Universiteit Utrecht

GP Post Call Wait Times Predictions Using LSTM

by

Yaqin Tian

8558299

Submitted to the Faculty of Utrecht University

In Partial Fulfillment of the Requirements for the degree of

Applied Data Science

Supervisors: Gerard Vreeswijk, Palina Salanevich

External Supervisor: Daan Ooms

July 2023

TABLE OF CONTENTS

ABSTRACT

This study investigates the application of Long Short-Term Memory (LSTM) networks to predict call waiting times at General Practitioner (GP) posts in the Netherlands, a critical facet of healthcare service quality and patient experience. The LSTM model showed a better performance over ARIMAX forecasting methods (Barteková, 2023), exhibiting a Mean Absolute Error (MAE) as low as 1.76 and a Root Mean Squared Error (RMSE) of 2.82 for one-day-ahead forecasts.

The findings underline the promise of LSTM networks in healthcare applications. Notably, the developed LSTM model could be integrated into web-based platforms to facilitate optimal healthcare resource allocation, thereby enhancing service quality and patient satisfaction. Furthermore, these results inspire confidence in the potential of deep learning methods to address complex forecasting problems in a variety of fields.

1. Introduction

Waiting to receive consultation from healthcare professionals can be quite stressful, especially in situations when time plays a vital role in urgent cases. In addition, a survey conducted by Hill and Joonas, involving 200 patients, revealed that the length of the wait time significantly influences patients' perceptions of the care quality and overall likeability of the healthcare provider’s office (Hill & Joonas, 2006). Further study shows that long wait time not only frustrates the patients but also decreases their personal productivity (Oostrom, Einav, & Finkelstein, 2017). Along the same line, another researcher sampled 5030 patients and measured the patient satisfaction ratings in the US and found that the “longer waiting times were associated with lower patient satisfaction (Anderson, Camacho, & Balkrishnan, 2007). Given these factors, having accurate wait time estimations becomes critical. With the wait time prediction, healthcare providers can proactively manage patient traffic and coordinate staff schedules. Patients, concurrently, can utilize the information to strategically plan their visits ensuring the healthcare resources are more available for genuine emergencies. Forecasting wait time predictions serves the interests for both the healthcare providers and patients. By setting better expectations and managing potential wait times, healthcare providers can enhance patient satisfaction, improve overall healthcare experiences, and ultimately contribute to better patient outcomes.

In the Netherlands, the General Practitioner (GP), is usually the first point of contact when someone has any health concerns before going to any specialized medical care (Schellevis, Westert et al.). GP plays a crucial role in the Dutch health-care system. Although not mandatory, the majority of citizens in the Netherlands are registered with a GP and benefit from the services provided from them. They ensure quality care for the general public and are expected to be easily accessible for the patients (Tikkanen et al., 2020). Outside of regular GP office hours, the “GP posts" are established by a group of GPs to serve the local communities. They aim to provide after-hours care, specifically from 5 pm to 8 am during weekdays and full 24 hours during the weekends (Uden et al., 2006). A single regional telephone number to contact the GP is available for the public. Patients are strongly advised to phone the GP posts prior to visiting in person. Patients who call the GP posts undergo a triage process facilitated by triage nurses (Uden et al., 2006). This process adheres to a protocol established by the Dutch Association of General Practitioners and is conducted under the supervision of a GP (Kool, Homberg, & Kamphuis, 2008). As such, it is common for patients to experience some wait time before they are being triaged and further consulted by a triage nurse for their symptoms.

Waiting time holds varying definitions across different contexts. In healthcare, it typically refers to the duration from when a health issue is first identified to when it is subsequently treated (Fogarty & Cronin, 2008). However, within a call center setting, "waiting time" represents the period before an incoming call is answered (Zhan & Ward, 2014). The focus of this thesis lies in examining the call waiting time related to the GP post in the Netherlands. Here, a patient's phone interaction can be broken down into three phases: (I) the span from when the patient makes the call until it is picked up, (II) the point from which the patient is welcomed and connected to the appropriate triage nurse, and (III) the actual consultation period with the nurse. The primary goal of this research is to predict the duration of the first phase: the waiting time from when a patient places a call until it is answered.

Currently, a considerable number of academic writings exist on wait time predictions at one stop service providers such as banks or post offices using advanced analytics and machine learning algorithms (Sanit-in & Saikaew, 2019). In a study conducted by Kyritsis & Deriaz (2019), they investigated the application of machine learning in banks to predict client waiting times, and its generalization in various industries. Furthermore, in the healthcare domain, the scholarly community tends to focus on exploring predictive analysis for emergency rooms (Arha, 2017) and outpatient visits (Lin et al., 2019). Several studies also emphasized on identifying the main cause of the wait time and potential interventions to address them (Park & Kwag, 2009). However, research directly related to the wait time predictions for the triage call system in General Practitioner (GP) posts or equivalent primary care settings are relatively sparse. We hope our work can contribute to this under explored area and bring insights for future work.

1.1 Research Question

Traditionally, queueing theory has been the go-to model for predicting wait times across various settings. This mathematical model is, however, anchored on several specific assumptions, such as the statistical distributions of arrival and service mechanisms, and the constant mean arrival rate over time (Adan & Resing, 2002). While these assumptions can simplify the modeling process, they may not always be reflective of real-world conditions, which often possess a more complex dynamism. As a potential solution, Kyritsis & Deriaz (2019) demonstrated the feasibility of machine learning as a practical substitute for queueing theory in estimating waiting times. This research underpins a shift towards more flexible methodologies.

Time series as defined by William Wei (2013), is a sequence of data points collected over time, and time series analysis involves analyzing these data points focusing on either their description or their potential for forecasting and inferences (Wei, 2013). Our study aims to utilize historical data for the prediction of future wait times, a task inherently based on a chronological sequence. Thus, we delve into state-of-the-art machine learning algorithms for time series analysis. As discussed in the research by De Gooijer and Hyndman (2006), the time series analysis encompasses a wide range of techniques, from linear approaches like ARIMA to nonlinear models such as the artificial neural networks (ANNs). While my colleague Barteková (2023) investigates the use of linear ARIMA models in predicting wait times on calls at the GP post, this study focuses on the application of the nonlinear approach which is expected to be advantageous when the data has an unknown functional relationship and are difficult to fit (Darbellay & Slama, 2000). Hence, we hypothesize that using the neural network approach may prove effective in predicting future wait times based on historical data. Given these considerations, we intend to compare the effectiveness of the linear regression prediction model, such as ARIMA (Barteková, 2023), with Recurrent Neural Network algorithms like LSTM. Thus, the focus of our inquiry narrows down to the following pivotal research question:

- Can the wait time on calls at the GP post be predicted using a LSTM model?

In translating our research question into a data science context, we approached the task by identifying which historical variables from the GP post call data can effectively be used to train a LSTM model, and then determining if the model can produce accurate forecasts for future wait times when tested on unseen data.

This thesis is organized into six main sections. The first section covers the context of the study and the research question. Secondly, we will review relevant academic literature, focusing on predictive analysis in healthcare, and the Long Short-Term Memory (LSTM) networks as one type of the Recurrent Neural Network (RNN). The third section describes the data, and the fourth section describes the method used including data preprocessing and fine-tuning of the model. In the fifth section, we present the results with performance metrics and compare findings. The final section provides a discussion of these results, highlighting their implications and limitations.

In adherence to ethical guidelines and to protect the privacy of the patients, all data used in this research do not contain any Personally Identifiable Information (PII) or Sensitive Personal Identifiable Information (SPII). The GP post remains anonymous, and access to the predictive modeling code is restricted and remains on the server of the host company to comply with data security and confidentiality protocols.

2. Literature Review

In this section, we investigate several key areas that formed the backbone of our case study: predicting call wait times for a GP Post in the Netherlands. Given that our case sits in the intersection of wait time predictions in healthcare and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, we will further look into the following topics. First, we examine the role and importance of predictive analysis in healthcare, considering its applications on both patient outcomes and system efficiency. Next, we focus on the specific domain of wait time predictions, highlighting the challenges and opportunities therein. We then offer a discussion on RNNs and their effectiveness in sequence prediction problems. Lastly, we will give an overview of the LSTM networks, a special type of RNN. By exploring these topics, we aim to provide a general overview of the current state of the art academic work and its relevance to our case.

2.1 Predictive Analysis in Healthcare and Wait Time Predictions

From student performance prediction (Albreiki et al., 2021) to demand forecasting for inventory planning (Tiwari et al.,2018), numerous areas have extensively leveraged the application of predictive analysis. As defined by Mishra & Silakari (2012), predictive analysis uses diverse statistical and analytical methodologies to construct models that can anticipate future events based on past data. In healthcare, the use of predictive analysis spans a range of use cases, from detecting brain tumors (Sapra et al., 2013) to assessing stroke risks associated with type 2 diabetes (Kothari et al., 2002). Beyond its direct applications in patient diagnosis and treatment, some research has also sought to enhance healthcare resource allocations to optimize patient experiences (Bates et al., 2018). For example, a bed management prediction model was developed by Kumar et al (2008) which helped the hospital planners to better anticipate bed demand based on historical bed occupancy data. Further, in recent years, machine learning has emerged as a powerful tool for modeling and predicting wait times in emergency rooms

(Gonçalves et al., 2019; Ameur et al., 2023), and outpatient clinics (Joseph et al., 2022; Li et al., 2021). Lin et al. (2021) utilized the data collected from a pediatric ophthalmology outpatient clinic and used several machine learning models, including random forest, elastic net, gradient boosting machine, support vector machine to predict wait time. Their study demonstrated the possibility to use machine learning models for improved predictions in outpatient clinics.

Given our case for GP posts is more similar to emergency departments due to the "after hours operating" nature, we took a closer look at urgent care wait time prediction. Kuo et al. (2020) investigated four machine learning algorithms for emergency department wait time prediction. The study revealed that these algorithms, compared to a baseline multiple linear regression model, reduced the mean square error by approximately 20%. In a similar context, Gonçalves et al. (2018) conducted a study using the Random Forest algorithm to predict the emergency waiting times, utilizing data from a Portuguese hospital and concluded on the high effectiveness of the algorithm.

2.2 Recurrent Neural Network and Long Short-Term Memory (LSTM) models

2.2.1 RNN

While traditional machine learning models have shown promising results on wait time predictions, the recent advancements in deep learning have shown to substantially improve prediction modelings. In a systematic review conducted by with Hewamalage el at. (2021), they confirmed that Recurrent Neural Network (RNNs) outperforms statistical benchmarks in many forecasting situations and a great option for practitioners. In the context of wait time prediction, Kyritsis & Deriaz (2019) developed a fully connected neural network that achieved a mean absolute error of 3.35 minutes for wait time predictions and confirmed that such algorithm can be used as an alternative to queueing theory for wait time prediction. In addition, Cheng & Kuo (2020) utilized Long Short-Term Memory (LSTM) recurrent neural networks to predict Emergency Department wait times predictions and found the model reduced the average mean error by 3 minutes compared to a linear regression model.

As illustrated by LeCun, Bengio, and Hinton (2015) with their noteworthy publication on deep learning, RNNs can be conceptualized as a special subclass of neural network with feedback loop-like recurrent units within its hidden states. The recurrent nature allows the RNN structure to be more suitable for sequential data.

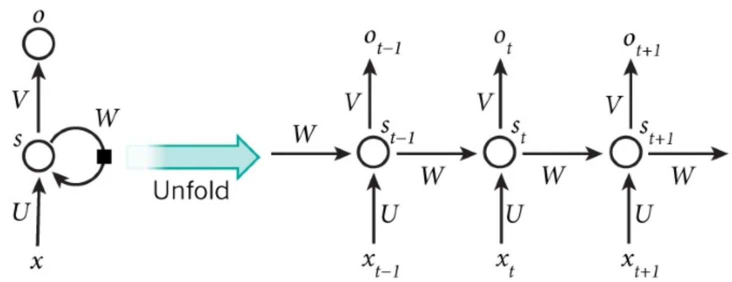


Figure 1 - taken from Deep Learning by LeCun, Bengio, and Hinton (2015).
X is the input; S is the hidden state and O is the output. t-1, t, t+1 are timestamps denoting the sequence. Xt is the input at time t, and Ot is the output at time t. St can be updated by all the information combined from the previous layer S_{t-1} and X_t of the current layer.

When we unfold the RNN, it can be taken as a deep forward neural network. The S_t can be calculated by applying f() on the weighted sum of these inputs noted as $f(WS_{t-1}+UX_t)$. The f() is known as the activation function. Common activation functions are Relu, Sigmoid(σ), and tangent hyperbolic function (Sharma et al., 2017). The values of W, U and V are weights and biases shared across all layers and they are iteratively optimized and learned using the gradient backpropagation method (Narendra & Parthasarathy, 1991). However, RNNs have been historically noted for their issues with gradient explosion and vanishing. In other words, as the network increases in time steps, the gradients tend to either expand excessively or diminish substantially through training (Bengio et al. 1994)

2.2.2 LSTM

The Long Short-Term Memory (LSTM) model, an evolution of Recurrent Neural Networks (RNNs), was designed to handle long-term dependencies and address the gradient problems inherent in RNNs (Hochreiter & Schmidhuber, 1997). In addition to a hidden state (h_t in Figure 2 below), LSTMs introduce a cell state (C_t) for managing and preserving long term memory throughout the learning process. LSTM also uses the gate mechanisms to decide which information to retain or discard. (f_t, i_t and o_t in Figure 2)

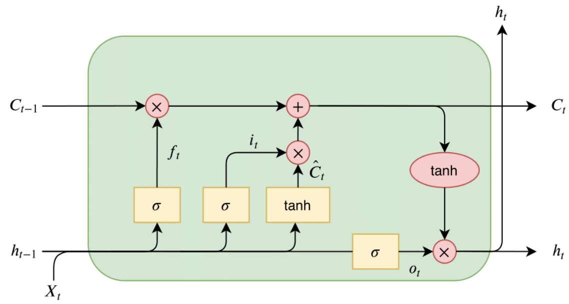


Figure 2: LSTM Unit (Ingolfsson, 2021)

The hidden state from the previous timestamp is h_{t-1} and the X_t from the current timestamp are combined before passing through gates with activation functions noted as Sigmoid(σ) and the tangent hyperbolic function (tanh) and with various weight (w) and bias (b) terms. They are calculated as:

Forget Gate (*ft*)*:* this gate operates based on the sigmoid function which ranges between 0 to 1. A sigmoid output near 1 signifies high retention of information, while a value close to 0 implies information should be discarded from the internal cell state (Ĉt*).*

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f)$$

Input gate(i_t): Similarly, this gate controls what information should be added to Ĉt with the sigmoid function, but with a different set of W_i and B_i.

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i)$$

The internal cell state (Ĉt), often referred to as the "candidate gate," uses the tanh function to normalize the combined data. This data consists of the previous combined hidden state (h_{t-1}) and the current input X_t data adjusted by corresponding weights and bias terms. The output from this process, due to the tanh function, falls between -1 and 1 and it represents the potential new information that could be added to the cell state. The extent of this addition is determined by the output from the input gate (i_t) by the point wise multiplication (red ⊗ in Figure 2)

$$\hat{C}_t = \tanh(W_C \cdot [h_{t-1}, X_t] + b_C)$$

The latest cell state C_t is updated from a combination of the previous cell state, C_{t-1}, regulated by the output from the forget gate (f_t), and the product of the input gate (i_t) with the previous cell state (C_{t-1}). It allows the model to determine what information to keep from the prior cell and what new information to be integrated from the current input.

$$C_t = i_t \cdot \hat{C}_t + f_t \cdot C_{t-1}$$

Output Gate (o_t): It goes through a similar mechanism with the sigmoid activation function and determines much information to pass down to the hidden state (h_t).

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o)$$

After the latest cell state C_t passes through the tanh activation function and multiplied by the results from O_t, the hidden state (h_t) is updated.

$$h_t = o_t \times \tanh(C_t)$$

Figure 3 shows how the aforementioned LSTM unit is chained together in the network sequences compared to the plain RNN structure.

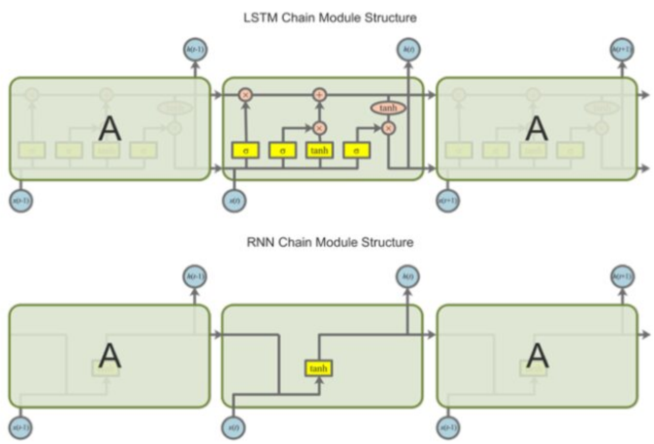


Figure 3 LSTM and RNN comparison (Zhang et al., 2020)

3. Data Description

The data we used in this study was collected from a local GP post office, and it consisted of two primary datasets - 'shift data' and 'call data'. The 'shift data' provided information about shift types and their respective start and end times over a few months. The 'call data', covering a longer duration, recorded various metrics, including line type (normal or emergency), call start and end times, answer time, wait time, call length, and urgency level as determined by the triagist. Some records in the 'call data' file contained missing values in the 'wait time' column.

Additionally, an external 'calendar data' was also used, providing detailed time information (month, day, hour, second), classification of days into weekdays or weekends, and indication of holidays.

4. Method

4.1 Data Preparation and Feature Selection

4.1.1 Data Preprocessing

We enhanced the dataset by adding the calendar data with the intention to build a multivariate model (Barteková, 2023). This strategy has been suggested for improving model prediction accuracy by Zhao et al., 2022 for their research on predicting hospital visits with deep learning models. Furthermore, for the practical application of our LSTM model, we mapped the shift schedule data onto the call data based on the start and end time intervals derived from the shift table. The merged dataset resulted in the addition of a "Number of Shifts" variable, reflecting the number of shifts per hour for overlapping records.

In our approach to handle missing values for the “wait time” column (Barteková, 2023), we first derived an estimated 'wait time' from the difference between the 'start call' and 'call answered' columns, utilizing this estimate to fill in the missing records. Subsequently, we compared this estimated 'wait time' with the recorded 'wait time' to compute the “delay”. After calculating the average delay value, we adjusted our initial estimates for the missing 'wait time' entries by this average delay, thereby refining our imputation process.

In addition, we identified any wait time values that were less than 1 second or exceeded 2 hours as outliers (Barteková, 2023). Given that these outliers did not indicate any underlying phenomena or provide additional insights that would be lost upon their removal, we decided to exclude them from our dataset.

Furthermore, we utilized 'start_call' as the foundation for our timestamps. However, considering these timestamps merely marked the occurrence of random incoming calls, we adjusted them to an hourly basis (Barteková, 2023). This adjustment was not only integral to facilitating model building and training, but also crucial in making the prediction outcomes more meaningful and interpretable for end users.

After these preprocessing steps, we conducted some exploratory analysis to understand the characteristics and patterns in our processed dataset better. Details of this exploratory analysis can be found in the appendix (Fig 10-12).

4.1.2 Feature Engineering

Feature engineering, as outlined by Verdonck et al. (2021), is an indispensable process in machine learning that deals with transforming raw data into a format that is both comprehensible and actionable for algorithms. It involves creating insightful features from raw variables that can better expose underlying patterns, thus enhancing the algorithm's predictive accuracy and interpretability.

Recognizing that the GP post only operates during specific hours and days, we created an 'open' variable (Barteková, 2023) that indicates whether the post is open or closed at a particular hour. To ensure the completeness of our time series data, we filled in the gaps in the timestamps for these hours when the GP post is closed, resulting in a record for every hour from March 1, 2023, to June 1, 2023. This additional feature provides contextual information that can help improve our model's understanding and prediction of the data.

Furthermore, to enhance data comprehension and representation, we converted the wait time from seconds to minutes. As a result of these feature engineering steps, we obtained a clean dataframe with 2232 records ready for model training.

The preprocessing and feature engineering stages of this research were significantly contributed by my colleague, Katarína Barteková. Her extensive work, including parallel research on a family of ARIMAX models (Barteková, 2023), played a crucial role in shaping our approach.

4.1.3 Data Split

The final dataset was then divided into training, validation, and testing sets. Specifically, as shown in figure 5, we allocated 55% of the data for model training, 10% for model validation during the training process, and the remaining 35% for final model evaluation. This rigorous preprocessing approach ensured that our LSTM model was trained on a robust, high-quality dataset, which ultimately enhances the reliability and accuracy of its predictions.

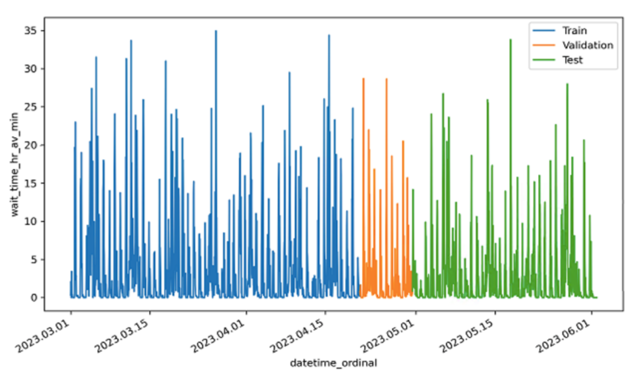


Figure 4: Distribution of wait time length, represented in minutes, per hourly interval highlighting the data split.

4.2 Model Development

4.2.1 Model Architecture

Guided by the findings from a study conducted by Yadav, A., Jha, C. K., & Sharan, A. (2020), which experimented with one to seven hidden layers for LSTM models, we opted for a single-layer architecture for our model. Yadav et al. found that a single hidden layer is generally sufficient, with additional layers needed only for exceptionally complex problems. Considering the nature of our dataset and the task at hand, a one-layer architecture seemed fitting for our study. The architecture of our LSTM model is described as follows:

- The **input layer** accepts a sequence of 5 time steps back from the past, each with 7 features. These 7 features were selected based on their relevance to wait time and the domain knowledge provided by the host company. These 7 features include the average hourly wait time in minutes, hour of the day, day of the week, the season, whether it's a holiday, whether the post office is open and the number of shifts scheduled for the hour. A detailed description of each can be found in the appendix (Table3).
- The one **hidden LSTM layer** has 50 units, also known as "neurons" that help capture and store the contextual information from the sequence data for the LSTM layer.
- The **output layer** has one neuron which provides the final output indicating the predicted average waiting time for that hourly interval.

4.2.2 Evaluation Metrics

Botchkarev (2018) reviewed the performance metrics in machine learning regression and forecasting in the past 25 years and highlighted the predominance of Mean Square Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) metrics being the most popular. Given that our predicted outcomes will contain values that are zero or near-zero, leading to infinitely large MAPE, we chose to exclude it from our analysis. Instead, we chose MSE as our model's loss function, which is highly sensitive to large errors and effectively penalizes and reduces them during training. Additionally, we used RMSE and MAE for performance evaluation. These metrics provide critical insights into the average magnitude of errors and are expressed in the same units as our target variable. Despite the availability of numerous metrics, our decision to use these three is justified by their widespread use in the field and their unique ability to provide better insight into prediction accuracy.

- **MSE:** the average of squared difference between the predicted $\hat{y}_i$ and actual values y_i (Marmolin, 1986) $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$
- **RMSE:** the square root of the MSE (Willmott & Matsuura, 2005)

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- **MAE:** the average of the absolute difference between model prediction y_i and actual value x_i (Willmott & Matsuura, 2005)

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

4.2.3 Hyperparameter Tuning

The model's hyperparameters were optimized using a random search of the Keras Tuner library. This process involved testing randomly selected hyperparameter combinations and evaluating them based on the model's loss function (Bakhashwain & Sagheer, 2021).

The learning rate and number of units (“neurons”) in each layer are crucial hyperparameters in training the neural network. The learning rate determines how much the weights are updated during each iteration of training. Too small can make the convergence too slow while too large can overshoot the optimal values, leading to the divergent training (Smith, 2017).

In their work, Siami-Namini, Tavakoli, & Namin (2018) define an epoch as the entire iteration through a given dataset for training. This process involves adjusting the model's weights to improve prediction accuracy and minimize cost, and is repeated multiple times. We ran our LSTM model for 50 epochs in search of the optimal combination of learning rate and units, with the aim of achieving the smallest MSE. The selected combination was then utilized for the final model construction. Figure 5 displays a scatter plot of the tuning process outcomes for the top 20 trials. A more detailed summary of the tuning results can be found in the appendix (*Table 4*).

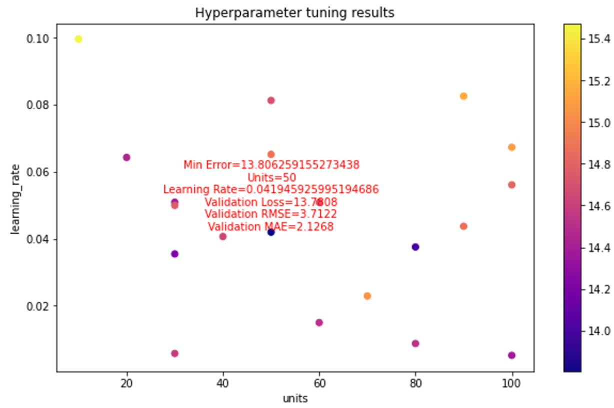


Figure 5 Hyperparameter Tuning Results Visualization

4.2.4 Model Evaluation

In this subsection, we evaluate the performance of our model using two key statistical metrics, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The results from this evaluation, as well as our interpretations of these results, are detailed here.

The LSTM model's performance was assessed using the test dataset, which comprised 35% of the total data obtained from the data splitting process. The test data was hidden during model training thus serves as a good indicator of the model's generalization capabilities. The resulting MSE from the loss function is 14.04, with an MAE of 2.19 and RMSE of 3.75. Given our target outcome is measured in minutes, these values are represented with two decimal places for practical significance. This level of precision provides ample detail for our purposes while ensuring the results remain meaningful and easy to interpret.

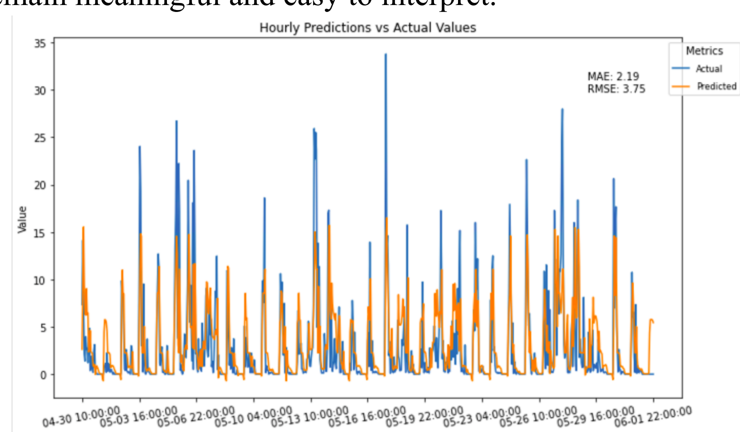


Figure 6 presents a comparison of the predicted hourly wait times and actual values in the test dataset.

5. Results and Discussion

5.1 Answering the Research Question

Our LSTM model with a single layer demonstrated an effective performance in predicting the call wait time at the GP post. The model exhibited a significant ability to capture the inherent complexity and temporal dependencies within the data. These findings indicate that the LSTM model is a promising deep learning approach for forecasting future wait times, offering practical implementation potential.

5.2 Results and Discussions

Our models, designed to predict wait times one day, one week, and one month in advance, aimed to support efficient staff scheduling at the GP post. These forecasting scenarios were simulated and evaluated against actual data up until June 1, 2023. For the one-day-ahead prediction, the model was trained with data up to 24 hours prior to June 1. Similarly, for a week-ahead forecast, the model was trained with data from 168 hours (24 hours * 7 days) prior, and for the one-month-ahead prediction, data from 730 hours (24 hours * 30 days) back were used.

The performance of each scenario, evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), is summarized in the *Table1* below:

Prediction Scenario	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Prediction vs Actual Visualization
One Day Into the future	1.76	2.82	*Fig 7*
One Week into the Future	2.52	4.15	*Fig 8*
One MonthInto the Future	2.20	3.80	*Fig 9*

Table 1 showing the performance metrics of each prediction

Figure 7 - One Day into the Future

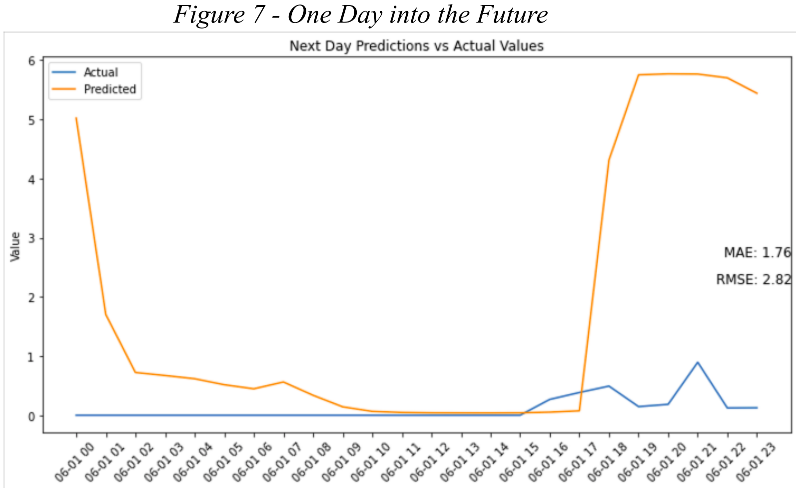


Figure 8 - One Week into the Future

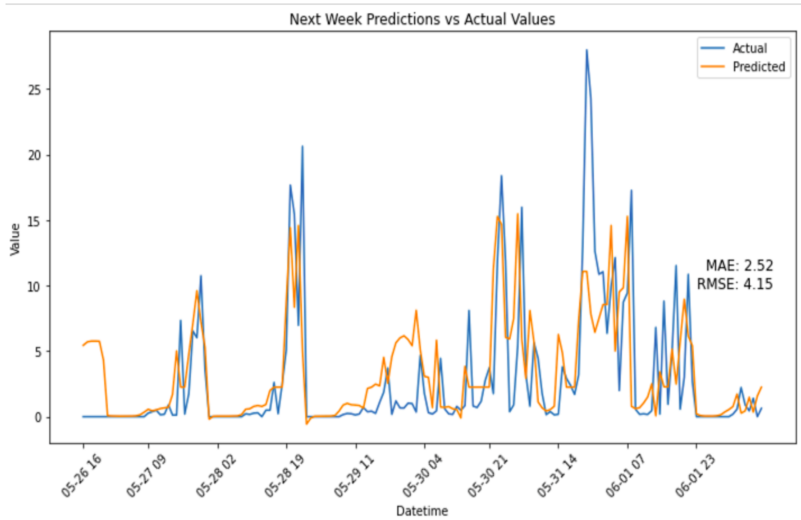
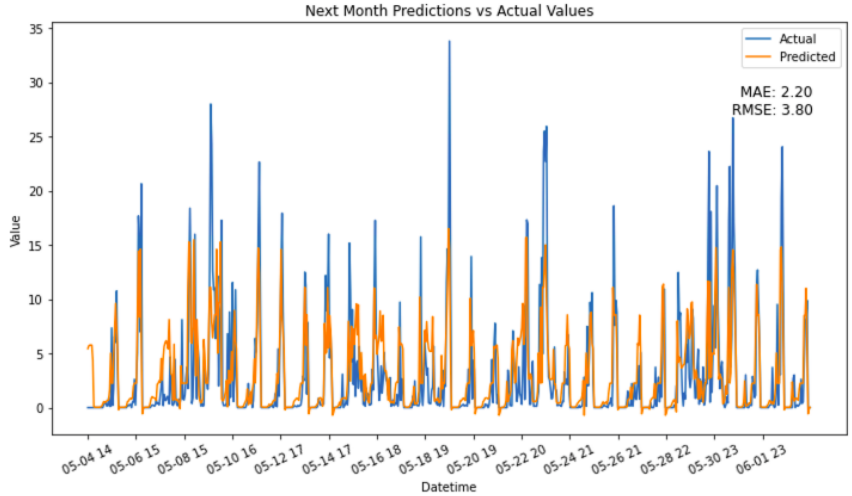


Figure 9 - One Month into the Future



5.3 Comparison of models

The same dataset for predicting wait time has also been explored using a variety of ARIMA models, with the detailed methodology and results available in Katarína's work (2023). The performance metrics of the LSTM models for the three scenarios mentioned previously are presented alongside those of the best performing SARIMA model detailed below (*Table 2*) for reference.

One Day into the Future		
Model	**Mean Absolute Error (MAE)**	**Root Mean Squared Error (RMSE)**
SARIMAX	2.95	3.17
LSTM	**1.76**	**2.82**
One Week into the Future		
Model	**Mean Absolute Error (MAE)**	**Root Mean Squared Error (RMSE)**
SARIMAX	3.65	4.77
LSTM	**2.52**	**4.15**
One Month into the Future		
Model	**Mean Absolute Error (MAE)**	**Root Mean Squared Error (RMSE)**
SARIMAX	3.96	5.36
LSTM	**2.20**	**3.80**

Table 2 performance comparison with SARIMAX

In this project, while the "number of calls" feature was collected historically along with other features and used in training a separate model (the details of which can be found in the appendix Fig 13-16), it was not included in the primary model presented here. This is due to the practical limitations in real-world scenarios: at the time of making future predictions (such as one day, one week, or one month into the future), the number of calls could be unknown.

6. Conclusion and Future Work

This study demonstrates the LSTM model's capabilities in predicting patient wait times at a GP post in the Netherlands. Notably, the 'emergency level' of calls wasn't used in training due to the dataset's hourly intervals and the varying types of emergencies within each hour. Including this variable could potentially refine wait time predictions, The model also limits to a single GP post and potential variance in performance could occur due to its distinct conditions across different locations.

In the pursuit of model improvement, a more comprehensive dataset especially if data collection expands both in terms of timeframe and number of GP posts, exploration of RNN variant with deeper architectures, as suggested by Zhao et al., 2022, could be potential directions. These enhancements might lead to better performance.

In conclusion, this study underscores the potential of LSTM applicability in healthcare resource planning. With further research and development, these models could be refined and extended to address a broader range of predictive scenarios, bringing us closer to improving healthcare operations management with more data-driven techniques.

APPENDIX

Author contributions:

Data provider: Esculine (external company), Data preprocessing (performed for original dataset, external dataset for comparison): Katarína Barteková, Feature Engineering: feature Number of shifts: Jackie Tian, all other features (time index, average waiting time per hour, season, week day, holiday, hour, minute, open/not open, number of incoming calls): Katarína Barteková, LSTM model development: Jackie Tian, ARIMA-family models: Katarína Barteková

Exploratory Data Analysis:

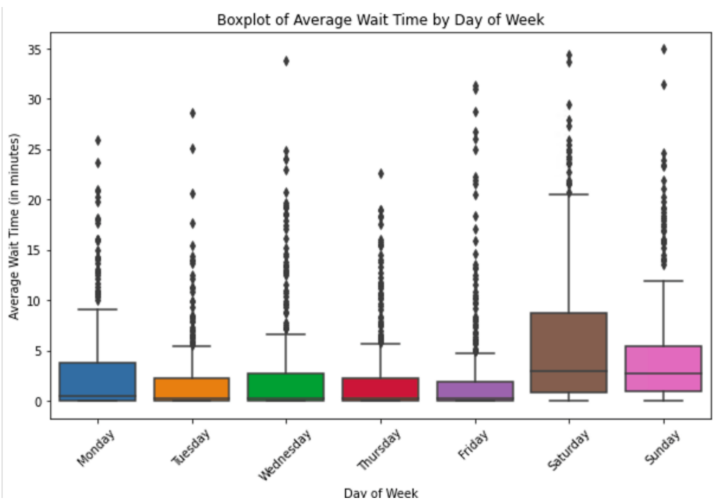


Figure 10

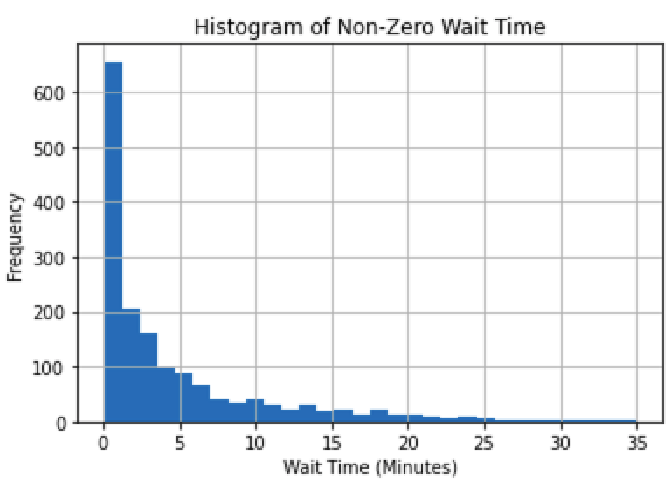


Figure 11

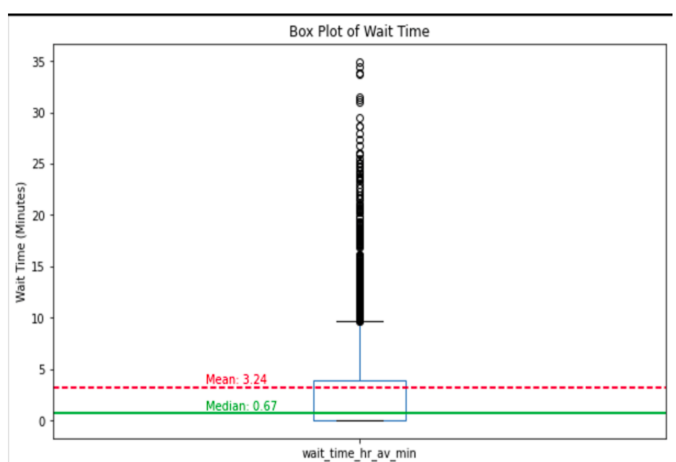


Figure 12

Employing a variety of visualization techniques, we examined the patterns and trends within the wait time data, our target variable, in the dataset that had undergone preprocessing. These visualizations are based specifically on the more recent three months' data. Figure 10 provides a boxplot, revealing the distribution of these wait times across different days of the week. It notably indicates that Saturdays are particularly busy.

Our histogram in Figure 11 reveals that most wait times generally cluster within the 0-5 minute range. Yet, despite the average wait time being approximately 3.24 minutes, we observe a wide spread in the data. This spread signifies a considerable degree of variability in wait times. The median wait time, more resistant to outliers, stands at 0.67 minutes. The contrast between the mean and median, combined with the broad spread visible in our boxplot (Figure 12), indicates the presence of outliers and a skewed distribution. These aspects were taken into account with the construction of our predictive model. Note, the statistics such as mean and spread reported in these figures are specific to this preprocessed, recent dataset.

Table 3: Detailed Description of the features:

Feature Name	Description
wait_time_hr_av_min	The average wait time in the corresponding hourly interval. This is our target feature.
week_day	Indicates what day of the week it is. 0 stands for Monday, 6 means Sunday
holiday	1 indicates holiday, 0 indicates non-holiday
hour	Denotes the hour during the 24-hour day period.
season	Indicates the season. Winter is denoted as 1.
open	Indicates if it's the operating hours for the GP post
number of shifts	Number of shifts scheduled for the hour
number of calls*	Shows the number of recorded incoming calls occurred in that hourly interval.

Table 4: Hyperparameter Random Search Summary Table

Trial Index	Units	Learning Rate	Score
1	50	0.041945926	13.80625916
2	80	0.037508816	14.00577354
3	30	0.035440103	14.2163868
4	100	0.005170737	14.3828907
5	30	0.050875152	14.39771843
6	20	0.064230593	14.45834446
7	60	0.014947735	14.51144648
8	80	0.00869646	14.5292449
9	30	0.005735792	14.58647633
10	40	0.040621178	14.6414485
11	50	0.081246716	14.69742632
12	60	0.050828452	14.6996789
13	30	0.049933156	14.75982189
14	100	0.056062704	14.82434988
15	90	0.043690736	14.86891747
16	50	0.065145506	14.89227343
17	70	0.022890692	15.05151224
18	100	0.067279485	15.09651375
19	90	0.082536401	15.16188622
20	10	0.099558951	15.47246981

Model with Calls

Figure 13: Performance on testing dataset:

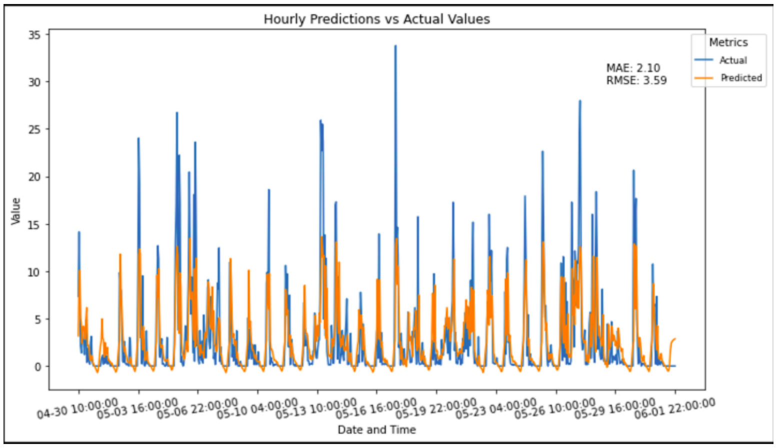


Figure 14: One day into the Future:

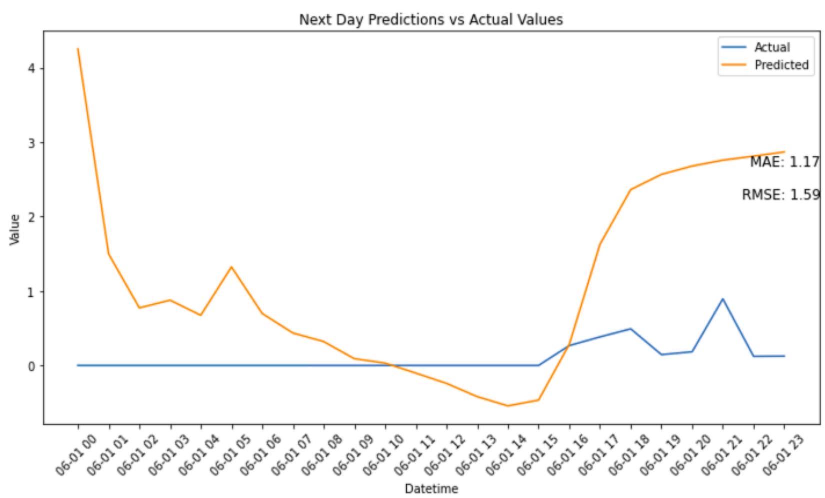


Figure 15: One Week into the Future

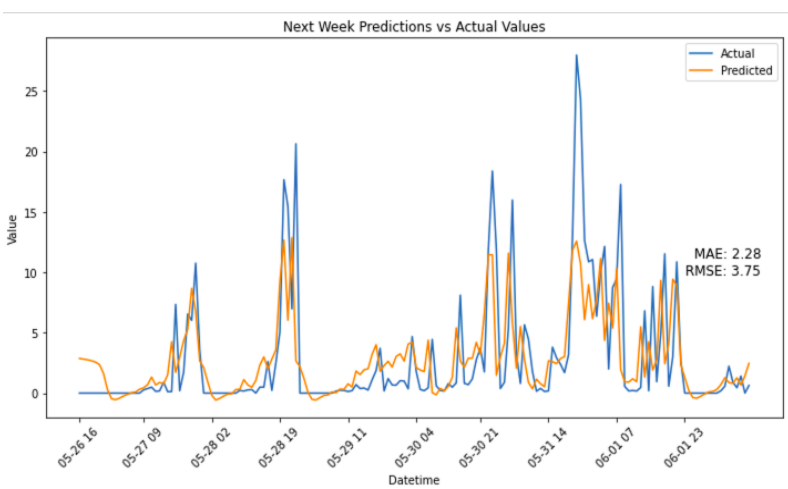
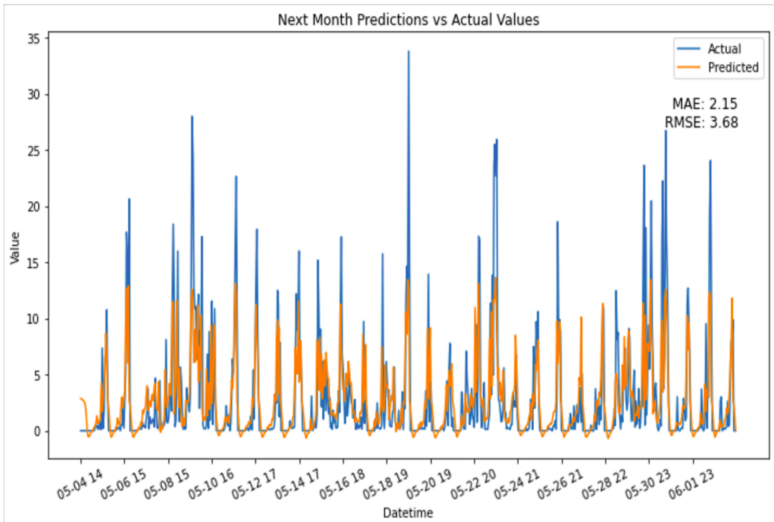


Figure 16:One Month into the Future:



Performance Comparison of the Model With the Number of Calls vs SARIMAX

Table 5: One Day into the Future:

Model	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
SARIMAX	1.60	1.80
LSTM	**1.17**	**1.59**

Table 6: One Week into the Future:

Model	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
SARIMAX	3.23	4.50
LSTM	**2.28**	**3.75**

Table 7: One Month into the Future:

Model	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
SARIMAX	3.55	5.08
LSTM	**2.15**	**3.68**

Packages Used in the Analysis:

Tensorflow: 2.7.0, Keras: 2.7.0, pandas: 1.2.4, numpy: 1.22.4, sklearn: 1.0.2

REFERENCES

Adan, I., & Resing, J. (2002). Queueing theory.

Arha, G. (2017). Reducing wait time prediction in hospital emergency room: lean analysis using a random forest model.

Ameur, N. B., Lahyani, I., Thabet, R., Megdiche, I., Steinbach, J. C., & Lamine, E. (2023). Predicting Patient's Waiting Times in Emergency Department: A Retrospective Study in the CHIC Hospital Since 2019. In Advances in Model and Data Engineering in the Digitalization Era: MEDI 2022 Short Papers and DETECT 2022 Workshop Papers, Cairo, Egypt, November 21–24, 2022, Proceedings (pp. 44-57). Cham: Springer Nature Switzerland.

Anderson, R.T., Camacho, F.T., & Balkrishnan, R. (2007). Willing to wait?: The influence of patient wait time on satisfaction with primary care. BMC Health Services Research, 7, 31 - 31.

Bakhashwain, N., & Sagheer, A. (2021). Online tuning of hyperparameters in deep LSTM for time series applications. International Journal of Intelligent Engineering and Systems, 14(1), 212-220

Bates, D. W., Heitmueller, A., Kakad, M., & Saria, S. (2018). Why policymakers should care about "big data" in healthcare. Health Policy and Technology, 7(2), 211-216.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), 157-166.

Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. arXiv preprint arXiv:1809.03006.

Cheng, N., & Kuo, A. (2020). Using Long Short-Term Memory (LSTM) Neural Networks to Predict Emergency Department Wait Time. Studies in health technology and informatics, 272, 199-202.

Darbellay, G. A., & Slama, M. (2000). Forecasting the short-term demand for electricity—Do neural networks stand a better chance?. International journal of forecasting, 16(1), 71-83.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. International journal of forecasting, 22(3), 443-473.

Fogarty, C., & Cronin, P. (2008). Waiting for healthcare: a concept analysis. Journal of advanced nursing, 61(4), 463-471.

Gonçalves, F., Pereira, R., Ferreira, J., Vasconcelos, J. B., Melo, F., & Velez, I. (2019). Predictive analysis in healthcare: Emergency wait time prediction. In Ambient Intelligence–Software and Applications–, 9th International Symposium on Ambient Intelligence (pp. 138-145). Springer International Publishing.

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. International Journal of Forecasting, 37(1), 388-427.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688.

Ingolfsson, T.M. (2021) Insights into LSTM architecture, Thorir Mar Ingolfsson. Available at: https://thorirmar.com/post/insight_into_lstm/ (Accessed: 03 June 2023).

Barteková, K. (2023). Predicting Patient Waiting Times at an On-Call Emergency Line at a Dutch GP Office Using ARIMA-family Time Series Forecasting (Unpublished master's thesis). [Utrecht University].

Kool, R.B., Homberg, D.J. & Kamphuis, H.C. Towards integration of general practitioner posts and accident and emergency departments: a case study of two integrated emergency posts in the Netherlands. BMC Health Serv Res 8, 225 (2008). https://doi.org/10.1186/1472-6963-8-225

Kuo, Y., Chan, N.B., Leung, J., Meng, H.M., So, A.M., Tsoi, K.K., & Graham, C.A. (2020). An Integrated Approach of Machine Learning and Systems Thinking for Waiting Time Prediction in an Emergency Department. International journal of medical informatics, 139, 104143.

Kyritsis, A.I., & Deriaz, M. (2019). A Machine Learning Approach to Waiting Time Prediction in Queueing Scenarios. 2019 Second International Conference on Artificial Intelligence for Industries (AI4I), 17-21.

Lin, W. C., Goldstein, I. H., Hribar, M. R., Sanders, D. S., & Chiang, M. F. (2019). Predicting wait times in pediatric ophthalmology outpatient clinics using machine learning. In AMIA Annual Symposium Proceedings (Vol. 2019, p. 1121). American Medical Informatics Association.

Marmolin, H. (1986). Subjective MSE measures. IEEE transactions on systems, man, and cybernetics, 16(3), 486-489.

Oostrom, T., Einav, L., & Finkelstein, A.N. (2017). Outpatient Office Wait Times And Quality Of Care For Medicaid Patients. Health affairs, 36 5, 826-832.

Park, C., & Kwag, E. (2009). A Case Study about Managing Waiting Time for Raising Customer's Satisfaction in the Medical Service.

Sanit-in, Y., & Saikaew, K. R. (2019). Prediction of waiting time in one stop service. International Journal of Machine Learning and Computing, 9(3), 322-327.

Sapra, P., Singh, R., & Khurana, S. (2013). Brain tumor detection using neural network. International Journal of Science and Modern Engineering (IJISME) ISSN, 2319-6386.

Schellevis, F.G., Westert, G.P. & De Bakker, D.H. The actual role of general practice in the Dutch health-care system. J Public Health 13, 265–269 (2005). https://doi.org/10.1007/s10389-005-0124-z

Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. Towards Data Sci, 6(12), 310-316.

Shetty, S. A., Padmashree, T., Sagar, B. M., & Cauvery, N. K. (2021). Performance analysis on machine learning algorithms with deep learning models for crop yield prediction. In Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2020 (pp. 739-750). Springer Singapore.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018, December). A comparison of ARIMA and LSTM in forecasting time series. In 2018 17th IEEE international conference on machine learning and applications (ICMLA) (pp. 1394-1401). IEEE.

Smith, L. N. (2017, March). Cyclical learning rates for training neural networks. In 2017 IEEE winter conference on applications of computer vision (WACV) (pp. 464-472). IEEE.

Tikkanen, R., Osborn, R., Mossialos, E., Djordjevic, A., & Wharton, G. A. (2020, June 5). Netherlands. The Commonwealth Fund. Retrieved June 13, 2023, from https://www.commonwealthfund.org/international-health-policy-center/countries/netherlands#ensuring-quality-of-care

Tiwari, S., Wee, H. M., & Daryanto, Y. (2018). Big data analytics in supply chain management between 2010 and 2016: Insights to industries. Computers & Industrial Engineering, 115, 319-330.

Uden, C. V., Giesen, P. H., Metsemakers, J. F., & Grol, R. P. (2006). Development of out-of-hours primary care by general practitioners (GPs) in The Netherlands: from small-call rotations to large-scale GP cooperatives.

Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2021). Special issue on feature engineering editorial. Machine Learning, 1-12.

Wei, W. W. S. (2013). Time series analysis. In T. D. Little (Ed.), The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis, Oxford Library of Psychology (Online ed.). Oxford Academic. https://doi.org/10.1093/oxfordhb/9780199934898.013.0022

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

Yadav, A., Jha, C. K., & Sharan, A. (2020). Optimizing LSTM for time series prediction in Indian stock market. Procedia Computer Science, 167, 2091-2100.

Zhan, D., & Ward, A. R. (2014). Threshold routing to trade off waiting and call resolution in call centers. Manufacturing & Service Operations Management, 16(2), 220-237.

Zhang, L., Yang, L., Gu, C., & Li, D. (2020). Lstm-based short-term electrical load forecasting and anomaly correction. In E3S Web of Conferences (Vol. 182, p. 01004). EDP Sciences.

Zhao, X., Lai, J. W., Ho, A. F. W., Liu, N., Ong, M. E. H., & Cheong, K. H. (2022). Predicting hospital emergency department visits with deep learning approaches. Biocybernetics and Biomedical Engineering, 42(3), 1051-1065.