 to 365] |
| date_mweek | The variable represents the week of the month, ranging from 1 to 5, for a given date (e.g., January 7 = 1; January 15 = 3) | Categorical [1 to 5] |
| date_week | The variable represents the week number of the year (considering the first week starts on the first Sunday). Thus, in a year where January 1st falls on a Tuesday, this week is designated as week 53 of the previous year. Week 1, in turn, begins on January 6th | Categorical [1 to 53] |
| date_week2 | The variable is a binary indicator representing the biweekly frequency module. The term "module" refers to the number that represents two possible states in each two-week cycle, taking values of 1 or 0 (e.g., January 7 = 1; January 14 = 0; January 15 = 1) | Binary [1 or 0] |
| date_week3 | The variable represents the three-week frequency module. The variable can take on values of 1, 2, or 0 (e.g., January 7 = 1; January 14 = 2; January 15 = 0; and January 22 = 1) | Categorical [0 to 2] |
| date_week4 | The variable represents the quadriweekly frequency module, with values ranging from 0 to 3 (e.g., January 7 = 1; January 14 = 2; January 15 = 3; January 22 = 0; and January 29 = 1) | Categorical [0 to 3] |
| date_mday7 | The variable is used to order each weekday occurrence within a month, starting from 1 (e.g., the first Saturday of the month will have mday7 = 1, the second mday7 = 2, and so on; the same applies to other weekdays) | Categorical [1 to 5] |

variables for each ED. By conducting variable selection, we can analyze the effect of FE on the six predictive models. The selected subsets of predictors give the most important predictors within each dataset.

Thirty-four candidate predictor variables were used, including FE variables. A variable selection step based on RF was performed to identify the most important variables, considering that some of the FE variables may not result significant in describing the dependent variable (number of patients).

### Proposed method

Figure 1 displays the flowchart of the proposed method, with steps detailed in the following subsections.

#### *Step 1*

In the first step, the 11 datasets are divided into training and testing sets, considering the temporal structure of each dataset. Two test sets are generated for each dataset, containing 7 and 45 observations in each cross-validation fold. The test sets are created in a sliding manner, thus respecting the temporal order of the 5 folds of time-series split cross-validation (TSCV) used in all ML algorithms. This split of training and testing is repeated several times for the 7 and 45-day test sets, corresponding to approximately 97%/3% and 80%/20% splits. The 7-day test horizon was chosen as it has been widely used in previous studies (see Table 1), facilitating performance comparison with existing research in the field. The 45-day horizon corresponds to an 80% training and 20% testing data split, as recommended by Hyndman and Athanasopoulos [63].

#### *Step 2*

In step 2, prior to training the ML algorithms, the datasets are first pre-processed following the stages below:

Stage 1: Feature engineering [30, 64] involves the creation of additional variables related to the calendar and derived from the patient's visit date to the ED, corresponding to the time series signature [65]. Feature-engineered variables were created with the aim of enhancing the performance of ML algorithms, as recommended by Verdonck et al. [30]. The complete set of predictors considered in the analysis consists of the feature-engineered variables (Table 3) and the original variables (Table 2).

The FE variables were created using the timetk package [65] in R, utilizing the step_timeseries_signature() function. The function automatically generates a set of variables based on the information in the date column regarding the number of patients arrivals, and is used to create a "time series signature," which decomposes a temporal variable (such as dates or timestamps) into several derived variables that represent different components of the date.

A step-by-step description of the function's process is as follows: (i) A temporal data column is provided as input to the function, which can be in either Date or POSIXct format (date and time). The column must be specified within the data frame to which the transformation will be applied. (ii) the function decomposes the time variable into several derived variables, including various components of the timestamp, such as year, month, day of the month (mday), day of the year (yday), week of the year (week), quarter, and others. (iii) the variables generated are then added to the data frame as additional columns, each representing a specific temporal feature. This allows the original data frame to contain all the derived time characteristics, which can then be used in ML models. For more details, readers can refer to [66, 67].

Stage 2: Min–max normalization of continuous predictors. The stability and prediction performance of ML algorithms depends on the quality of input data [68]. Observations of all continuous predictor variables (except for dummy variables) were rewritten in the [0,1] interval to eliminate scale effects, using the expression:

$$x_{norm} = (x - x_{min})/(x_{max} - x_{min}) \quad (1)$$

where $x$ and $x_{norm}$ are the observed and normalized observations, and $x_{max}$ and $x_{min}$ are the variable's maximum and minimum observed values.

Stage 3: Variable selection plays a key role in ML workflows, contributing to faster training, increased accuracy, and easier analysis of the modeled phenomenon's mechanisms [69]. The use of ML algorithms is often criticized due to the difficulty in visualizing the impact of predictors on the outcome of interest [70], which is intensified in the case of a large number of predictors. To assess the importance of variables, a variable selection procedure based on an RF was adopted [71].

The complete set of predictors from the 11 analyzed datasets was submitted to the RF variable selection method. Two approaches are commonly used to assess the importance of a variable in the method: permutation importance and impurity-based importance [72]. Our study adopted the permutation importance approach, and the selection was conducted in four steps: (i) train an RF model using the complete dataset; (ii) employ the

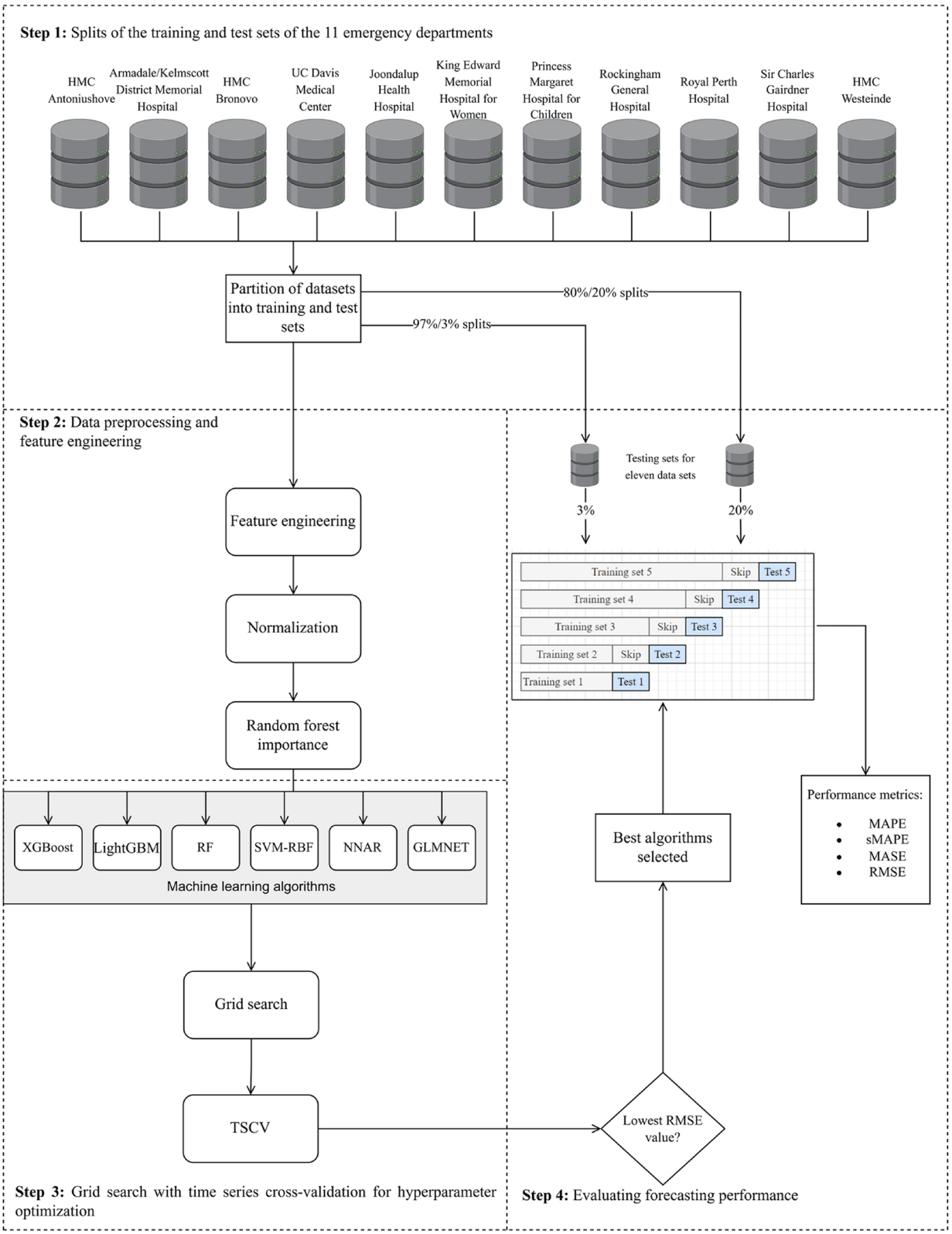


**Fig. 1** Flowchart of methodological steps

permutation importance method to calculate importance scores for each variable; (iii) rank variables based on their importance scores and organize them in descending order of importance; and (iv) establish a cutoff threshold for selecting a subset of top-performing predictor variables to be used as a reduced dataset in subsequent analyses.

The permutation importance method begins by creating a prediction model using the complete dataset and recording its accuracy. Subsequently, one of the variables in the dataset is chosen, and its values are randomly shuffled while the other variables remain unchanged. This process eliminates any existing relationship between the dependent variable and the shuffled variable, and that will be reflected in the model accuracy if the original relationship is significant [73]. The next step involves recording the difference in accuracy between the initial model and the model with the shuffled variable, which becomes the variable's importance score. The larger the score, the more important the variable is in the prediction [74]. After shuffling for all 34 predictor variables in the datasets, the obtained scores are arranged in descending order, and only variables with scores above the threshold value are retained in the reduced datasets.

The RF variable selection method was chosen for two reasons: (i) Bommert et al. [72] conducted an analysis of 22 variable selection methods on 16 high-dimensional datasets across various domains, concluding that RF displayed the highest accuracy compared to other methods; and (ii) the technique has not been adopted in previous studies on predicting patient arrivals in EDs. It is important to note that the GLMNET regressor has a variable selection step imbedded in the model, which was deactivated. Thus, RF variable selection was employed in all algorithms, enabling a direct comparison of their performances.

#### *Step 3*

Once the ideal subset of input variables is defined, it is necessary to refine ML methods by optimizing their hyperparameters. To achieve that, a grid search [31] was employed. The method involves defining ranges of candidate values for the tuning parameters [75] and evaluating combinations of values that result in models that better fit the data. RMSE is the most commonly used metric for this purpose [31, 75]; see Eq. (5).

In our application, we conducted a grid search with fivefold cross-validation for hyperparameter optimization of all tested algorithms, as recommended by Kuhn and Johnson [31]. We used the "tune" package [76] in conjunction with the tidymodels framework [77] to compute the best parameter combinations across 25 candidate models for each ML algorithm. The optimal tuning parameters yielded the lowest RMSE, calculated using the training portion of the datasets. Extensive reports of the grid search are available upon request from the authors.

#### *Step 4*

According to Hyndman and Athanasopoulos [63] and Kuhn and Johnson [64], the most suitable method for assessing the performance of modeling datasets with temporal dependence is the TSCV. This type of validation captures the effects of trends, seasonality, and other aspects that may be present in time series [63, 64]. Unlike classical cross-validation techniques such as *k*-fold, which assume independence and identical distribution of observations, TSCV avoids random splitting between training and test sets, respecting the temporal sequence of the data [63, 64]. TSCV involves the following stages:

1. Split the Dataset: The data is initially divided into training and test subsets, considering their temporal sequence. Older observations are used for training, while more recent observations are allocated for testing [63, 64].
2. Define a Moving Window: A fixed-size moving window is defined to create subsequent folds; it dictates the extent of the training and test data in each iteration [63]. In our study, 7 and 45-day windows were tested.
3. Iterate: At each iteration, the moving window is shifted forward along the time series, i.e., the second resampling uses the test set from the initial split (referred to as skip 1) as part of the training set. Consequently, the size of the training sets in each fold is not the same since they grow cumulatively as the moving window progresses [64].
4. Evaluate performance: After five iterations, performance is assessed by calculating the average of the error metric values for the five test sets.

Figure 2 illustrates the stages above for the 45-day moving window. Four recent studies ( [20, 40, 50, 78]) also used TSCV when predicting patient arrivals.

### Overview of selected machine learning algorithms and performance metrics

In this section, we justify the choice of ML algorithms tested in our comparative analysis. In Appendix Table A2, algorithms are grouped by similarity and summarized, with primary references provided for readers to access detailed information.

XGBoost is a Gradient Boosting (GB) algorithm that builds trees iteratively to predict residuals and combines them for final predictions [79]. It efficiently handles large

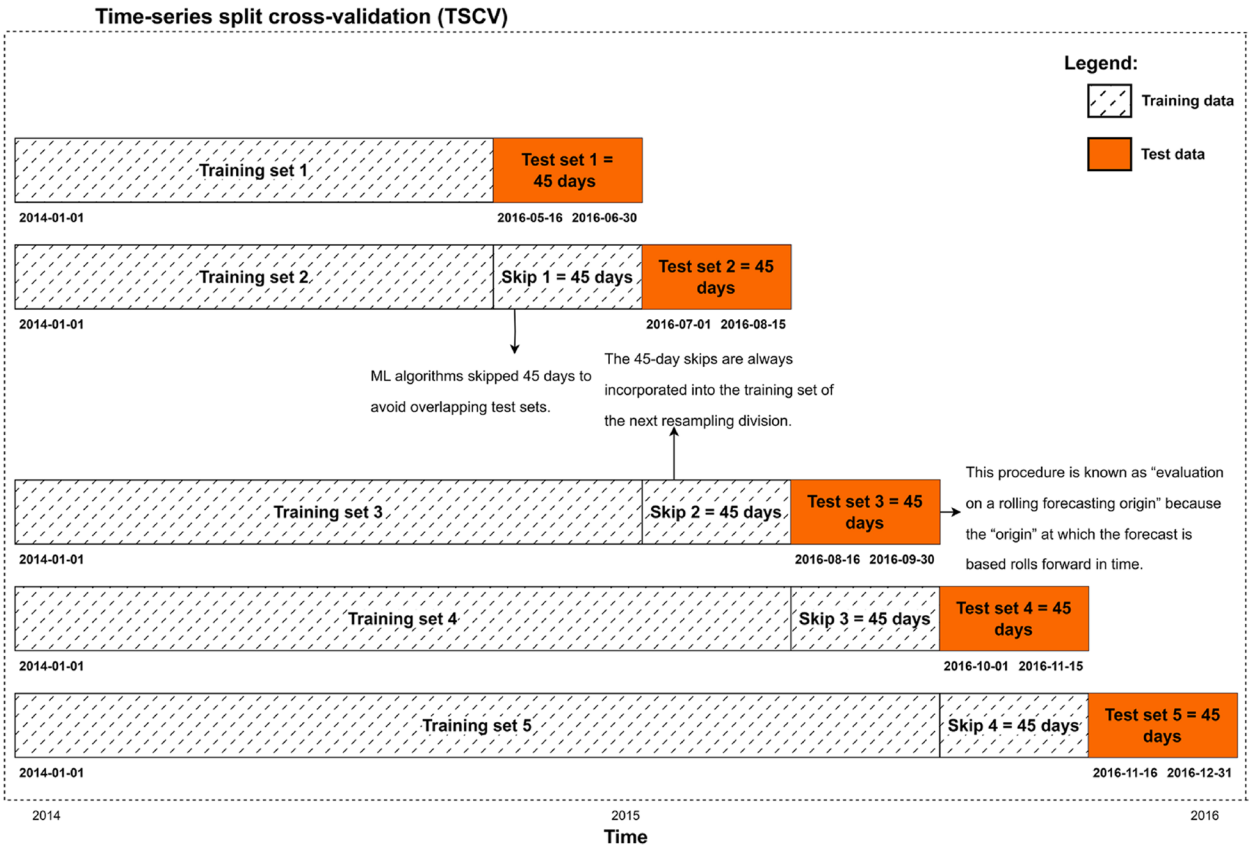


**Fig. 2** Example of a TSCV implementation on a generic dataset considering a 45-day moving window

datasets by adding weak learners and transforming them into a strong model [79]. LightGBM is a GB variant that uses a histogram-based and leaf-wise approach to optimize tree construction [80]. LightGBM speeds up training and reduces memory usage with techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). RF is an ensemble method that builds multiple decision trees on random data subsets, reducing overfitting and increasing accuracy. SVM-RBF solves regression problems by using a kernel function to maximize margins and minimize prediction errors [58]. The RBF kernel measures similarity between instances, and key parameters such as gamma and $C$ are optimized to improve predictive performance [58]. NNAR is a neural network model that uses lagged values of a time series as inputs, forming a model analogous to ARIMA [63]. NNAR has input, hidden, and output layers, with the number of hidden nodes determined by lag order, and parameters optimized [63]. GLMNET combines linear regression with regularization (LASSO and Ridge) to handle high-dimensional data [81]. It uses the elastic-net penalty ($\alpha$) and regularization parameter ($\lambda$) to balance variable selection and shrinkage, with optimal values determined via cross-validation [81].

The selection of these algorithms was motivated by the following reasons: (i) XGBoost [40, 55], RF [7, 20], and GLMNET [20] are reported as presenting superior performance in patient arrival prediction; (ii) to the best of our knowledge, LightGBM, SVM-RBF, and NNAR are being used for the first time in such context; (iii) XGBoost, LightGBM, and GLMNET are the fastest ML methods in terms of execution speed and computational efficiency [79–81]. Given that predictions were made on 11 datasets, the computational time of the algorithms becomes a crucial factor; and (iv) the use of the three decision tree-based algorithms (XGBoost, LightGBM, and RF) in the same study is justified by differences in their tree construction strategies (e.g., XGBoost adopts a level-wise strategy while LightGBM uses a leaf-wise strategy [80]) and distinct approaches to handling overfitting (e.g., XGBoost employs Gradient Boosting to mitigate overfitting, whereas LightGBM utilizes GOSS to address the issue), as explained in Table A2.

The performance of the models presented in Table A3 in predicting ED arrivals in horizons of 7 and 45 days

was determined by analyzing four metrics: (i) MAPE in Eq. (2); (ii) Symmetric Mean Absolute Percentage Error (sMAPE) in Eq. (3); (iii) Mean Absolute Scaled Error (MASE) in Eq. (4); and (iv) RMSE in Eq. (5).

In Sect. " Background", we demonstrated that the most commonly used metrics for evaluating patient arrival predictions are MAPE and RMSE. Both metrics enable direct comparison since they are scale-independent, allowing for the assessment of predictions across different scenarios [36, 78]. MAPE presents results in percentage form, which is more easily interpretable [59]. sMAPE also provides results in percentage form, being considered a suitable choice in time series with zero entries [82], which can occur when there are no patient arrivals on certain days. According to Hyndman and Koehler [82], MASE is a superior metric for assessing forecasting model accuracy since it is less sensitive to outliers, less variable in small samples, scale-independent, easy to interpret, and can be used to compare the accuracy of various time series [82]. MASE values smaller than 1 (the closer to 0, the better) indicate that the model's forecast is better than the Naive model [82]. RMSE is a measure that represents the squared differences between predicted and actual values. A RMSE value close to zero indicates a better model fit to the data, as it signifies smaller differences between predicted and actual values.

$$MAPE = T^{-1}\sum_{t=1}^{T}\left|\frac{Y_{(t)}-\hat{Y}_{(t)}}{Y_{(t)}}\right| \tag{2}$$

$$sMAPE = \frac{100}{N}\sum_{t=1}^{T}\frac{\left|\widehat{Y}_{(t)}Y_{(t)}\right|}{\left(\left|Y_{(t)}\right|+\left|\widehat{Y}_{(t)}\right|\right)/2} \tag{3}$$

$$MASE = \frac{MAE}{\frac{1}{T-1}\sum_{t=2}^{T}\left|Y_{(t)-1}\right|} \tag{4}$$

$$RMSE = \left(T^{-1}\sum_{t=1}^{T}\left(Y_{(t)}-\widehat{Y}_{(t)}\right)^2\right)^{\frac{1}{2}} \tag{5}$$

In all equations above, $Y_{(t)}$ denotes the observed value of the series at time $t$, $\widehat{Y}_{(t)}$ denotes the predicted value of the series at time $t$, and $T$ is the total number of observations in the time series.

## Results

### Exploratory data analysis

Figure 3 displays the time series of daily patient arrivals in the analyzed datasets. Time series graphs allow

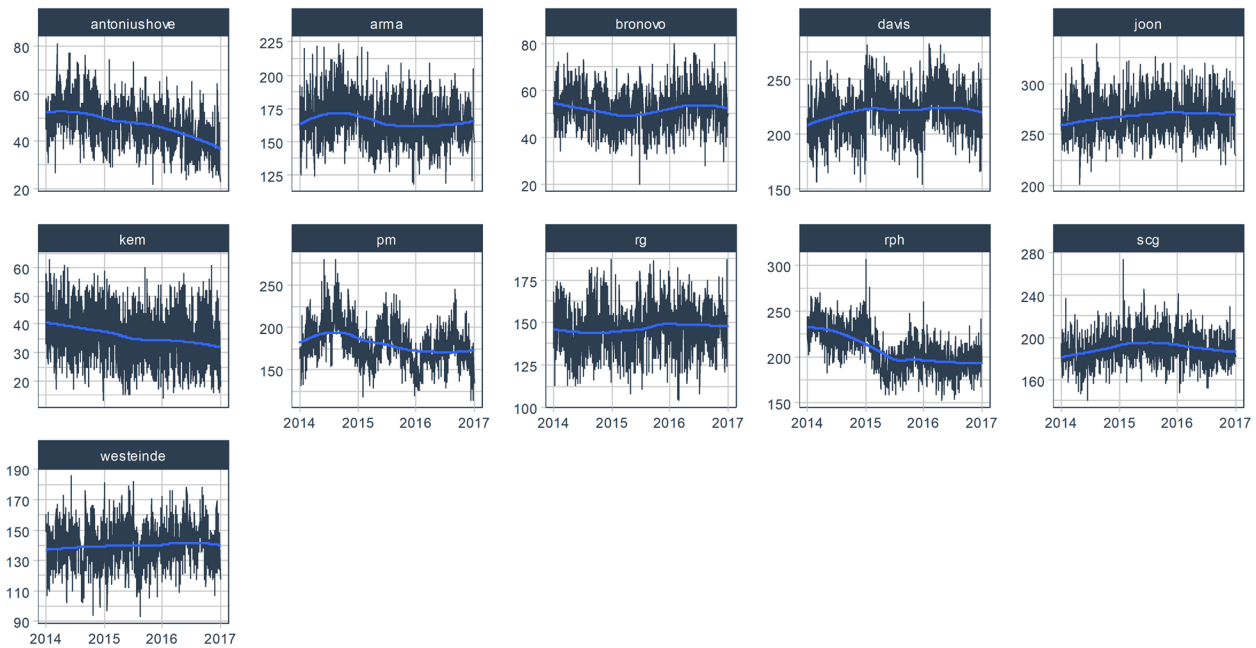


**Fig. 3** Series of daily arrivals to EDs for selected hospitals in Australia, USA and the Netherlands

observing data behavior over time, as indicated by [63]. Figure 3 reveals distinct arrival patterns and fluctuations. Most ED series exhibit periodic variations, indicating seasonality, high pointwise variability over the years, and nonlinear trends. Such characteristics justify the choice of ML models to describe patients' arrival patterns, as they are more suitable to capture complex non-linear data behaviors [7].

Fig. 4 displays box plots stratifying patients' daily arrivals by day of the week (top) and month of the year (bottom) for each ED dataset. Mondays and Sundays are the busiest days, totaling 363,874 (15.02%) and 360,578 (14.88%) of the arrivals, respectively. In opposition, Wednesdays had the lowest arrival numbers (333,242 or 13.75%). Mondays displayed significant variation in arrivals across EDs, consistent with previous findings reported by [16, 20] and [21]. In general, patient arrivals peaked on Mondays, decreasing through the week, reaching the lowest point on Wednesdays, and starting to increase again on Fridays, reaching another peak on Sundays. Prior studies have also noted increased arrivals on Fridays [3, 6, 15]. Monthly analysis indicates August as the busiest month (213,936 arrivals or 8.83%), followed by March (211,986 arrivals or 8.75%); in opposition, November had the lowest number of arrivals (181,020 or 7.47%). The strong seasonal pattern identified in the datasets across days of the week and months of the year indicates that the FE approach for extracting temporal features was appropriate for the analysis.

#### Forecasting performance

To assess performance, metrics were computed using eqns. (2) to (5) in each cross-validation resampling set. The average performance was analyzed across two test sets, with durations of 7 and 45 days. In Table 4 present the performance results for the algorithms tested. Results demonstrated that the application of FE contributed to enhancing the algorithms' performance.

XGBoost achieved the best performance in five out of the eleven analyzed datasets, displaying MAPE values ranging from 5.08% to 21.37%, sMAPE values ranging from 4.96% to 7.22% and RMSE values from 7.03 to 24.14 for a 7-day prediction horizon. XGBoost used different combinations of ten variables for each dataset, including index.num, yday, week, half, quarter, mday, qday, minimum, mean, and maximum temperatures, day of the week (Monday, Tuesday, Friday, Saturday, and Sunday), and month of the year (August).

For the 45-day test sets, XGBoost and NNAR displayed the best performance across ten of the datasets, with MAPE values ranging from 5.08% to 19.41%, sMAPE values ranging from 5.11% to 6.16%, and RMSE from 7.89 to 25.14. The variables used consisted of different combinations of the following predictors: index.num, yday, week, half, quarter, mday, qday, minimum, mean, and maximum temperatures, day of the week (Sunday, Monday, Wednesday, Thursday, and Friday), and month of the year (August and November). Table A3 allows visualizing algorithms that exhibited superior performance across all datasets and the most important predictors used in the forecasts.

Fig. 5 illustrates the comparisons of the fivefold cross-validation predictions and the fivefold average, represented by the vertical line in the graphs, for six algorithms across two horizons.

#### Feature importance

Since the literature does not indicate a universal cutoff threshold value for feature selection, we tested different values (0.60, 0.70, 0.80, 0.90) across all datasets. The final choice of threshold considered two main criteria: (*i*) finding the optimal trade-off between quantity of variables, model simplification, and model performance; and (*ii*) the authors' expertise in modeling medical datasets.

During the testing phase, adopting a threshold value of 0.60 resulted in retaining a large number of predictor variables with importance scores below 10%, while the threshold of 0.90 retained only four candidate variables across all datasets, eliminating several with high importance scores and undermining model performance. A final threshold value of 0.70 was chosen as it provided the best balance between the number of variables retained and model performance.

Considering the different sets of variables retained in each dataset and the varying importance weights assigned to them by the ML prediction algorithms, we conclude that the most important variables in the analyzed datasets were index.num, yday, week, qday, quarter, minimum, mean, and maximum temperatures, and dummy variables representing the day of the week (mainly Monday, Wednesday, Friday, and Sunday). The results of the permutation importance scores via RF for the top ML algorithms can be viewed in Fig. 6. Table A4 summarizes the top 10 most important predictors across all datasets for the best ML algorithms in each dataset. The FE variables index.num, yday, and week displayed higher importance scores in the ML models for the majority of datasets analyzed. This result demonstrates that the FE approach can enhance the performance of ML in predicting daily patient arrivals in EDs.

## Discussion

This study introduces some methodological innovations for predicting patient arrivals using ML algorithms, addressing certain limitations reported in the literature. First, we adopted a grid-search with fivefold

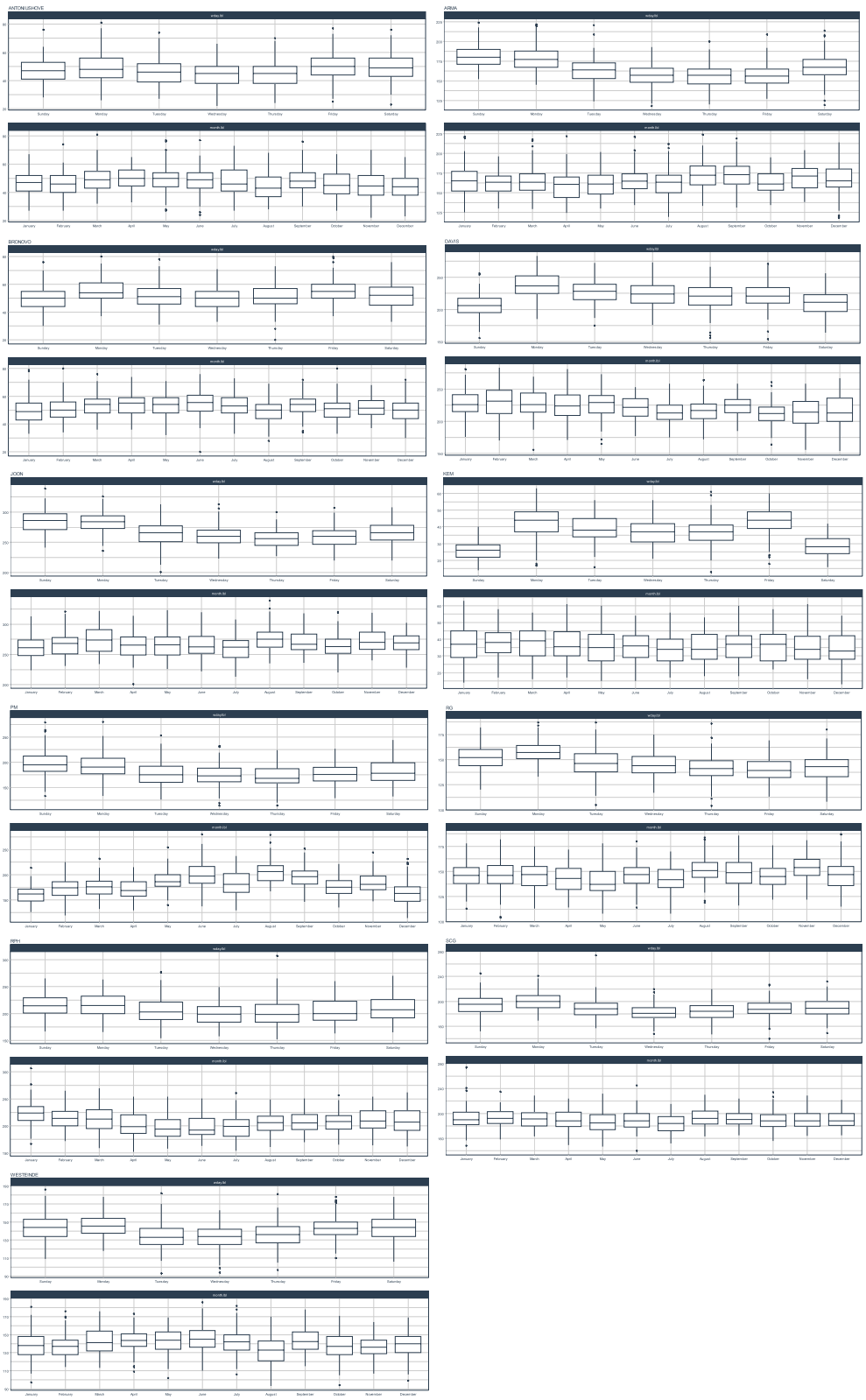


**Fig. 4** Box plots of daily arrival volumes by EDs during the days of the week, stratified by month of the year

**Table 4** Results of average performance metrics using grid-search with fivefold cross-validation

| **ED ANTONIUSH-OVE** | | | | | | **ED ARMA** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** | **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** |
| **7-day test set** | | | | | | | | | | | |
| XGBoost | Mean | 19.15 | 18.45 | 0.89 | 8.13 | XGBoost | Mean | 5.74 | 5.61 | 0.80 | 11.72 |
| LightGBM | | 22.89 | 20.21 | 1.06 | 9.29 | LightGBM | | 5.90 | 5.78 | 0.85 | 11.54 |
| RF | | 23.76 | 20.42 | 1.11 | 9.68 | RF | | 5.84 | 5.70 | 0.82 | 11.96 |
| SVM-RBF | | 20.59 | 18.49 | 0.93 | 8.45 | SVM-RBF | | 5.52 | 5.40 | 0.78 | 11.69 |
| NNAR | | 19.45 | 17.87 | 0.90 | 8.10 | NNAR | | 6.01 | 5.91 | 0.85 | 12.05 |
| GLMNET | | 20.18 | 18.15 | 0.93 | 8.26 | GLMNET | | 5.48 | 5.35 | 0.77 | 11.44 |
| **45-day test set** | | | | | | | | | | | |
| XGBoost | Mean | 16.29 | 15.95 | 0.79 | 7.89 | XGBoost | Mean | 5.90 | 5.89 | 0.73 | 12.62 |
| LightGBM | | 17.40 | 15.92 | 0.79 | 7.95 | LightGBM | | 6.05 | 6.00 | 0.74 | 12.74 |
| RF | | 21.04 | 18.33 | 0.93 | 9.06 | RF | | 6.08 | 6.04 | 0.74 | 12.60 |
| SVM-RBF | | 18.35 | 16.49 | 0.82 | 8.26 | SVM-RBF | | 5.97 | 5.94 | 0.73 | 12.70 |
| NNAR | | 17.83 | 16.22 | 0.81 | 8.04 | NNAR | | 6.25 | 6.18 | 0.76 | 13.00 |
| GLMNET | | 18.47 | 16.56 | 0.83 | 8.24 | GLMNET | | 5.90 | 5.90 | 0.73 | 12.64 |
| **ED BRONOVO** | | | | | | **ED DAVIS** | | | | | |
| **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** | **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** |
| **7-day test set** | | | | | | | | | | | |
| XGBoost | Mean | 14.10 | 13.37 | 0.63 | 8.49 | XGBoost | Mean | 9.90 | 9.90 | 0.90 | 24.14 |
| LightGBM | | 15.50 | 14.40 | 0.69 | 8.91 | LightGBM | | 10.91 | 10.75 | 0.96 | 27.05 |
| RF | | 15.19 | 14.71 | 0.69 | 8.97 | RF | | 10.99 | 11.17 | 0.98 | 27.87 |
| SVM-RBF | | 14.01 | 13.57 | 0.63 | 8.53 | SVM-RBF | | 10.31 | 10.17 | 0.90 | 25.36 |
| NNAR | | 13.64 | 13.11 | 0.61 | 8.28 | NNAR | | 10.12 | 10.09 | 0.90 | 25.31 |
| GLMNET | | 14.10 | 13.64 | 0.63 | 8.38 | GLMNET | | 10.46 | 10.30 | 0.91 | 25.79 |
| **45-day test set** | | | | | | | | | | | |
| XGBoost | Mean | 13.22 | 12.74 | 0.76 | 8.54 | XGBoost | Mean | 10.99 | 11.06 | 0.96 | 26.94 |
| LightGBM | | 13.25 | 12.76 | 0.76 | 8.57 | LightGBM | | 10.88 | 10.76 | 0.93 | 26.24 |
| RF | | 13.15 | 12.81 | 0.76 | 8.60 | RF | | 11.10 | 11.25 | 0.98 | 27.81 |
| SVM-RBF | | 12.75 | 12.68 | 0.75 | 8.56 | SVM-RBF | | 10.51 | 10.37 | 0.91 | 25.77 |
| NNAR | | 12.85 | 12.45 | 0.74 | 8.37 | NNAR | | 10.29 | 10.23 | 0.90 | 25.14 |
| GLMNET | | 12.61 | 12.58 | 0.75 | 8.56 | GLMNET | | 10.57 | 10.41 | 0.92 | 26.01 |
| **ED JOON** | | | | | | **ED KEM** | | | | | |
| **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** | **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** |
| **7-day test set** | | | | | | | | | | | |
| XGBoost | Mean | 5.03 | 4.96 | 0.93 | 16.50 | XGBoost | Mean | 21.37 | 19.60 | 0.50 | 7.03 |
| LightGBM | | 4.73 | 4.66 | 0.88 | 15.96 | LightGBM | | 23.39 | 20.10 | 0.52 | 7.59 |
| RF | | 4.84 | 4.80 | 0.90 | 16.84 | RF | | 24.41 | 20.70 | 0.52 | 7.62 |
| SVM-RBF | | 4.61 | 4.52 | 0.85 | 16.22 | SVM-RBF | | 22.87 | 20.32 | 0.52 | 7.27 |
| NNAR | | 5.09 | 5.02 | 0.94 | 16.68 | NNAR | | 21.62 | 18.98 | 0.48 | 7.10 |
| GLMNET | | 4.91 | 4.76 | 0.88 | 16.85 | GLMNET | | 23.21 | 20.00 | 0.51 | 7.53 |
| **45-day test set** | | | | | | | | | | | |
| XGBoost | Mean | 5.11 | 5.11 | 0.78 | 17.49 | XGBoost | Mean | 19.41 | 17.68 | 0.51 | 7.15 |
| LightGBM | | 5.08 | 5.06 | 0.77 | 17.45 | LightGBM | | 20.77 | 18.65 | 0.54 | 7.37 |
| RF | | 5.01 | 4.98 | 0.76 | 17.43 | RF | | 20.20 | 18.43 | 0.53 | 7.40 |
| SVM-RBF | | 4.98 | 4.97 | 0.76 | 17.25 | SVM-RBF | | 19.15 | 18.02 | 0.52 | 7.18 |
| NNAR | | 5.13 | 5.04 | 0.77 | 17.67 | NNAR | | 18.92 | 17.42 | 0.50 | 6.92 |
| GLMNET | | 5.36 | 5.27 | 0.80 | 18.10 | GLMNET | | 19.47 | 18.21 | 0.53 | 7.37 |
| **ED PM** | | | | | | **ED RG** | | | | | |
| **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** | **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** |

**Table 4** (continued)

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **7-day test set** | | | | | | | | | | | | |
| XGBoost | Mean | 9.72 | 9.54 | 0.71 | 18.61 | | XGBoost | Mean | 7.21 | 7.22 | 0.81 | 12.96 |
| LightGBM | | 10.50 | 10.17 | 0.76 | 20.01 | | LightGBM | | 6.81 | 6.79 | 0.74 | 12.87 |
| RF | | 9.89 | 9.68 | 0.71 | 18.66 | | RF | | 6.55 | 6.51 | 0.71 | 11.93 |
| SVM-RBF | | 11.22 | 10.44 | 0.77 | 20.77 | | SVM-RBF | | 7.37 | 7.35 | 0.81 | 13.67 |
| NNAR | | 9.58 | 9.20 | 0.67 | 18.32 | | NNAR | | 7.55 | 7.44 | 0.83 | 13.66 |
| GLMNET | | 11.94 | 11.01 | 0.81 | 22.01 | | GLMNET | | 7.83 | 7.65 | 0.86 | 14.05 |
| **45-day test set** | | | | | | | | | | | | |
| XGBoost | Mean | 8.67 | 8.58 | 0.87 | 18.68 | | XGBoost | Mean | 6.23 | 6.16 | 0.69 | 11.42 |
| LightGBM | | 8.85 | 8.63 | 0.87 | 18.85 | | LightGBM | | 6.31 | 6.17 | 0.69 | 11.54 |
| RF | | 8.83 | 8.63 | 0.87 | 18.83 | | RF | | 6.28 | 6.15 | 0.69 | 11.47 |
| SVM-RBF | | 9.08 | 8.91 | 0.90 | 19.64 | | SVM-RBF | | 6.77 | 6.65 | 0.75 | 12.07 |
| NNAR | | 9.02 | 8.96 | 0.91 | 19.85 | | NNAR | | 6.77 | 6.64 | 0.75 | 12.15 |
| GLMNET | | 9.73 | 9.54 | 0.96 | 21.06 | | GLMNET | | 6.72 | 6.63 | 0.75 | 12.03 |
| **ED RPH** | | | | | | **ED SCG** | | | | | | |
| **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** | | **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** |
| **7-day test set** | | | | | | | | | | | | |
| XGBoost | Mean | 5.90 | 5.96 | 0.94 | 14.89 | | XGBoost | Mean | 5.08 | 5.08 | 0.85 | 11.93 |
| LightGBM | | 6.24 | 6.37 | 0.99 | 15.75 | | LightGBM | | 5.35 | 5.33 | 0.94 | 12.17 |
| RF | | 6.54 | 6.49 | 1.03 | 15.72 | | RF | | 5.25 | 5.24 | 0.90 | 11.96 |
| SVM-RBF | | 6.21 | 6.28 | 0.98 | 15.45 | | SVM-RBF | | 5.22 | 5.23 | 0.87 | 12.12 |
| NNAR | | 6.37 | 6.43 | 1.00 | 15.63 | | NNAR | | 5.21 | 5.19 | 0.90 | 11.73 |
| GLMNET | | 6.79 | 6.97 | 1.07 | 16.86 | | GLMNET | | 5.85 | 5.72 | 0.99 | 12.75 |
| **45-day test set** | | | | | | | | | | | | |
| XGBoost | Mean | 6.04 | 5.96 | 0.84 | 14.57 | | XGBoost | Mean | 5.64 | 5.60 | 0.82 | 12.61 |
| LightGBM | | 6.06 | 5.98 | 0.84 | 14.67 | | LightGBM | | 5.72 | 5.68 | 0.82 | 12.71 |
| RF | | 6.01 | 5.96 | 0.84 | 14.44 | | RF | | 5.69 | 5.60 | 0.81 | 12.56 |
| SVM-RBF | | 5.71 | 5.74 | 0.81 | 14.37 | | SVM-RBF | | 5.69 | 5.60 | 0.81 | 12.56 |
| NNAR | | 5.85 | 5.86 | 0.83 | 14.25 | | NNAR | | 6.06 | 5.89 | 0.85 | 13.37 |
| GLMNET | | 6.63 | 6.85 | 0.98 | 16.61 | | GLMNET | | 6.35 | 6.13 | 0.88 | 13.71 |
| **ED WESTEINDE** | | | | | | | | | | | | |
| **Methods** | **Resampling** | **MAPE** | **sMAPE** | **MASE** | **RMSE** | | | | | | | |
| **7-day test set** | | | | | | | | | | | | |
| XGBoost | Mean | 8.49 | 8.56 | 0.79 | 14.01 | | | | | | | |
| LightGBM | | 8.47 | 8.43 | 0.78 | 14.00 | | | | | | | |
| RF | | 8.74 | 8.75 | 0.81 | 14.18 | | | | | | | |
| SVM-RBF | | 8.18 | 8.20 | 0.76 | 13.71 | | | | | | | |
| NNAR | | 8.21 | 8.29 | 0.77 | 13.88 | | | | | | | |
| GLMNET | | 8.27 | 8.39 | 0.78 | 14.01 | | | | | | | |
| **45-day test set** | | | | | | | | | | | | |
| XGBoost | Mean | 7.12 | 7.15 | 0.81 | 12.57 | | | | | | | |
| LightGBM | | 7.30 | 7.29 | 0.83 | 12.77 | | | | | | | |
| RF | | 7.62 | 7.64 | 0.87 | 13.27 | | | | | | | |
| SVM-RBF | | 7.44 | 7.44 | 0.85 | 12.94 | | | | | | | |
| NNAR | | 7.38 | 7.31 | 0.83 | 13.09 | | | | | | | |
| GLMNET | | 7.55 | 7.53 | 0.86 | 13.16 | | | | | | | |

cross-validation for hyperparameter optimization across all algorithms. Overfitting of the ML algorithms was avoided using the TSCV method, and the results obtained are generalizable because we used datasets from eleven EDs in different countries. As of the literature review, none of the studies listed in Table 1 had


**Fig. 5** Illustrates the comparison of the fivefold cross-validation prediction and the fivefold average represented by the vertical line in the graphs for six algorithms at two horizons. Note: The graphs present the performance of resampling predictions using a variable selection step for the 7-day and 45-day test sets in each ED

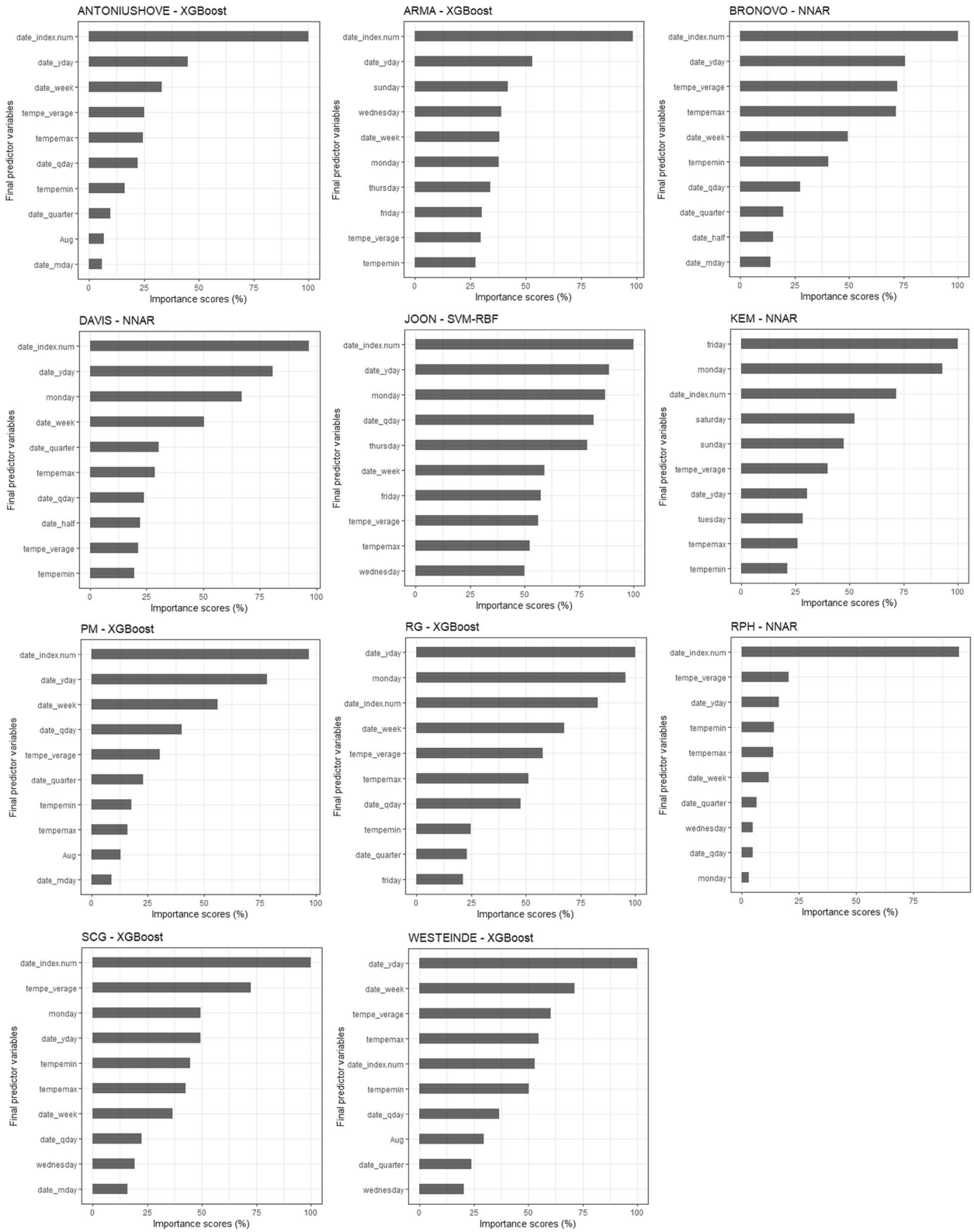


**Fig. 6** Features selection using a RF permutation importance score. Notes: The figure displays bar charts of the top 10 predictors for patient arrivals, using the 11 datasets for the best models of ML

used grid-search for hyperparameter tuning. Second, we adopted an FE approach to improve algorithm performance, which is new in the ED patient arrival prediction literature. Third, we incorporated a feature selection step based on RFs. These methodological innovations enabled more precise and reliable results in patient arrival prediction. Our approach stands out compared to similar works in terms of prediction performance.

Sudarshan et al. [7] compared three ML algorithms in forecasting daily arrivals for a 7-day horizon. The LSTM algorithm, incorporating six meteorological and seven calendar variables, achieved average MAPEs of 9.31% and 8.91%. Xu [18] employed six ML and hybrid methods to predict daily arrivals for a 7-day horizon. The methods incorporated variables such as day of the week, month of the year, holidays, school vacations, and temperatures. The ARIMA-LR with smoothing achieved MAPEs ranging from 6.1% to 12.9% and RMSEs from 5.33 to 147.

Vollmer et al. [20] compared the performance of eight statistical and ML models in forecasting daily arrivals considering 1 to 7-day horizons. They used variables such as day of the week, month of the year, public holidays, school vacations, temperatures, and precipitation. The GLMNET model achieved the best performance, with MAPE values of 6.8% and 8.6%. Yousefi et al. [21], in forecasting daily arrivals using the LSTM model for horizons of 1 to 7 days, incorporated predictors such as football game events, weekends, and holidays, reporting an average MAPE of 5.55%.

Pekel et al. [39] compared three hybrid ML algorithms in forecasting daily arrivals, using variables such as month of the year, day of the week, holidays, and maximum temperature. The PSO-ANN model achieved the best performance, with MAPE values of 6% and RMSE of 53.29. Zhang et al. [41] compared nine ML models in forecasting daily arrivals for a 90-day horizon, incorporating seven calendar variables and eight meteorological variables. The SVR model displayed the best performance, with a MAPE of 8.81% and RMSE of 26.84.

In forecasting daily arrivals for a 53-day horizon, Zhao et al. [49] compared eight ML and statistical methods, testing variables such as day of the week, temperature, and relative humidity. The DLSTM algorithm was the best-performing, with a MAPE of 5.67% and RMSE of 25.29. Petsis [40] produced daily arrival forecasts for 1 and 2 days ahead, incorporating nine calendar and five meteorological variables using XGBoost. The obtained results showed MAPEs of 6.5% and 6.91%, along with RMSE values of 22.96 and 23.9. Finally, Rocha and Rodrigues [5] compared ten ML algorithms and hybrid methods in forecasting daily arrivals using calendar variables such as year, month of the year, day of the week, time of day, and holidays. The RNN-1L model displayed the best performance, with RMSE values ranging from 4.8 to 26 and sMAPE values from 4.3% to 21.3%.

Our study compared six ML algorithms using the FE approach in forecasting daily arrivals using datasets from eleven EDs for horizons of 7 and 45 days, displaying performance improvements compared to previous studies. Table 5 provides a horizon-based comparison between our study and the best-performing methods in the literature.

We incorporated the FE approach into the ML workflow to predict daily patient arrivals in EDs, which we believe represents a significant contribution to the research field. The results obtained across eleven EDs indicate that FE variables were informative in forecasting daily patient arrivals. Studies on ML for predicting patient arrivals typically rely solely on meteorological variables and traditional calendar variables, such as day of the week and month of the year [5, 7, 18, 20, 21, 39–41, 49]. The performance of the prediction algorithms was positively impacted by the inclusion of FE variables, especially index.num, yday, week, and qday. Predicting patient arrivals in EDs is not a new research topic. However, this is the first study that systematically investigates whether FE variables are relevant predictors for daily patient arrival forecasts.

The scope of this study, encompassing 11 different EDs, allowed us to assess the importance of using FE to improve the generalizability of our results. Verdonck et al. [30] recommended the use of FE in ML-based analysis workflows to enhance algorithm performance. The FE approach employed in this study created a set of new predictor variables based on arrival timestamps. A recent systematic review on ED arrival forecasting [36] suggests, for future studies, that exploring new variables with the potential to become significant and reliable predictors is an underexplored area requiring further research. Our study addresses this demand.

Specifically in the EDs of Antoniushove, ARMA, Davis, Joon, PM, RG, RPH, and Westeinde hospitals, it was observed that the FE variables index.num, yday and week displayed high levels of importance (Fig. 6), surpassing those attributed to meteorological variables, which are typically considered informative in predicting patient arrivals in EDs, e.g., [3, 6, 7, 41, 42, 46].

The systematic review by Wangon et al. [37] concluded that calendar variables hold greater importance than meteorological variables in predicting patient arrivals in EDs, aligning with our findings. Another systematic review by Jiang et al. [38] supported this conclusion, indicating that traditional calendar variables are more frequently used compared to other types of predictor variables. Our study also demonstrated the importance of predictors associated with temperatures and days of

**Table 5** Comparison of the forecasting performance achieved with related works stratified by forecasting horizon

| Forecasting Horizon | Reference and year | Method(s) used | MAPE(%) | RMSE |
|---|---|---|---|---|
| 3 up to 7 days ahead | Marcilio et al. [42] 2013 | GLM and GEE | 4,5–9,9 | not applied |
| | Xu [18] 2016 | ARIMA-LR (smoothing), ARIMA-LR and GLM | 6,8–9,6 | 70,5–104 |
| | Calegari et al. [46] 2016 | SARIMA, SS and SMHW | 10,67–12,01 | not applied |
| | Asheim et al. [48] 2019 | Poisson time-series regression model | 31–38 | not applied |
| | Jilani et al. [15] 2019 | NN e FTS | 3.03–7,42 | 6,16–16,55 |
| | Whitt et al. [3] 2019 | SARIMAX | 8,4–10,59 | not applied |
| | Zhang et al. [35] 2019 | ARIMA, SVR and ARIMA-SVR | 7,02–7,36 | 19,20–20,34 |
| | Choudhury and Urena [1] 2020 | ARIMA, HW and NN | not applied | 1,55–27,86 |
| | Yousefi et al. [21] 2020 | LSTM | 5,59–6,31 | not applied |
| | Erkamp et al. [6] 2021 | MLR | 8.68–12.20 | not applied |
| | Rocha and Rodrigues [5] 2021 | RNN, XGBoost and RNN-XGBoost | not applied | 4,7–4,9 |
| | Vollmer et al. [20] 2021 | GLMNET, LM and GBM | 6,7–8,6 | not applied |
| | Sudarshan et al. [7] 2021 | RF, LSTM and CNN | 8,91–10,69 | not applied |
| | Cheng et al. [61] 2021 | SARIMAX, HW and VAR | 5–15,3 | not applied |
| | Murtas et al. [83] 2022 | ARIMA | 6,6–11,2 | not applied |
| | Petsis et al. [40] 2022 | XGBoost | 6,5–6,91 | 22,96–23,9 |
| | Tuominen et al. [54] 2022 | ARIMAX, RLS-FS and RLS-SA | 6,6–6,9 | not applied |
| | Tello et al. [57] 2022 | ARIMA and SVR | 3,34–5,17 | 14,10–20,57 |
| | Zhang et al. [41] 2022 | SVR, RF and KNN | 8,81–9,63 | 26,84–30,23 |
| **Current study** 7-day test set | ED ARMA | GLMNET and SVM-RBF | 5,48–5,52 | 11,44–11,69 |
| | ED JOON | SVM-RBF and LightGBM | 4,61–4.73 | 15,95–16,22 |
| | ED RG | RF and LightGBM | 6,55–6,81 | 11,93–12,87 |
| | ED RPH | XGBoost and SVM-RBF | 5,90–6,21 | 14,89–15,45 |
| | ED SCG | XGBoost and NNAR | 5,08–5,21 | 11,73–11,93 |
| 8 up to 45 days ahead | Marcilio et al. [42] 2013 | GLM and GEE | 8,7–12,8 | not applied |
| | Bergs et al. [45] 2014 | ETS | 2,63–4,76 | not applied |
| | Calegari et al. [46] 2016 | SARIMA, SS and SMHW | 11,35–12,29 | not applied |
| | Juang [19] 2017 | ARIMA | 8,91 | not applied |
| | Carvalho-Silva et al. [23] 2018 | ARIMA | 5,22–9,29 | not applied |
| | Jilani et al. [15] 2019 | NN e FTS | 2,01–2,81 | 57,30–167,89 |
| | Khaldi et al. [27] 2019 | EEMD-ANN, DWT-ANN and ANN | not applied | 52,86–149,23 |
| | Vollmer et al. [20] 2021 | GLMNET, LM and GBM | 6,8–8,9 | not applied |
| | Pekel et al. [39] 2021 | PSO-ANN, Bayesian ANN and GA-ANN | 6–8,8 | 53.29–83.85 |
| | Tuominen et al. [54] 2022 | ARIMAX, RLS-FS and RLS-SA | 7,4–7,8 | not applied |
| | Susnjak et al. [78] 2023 | Voting regressor | 8,9–12,8 | 10,60–15,9 |
| | Gafni-Pappas et al. [50] 2023 | RF and XGBoost | not applied | 18,94–18,96 |
| **Current study** 45-day test set | ED ARMA | XGBoost and GLMNET | 5,90–5,90 | 12,62–12,64 |
| | ED JOON | SVM-RBF and RF | 4.98–5.01 | 17,25–17,43 |
| | ED RG | XGBoost and RF | 6,23–6,28 | 11,42–11,47 |
| | ED RPH | SVM-RBF and NNAR | 5,71–5,85 | 14,25–14,37 |
| | ED SCG | XGBoost and RF | 5,64–5,69 | 12,61–12,56 |

*ANN* Artificial Neural Networks, *ARIMA* Autoregressive Integrated Moving Average, *ARIMAX* Autoregressive Integrated Moving Average with Explanatory Variable, *ARIMA-LR* ARIMA-Linear regression, *CNN* Convolutional Neural Networks, *DBN* Deep Belief Network, *EEMD-ANN* Artificial Neural Networks with Ensemble Empirical Mode decomposition, *XGBoost* Extreme Gradient Boosting, *RLS-FS* Floating Search with Recursive Least Squares, *FTS* Fuzzy Time Series, *CNN-GRU* Gated recurrent unit with convolutional neural networks, *GA-ANN* Genetic Algorithm-based ANN, *GLMNET* Generalized Linear Models via Coordinate Descent, *GLM* generalized linear model, *GBM* Gradient Boosting Machines, *HW* Holt-Winters, *KNN* k-nearest neighbours, *LM* Linear model, *LSTM* Long Short-Term Memory, *MLR* Multiple Linear Regression, *MSARIMA* Multivariate Autoregressive Integrated Moving Average, *NN* Neural Network, *NNAR* Neural Network Autoregression, *PSO-ANN* Particle Swarm Optimization algorithm-based ANN, *RF* Random Forest, *RNN* Recurrent Neural Networks, *RBM* Restricted Boltzmann machines, *SARIMA* Seasonal Autoregressive Integrated Moving Average, *SARIMAX* Seasonal Autoregressive Integrated Moving Average with external variables, *SS* Simple Seasonal Exponential Smoothing, *RLS-SA* Simulated Annealing with Recursive Least Squares, *SVM-RBF* Support Vector Machine with Radial Basis Function, *SVR* Support Vector Regression, *VAR* Vector Autoregression Model

the week, with results consistent with other similar studies, e.g., [41]–[43] for temperatures, and [3, 6, 7, 47] for days of the week.

Notably, both NNAR and XGBoost achieved better forecasting results. In addition to the use of feature-engineered variables, there are reasons related to model structure that justify their better performance. NNAR is effective at capturing nonlinear and complex patterns in time series, as it combines the flexibility of neural networks with an autoregressive approach, allowing the model to learn from past dependencies in the data to make more accurate predictions [63]. This ability to model nonlinear complexities is particularly advantageous in-patient arrival forecasting, where the volume of visits can be influenced by a combination of seasonal factors, such as weather variations, holidays, and epidemic events. Nonlinearities, if present, are also more easily captured by the large number of feature-engineered variables derived from arrival timestamps.

On the other hand, the XGBoost algorithm, based on decision-trees, offers several advantages that are particularly valuable in patient arrival forecasting, namely: (i) ability to avoid overfitting, which is essential in emergency demand forecasting where data variability can be significant, (ii) capacity to generalize results with large volumes of data, and (iii) iterative learning mechanism that corrects errors from previous decision trees by adjusting the residuals. These characteristics allows the model to efficiently learn from the data, adapting to different patient arrival patterns, such as seasonal variations or demand peaks.

#### Managerial implications and practical implementation

The creation of calendar features through FE has proven highly effective in enhancing the predictive performance of ML models, particularly in forecasting patient arrivals across multiple EDs in three countries. This approach demonstrates that timestamps associated with time series data capture fundamental seasonal patterns and trends essential for accurate forecasting. Temporal components such as the day of the month and week of the year enable models to recognize recurring behaviors and underlying dynamics of patient flow, which are key for predicting daily patient volumes.

Incorporating FE variables into ML models allows for better identification of these recurring patterns, thereby improving the accuracy of predictions. Such enhanced performance has significant practical implications for ED management. Accurate forecasts enable better staffing decisions, ensuring adequate healthcare provider availability, especially during peak periods such as weekends and Mondays. Additionally, understanding temporal patterns aids in optimizing scheduling strategies, reducing wait times, and improving patient care. Furthermore, precise predictions facilitate efficient management of limited resources, such as physical space and medical supplies, by allowing for proactive decision-making and reducing pressure on the healthcare system.

Consider two hypothetical ED scenarios to illustrate the impact of accurate predictions. The first scenario involves staff allocation. With accurate forecasts predicting increased patient volume during weekends or holidays, emergency managers can adjust work schedules or hire additional staff, thereby avoiding overload and maintaining care quality.

The second scenario focuses on bed and equipment management. If forecasts indicate a significant increase in patient arrivals, hospitals can proactively manage resources by adjusting internal logistics and prioritizing discharges or transfers. Inaccurate predictions could lead to bed shortages, resulting in patients being placed on stretchers in hallways and increasing health risks. Advance knowledge of demand patterns also enables hospitals to redirect excess patients to other facilities within public health networks.

Hospitals can implement automated systems that leverage FE data directly in the decision-making process. For example, integrating predictions from ML models into hospital management software could trigger automatic alerts recommending staff scheduling adjustments based on predicted arrivals. Such systems are particularly valuable in urban centers with variable demand. Additionally, accurate predictions can help reduce operational costs by optimizing the management of supplies and medications, such as anticipating and adjusting stock levels for seasonal demands.

The main cost–benefit aspects of implementing ML models for predicting patient arrivals in EDs are: (i) reduction in operational costs from more accurate predictions that enable staffing adjustments, such as lowering overtime and temporary hire costs; (ii) reduction in emergency purchasing costs through better demand predictions that optimize stock management, avoiding waste or shortages during critical times; (iii) improved management of bed occupancy and equipment use, preventing overcrowding and improving patient flow, thereby reducing emergency care costs; and (iv) integration of ML models into ED management systems, enabling automated alerts for real-time decisions, optimizing responses to demand fluctuations and alleviating overload during peak periods. These benefits are particularly

relevant in resource-limited settings, where accurate predictions help prevent unnecessary expenses, improve resource allocation, and support the healthcare system's sustainability while ensuring quality care for all patients.

## Conclusion

In this paper, we compared the performance of six ML algorithms across two forecast horizons for predicting daily arrivals in EDs. We used both traditional meteorological and calendar predictors alongside feature-engineered variables. The algorithms were optimized using hyperparameter tuning via fivefold cross-validated grid-search. Variable selection was conducted using a random-forest method, identifying key predictors such as index.num, yday, week, qday, minimum, mean, and maximum temperatures, and the day of the week. Performance evaluation employed four error metrics within a fivefold cross-validation framework.

Our results surpass many existing studies in the literature, demonstrating superior predictive accuracy crucial for effective resource management in EDs, reducing patient waiting times and lengths of stay. Notably, XGBoost consistently outperformed other models across all forecast horizons, with FE significantly enhancing the predictive capabilities of all ML algorithms.

Unlike typical studies on ED patient arrival prediction, our findings are robust and can be readily applied and replicated in other ED settings. We have provided comprehensive R code for all methodological steps and used publicly accessible datasets, facilitating easy adaptation and extension with additional predictor variables as needed.

This study presents some limitations. The first is associated with the databases analyzed. The advantages of including additional informative predictor variables, such as wind speed, air quality, precipitation, holidays, and special or epidemic events, in improving prediction quality were not explored. Such variables could be valuable for enhancing ML performance and represent a promising direction for future research. Future studies could also apply FE to other variables, such as meteorological data, to assess potential performance gains. Other limitations include the use of only one variable selection method, based on the RF technique via permutation importance, which may introduce bias into the results. By relying on a single selection method, such as RF, the study may prioritize variables based on a specific criterion, potentially overlooking other relevant variables that could be identified using alternative selection methods. Additionally, the adoption of computationally intensive ML algorithms, such as SVM-RBF and NNAR, with hyperparameter tuning, may pose challenges in hospital settings with limited computational resources. These algorithms require high processing power, which may hinder their implementation in hospitals with constrained infrastructure.

### Reproducibility

The R code for replicating all the results obtained in this study is available in the GitHub repository (https://github.com/forecastingEDs/Feature-engineering-and-machine-learning-algorithms).

#### Abbreviations

| | |
|---|---|
| ANN | Artificial neural networks |
| ARIMA | Autoregressive Integrated Moving Average |
| ARIMA-ANN | Autoregressive Integrated Moving Average with Artificial Neural Network |
| ARIMA-LR | Autoregressive Integrated Moving Average with Linear Regression |
| ARIMA-SVR | Autoregressive Integrated Moving Average with Support Vector Regression |
| ARIMAX | Autoregressive Integrated Moving Average with Explanatory Variable |
| BiLSTM | Bidirectional Long Short-Term Memory |
| CNN | Convolutional Neural Networks |
| DLSTM | Deep Stacked Architecture with Long Short-Term Memory |
| DNNs | Deep Neural Networks |
| ED | Emergency department |
| EFB | Exclusive Feature Bundling |
| ES | Exponential Smoothing |
| ETS | Error-trend-seasonal |
| FE | Feature engineering |
| GB | Gradient Boosting |
| GLMNET | Lasso and Elastic-Net Generalized Linear Model |
| GOSS | Gradient-based One-Side Sampling |
| HW | Holt-Winters |
| KNN | K-nearest neighbours |
| LightGBM | Light Gradient Boosting Machine |
| LR | Logistic regression |
| LSTM | Long Short-Term Memory |
| MAE | Mean Absolute Error |
| MAPE | Mean Absolute Percentage Error |
| MASE | Mean Absolute Scaled Error |
| ML | Machine learning |
| MLP | Multilayer Perceptron Neural Network |
| NNAR | Neural Network Autoregression |
| PSO-ANN | Particle Swarm Optimization algorithm-based ANN |
| RF | Random forest |
| RMSE | Root Mean Square Error |
| RNN | Recurrent Neural Networks |
| RNN-1L | Recurrent Neural Network with One Layer |
| SARIMA | Seasonal Autoregressive Integrated Moving Average |
| SARIMAX | Seasonal Autoregressive Integrated Moving Average with external variables |
| Snaive | Seasonal Naive |
| sMAPE | Symmetric Mean Absolute Percentage Error |
| SVM-RBF | Support Vector Machine with Radial Basis Function |
| SVR | Support Vector Regression |
| TSCV | Time-series split cross-validation |
| XGBoost | eXtreme Gradient Boosting |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02788-6.

Supplementary Material 1.

**Acknowledgements**
We sincerely thank the editor and reviewers for their time and effort in reviewing this article.

**Authors' contributions**
BP and FF collaboratively devised and structured the research plan. BP developed the computational programming, data analysis, and composed the manuscript. BP and FF oversaw the study's design and execution, evaluating data analysis procedures. FF extended methodological support/advice, writing, and critically reviewed the manuscript. All authors contributed to data interpretation, provided feedback on earlier drafts, and approved the final version of the manuscript.

**Funding**
Not applicable.

**Data availability**
Dataset is publicly available and was acquired from the Harvard Dataverse database [60].The R codes for all methodological steps developed are provided at the following GitHub repository: (https://github.com/forecastingEDs/Feature-engineering-and-machine-learning-algorithms).

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 23 January 2024 Accepted: 26 November 2024
Published online: 18 December 2024

**References**

1. Choudhury A, Urena E. Forecasting hourly emergency department arrival using time series analysis. Br J Heal Care Manag. 2020;26(1):34–43. https://doi.org/10.12968/bjhc.2019.0067.
2. L. He, S. Chalil Madathil, A. Oberoi, G. Servis, and M. T. Khasawneh, "A systematic review of research design and modeling techniques in inpatient bed management," Comput Ind Eng. 2019;127(October 2018):451–466.
3. Whitt W, Zhang X. Forecasting arrivals and occupancy levels in an emergency department. Oper Res Heal Care. 2019;21:1–18. https://doi.org/10.1016/j.orhc.2019.01.002.
4. Yucesan M, Gul M, Celik E. A multi-method patient arrival forecasting outline for hospital emergency departments. Int J Healthc Manag. 2018;13(S1):283–95. https://doi.org/10.1080/20479700.2018.1531608.
5. Rocha CN, Rodrigues F. Forecasting emergency department admissions. J Intell Inf Syst. 2021;56(3):509–28. https://doi.org/10.1007/s10844-021-00638-9.
6. Erkamp NS, van Dalen DH, de Vries E. Predicting emergency department visits in a large teaching hospital. Int J Emerg Med. 2021;14(1):1–12. https://doi.org/10.1186/s12245-021-00357-6.
7. Sudarshan VK, Brabrand M, Range TM, Wiil UK. Performance evaluation of Emergency Department patient arrivals forecasting models by including meteorological and calendar information: A comparative study. Comput Biol Med. 2021;135(January):104541. https://doi.org/10.1016/j.compbiomed.2021.104541.
8. American College of Emergency Physicians (ACEP). Crowding. Policy statement. Ann Emerg Med. 2013;61(6):726–7. https://doi.org/10.1016/j.annemergmed.2013.03.037.
9. Ortíz-Barrios MA, Alfaro-Saíz JJ. Methodological approaches to support process improvement in emergency departments: A systematic review. Int J Environ Res Public Health. 2020;17(8):2–41. https://doi.org/10.3390/ijerph17082664.
10. Rasouli HR, Aliakbar Esfahani A, and Abbasi Farajzadeh M, "Challenges, consequences, and lessons for way-outs to emergencies at hospitals: a systematic review study." BMC Emerg Med. 2019;19(1):62. https://doi.org/10.1186/s12873-019-0275-9.
11. Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB. On patient flow in hospitals: A data-based queueing-science perspective. Stoch Syst. 2015;5(1):146–94. https://doi.org/10.1214/14-ssy153.
12. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: A systematic review of causes, consequences and solutions. PLoS ONE. Aug.2018;13(8): e0203316. https://doi.org/10.1371/journal.pone.0203316.
13. Gul M, Celik E. An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. Heal Syst. 2018;00(00):1–22. https://doi.org/10.1080/20476965.2018.1547348.
14. Moukarzel A, et al. Burnout syndrome among emergency department staff: Prevalence and associated factors. Biomed Res Int. 2019;2019:2–10. https://doi.org/10.1155/2019/6462472.
15. Jilani T, Housley G, Figueredo G, Tang PS, Hatton J, Shaw D. Short and Long term predictions of Hospital emergency department attendances. Int J Med Inform. 2019;129(May):167–74. https://doi.org/10.1016/j.ijmedinf.2019.05.011.
16. Harrou F, Dairi A, Kadri F, Sun Y. Forecasting emergency department overcrowding: A deep learning framework. Chaos, Solitons Fractals. Oct.2020;139: 110247. https://doi.org/10.1016/J.CHAOS.2020.110247.
17. Chen C-F, Ho WH, Chou HY, Yang SM, Te Chen I, Shi H-Y. Long-term prediction of emergency department revenue and visitor volume using autoregressive integrated moving average model. Comput Math Methods Med. 2011;2011:2–7. https://doi.org/10.1155/2011/395690.
18. Xu Q, Tsui KL, Jiang W, Guo H. A Hybrid Approach for Forecasting Patient Visits in Emergency Department. Qual Reliab Eng Int. 2016;32(8):2751–9. https://doi.org/10.1002/qre.2095.
19. Juang WC, Huang SJ, Huang FD, Cheng PW, Wann SR. Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. BMJ Open. 2017;7(11):1–7. https://doi.org/10.1136/bmjopen-2017-018628.
20. Vollmer MAC, et al. A unified machine learning approach to time series forecasting applied to demand at emergency departments. BMC Emerg Med. 2021;21(1):1–14. https://doi.org/10.1186/s12873-020-00395-y.
21. Yousefi M, Yousefi M, Fathi M, Fogliatto FS. Patient visit forecasting in an emergency department using a deep neural network approach. Kybernetes. 2020;49(9):2335–48. https://doi.org/10.1108/K-10-2018-0520.
22. Boyle J, et al. Predicting emergency department admissions. Emerg Med J. 2012;29(5):358–65. https://doi.org/10.1136/emj.2010.103531.
23. Carvalho-Silva M, Monteiro MTT, de Sá-Soares F, Dória-Nóbrega S. Assessment of forecasting models for patients arrival at Emergency Department. Oper Res Heal Care. 2018;18:112–8. https://doi.org/10.1016/j.orhc.2017.05.001.
24. Harrou F, Dairi A, Kadri F, Sun Y. Effective forecasting of key features in hospital emergency department: Hybrid deep learning-driven methods. Mach Learn with Appl. Mar.2022;7: 100200. https://doi.org/10.1016/j.mlwa.2021.100200.
25. Lucini FR, et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. Int J Med Inform. 2017;100:1–8. https://doi.org/10.1016/j.ijmedinf.2017.01.001.
26. Lucini FR, et al. Man vs. machine: Predicting hospital bed demand from an emergency department. PLoS One. 2020;15(8):1–11. https://doi.org/10.1371/journal.pone.0237937.
27. Khaldi R, El Afia A, Chiheb R. Forecasting of weekly patient visits to emergency department: Real case study. Procedia Comput Sci. 2019;148:532–41. https://doi.org/10.1016/j.procs.2019.01.026.
28. Isken MW, Aydas OT, Roumani YF. Queueing inspired feature engineering to improve and simplify patient flow simulation metamodels. J Simul. Feb.2023;00(00):1–18. https://doi.org/10.1080/17477778.2023.2181716.
29. Bojer CS, Meldgaard JP. Kaggle forecasting competitions: An overlooked learning opportunity. Int J Forecast. Apr.2021;37(2):587–603. https://doi.org/10.1016/j.ijforecast.2020.07.007.
30. Verdonck T, Baesens B, Óskarsdóttir M, vanden Broucke S. Special issue on feature engineering editorial. Mach Learn. 2024;113(7):3917–28. https://doi.org/10.1007/s10994-021-06042-2.

31. Kuhn M, Johnson K. Feature Engineering and Selection: A Practical Approach for Predictive Models. Taylor & Francis Group; 2019. [Online]. Available: https://bookdown.org/max/FES/.
32. Butcher B, Smith BJ. Feature Engineering and Selection: A Practical Approach for Predictive Models. Am Stat. Jul.2020;74(3):308–9. https://doi.org/10.1080/00031305.2020.1790217.
33. Petropoulos F, et al. Forecasting: theory and practice. Int J Forecast. Jul.2022;38(3):705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001.
34. Ejohwomu OA, et al. Modelling and Forecasting Temporal PM2.5 Concentration Using Ensemble Machine Learning Methods. Buildings. 2022;12(1):46. https://doi.org/10.3390/buildings12010046.
35. Zhang Y, Luo L, Yang J, Liu D, Kong R, Feng Y. A hybrid ARIMA-SVR approach for forecasting emergency patient flow. J Ambient Intell Humaniz Comput. 2019;10(8):3315–23. https://doi.org/10.1007/s12652-018-1059-x.
36. Silva E, Pereira MF, Vieira JT, Ferreira-Coimbra J, Henriques M, Rodrigues NF. Predicting hospital emergency department visits accurately: A systematic review. Int J Health Plann Manage. Jul.2023;38(4):904–17. https://doi.org/10.1002/hpm.3629.
37. Wargon M, Guidet B, Hoang TD, Hejblum G. A systematic review of models for forecasting the number of emergency department visits. Emerg Med J. 2009;26:395–9. https://doi.org/10.1136/emj.2008.062380.
38. Jiang S, Liu Q, Ding B. A systematic review of the modelling of patient arrivals in emergency departments. Quant Imaging Med Surg. 2023;13(3):1957–19. https://doi.org/10.21037/qims-22-268.
39. Pekel E, Gul M, Celik E, Yousefi S. Metaheuristic Approaches Integrated with ANN in Forecasting Daily Emergency Department Visits. Math Probl Eng. 2021;2021:1–14. https://doi.org/10.1155/2021/9990906.
40. Petsis S, Karamanou A, Kalampokis E, Tarabanis K. Forecasting and explaining emergency department visits in a public hospital. J Intell Inf Syst. 2022;59(2):479–500. https://doi.org/10.1007/s10844-022-00716-6.
41. Zhang Y, Zhang J, Tao M, Shu J, Zhu D. Forecasting patient arrivals at emergency department using calendar and meteorological information. Appl Intell. 2022;2021:11232–43. https://doi.org/10.1007/s10489-021-03085-9.
42. Marcilio I, Hajat S, Gouveia N. Forecasting daily emergency department visits using calendar variables and ambient temperature readings. Acad Emerg Med. 2013;20(8):769–77. https://doi.org/10.1111/acem.12182.
43. Menke NB, Caputo N, Fraser R, Haber J, Shields C, and Menke MN, "A retrospective analysis of the utility of an artificial neural network to predict ED volume." Am J Emerg Med. 2014;32(6):614–617. https://doi.org/10.1016/j.ajem.2014.03.011.
44. Kadri F, Harrou F, Chaabane S, Tahon C. Time series modelling and forecasting of emergency department overcrowding. J Med Syst. 2014;38(107):2–20. https://doi.org/10.1007/s10916-014-0107-0.
45. Bergs J, Heerinckx P, Verelst S. Knowing what to expect, forecasting monthly emergency department visits: A time-series analysis. Int Emerg Nurs. Apr.2014;22(2):112–5. https://doi.org/10.1016/j.ienj.2013.08.001.
46. Calegari R, Fogliatto FS, Lucini FR, Neyeloff J, Kuchenbecker RS, Schaan BD. Forecasting daily volume and acuity of patients in the emergency department. Comput Math Methods Med. 2016;2016:2–8. https://doi.org/10.1155/2016/3863268.
47. Hertzum M. Forecasting Hourly Patient Visits in the Emergency Department to Counteract Crowding. Ergon Open J. 2017;10(1):1–13. https://doi.org/10.2174/1875934301710010001.
48. Asheim A, Bache-Wiig Bjørnsen LP, Næss-Pleym LE, Uleberg O, Dale J, Nilsen SM. Real-time forecasting of emergency department arrivals using prehospital data. BMC Emerg Med. 2019;19(1):42. https://doi.org/10.1186/s12873-019-0256-z.
49. X. Zhao, J. W. Lai, A. F. Wah Ho, N. Liu, M. E. Hock Ong, and K. H. Cheong, "Predicting hospital emergency department visits with deep learning approaches," Biocybern. Biomed Eng. 2022;5537(August):127–133. https://doi.org/10.1016/j.bbe.2022.07.008.
50. Gafni-Pappas G, Khan M. Predicting daily emergency department visits using machine learning could increase accuracy. Am J Emerg Med. Mar.2023;65:5–11. https://doi.org/10.1016/j.ajem.2022.12.019.
51. Hu Y, et al. Use of Real-Time Information to Predict Future Arrivals in the Emergency Department. Ann Emerg Med. 2023;81(6):728–37. https://doi.org/10.1016/j.annemergmed.2022.11.005.
52. Reboredo JC, Barba-Queiruga JR, Ojea-Ferreiro J, Reyes-Santias F. Forecasting emergency department arrivals using INGARCH models. Health Econ Rev. Oct.2023;13(1):51. https://doi.org/10.1186/s13561-023-00456-5.
53. Rostami-Tabar B, Browell J, Svetunkov I. Probabilistic forecasting of hourly emergency department arrivals. Heal Syst. May2023;00(00):1–17. https://doi.org/10.1080/20476965.2023.2200526.
54. Tuominen J, et al. Forecasting daily emergency department arrivals using high-dimensional multivariate data: a feature selection approach. BMC Med Inform Decis Mak. 2022;22(1):1–12. https://doi.org/10.1186/s12911-022-01878-7.
55. De Hond A, et al. Machine learning for developing a prediction model of hospital admission of emergency department patients: Hype or hope? Int J Med Inform. Aug.2021;152: 104496. https://doi.org/10.1016/J.IJMEDINF.2021.104496.
56. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PLoS ONE. Jul.2018;13(7): e0201016. https://doi.org/10.1371/journal.pone.0201016.
57. Tello M, et al. Machine learning based forecast for the prediction of inpatient bed demand. BMC Med Inform Decis Mak. 2022;22(1):1–13. https://doi.org/10.1186/s12911-022-01787-9.
58. Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer New York; 2013. https://doi.org/10.1007/978-1-4614-6849-3.
59. Makridakis S. Accuracy concerns measures: theoretical and practical concerns. Int J Forecast. 1993;9(4):527–9. https://doi.org/10.1016/0169-2070(93)90079-3.
60. Van der Linden N. "ED visits and temperature," Harvard Dataverse, V1. Emergency department visits and temperature for a selection of hospitals in the Netherlands, USA, Botswana, Pakistan, and Australia. 2019. https://doi.org/10.7910/DVN/QHPZOX.
61. Cheng Q, Tanik N, Scott C, Liu Y, Platts-mills TF, Ziya S. Forecasting emergency department hourly occupancy using time series analysis. Am J Emerg Med. 2021;48:177–82. https://doi.org/10.1016/j.ajem.2021.04.075.
62. Makridakis S, Assimakopoulos V, Spiliotis E. Objectivity, reproducibility and replicability in forecasting research. Int J Forecast. Oct.2018;34(4):835–8. https://doi.org/10.1016/j.ijforecast.2018.05.001.
63. Hyndman RJ, Athanasopoulos G. Forecasting: Principles and Practice. 3rd ed. Melbourne: OTexts; 2021. [Online]. Available: https://otexts.com/fpp3/.
64. Kuhn M, Johnson K. 3.4 Resampling. In: Feature Engineering and Selection: A Practical Approach for Predictive Models. Taylor & Francis Group; 2019. [Online]. Available: https://bookdown.org/max/FES/resampling.html#rolling-origin-forecasting.
65. Dancho M, Vaughan D. timetk: A Tool Kit for Working with Time Series. R Package; 2023. [Online]. Available: https://cran.r-project.org/package=timetk.
66. M. Dancho, "Calendar Features," Comprehensive R Archive Network CRAN, 2024. https://business-science.github.io/timetk/articles/TK01_Working_With_Time_Series_Index.html#time-series-signature (accessed 12 Dec 2022).
67. M. Dancho, "Working with the Time Series Index Using Timetk," 2017. http://cran.nexr.com/web/packages/timetk/vignettes/TK01_Working_With_Time_Series_Index.html (accessed 12 Dec 2022).
68. Zhu X, Hu J, Xiao T, Huang S, Wen Y, Shang D. An interpretable stacking ensemble learning framework based on multi-dimensional data for real-time prediction of drug concentration: The example of olanzapine. Front Pharmacol. 2022;13(September):1–20. https://doi.org/10.3389/fphar.2022.975855.
69. Li J, et al. Feature Selection: A Data Perspective. ACM Comput Surv. 2016;50(6). https://doi.org/10.1145/3136625.
70. Greenwell BM, Boehmke BC. Variable Importance Plots—An Introduction to the vip Package. R J. 2020;12(1):343. https://doi.org/10.32614/RJ-2020-013.
71. Pawley S, Kuhn M, Jacques-Hamilton R. colino: Recipes Steps for Supervised Filter-Based Feature Selection. R Package; 2023. [Online]. Available: https://stevenpawley.github.io/colino.
72. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Stat Data Anal. 2020;143:106839. https://doi.org/10.1016/j.csda.2019.106839.
73. Gómez-Ramírez J, Ávila-Villanueva M, Fernández-Blázquez MÁ. Selecting the most important self-assessed features for predicting conversion

to mild cognitive impairment with random forest and permutation-based methods. Sci Rep. Nov.2020;10(1):20630. https://doi.org/10.1038/s41598-020-77296-4.
74. Makungwe M, Chabala LM, Chishala BH, Lark RM. Performance of linear mixed models and random forests for spatial prediction of soil pH. Geoderma. 2021;397(April):115079. https://doi.org/10.1016/j.geoderma.2021.115079.
75. Kuhn M, Silge J. Tidy Modeling with R: A Framework for Modeling in the Tidyverse. 1st ed. O'Reilly Media; 2022. [Online]. Available: https://www.tmwr.org/grid-search.html.
76. M. Kuhn, "tune: Tidy Tuning Tools." 2023. [Online]. Available: https://cran.r-project.org/package=tune
77. M. Kuhn and H. Wickham, "Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles." 2020. [Online]. Available: https://www.tidymodels.org
78. T. Susnjak and P. Maddigan, "Forecasting patient demand at urgent care clinics using explainable machine learning," CAAI Trans. Intell. Technol., pp. 1–22, Jul. 2023, https://doi.org/10.1049/cit2.12258.
79. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
80. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, Long Beach, CA, USA, 2017, pp. 3147–3155.
81. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22. https://doi.org/10.18637/jss.v033.i01.
82. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast. 2006;22:679–88. https://doi.org/10.1016/j.ijforecast.2006.03.001.
83. Murtas R, Tunesi S, Andreano A, Russo AG. Time-series cohort study to forecast emergency department visits in the city of Milan and predict high demand: a 2-day warning system. BMJ Open. 2022;12(4): e056017. https://doi.org/10.1136/bmjopen-2021-056017.

## Publisher's Note