Time Series Forecasting for Patient Arrivals in Online Health Services

Syed Kazmi
Data Science Lab
Ryerson University
Toronto, Ontario, Canada
sakazmi@ryerson.ca

Aysun Bozanta
Data Science Lab
Ryerson University
Toronto, Ontario, Canada
aysun.bozanta@ryerson.ca

Mucahit Cevik
Data Science Lab
Ryerson University
Toronto, Ontario, Canada
mcevik@ryerson.ca

ABSTRACT

The prediction of the patient volume is an important task in online healthcare services as it affects several operational decisions including staffing and scheduling of the medical personnel. The aim of this study is to investigate the performance of various statistical and machine learning models for the task of predicting the volume of patients requesting consultation from a particular online health service over a given time period. For this purpose, we consider a dataset adopted from an online healthcare application and apply both statistical forecasting techniques (e.g., exponential smoothing and ARIMA) and machine-learning models (e.g., random forest and long short term memory networks) to predict the patient volume. We evaluate the performances of alternative prediction methods by using MAE, NRMSE, and ND metrics. Our results show that the long short-term memory networks and ARIMA outperform other algorithms, however, the performances of the prediction models are highly impacted by the dataset size. While our results are in line with those reported in previous studies, we concur that more data and informative external covariates can be useful for further improving the prediction performance.

KEYWORDS

Time Series; Machine learning; Forecasting; Healthcare Applications

ACM Reference Format:

Syed Kazmi, Aysun Bozanta, and Mucahit Cevik. 2021. Time Series Forecasting for Patient Arrivals in Online Health Services. In *Proceedings of 31st Annual International Conference on Computer Science and Software Engineering (CASCON'21).* ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

The healthcare system has been experiencing a significant digital transformation in recent years. This transformation has been further accelerated by the uncontrollable spread of Covid-19 since more and more people started to prefer getting health services online. In this regard, carefully designed and maintained online platforms not only reduce the burden of the hospitals but also help users receive timely medical advice [2].

Several digital healthcare applications emerged in recent years, which enable their users to chat in real-time with a professional and experienced doctor, send their photos or videos, and get advice about their health problems in minutes. These applications eliminate the barrier that keeps many people from accessing healthcare for reasons such as having to take time off of work, high medical costs, and exposure to germs in a hospital setting.

Increasing demand for an online service creates certain challenges for service providers including staffing and scheduling of the medical personnel [3]. In this regard, it can be very important to accurately predict future demand (i.e., the number of patient requests) in advance. An accurate demand prediction would make efficient resource allocation possible by predicting the necessary workload. It not only yields cost savings, but also reduces waiting times, and increases customer satisfaction by assigning the optimal number of doctors in each hour of the day. In addition, increased customer satisfaction helps to enlarge the customer base in the long run, hence contributes to profitability.

Online healthcare services operate either in hybrid settings, i.e., medical services are integrated across online and offline channels, where a physician may consult a patient online and later schedule an offline visit [13], or entirely online, where the patients may not even necessarily have to book an appointment, and are not restricted to specific geographic regions or visiting hours. In our study, we consider the latter case, where an online medical service connects the patients to the doctors. A unique challenge to healthcare platforms that operate entirely online is that, since they are accessible to a wider audience (e.g., a whole country or worldwide) for 24 hours a day, it is difficult to identify any daily or weekly trends due to the increased randomness in the system. In terms of scheduling the doctors (e.g., on an hourly or 2-hourly basis), such platforms require long prediction horizons to be able to perform the scheduling tasks in advance.

In this paper, we conduct an empirical study to assess the performances of various statistical and machine learning models to predict the number of patients requesting doctor consultation in an online medical service. We experiment with the various statistical models (e.g., ARIMA), ensemble methods (e.g., random forest), and deep learning methods (e.g., recurrent neural networks). We also consider feature engineering strategies and data transformations to improve predictive performance.

The main contribution of this study can be summarized as follows:

CASCON'21, Nov 22-25, 2021, Online/Virtual

- The main novelty of our work is due to the application problem. That is, few other studies considered time series forecasting approaches for predicting volume of patients for online medical services, which has different dynamics than patient arrivals to the hospitals/medical clinics.
- We treat this problem as a regression problem (i.e., predicting the number of patients), and leverage popular time series forecasting methods such as ensemble models and deep learning methods. We provide a detailed numerical study using two distinct datasets, which contribute to the understanding of the strengths and limitations of various forecasting methods in an important practical problem.

The rest of the paper is organized as follows. Section 2 provides a brief discussion on the relevant literature on time series forecasting algorithms, and the application of time series forecasting for predicting patient arrivals. It is followed up by the description of datasets, algorithms, and experimental setup in Section 3. Section 4 presents results for our numerical study which involves two distinct datasets. Finally, Section 5 describes the threats to validity, and Section 6 provides concluding remarks along with future research directions.

2 LITERATURE REVIEW

We review the related literature under two subheadings. First, we discuss the time series forecasting algorithms and their strength and weaknesses under certain conditions. Then, we present the existing studies that apply time series forecasting for patient arrivals.

2.1 Time series forecasting methods

A time series forecasting problem involves building models based on historic data points to predict future data points [6]. Traditional time series forecasting models include Exponential Smoothing (ES), Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA) [6]. These traditional models are typically linear, and simple to use for time series forecasting. For instance, an AR model formulates future predictions as a linear function of its past observations whereas an MA model formulates future predictions as a linear function of the residual errors [20]. The ARMA model combines the capabilities of AR and MA to predict future values as a linear function of the observations and residual errors [6]. All three models require the time series to not include trend and seasonal components [6]. On the other hand, ARIMA and SARIMA models can be particularly suitable for non-stationary time series that exhibit trend and seasonality. However, all of these methods fail to capture nonlinear relationships in the time series [6].

Sophisticated and mathematically complex models such as Deep Neural Networks (DNN) have emerged as state-of-the-art models for capturing the complex, nonlinear behavior for certain time series prediction tasks [20]. Results from previous studies indicate that DNN models are capable of modeling seasonality and trend with high accuracy, making such models appropriate for solving complex real-world problems. Commonly used DNN models for time series forecasting problems include Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), and Long Short-Term Memory Networks (LSTM) [20]. In addition, when the available data is abundant, more complex deep learning architectures such as DeepAR and DeepState can be used for generating accurate probabilistic forecasts through global forecasting [1, 26].

MLPs can be considered to be among the simplest neural network architectures, hence they are relatively robust to noise and missing data. CNN models hold similar benefits as MLP models and in addition, are capable of extracting important features from the input data. LSTM models, a member of Recurrent Neural Networks (RNN) family, are designed to handle sequential data. Despite the adoption of DNN models for time series forecasting problems, there are several noted drawbacks associated with DNNs. Empirical studies indicate that DNN models yield mixed results, and are not capable of modeling seasonal patterns or trends accurately in certain cases (e.g., for small datasets) [20]. Such methods are also difficult to train and tune. In addition, due to the complex nature of the RNN models, the training time is typically much longer than commonly used statistical models and standard supervised learning techniques. Lastly, black-box nature of DNN models make their predictions difficult interpret, and hence might limit their adoption due to lack of trust towards the model.

2.2 Time series forecasting for patient arrivals

Time series forecasting and analysis are discussed in numerous application areas. The application of time series forecasting techniques can be seen in sales forecasting, climate analysis [22], demand forecasting [5], and stock market analysis [18]. Countless industries adopt time series models for strategic, tactical, and operational decisions. The application of time series forecasting in the medical domain ranges from resource planning to cancer detection. In particular, accurately predicting the volume of patients to a medical service can have significant benefits. With the substantial increase in aging population, hospital crowding is an important concern, and can affect clinical health outcomes and patient satisfaction.

Several studies have been conducted over the years to predict the number of patients arriving at hospitals and emergency services. Batal et al. [4] predicted daily urgent care patient volume using step-wise linear regression, and they reported an R^2 value of 0.786. Jones et al. [15] forecasted daily emergency department patient volumes using SARIMA, time series regression, exponential smoothing, and artificial neural network models. They obtained better results using external covariates such as calendar information and site-specific special-day effects. McCarthy et al. [21] predicted hourly emergency department patient volumes using Poisson regression considering temporal, climatic, and patients' demographic information and obtained nearly identical results with the observed data. Kam et al. [16] predicted the daily emergency department patient volumes using various statistical forecasting methods, and reported that multivariate SARIMA performed better than all other algorithms. Kim et al. [17] predicted 4-hourly patients volume using ARIMA, SARIMA, and GARCH methods over 30-day forecast periods. The ARIMA model outperformed all other models by obtaining MAPE values of 6%-17.2% in one day ahead forecast, and 8.8% in a month ahead forecast. Zhou et al. [29] developed a hybrid approach using the SARIMA and the nonlinear autoregressive neural network (NARNN) models. They compared the performance of this hybrid

approach with those of traditional forecasting methods using MAE, RMSE, and MAPE metrics. Their results showed that the hybrid approach provides minor performance gains over other methods. Choudhury [9] implemented ARIMA, Holt-Winters, TBATS, and neural network methods to forecast hourly ED patient arrivals and reported that ARIMA produced the best performance over various for performance metrics [9].

Sudarshan et al. [27] conducted an empirical analysis for predicting the number of patients arriving in an emergency department. They used covariates such as weather forecasts for the 7-day and 3-day prediction horizons, and actual weather information from the past 3 days. The authors concluded that neural network-based models, namely LSTM and CNN, outperform Random Forest (RF) on the MAPE metric for both short-term as well as long-term predictions. Piccialli et al. [24] presented a multi-source time series fusion and forecasting framework using deep learning models. The proposed architecture uses an ensembling technique consisting of machine learning models and a hybrid neural network, that intakes medical booking time series as well as exogenous features (e.g., weather and air quality) to predict the number of bookings for medical examinations using a 7-day prediction horizon. The paper emphasized the value of extracted features from the time sequences when used coherently alongside the external features.

3 METHODS

In this section, we describe our datasets and summarize the forecasting methods used in our analysis.

3.1 Data

We performed our experiments using two distinct datasets. We first experimented with the hospital admission dataset provided by Zhou et al. [29], which is publicly available. Secondly, we used a dataset consisting of the number of patients arriving to an online health services over a given period of time. This dataset is obtained from Your Doctors Online[1], which is an online application that connects patients with doctors. While both of these datasets involve patient arrivals, there are certain differences in patient arrival patterns and dataset sizes. In the online health services dataset, we only considered the patients from North America, which is the total number of patient requests from the US and Canada. The reason for choosing these two countries is their high volume of requests compared to other countries. In addition, due to being from the same time zone, incoming requests from North America are more likely to show patterns based on hour of the day.

Table 1 presents the descriptive statistics of these two datasets. The hospital admission data is daily, and consists of patient admissions over 273 days, does not contain any missing data. The data starts from Jan 4, 2016, and ends on Oct 2, 2016, with a minimum of 40 and a maximum of 685 daily patient admissions. On the other hand, online health service data is hourly, and there are 9559 hours in total. The data was collected between July 10, 2019, and Nov 10, 2020, with a minimum of 0 and a maximum of 22 patients.

We consider hourly and two-hourly versions of online health service dataset. Having a two-hourly transformation of the dataset can be useful for aggregating the volume of the arriving patients, since the patient volume can be very low (i.e., zero or near-zero) at certain times. In addition, it can introduce patterns to the time series, which might benefit forecasting model performance. Figure 1 presents the patient volume between the dates Nov 6 - Nov 10, 2020, for hourly and two-hourly versions of the data.

[1] https://yourdoctors.online

3.2 Forecasting algorithms for the regression analysis

In order to predict the number of patients, we consider different algorithms, including ES, MA, ARIMA, RF regressor, and LSTM. The brief explanations for each algorithm are provided below.

1. **Naive Baseline:** We choose the naive baseline based on the assumption that the number of patients arriving next week will be the same as the week before.

2. **ES:** In the exponential smoothing algorithm, the predictions are generated by the weighted averages of past data points, and the weights decrease exponentially as the data points get farther away from the prediction window [11]. ES is a particularly effective method for the datasets having a linear trend and seasonality.

3. **MA:** In the moving average algorithm, averages from various subsets of the historic data are used to predict future data points.

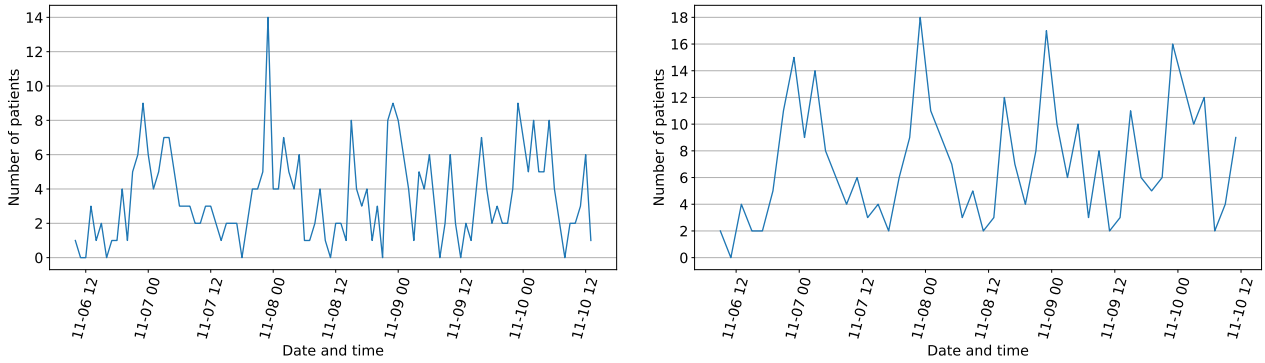
4. **ARIMA:** $ARIMA(p, d, q)$ model consists of parameters p, d, and q, where p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model [25]. If seasonality exists in the dataset, Seasonal ARIMA (SARIMA) can be used as an alternative to ARIMA.

5. **RF:** Random forests is an ensemble method that fits different decision trees on various sub-samples of the dataset and calculates the prediction through averaging [7]. Training data for an RF model can be constructed by extracting certain features from the time series and using those alongside the available external covariates and lag values. We consider RF as a representative supervised learning model in our analysis. Alternatively, machine learning models such as gradient boosted regressors (GBR) and support vector regressor (SVR) can also be used for forecasting.

6. **LSTM:** Recurrent neural networks (RNNs) such as LSTMs and Gated Recurrent Units (GRUs) have been frequently used for different time series forecasting problems [12]. Through memory cells and various gates, LSTMs have the ability to capture long term dependencies, which makes them particularly suitable for the forecasting task. By stacking multiple LSTMs, a stacked LSTM network can be created, which can be a useful strategy for time series forecasting as this allows hidden states at different layers to operate at different time scales [23]. In our analysis, we consider different configurations of LSTM networks (e.g., with different number of layers and hidden units) to identify the best performing LSTM model for our forecasting task.

Table 1: Dataset properties

Dataset	Length	Domain	Granularity	Missing data	Start	End	Min	Max
Hospital Admission [29]	273	$\mathbb{R}_+$	Daily	No	01/4/16	10/2/16	40	685
Online Health Service	9559	$\mathbb{R}_+$	Hourly	No	07/10/19	11/10/20	0	22



(a) Hourly Data **(b) 2-Hourly Data**

Figure 1: Patient volume in the online health service between the dates Nov 6 - Nov 10, 2020

Note that we only consider some of the most popular algorithms found in the literature as representatives of the mainstream forecasting methodologies (e.g, ARIMA for statistical forecasting, RF for supervised learning, and LSTM for RNN-based deep learning). ES and MA perform well when a certain degree of continuity between the past and future values can be assumed, and seasonal or cyclical variations do not exist. Accordingly, they serve as a benchmark for measuring performances with respect to the quality and interpretability of the given data. ARIMA/SARIMA combines autoregression and MA to make stronger predictions, but its drawback is that it assumes linear relationships between independent and dependent variables, as well as a constant standard deviation in errors in the model over time [19]. RF is widely popular due to its ability to learn from nonlinear decision boundaries without requiring any feature modifications (i.e., feature selection, scaling and normalization). However, it requires high computational power, and tends to be slow to train as the number of features increase. Also, unlike simpler linear models, it is often difficult to interpret RF predictions, therefore, RF is typically regarded as a "black-box" model [8]. LSTM is a widely used RNN architecture due to its ability to capture long term dependencies, and utilize long sequences. However, due to its structure, it is less computationally efficient when compared to some of the other deep learning architectures such as CNNs and GRU [10]. We also note that, in our preliminary analysis, we experimented with a wide range of forecasting methods (including other RNN architectures such as GRU and standard machine learning methods such as gradient boosted regressor) for the patient arrival datasets, however, we did not observe any noticeable performance improvements over the methods presented in this study.

3.3 Experimental setup

We consider following evaluation metrics to compare the performances of different algorithms:

- Normalized Deviation (ND)

$$\mathrm{ND}(y,\hat{y}) = \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i|}{\sum_{i=1}^{N} |y_i|} \tag{1}$$

- Normalized Root Mean Squared Error (NRMSE)

$$\mathrm{NRMSE}(y,\hat{y}) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}}{\frac{1}{N}\sum_{i=1}^{N} |y_i|} \tag{2}$$

- Mean Absolute Error (MAE)

$$\mathrm{MAE}(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{3}$$

where $y = [y_1, \ldots, y_N]$ and $\hat{y} = [\hat{y}_1, \ldots, \hat{y}_N]$ represent ground truth and predicted values over a prediction horizon of N, respectively. Note that ND and NRMSE values are normalized, therefore, they enable analyzing the prediction performance independently of the scale of the data. On the other hand, metrics such as RMSE and MAE can provide more problem specific insights. We do not consider mean absolute percentage error (MAPE), which is commonly used metric for forecasting, as online health service data contains zeros in the ground truth.

The parameter setting for each model is specified in Table 2. Note that these parameters are identified through extensive hyperparameter tuning using both datasets. While the models such as RF and GBR are fairly insensitive to the hyperparameters, deep learning

model performance can be significantly enhanced through hyperparameter tuning. Our LSTM model contains two hidden layers with 32 and 16 hidden units, and a fully connected dense layer. During LSTM training, the model is trained until the validation accuracy does not improve any further. For all the machine learning models, lookback (i.e., lag) value of 24 time steps is used.

We considered various feature engineering strategies to improve the performance of the prediction methods. In addition to the lag values, we used cyclic day-of the week and hour-of-day features, which are incorporated using sinusoidal transformation as follows [14]:

$$\text{day of week}_{sin} = \sin\left(dow * \frac{2\pi}{7}\right) \quad \text{day of week}_{cos} = \cos\left(dow * \frac{2\pi}{7}\right)$$

$$\text{hour of day}_{sin} = \sin\left(hod * \frac{2\pi}{24}\right) \quad \text{hour of day}_{cos} = \cos\left(hod * \frac{2\pi}{24}\right)$$

Table 2: The hyperparameter settings used in the experiments for the employed models (identical for both datasets)

Model	Hyperparameters
LSTM	*hidden units:* {32, 16}, *optimizer:* adam, *loss:* MAE, *batch size:* 32, *lookback:* 24
RF	*# of trees:* 100, *splitting criterion:* MSE, *max depth:* ∞, *lookback/lag:* 24
GBR	*loss:* least square, *criterion:* MSE, *learning rate:* 0.1, *lookback/lag:* 24

Various training strategies can be employed for the time series forecasting models [28]. The multiple input - multiple output (MIMO) strategy is particularly suited for neural network models where a single model is trained to predict multiple time steps. However, the MIMO strategy is not directly applicable to other machine learning models such as RF and GBR. Therefore, we employ a direct strategy for these models where a separate model is trained to predict each time step in the forecasting horizon. For testing the forecasting models over the online health services dataset, we apply augmented out-of-sampling method, which enable updating/retraining the model after each test sample prediction [14]. In this technique, after each test, the prediction model is updated to include the test data for training purposes. Typically, a sliding window approach should be used when old data should be disposed of. With augmented out of sampling technique, instead of full model training, the model is updated on a small portion of data, which is more efficient for computationally expensive models such as LSTMs and GRUs.

4 RESULTS

We next provide results for two time series datasets involving patient arrivals. We first examine the performance of various forecasting methods for the hospital admission dataset. As this dataset is obtained from a recent study by Zhou et al. [29], as a sanity check, we first directly compare the model performances against the reported results in [29]. Then, we provide more detailed results with hospital admission dataset. Next, we present the results obtained for the online health services dataset, and discuss the practical implications of our findings.

4.1 Forecasting results for hospital admission data

In their study, Zhou et al. [29] used SARIMA and the nonlinear autoregressive neural network (NARNN) models to predict the monthly and daily number of new admission inpatients. We extended their analysis by considering RF, LSTM, and GBR models on the same dataset to predict a daily number of new admission inpatients. We used RMSE, MAE, MAPE, NRMSE, and ND to evaluate the performances of these forecasting models. Table 3 presents the testing errors of those algorithms for two testing samples, one week (7-day prediction horizon) and four week (28-day prediction horizon), which is the adopted approach by Zhou et al. [29]. All algorithms outperformed the naive baseline. The errors for one week prediction period are significantly lower than four week prediction period for all the algorithms which is intuitive. The results in terms of error values are very close to each other among different algorithms. For one week testing period, the NARNN algorithm outperforms in terms of RMSE, MAE, and MAPE values and the RF algorithm follows it with the second-lowest error values. For the four-week testing period, the RF obtained the best RMSE and NRMSE results, while NARNN obtained the lowest MAE and MAPE values. Figure 2 shows the number of patients estimated by different algorithms and the actual number of patients for one-week and four-week predictions. These results show that, especially for one-week predictions, the forecasting model performances are comparable.

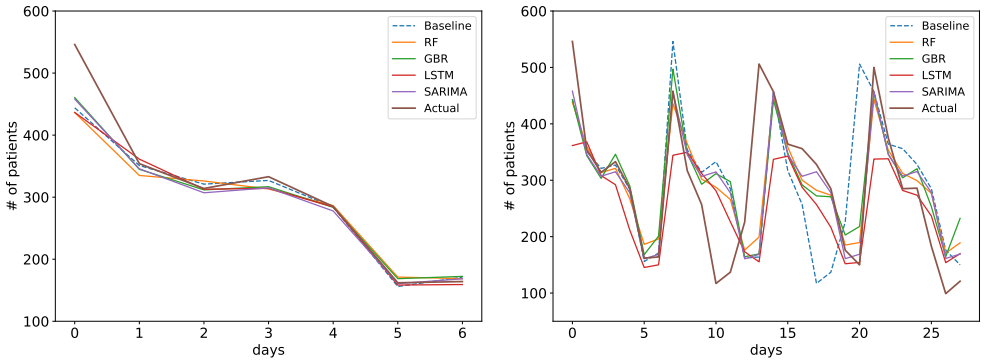
Table 4 presents the performances of the naive baseline, SARIMA, RF, LSTM, and GBR for four week testing period, where we applied the rolling horizon technique over the provided 28-day test data to generate 21 test sets each with one-week prediction window. These results show that assessing the model performances over single test sets can be very misleading as the performance values are significantly worse compared to their counterparts in Table 3. The error values of RF and SARIMA are very close to each other and lower than all other algorithms. Figure 3 presents the actual and predicted values by the RF algorithm for a consecutive four-week period using the rolling horizon technique (i.e., non-overlapping four of 21 test samples are illustrated). Especially for week 1 and week 4, the RF algorithm-generated very close predictions to the actual values, however, for week 2 and 3, the prediction performance is very poor, which shows the importance of conducting multiple tests to assess the model performance.

4.2 Forecasting results for online health service data

We applied various algorithms to predict the patient volume for the online health service. We compared the performances of different algorithms, namely, ES, MA, ARIMA, RF, and LSTM to find the best performing algorithm on this dataset. We evaluated the performances of these algorithms using ND, NRMSE, and MAE metrics. Table 5 presents the mean and standard deviations of evaluation metrics from hourly and 2-hourly datasets for the 6, 12, 24, and 48-hour prediction horizons. Ideally, longer prediction horizons

Table 3: Forecasting performance over a single test sample for the hospital admission dataset (results for NARNN model are directly adopted from [29])

Testing Period	Model	RMSE	MAE	MAPE	NRMSE	ND
1 week	NARNN [29]	**24.54**	**14.12**	**3.44**	-	-
4 weeks		**86.88**	**49.24**	**23.14**	-	-
1 week	Naive baseline	38.88	18.86	5.16	0.13	0.06
4 weeks		126.23	84.68	36.58	0.43	0.29
1 week	SARIMA	34.34	19.17	4.59	0.11	0.06
4 weeks		86.91	50.20	23.70	0.30	0.17
1 week	RF	33.35	18.48	4.78	0.11	0.06
4 weeks		82.62	53.59	25.22	0.28	0.18
1 week	LSTM	42.14	20.84	4.79	0.14	0.17
4 weeks		101.01	69.93	26.35	0.35	0.24
1 week	GBR	43.15	25.12	6.33	0.14	0.08
4 weeks		91.90	59.99	28.84	0.32	0.21



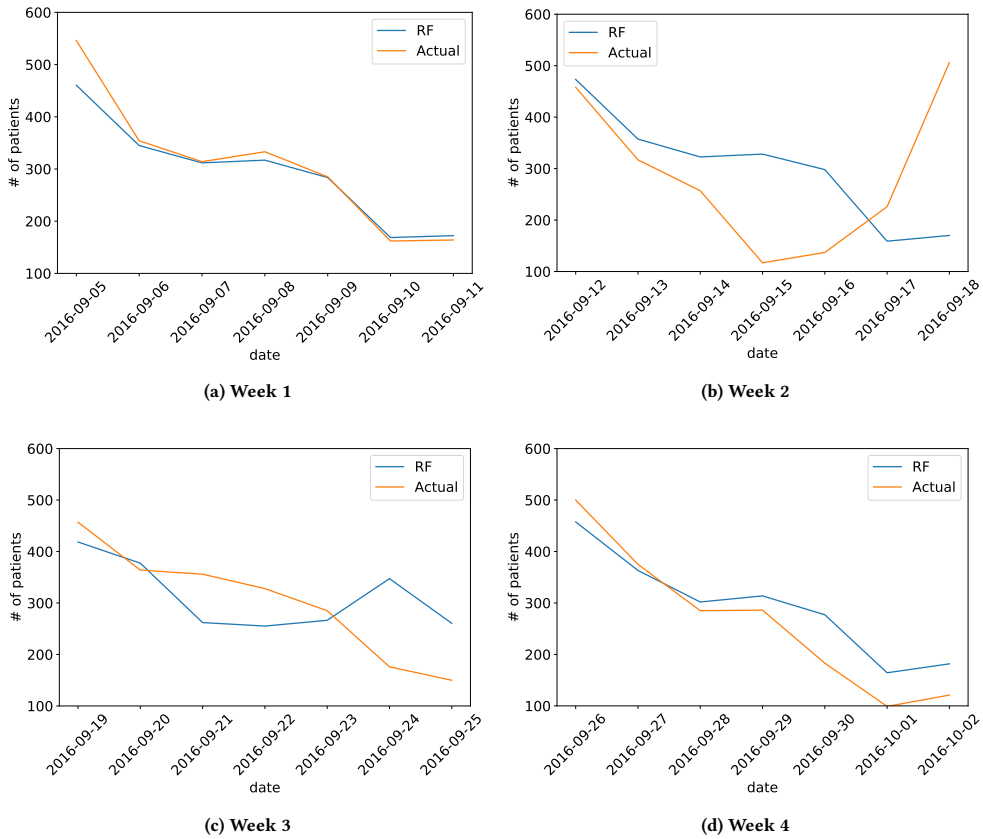
(a) 1 week prediction **(b) 4 week prediction**

Figure 2: Single test sample results for the hospital admission dataset

Table 4: Forecasting performance over 21 test samples for the hospital admission dataset (test samples are generated using rolling window method, prediction horizon is one week, "average error ± stdev of error" is provided)

Testing Period	Model	RMSE	MAE	MAPE	NRMSE	ND
1 week rolling	Naive baseline	128.7 ± 53.2	95.3 ± 42.9	43.8 ± 20.7	0.44 ± 0.17	0.32 ± 0.14
1 week rolling	SARIMA	87.3 ± 57.4	**66.4 ± 43.2**	**28.9** ± 19.5	0.30 ± 0.19	0.23 ± 0.14
1 week rolling	RF	**86.9 ± 42.6**	67.5 ± 33.4	31.5 ± 17.8	**0.30 ± 0.15**	**0.23 ± 0.12**
1 week rolling	LSTM	117.2 ± 46.4	85.7 ± 36.8	39.9 ± 20.3	0.40 ± 0.16	0.29 ± 0.13
1 week rolling	GBR	100.8 ± 45.3	76.9 ± 35.8	35.2 ± 17.8	0.34 ± 0.15	0.26 ± 0.12

(e.g., one week = 168 hours) are better for practical purposes as it allows the users to plan ahead, however, extending the prediction horizon further with a limited dataset would lead to a significant deterioration in prediction performance [14]. For the 6-hour and 48-hour prediction horizons, LSTM obtained the lowest error values



(a) Week 1 **(b) Week 2**

(c) Week 3 **(d) Week 4**

Figure 3: Weekly rolling horizon predictions for the hospital admission dataset

and outperform other algorithms. For the 24-hour prediction horizon, LSTM is the best performing algorithm for hourly data, while ARIMA performed better on the two-hourly version of the data. For the 12-hour prediction horizon, LSTM and ARIMA performed very closely on the hourly data, however, LSTM is better in terms of the NRMSE (0.59), and MAE (1.57) metrics. ARIMA generated the lowest error values in terms of all metrics on hourly data for the 12-hour prediction horizon. However, it is important to note that the performance values are very similar for RF, ARIMA and LSTM methods, whereas, naive baseline (and simple statistical methods) consistently performed worse, indicating the overall difficulty of the prediction task. It is also important to note that, the results for 2-hourly data are considerably better than those of hourly data, which shows that this simple data aggregation can improve the prediction performance. When compared to the results obtained from hospital admission dataset (see Table 4), ND/NRMSE values for 2-hourly dataset are relatively worse, however, not by a significant margin.

Figure 4 presents the predicted values by the LSTM and ARIMA algorithms, which are the best-performing ones, and the actual number of hourly arriving patients. Although the models capture the overall pattern of the hourly patient volume and obtained close results, they are both unable to capture the data trends. Similarly, Figure 5 shows the predicted values by the LSTM and ARIMA algorithms for the two-hourly version of the data. In practice, assuming that the medical personnel in the online health service frequently works in 2-hour shifts, these predictions can be used to make high level schedules. However, high error rates might prevent further adoption of these forecasting methods. Note that we expect an improved forecasting accuracy when the dataset is further aggregated (e.g., to daily arrivals). However, performing predictions with further aggregated data would not be useful for practical purposes (e.g., to determine the schedules for doctors in an online medical service), and hence omitted from our analysis.

5 THREATS TO VALIDITY

We carefully designed our numerical study with six different forecasting methods and two distinct datasets to ensure valid research outcomes. For the online health services dataset, for validation, we selected the last 25 batches consisting of 48-hour windows from the dataset. Prediction windows for 6-hour, 12-hour and 24-hour predictions are selected from this 48-hour window, so that the performance values would be comparable across different prediction windows. The same 25 batches were also used for the validation of 2-hourly predictions as well.

Table 5: Comparison of different algorithms on hourly and 2-hourly data for online health services data ("average error ± stdev of error" is provided)

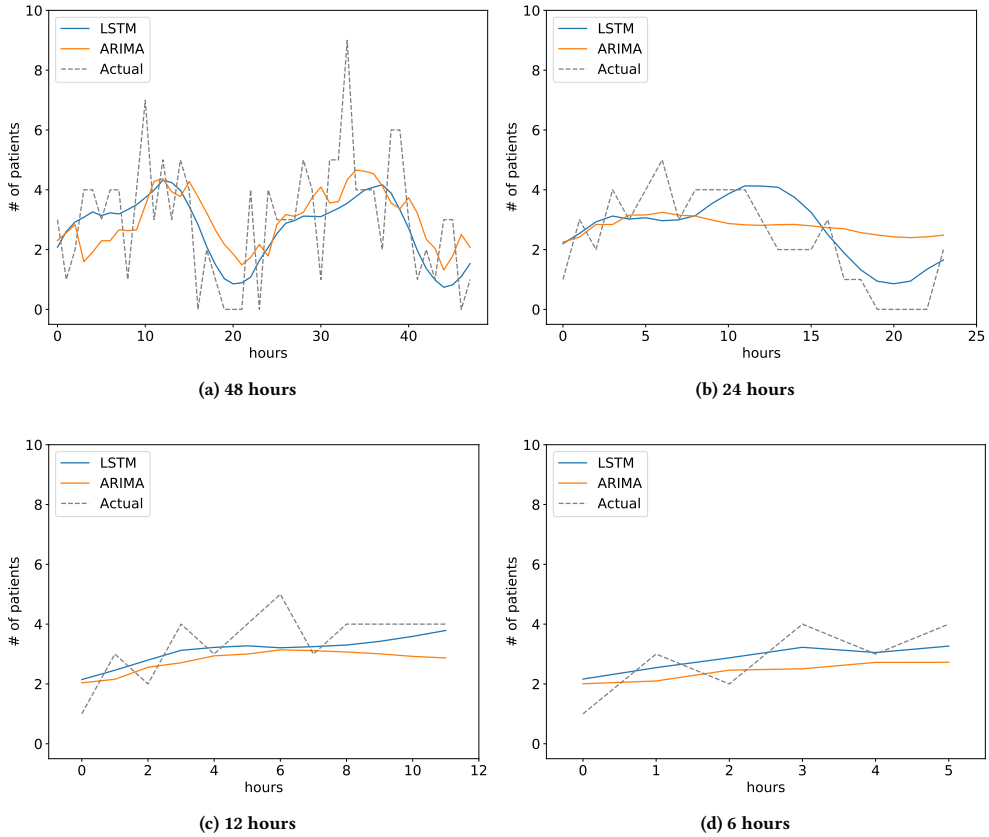
		Hourly			2-Hourly		
Model	Pred.	ND	NRMSE	MAE	ND	NRMSE	MAE
LSTM	6	**0.51 ± 0.26**	**0.62 ± 0.29**	**1.39 ± 0.53**	**0.43 ± 0.27**	**0.49 ± 0.3**	**2.03 ± 0.93**
RF	6	0.55 ± 0.28	0.66 ± 0.32	1.46 ± 0.61	0.46 ± 0.31	0.54 ± 0.35	2.14 ± 1.04
ARIMA	6	0.55 ± 0.26	0.66 ± 0.28	1.48 ± 0.48	0.47 ± 0.29	0.55 ± 0.32	2.26 ± 0.88
MA	6	0.53 ± 0.25	0.64 ± 0.27	1.44 ± 0.47	0.46 ± 0.27	0.53 ± 0.29	2.19 ± 0.92
ES	6	0.53 ± 0.21	0.67 ± 0.27	1.48 ± 0.54	0.52 ± 0.36	0.57 ± 0.37	2.39 ± 1.19
Naive Baseline	6	0.77 ± 0.33	0.91 ± 0.36	2.13 ± 0.71	0.63 ± 0.38	0.72 ± 0.40	3.01 ± 1.57
LSTM	12	**0.46 ± 0.13**	**0.59 ± 0.15**	**1.57 ± 0.41**	0.37 ± 0.16	0.44 ± 0.17	2.34 ± 0.8
RF	12	0.47 ± 0.15	0.60 ± 0.17	1.62 ± 0.41	0.37 ± 0.16	0.44 ± 0.17	2.35 ± 0.83
ARIMA	12	**0.46 ± 0.10**	0.60 ± 0.13	1.60 ± 0.42	**0.35 ± 0.14**	**0.43 ± 0.15**	**2.26 ± 0.73**
MA	12	0.48 ± 0.10	0.63 ± 0.13	1.69 ± 0.44	0.40 ± 0.10	0.51 ± 0.15	2.64 ± 0.75
ES	12	0.48 ± 0.11	0.64 ± 0.15	1.66 ± 0.35	0.40 ± 0.15	0.48 ± 0.18	2.57 ± 0.95
Naive Baseline	12	0.61 ± 0.18	0.76 ± 0.22	2.10 ± 0.50	0.45 ± 0.17	0.55 ± 0.19	2.89 ± 0.95
LSTM	24	**0.50 ± 0.09**	**0.66 ± 0.11**	**1.57 ± 0.31**	0.39 ± 0.11	0.48 ± 0.12	2.42 ± 0.7
RF	24	0.52 ± 0.12	0.67 ± 0.12	1.60 ± 0.26	0.40 ± 0.11	0.49 ± 0.11	2.47 ± 0.59
ARIMA	24	0.51 ± 0.08	0.66 ± 0.10	1.59 ± 0.25	**0.37 ± 0.08**	**0.47 ± 0.10**	**2.32 ± 0.56**
MA	24	0.56 ± 0.08	0.7 ± 0.09	1.74 ± 0.26	0.43 ± 0.08	0.54 ± 0.09	2.69 ± 0.63
ES	24	0.56 ± 0.07	0.72 ± 0.10	1.75 ± 0.28	0.45 ± 0.08	0.57 ± 0.09	2.84 ± 0.65
Naive Baseline	24	0.69 ± 0.16	0.87 ± 0.21	2.12 ± 0.38	0.51 ± 0.16	0.64 ± 0.19	3.15 ± 0.85
LSTM	48	**0.50 ± 0.06**	**0.65 ± 0.09**	**1.54 ± 0.21**	**0.37 ± 0.05**	**0.48 ± 0.09**	**2.33 ± 0.46**
RF	48	0.51 ± 0.06	0.66 ± 0.08	1.58 ± 0.17	0.44 ± 0.29	0.52 ± 0.33	2.06 ± 1.02
ARIMA	48	0.52 ± 0.06	0.67 ± 0.08	1.61 ± 0.21	0.38 ± 0.04	0.49 ± 0.08	2.36 ± 0.34
MA	48	0.52 ± 0.06	0.67 ± 0.08	1.62 ± 0.22	0.44 ± 0.05	0.57 ± 0.07	2.77 ± 0.44
ES	48	0.56 ± 0.05	0.72 ± 0.08	1.74 ± 0.22	0.48 ± 0.05	0.65 ± 0.04	2.82 ± 0.62
Naive Baseline	48	0.68 ± 0.08	0.88 ± 0.12	2.12 ± 0.25	0.52 ± 0.10	0.65 ± 0.12	3.19 ± 0.61

Each batch was evaluated using the metrics defined in section 3.3, and the final results provide the mean as well as the standard deviation of the results across the 25 batches. This allows us to capture the variance in the test set and get a true evaluation of our models.

We selected our parameters for the ARIMA models upon heuristic examination, and the seasonal components were selected after examining the seasonality trends. The ML models were allotted random states to avoid stochasticity, and all deep learning models were tuned based on batch size, learning rate, and the number of epochs. We note that further parameter tuning for the forecasting models might improve the predictive accuracy. In addition, while we explored many different models in our preliminary analysis, custom deep learning architectures and hybrid models might provide improved prediction performance for our forecasting task. Also, more datasets with different characteristics can provide new insights into the problem of forecasting patient arrivals to online health services. It is also important to note that some portion of our dataset coincides with spread of Covid-19, which potentially affected the patient arrival patterns to the particular healthcare application used as our data source. For instance, while we observe that there were some increase in patient arrivals at the beginning, there were also some fluctuations in demand due to recent increase in the competition in online health services area.

6 CONCLUSION

The purpose of this study is to provide an empirical analysis to compare commonly used forecasting methods for the task of predicting patient arrivals to online health services. We started our analysis by evaluating the employed forecasting methods over a closely related prediction task associated with hospital admissions,



(a) 48 hours **(b) 24 hours**

(c) 12 hours **(d) 6 hours**

Figure 4: Sample predictions for online health services data (hourly arrivals)

which was originally presented elsewhere. Our analysis with hospital admission dataset confirmed that high accuracy forecasts for patient arrivals might not be possible as the resulting normalized performance values (e.g., ND and NRMSE) are significantly worse than the ones obtained in other problems such as electricity or sales forecasting. Our results with online health services dataset showed that LSTM, ARIMA and RF models perform similarly, and ARIMA and LSTM models provide slightly better forecasts in certain data (e.g., hourly vs 2-hourly) and prediction horizon (e.g., 12-hour vs 24-hour) configurations. Moreover, we observed that aggregating the hourly data to 2-hourly data resulted in significant improvements in prediction accuracy.

The work presented in this paper can be further extended by collecting more data as well as performing a more detailed comparative analysis by including other forecasting models such as Poisson regression and convolutional neural networks. In addition, some external covariates such as weather information and economic indicators can be explored. The problem can also be treated as a multi-step time series classification problem, where patient arrivals can be classified into high, medium, and low categories. However, such an approach would require the dataset to be balanced in a manner that the time aspect of the class labels is preserved, as the class 'low' would have significantly high instances.

ACKNOWLEDGEMENTS

The authors would like to thank Your Doctors Online for providing funding and support for this research. This work was funded and supported by Mitacs through the Mitacs Accelerate Program.

REFERENCES

[1] Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, et al. 2020. GluonTS: Probabilistic and Neural Time Series Modeling in Python. *Journal of Machine Learning Research* 21, 116 (2020), 1–6.

[2] Joko Tri Atmojo, Wahyu Tri Sudaryanto, Aris Widiyanto, Arradini D Ernawati, and Dewi Arradini. 2020. Telemedicine, Cost Effectiveness, and Patients Satisfaction: A Systematic Review. *Journal of Health Policy and Management (2020)* (2020).

[3] Rashid Bashshur, Charles R Doarn, Julio M Frenk, Joseph C Kvedar, and James O Woolliscroft. 2020. Telemedicine and the COVID-19 pandemic, lessons for the future.

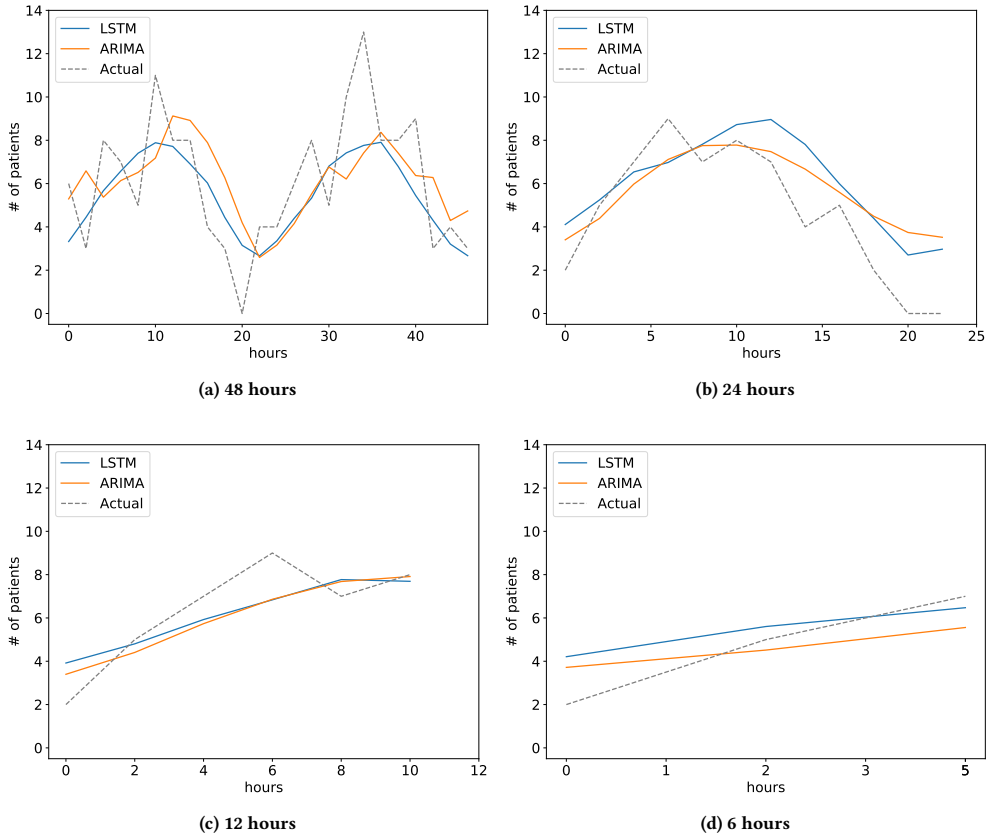
[4] Holly Batal, Jeff Tench, Sean McMillan, Jill Adams, and Phillip S Mehler. 2001. Predicting patient visits to an urgent care clinic using calendar variables. *Academic Emergency Medicine* 8, 1 (2001), 48–53.

[5] Joos-Hendrik Böse, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Dustin Lange, David Salinas, Sebastian Schelter, Matthias Seeger, and Yuyang Wang. 2017. Probabilistic demand forecasting at scale. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1694–1705.

[6] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.

[7] Leo Breiman. 1999. Random forests. *UC Berkeley TR567* (1999).

[8] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.



(a) 48 hours **(b) 24 hours**

(c) 12 hours **(d) 6 hours**

Figure 5: Sample predictions for online health services data (2-hourly arrivals)

[9] Avishek Choudhury. 2019. Hourly forecasting of emergency department arrivals: time series analysis. *arXiv preprint arXiv:1901.02714* (2019).

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[11] Everette S Gardner Jr. 1985. Exponential smoothing: The state of the art. *Journal of forecasting* 4, 1 (1985), 1–28.

[12] Yuxiu Hua, Zhifeng Zhao, Rongpeng Li, Xianfu Chen, Zhiming Liu, and Honggang Zhang. 2019. Deep learning with long short-term memory for time series prediction. *IEEE Communications Magazine* 57, 6 (2019), 114–119.

[13] Ni Huang, Zhijun Yan, and Haonan Yin. 2021. Effects of Online–Offline Service Integration on e-Healthcare Providers: A Quasi-Natural Experiment. *Production and Operations Management* (2021).

[14] Igor Ilic, Berk Gorgulu, and Mucahit Cevik. 2020. Augmented out-of-sample comparison method for time series forecasting techniques. In *Canadian Conference on Artificial Intelligence*. Springer, 302–308.

[15] Spencer S Jones, Alun Thomas, R Scott Evans, Shari J Welch, Peter J Haug, and Gregory L Snow. 2008. Forecasting daily patient volumes in the emergency department. *Academic Emergency Medicine* 15, 2 (2008), 159–170.

[16] Hye Jin Kam, Jin Ok Sung, and Rae Woong Park. 2010. Prediction of daily patient numbers for a regional emergency medical center using time series analysis. *Healthcare informatics research* 16, 3 (2010), 158.

[17] Kibaek Kim, Changhyeok Lee, Kevin O'Leary, Shannon Rosenauer, and Sanjay Mehrotra. 2014. Predicting patient volumes in hospital medicine: A comparative study of different time series forecasting methods. *Northwestern University, Illinois, USA, Scientific Report* (2014).

[18] Amaury Lendasse, Eric de Bodt, Vincent Wertz, and Michel Verleysen. 2000. Nonlinear financial time series forecasting-Application to the Bel 20 stock market index. *European Journal of Economic and Social Systems* 14, 1 (2000), 81–91.

[19] ZhiQiang Li, HongXia Zou, and Bin Qi. 2019. Application of ARIMA and LSTM in Relative Humidity Prediction. In *2019 IEEE 19th International Conference on Communication Technology (ICCT)*. IEEE, 1544–1549.

[20] Bryan Lim and Stefan Zohren. 2020. Time series forecasting with deep learning: A survey. *arXiv preprint arXiv:2004.13408* (2020).

[21] Melissa L McCarthy, Scott L Zeger, Ru Ding, Dominik Aronsky, Nathan R Hoot, and Gabor D Kelen. 2008. The challenge of predicting demand for emergency department services. *Academic Emergency Medicine* 15, 4 (2008), 337–346.

[22] Manfred Mudelsee. 2019. Trend analysis of climate time series: A review of methods. *Earth-science reviews* 190 (2019), 310–322.

[23] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2013. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026* (2013).

[24] Francesco Piccialli, Fabio Giampaolo, Edoardo Prezioso, David Camacho, and Giovanni Acampora. 2021. Artificial intelligence and healthcare: Forecasting of medical bookings through multi-source time-series fusion. *Information Fusion* (2021).

[25] Domenico Piccolo. 1990. A distance measure for classifying ARIMA models. *Journal of Time Series Analysis* 11, 2 (1990), 153–164.

[26] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2019. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* (2019).

[27] Vidya K. Sudarshan, Mikkel Brabrand, Troels Martin Range, and Uffe Kock Wiil. 2021. Performance evaluation of Emergency Department patient arrivals forecasting models by including meteorological and calendar information: A comparative study. *Computers in Biology and Medicine* (2021).

[28] Souhaib Ben Taieb, Gianluca Bontempi, Amir F Atiya, and Antti Sorjamaa. 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert systems with applications* 39, 8 (2012), 7067–7083.

[29] Lingling Zhou, Ping Zhao, Dongdong Wu, Cheng Cheng, and Hao Huang. 2018. Time series model for forecasting the number of new admission inpatients. *BMC medical informatics and decision making* 18, 1 (2018), 1–11.