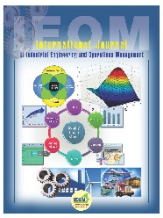
IEOM Society International

International Journal of Industrial Engineering and Operations Management (IJIEOM)
Volume 3, No. 1, October 2021
pp. 33 - 45



Predicting Patient Waiting Time in the Queue System Using Deep Learning Algorithms in the Emergency Room

Hassan Hijry
Department of Industrial Engineering
University of Tabuk,
Tabuk, 47512 Saudi Arabia
hhegri@ut.edu.sa, hhijry@oakland.edu

Richard Olawoyin
Department of Industrial and Systems Engineering
Oakland University, Rochester, MI, USA
olawoyin@oakland.edu
https://doi.org/10.46254/j.ieom.20210103

ABSTRACT

Many hospitals consider the length of time waiting in queue to be a measure of emergency room (ER) overcrowding. Long waiting times plague many ER departments, hindering the ability to effectively provide medical attention to those in need and increasing overall costs. Advanced techniques such as machine learning and deep learning (DL) have played a central role in queuing system applications. This study aims to apply DL algorithms for historical queueing variables to predict patient waiting time in a system alongside, or in place of, queueing theory (QT). We applied four optimization algorithms, including SGD, Adam, RMSprop, and AdaGrad. The algorithms were compared to find the best model with the lowest mean absolute error (MAE). A traditional mathematical simulation was used for additional comparisons. The results showed that the DL model is applicable using the SGD algorithm by activating a lowest MAE of 10.80 minutes (24% error reduction) to predict patients' waiting times. This work presents a theoretical contribution of predicting patients' waiting time with alternative techniques by achieving the highest performing model to better prioritize patients waiting in the queue. Also, this study offers a practical contribution by using real-life data from ERs. Furthermore, we proposed models to predict patients' waiting time with more accurate results than a traditional mathematical method. Our approach can be easily implemented for the queue system in the healthcare sector using electronic health records (EHR) data.

ARTICLE INFO

Submitted
15-Dec-2020

Revised
15-Dec-2020

Accepted
08-Jan-2021

KEYWORDS
Healthcare Management, Patient Priority, Waiting Time, Deep Learning, Queueing Theory

1. Introduction

Emergency rooms (ER) experience massive patient overcrowding at most hospitals because more than 50% of the patients who are admitted to the hospital come through the ER. The ER is a valuable part of hospitals, and as such, most sections require many resources due to long patient queues (Mor et al. 2015). In a setting such as healthcare, queueing is a risk factor because idle time can prove to be expensive for healthcare providers and unpleasant for patients. Moreover, it may also affect a patient's health conditions or life (Gupta and Denton 2008). Queuing theory (QT) is a traditional

mathematical technique that has been used to analyze queuing systems for decades (Gupta 2013). However, the traditional QT approach may not be sufficient in real life applications because the methodology is limited, for example, unrealistic assumptions of the time distribution it requires to do queueing analysis (Mahadevan 2015; Pianykh and Rosenthal 2015). Thus, alternative techniques, such as deep learning (DL) algorithms, are considered to significantly improve ER efficiency. DL algorithms are another class of machine learning method. Also, recent research reported that the methodology applied to predict the patient waiting time at ER has limited accuracy (Pak et al. 2020). Furthermore, DL algorithms can reduce human error and have even better accuracy compared with traditional methods (Shafaf and Malek 2019). The goal of this study was to deliver a novel and more accurate model for waiting time prediction and make an essential tool for reactive actions if ERs report long waiting times. This goal was motivated by high error rates in previous studies. The novel model produced in this study reduces error prediction compared to prior work on this topic.

From an applied perspective, a new algorithm was developed with DL to improve waiting time prediction accuracy for patients with low acuity using queueing predictor variables at the ER. The DL method was compared with traditional mathematical approaches. Realistic data was utilized from the triage monitoring system at an ER in Saudi Arabia which contained 30,909 patients between January and December of 2018.

Current studies have developed associations between the length of waiting time and the level of customer dissatisfaction (Abe 2019). Hence, they are encouraged to consider achieving adequate resource allocation to minimize waiting time in queues. Healthcare is one sector that is required to improve customer service in order to improve overall satisfaction and positive health outcomes. Extreme waiting times lead to poorer healthcare outcomes and are used as a measure of access to healthcare facilities (Liang 2010).

There are typically different techniques (e.g., mathematical analysis) to optimize queueing and resource utilization (Bittencourt et al. 2018). For example, queue models are frequently used to handle high demand by evaluating waiting times in hospital pharmacies and other multiple points of services. Similarly, other service industries that require security control, such as airports, also employ queueing models (Abe 2019). Moreover, waiting time in the queue is considered a measure of traffic control strategy performance. For instance, queuing delay is responsible for more than 90% of delays in travel time and traffic congestion at the airport (Peterson et al. 1995). Queuing models are also applicable in daily life where people wait in grocery stores or in restaurants for food. Studies have suggested that longer waiting times in any system may lead to higher consumption (Dong et al. 2019; Ülkü et al. 2020). When queues move slowly, it increases the waiting time and prominence thereof, and in turn requires the usage of more resources.

This study's contribution to the current literature is as follows: First, DL models were developed alongside or in place of queuing theory to predict waiting time in a queue using real data of the low patient acuity collected from electronic health records (EHR) at an ER in Saudi Arabia. Second, prediction error has been improved by implementing DL, resulting in a 24% reduction using MAE metric. Third, based on experiments performed in the research, this study provides a guideline for waiting time analysis in the queue—not only in healthcare, but also in other sectors, considering model understandability and the feature extraction process. We believe that the results will benefit practitioners and researchers who work on similar problems in different fields.

2. Literature Review

Previous research shows that long waiting times lead to patient frustration, anger, anxiety, and dissatisfaction (Curtis et al. 2018; Sun et al. 2000; Ward et al. 2017). Several studies have applied different methodologies to analyze ER waiting time predictions. For example, Kuo et al. (2020) combined machine learning with systems thinking to predict waiting time in ER. Stagge (2020) implemented a combination of approaches including machine learning and a simulation to predict patient waiting time; Arha (2017) used multi machine learning methods, such as Elastic Net and Random Forest, to predict the waiting time for low patient acuity in ER. Finally, Curtis et al. (2018) developed multiply machine learning algorithms including neural network to predict patient waiting time considering different predictors, such as patient arrival time, complete service time, and examination. Moreover, studies have designed predicting models to forecast the waiting time until treatment for patients with low acuity using algorithms, such as quantile regression (Pak, Gannon, and Staib 2020). Our study differs from the previous research on this topic because we implemented multi DL optimization algorithms to improve accuracy. Furthermore, we considered different predictors by extracting new parameters from the patient joining the queue (e. g., minute, hour, and day), waiting time in the queue, and departure time.

Many hospitals across the world face long waiting times and crowding in their ERs on a regular basis. The number of visits to ERs increases gradually every year in the United States (Di et al. 2015). In 2016, the National Center for Health

Statistics surveyed visits to the ER to be approximately 145.6 million annually visits (Kea et al. 2016). Not only have ER visits increased but waiting times in the ER have as well. For example, the Canadian Institute for Health Information reported in 2017 that the waiting times in ERs have noticeably increased since 2015. Evaluating the efficiency of ERs is a viable way to solve these problems (Rasouli et al. 2019).

Some hospitals are using queuing models to enhance optimal allocation of their staff by assessing patient arrival times (Kaushal et al. 2015; Sasanfar et al. 2020). A predicting model is becoming useful in the medical industry. Use of historical data to predict future patient waiting time is becoming a useful model for solving seasonal arrival and waiting times (Ruben et al. 2010; Cai et al. 2016). The data saved in EHR are the key to solve and examine healthcare problems that might contain hidden features. Other studies have focused on queuing system improvements in healthcare, mainly how it can be used in predictive models in the study of future behavior (Eiset et all 2019). Moreover, machine learning method has been applied in a study of projection of queuing behavior (Srivastava 2016; Stagge 2020). The two research studies rely on a predictive modeling approach through their research has been defective regarding time series analysis on queue data prediction. In the multi-hospital study conducted by Dong et al. (2019), historical waiting times were analyzed and findings showed that ER waiting time is one of the elements patients consider when deciding where to seek treatment services.

The previously presented information is helping to make operational decisions that result in waiting time and overcrowding reduction in the ER (Abir et al. 2019). Stintzing and Norrman (2017) compared artificial neural networks (ANN) as a prediction method with optimization through queueing theory in companies to predict queueing behaviors. The authors claimed that the results from ANN were positive and could be used for predicting the right amount of service each day. Various predicting techniques have attempted in queue analysis to optimize the waiting time in the queue (Moreno-Carrillo et al. 2019); however, our model with multi optimization algorithms can be applied in evaluating the low acuity patient waiting times in the ER. Consequently, the proposed model in this study can be used to provide insights to ER medical staff to determine patient waiting time in the queue by using EHR data.

3. Research Methodology

Deep learning techniques are implemented in this study to predict patient waiting time in queueing system alongside, or in place of, queueing theory (QT) using EHR data. Next, we compare the DL algorithms to find the best model with the lowest MAE. The model is presented, and a flowchart of the proposed methodology is shown in Figure 1. Each step is illustrated in the following subsections.

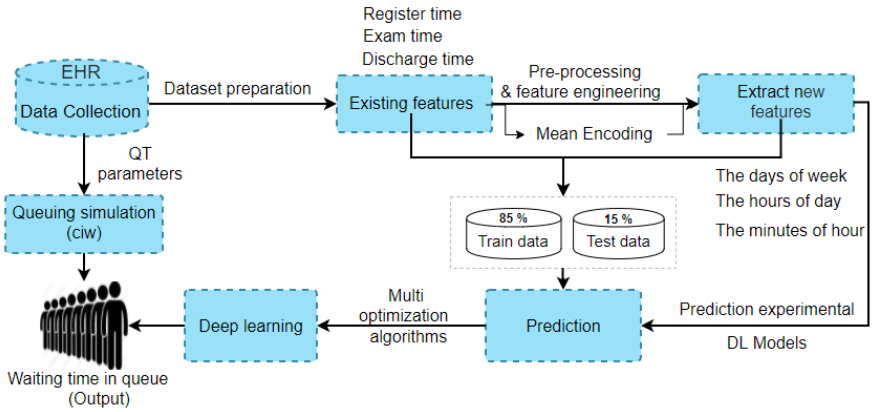


Figure 1. Proposed methodology for the study

3.1 Data Description and Preparation

The Triage Monitoring system is a national database established by the Ministry of Health of Saudi Arabia to ensure patient care quality. The data was extracted from the Triage Monitoring system, which includes data related to the formation and service of queues in hospitals at an ER between January and December of 2018. It recorded and monitored the patient flow time from when they registered until they left the hospital. These are the key referents for the application of machine learning, especially in the prediction of a new patient's waiting time upon joining a queue. These data included time of checking into the queue (arrival time/registration time), time taken in the queue (waiting time), time taken at server point (service time/exam by doctor), and sum of time taken in the entire system processes (length of stay).

The main data extracted from the EHR were entered randomly. Multiple steps were undertaken to clean, analyze, and finalize the data: **Step 1**, we converted our data using the arrival time/register time into weeks. **Step 2** used Step 1 data to create daily data. **Step 3**, the data were sorted based on arrival time to make the entries in a sequence of arrival. **Step 4**, missing and high values (manual data entered errors) in the data were excluded from our analysis. **Step 5** included only patients with level 3 through 5 acuities because more than 70% of the data that we collected included these types, which contained around 30,909 patients who were used in the training model after data cleaning. Also, irrelevant features, such as patient ID and names, were removed from the dataset.

The dataset reported one server (examination by doctor); hence, the triage service time was merged with waiting time (for time from arrival to first time being seen by doctor). Patients with these levels are considered less urgent or non-urgent. These types of patients do not need immediate care and are usually treated as first come first served. Therefore, the output variable in this model is the waiting time that the machine learning algorithms seek to predict. The mean waiting time in the dataset was 44.76 minutes, 39.0 minutes for median, and 20.23 minutes as standard deviation. Different input variables were analyzed, including service time, waiting time, and people waiting versus days of the week to give initial insights of our model data. The service time in our data is the time when the patient started getting treated by a physician until the treatment ended. Dataset (new features extracted) is used in this case to calculate patients in the queue and for every patient when they join the queue. To calculate the sum of people in the queue, every time patients departed from the queue, we found the sum of the waiting time and the arrival time before counting the number of people still in the queue when a new patient joined the queue. In machine learning, feature selection and data preparation are commonly used methods.

3.2 Pre-processing and Feature Engineering

The feature selection (selection of predictors) is an essential element in the machine learning model structure that determines the model's performance (Chandrashekar and Sahin 2014). There were main features extracted (e. g., minute, hour, and day) from the patient joining the queue in this study. Also, the patient's waiting time in the queue and leaving time were extracted. The following three are the main features:

1) Day was in the range of Monday (0) to Sunday (6).
2) Time in hours from 0 to 23rd hour.
3) Time in minutes starting at 0 minutes and continuing the 59th minute.

The categorical features were encoded using mean target encoding and extracting the new features from current features in the dataset. We adopted this method (feature extraction) as presented by Kyritsis and Michel (2019), which was applied in the bank. The mean target encoding was used to encode our data with the new features because it is a fast way to get most of the categorical variables encoded and gives higher cardinality features for regression problems (Pargent et al. 2019).

3.3 Prediction Experimental

The experiment on machine learning in this study used TensorFlow version 2.0.0.-beta1 and Python version 3.7.3. Also, different libraries were used to prepare and pre-process the data, such as Matplotlib, DateTime and Pandas. Accordingly, to validate and test the sensitivity of the model's performance, we split our dataset into two factions: the test set was 15%, and the training set was 85%. The test set was kept hidden throughout the training process. Moreover, by validating our model, it means that we used a test harness was used to give a fair estimation of the model's performance for making predictions on new data because it shows how sensitive the method is to applied data or new data that can be introduced to the model. Different optimization algorithms were used for this model in order to find the best with the lowest MAE. MAE is one of the metrics used to measure the machine learning model performance accuracy; it gives an idea of the magnitude of the error. It calculates all recorded means for the absolute errors by subtracting the prediction value from the actual value, as shown in the Eq. (1):

$$MAE = \frac{\sum_{i=1}^{n} abs(yi - \lambda(Xi)}{n} \quad (1)$$

We applied different optimizer algorithms, including Adam, Adagrad, RMSprop, and SGD for the iterative update of network weights based on our data training and to describe the math behind the algorithms; equations (2) to (12) below are cited and summarized from Ruder (2016). Stochastic gradient descent (SGD) optimization algorithm does not change during training for all weight updates and the learning rate and maintains a single learning rate (termed alpha). A learning rate is maintained for each network weight (parameter) and separately adapted as learning unfolds. In contrast SGD performs a parameter update for each training example $x^{(i)}$ and label $y^{(i)}$:

$$\theta = \theta - \eta \,.\nabla_\theta J(\theta; x^{(i)}; y^{(i)}) \tag{2}$$

Adam (adaptive moment estimation) is a combination of Root Mean Square Propagation (RMSprop) and momentum. It can also be used instead of the standard stochastic gradient descent procedure to update network weights iterative based on training data (Kingma et al. 2014). Adam can also be used like Adadelta and RMSprop when storing an exponentially decaying average of past squared gradient υ_t (Ruder 2016). Moreover, it keeps an exponentially decaying average of past gradients m_t, like momentum:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{3}$$

$$\upsilon_t = \beta_2 \upsilon_{t-1} + (1-\beta_2) g_t^2 \tag{4}$$

Where: m_t is estimates of the first moment (the mean) and υ_t is the second moment (the uncentered variance) of the gradients, respectively, hence the technique's name. m_t and υ_t are initialized as vectors of zero (as the authors of Adam observed that m_t and υ_t are biased towards zero, especially during the initial time steps and when the decay rates are low (e. g., β_1, and β_2 are close to one)). m_t and υ_t counteract these biases by computing bias corrected first and second-moment estimates:

$$\acute{m}_t = \frac{m_t}{1-\beta_1^t} \tag{5}$$

$$\acute{\upsilon}_t = \frac{\upsilon_t}{1-\beta_2^t} \tag{6}$$

Then, to update the parameters, we use these as shown in RMSprop, which yields the Adam update rule:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\acute{\upsilon}_t + \epsilon}} \acute{m}_t \tag{7}$$

RMSprop maintains per parameter learning rates which are adapted based on the average of recent magnitudes of the gradients for weight, such as how quickly it is changing. RMSprop is very effective but an unpublished, adaptive learning rate method proposed by Geoff Hinton in Lecture 6 slide 29 of his class (McMahan and Streeter 2014). In fact, RMSprop is identical to the first update vector derived from Adadelta, which is an extension of AaGrad optimization algorithm:

$$E[g^2]_t = 0.9E[g^2]_{t-1} + 0.1g_t^2 \tag{8}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t \tag{9}$$

Where: $E[g^2]_t$ *is* the decaying average over past squared gradients. Adaptive Gradient (AdaGrad) maintains a per-parameter learning rate that improves performance on problems with sparse gradients, such as computer vision and natural language problems (Brownlee 2020). For each parameter θ_i at every time step *t,* AdaGrad uses a different learning rate. Frist, AdaGrad's per-parameter is updated, which then is vectorized, for brevity; $g_{t,i}$ is set to be the gradient of objective function w.r.t. to the parameter θ_i at time step t:

$$g_{t,i} = \nabla_{\theta_t} J(\theta_{t,i}) \tag{10}$$

Stochastic gradient descent updates for each parameter θ_i at each time step *t* then becomes:

$$\theta_{t+1,i} = \theta_{t,i} - \eta . \theta g_{t,i} \quad (11)$$

AdaGrad modifies, in its update rule, the general learning rate η at each time step t for every parameter θi based on the past gradients that have been computed for θi:

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} \cdot g_{t,i} \quad (12)$$

Where: $G_t \in \mathbb{R}^{dxd}$ is a diagonal matrix where each diagonal element i, i is the sum of the squares of the gradients w.r.t. θi up to time step t^{11}. And $\in$: Is a smoothing term that avoids division by zero (usually on the order of $1e{-}8$). Then, The Rectified Linear Unit (ReLU) was used in hidden layers. The ReLU activation function is a linear function that will output the input directly if it is positive (x); otherwise, it will output zero. (that is, if it receives any negative output it will return zero. (Hara et al. 2015)). It is used in this model because it achieves better performance and is more comfortable to train when compared with other optimization functions (e.g., Sigmoid Function). ReLU can be written as Eq. (13):

$$f(x) = \max(0, x) \quad (13)$$

4. Results

In this study, we aimed to apply a DL approach with queueing theory. DL is one of the machine learning methods based on artificial neural networks. DL's power is the libraries built, such as Keras, which help to create extensive networks quickly and easily. Also, the simulation model for the queueing system was built to compare with the DL model using the Ciw library. Ciw is a discrete event simulation (DES) library supported by Python for queue networks (Palmer et al. 2019).

4.1 DL Models

Keras library by Python was used to apply DL mode and was trained with four input visible layers, 25 neurons for the first hidden layer, 18 neurons in the next hidden layer, and one output in the output layer. After 150 epochs of model training. Figure 2 shows the model predicted average waiting time against actual waiting time for the best outperform optimization algorithms (SGD). The blue color represents the real (actual) waiting time, and the orange color represents the predicted waiting time. It shows the predicted waiting time as being closest to the actual waiting time. The idea of what score a good/poor model can achieve only makes sense when it is interpreted in the situation of the skill scores of other models and trained on the same data. For this purpose, different optimization algorithms are compared and trained on the same data.

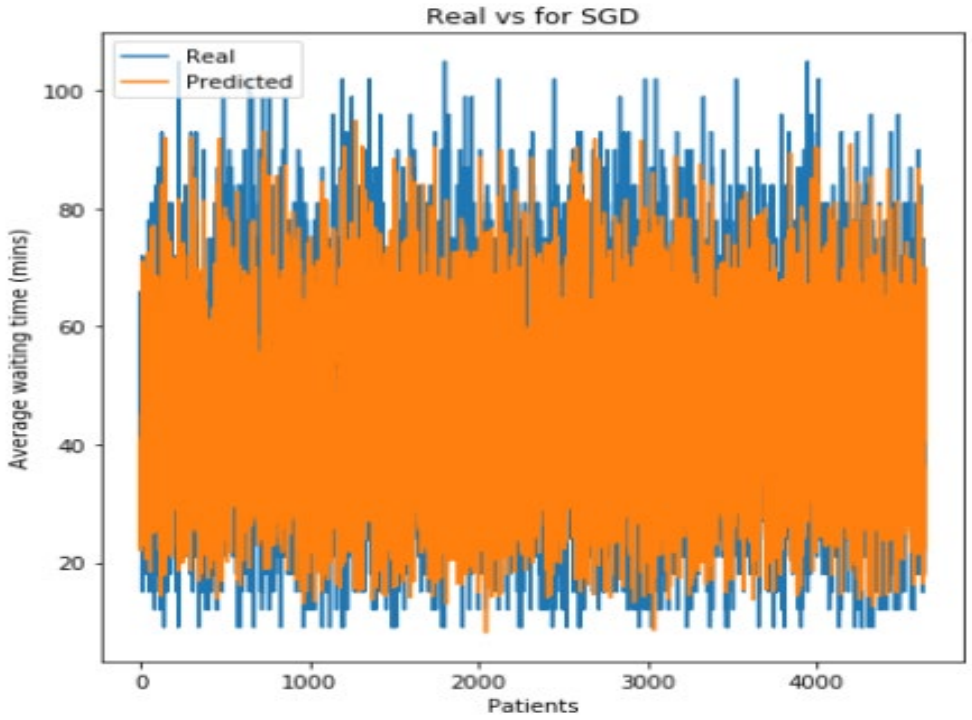


Figure 2. Waiting time predicted vs. actual for SGD algorithms

In our experiment, the four optimization algorithms are listed in order from high to low MAE in Table 1. The stochastic gradient descent (SGD) had the lowest MAE with 10.80 minutes, followed by RMSprop, and then Adam and AdaGrad optimization algorithm with around 12 minutes. After exhaustive tuning of all related hyperparameters used in the model, [25-18-1] values of the architecture were found to be suitable for the neural network for this model.

Table 1. Summary of the DL model (MAE results)

Optimization Algorithms	Network Architecture	Mean Absolute Error
AaGrad	[25-18-1]	12.78 minutes
Adam	[25-18-1]	11.16 minutes
RMSprop	[25-18-1]	11.14 minutes
SGD	[25-18-1]	10.80 minutes

On the plot of loss, the model has comparable performance on both training and validation datasets for all optimization algorithms as shown in Figure 3. The loss on both datasets may use this as a sign to stop training at an earlier epoch if these parallel plots start to depart consistently. Also, it shows the comparable skill on both train and validation datasets between the different optimization algorithms. The goal of using different optimizer algorithms is to change the attributes of our DL model, such as weights, learning rate, and to reduce the losses and reach the lowest MAE.

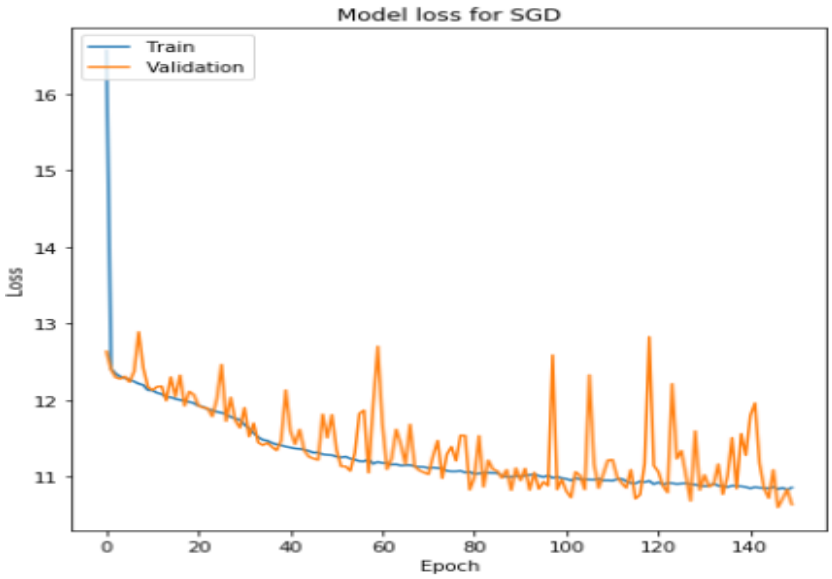


Figure 3. Plot of model loss on training and validation dataset

4.2 QT Models

The classic approach of queue theory was used to simulate the model in this study. The arrival rate (λ) and the service rate (μ) have been calculated using the same data applied in the DL model. Data was used for one day as shown in (Appendix A Table A), and is assumed the probability distribution of service time as an exponential distribution. The number of arriving patients per unit of time follows the Poisson distribution. Because patients with level 3 through 5 acuities were used in this study, the queueing model of M/M/1 system for simulating the analysis was used to reflect the ER system. The model initially runs for 1,440 minutes (one day). To ensure that the simulation reflects reality, the model runs for ten simulations in a loop and takes warm-up time and cool-down times of 100-time units. Different seeds were considered every time, so each trial yielded different results. Then, to achieve a more confident answer, the mean effect was taken over the trial results (68.24, 87.07, 59.377, 61.08, 63.64, 63.01, 70.75, 88.66, 63.64, 58.29 in minutes). Consequently, with each trial, the model ran for one day + 200 minutes (1,640 minutes). Patient mean waiting time resulting from the simulation model was 58.29 minutes, and the service time was 53.27 minutes. Comparing QT, the results to DL models in the dataset, the mean waiting time was about 44.74 minutes which is close to DL models prediction.

5. Discussion

Urgent and stochastic processes in the ER establish challenges to waiting time prediction. For example, the ER provides high acuity patients with effective emergency care but is typically not as efficient with patients seeking attention for non-urgent ailments. This leads to increased ER occupancy. While non-urgent patients wait in the ER, patients requiring highly urgent attention bypass waiting times, which may increase the waiting times for those who are non-urgent. Also, patients with non-urgent cases may vary from case to case due to patient-level attention needs (e.g., level, 3 to 5). However, it was established that low acuity care has an significant impact on overall ER waiting and service times for high acuity patients (Arha 2017).

The goal of this study was to deliver a more accurate model for waiting time prediction and create an essential tool for reactive actions if ERs report long waiting time. For instance, this model compares to other similar models for predicting waiting time in ER. Kuo et al. (2020) developed models with a mean-square error accuracy between 0.15 to 0.22. The model included different significant variables, such as Patient's triage categories, arrival time, number of doctors (within three hours of the Patient's arrival), number of patients in a queue (for triage, consultation, and departure upon the Patient's arrival). The proposed model, Kuo et al.'s model is limited by implementing a local triage system (Hong Kong). Also, a regional triage system by Arha (2017) estimated patient waiting time in an ER in Tennessee using a simple regressions model. Arha used similar predictor variables (e.g., time of day, day of week, and month of year), and a mean

square error as predictive accuracy (Arha 2017). To calculate this model variable and compare it with the proposed model requires collected clinical data. Pak et al. (2020) also developed a waiting time prediction model for low acuity patients assigned to the waiting room with an overall accuracy of 20% mean squared prediction error; the proposed models with SGD and RMSprop algorithms reduced the prediction errors by 24% compared to model improvement in Pak et al. (2020).

This study has some limitations, including data availability; not all ER information was included, such as a patient type of injury, X-ray process time, and laboratory test time. What is available in the dataset was extracted. Also, DL is known as data hunger; in this case, data was collected for only one year. As shown in the results, 10.80 minutes was reached as the lowest MAE, but this could decrease if the amount of data increases. In the experiment, 30,909 patients (level 3 through 5 acuities) were used in training after removing other levels (level 1 to 2 acuities) and missing data. Significant improvement was shown in waiting time prediction with available data when compared with a predicted average waiting time. Also, the model is simple enough to be implemented into an EHR system using relative information. The second limitation is the patient levels in the local triage system (assigned as levels 1 to 5) may differ geographically. For example, this data, levels 3 to 5 were set as low urgent, and levels 1 and 2 as high critical, but this may be different in other ER triage systems globally. The third limitation is this is a single location study, which could potentially impact the accuracy of the model, requiring more work to validate the model by using data from other ERs in different locations and with different populations.

6. Conclusion

This paper proposed a novel model to improve the accuracy of waiting time prediction for low acuity patients using DL techniques and ER data. The study used historic queueing variables to predict patient waiting time in a queuing system alongside, or in place of, traditional approaches (queueing theory). The traditional methods may not be sufficient in real life applications due to the limitations of the method, such as unrealistic assumptions of the time distribution required to do queueing analysis.

In the current literature, research reported that the methodology applied to predict patient waiting time in ERs has limited accuracy. Furthermore, DL algorithms can reduce human error and achieve better accuracy, when compared with traditional methods. Thus, alternative techniques, such as DL algorithms, must be used to significantly improve ER efficiency. For this purpose, a novel model for waiting time prediction was created and as an essential tool for reactive actions if ERs report long waiting times. Furthermore, four optimization algorithms, including SGD, Adam, RMSprop, and AdaGrad, were compared to find the best accuracy considering MAE metrics. Also, algorithms were compared with traditional mathematical approaches and data was utilized from the triage monitoring system in Saudi Arabia. The results showed that the DL model achieved better prediction accuracy than the traditional approach. Moreover, the novel model produced in this study resulted in a 24% error reduction when compared to prior work on this topic. The theoretical contribution of this paper is to predict patient waiting times with alternative techniques by achieving the highest performing model to better prioritize patient waiting in the queue. Also, this study offers a practical contribution by using real-life data from ERs. Furthermore, model have been proposed to predict patient waiting times with more accuracy than traditional mathematical models.

Future and extended work of this research could be as follows: more information from EHR could be implemented to the model such as different queueing predictor parameters. Moreover, different datasets from other hospitals and locations could be implemented. The service time of patients with the same acuity levels could be predicted. In addition, different machine learning algorithms could be applied to this model including linear and nonlinear regression. The model could be implemented on similar problems in different fields or sectors, including services and customer queueing. As part of future work, the model could be deployed as a web application to allow patients to join the queue prior to using EHR data.

Acknowledgements

The authors would like to acknowledge the Safety Engineering and Applications Laboratory (SEAL), School of Engineering and Computer Science (SECS), Oakland University for helpful comments and insights in this research work.

References

Abe, Y., Designing educative passenger journey by utilizing queueing and waiting times, *Masters Theses* Available: https://www.theseus.fi/handle/10024/265246, 2019.

Abir, M., Goldstick, J. E., Malsberger, R., Williams, A., Bauhoff, S., Parekh, V. I., Steven, K., and Jeffrey, S., Evaluating the impact of emergency department crowding on disposition patterns and outcomes of discharged patients, *International Journal of Emergency Medicine,* vol. 12, no. 1, pp. 1-11, 2019.

Arha, G., Reducing wait time prediction in hospital emergency room: lean analysis using a random forest model. *Masters Theses*, Available https://trace.tennessee.edu/utk_gradthes/4722/, 2017.

Bittencourt, O., Vedat, V., and Morty, Y., Hospital capacity management based on the queueing theory, *International Journal of Productivity and Performance Management*, vol. 67, no. 2, pp. 224-38, 2018.

Brownlee, J., Gentle introduction to the adam optimization algorithm for deep learning. machine learning mastery. Available: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/, 2020.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers and Electrical Engineering, 40(1), 16–28.

Cai, X., Oscar, P., Enrico, C., Fernando M., Richard D., David R., and Blanca G., Real-time prediction of mortality, readmission, and length of stay using electronic health record data, *Journal of the American Medical Informatics Association,* vol. 23, no. 3, pp. 553-61, 2016.

Chandrashekar, G., and Ferat, S., A survey on feature selection methods, *Computers and Electrical Engineering*, vol. 40, no. 1, pp.16-28, 2014.

Curtis, C., Chang, L., Thomas, J. B., and Oleg, S. P., Machine learning for predicting patient wait times and appointment delays, *Journal of the American College of Radiology*, vol. 15, no. 9, pp. 1310-1316, 2018.

Dong, J., Elad, Y., and Galit, B. Y., The impact of delay announcements on hospital network coordination and waiting times, *Management Science*, vol. 65, no. 5, pp. 1969-1994, 2019.

Di S. S., Paladino, L, V., Lalle, I., Magrini, L., and Magnanti, M., Overcrowding in emergency department: an international issue, *Internal and emergency medicine*, vol. 10, no. 2, pp. 171-175. 2015.

Eiset, A. H., Hans, K., and Mogens, E., Crowding in the emergency department in the absence of boarding - a transition regression model to predict departures and waiting time, *BMC Medical Research Methodology,* vol. 19, no. 1, pp. 68, 2019.

Gupta, D., *Queueing Models for Healthcare Operations*, handbook of healthcare operations management, Springer New York LLC, vol. 184, pp. 19–44, 2013.

Gupta, D., and Brian, D., Appointment scheduling in health care: challenges and opportunities, *IIE Transactions*, vol. 40, no. 9, pp. 800–819, 2008.

Hara, K., Daisuke, S., and Hayaru, S., Analysis of function of rectified linear unit used in deep learning, *Proceedings of the International Joint Conference on Neural Networks*, Killarney, Ireland, 12-17 July 2015.

Kaushal, A., Yuancheng, Z., Qingjin P., Trevor, S., Erin, W., Michael, Z., and Alecs, C., Evaluation of fast-track strategies using agent-based simulation modeling to reduce waiting time in a hospital emergency department, *Socio-Economic Planning Sciences*, vol. 50, pp. 18-31, 2015.

Kea, B., Rochelle, F., Robert, A. L., and Benjamin, C. S., Interpreting the national hospital ambulatory medical care survey: United States Emergency Department Opioid Prescribing, *Academic Emergency Medicine*, vol. 23, no. 2, pp. 159-165, 2006-2010

Kuo, Y. H., Nicholas, B. C., Janny, M. Y. L., Helen, M., Anthony, M. C. S., Kelvin, K. F. T., and Colin, A. G., An integrated approach of machine learning and systems thinking for waiting time prediction in an emergency department, *International Journal of Medical Informatics,* vol. 139, pp. 104-143, 2020.

Kyritsis, A. I. and Michel, D., A machine learning approach to waiting time prediction in queueing scenarios, *Proceedings of 2nd International Conference on Artificial Intelligence for Industries,* pp. 17-21, 2019.

Liang, T. K., Queueing for healthcare, Article *in Journal of Medical Systems,* vol. 36, no. 2, pp. 541-547, 2010.

Mor, A., Shlomo, I., Avishai, M., Yariv N. M., Yulia, T., Galit B. Y., On patient flow in hospitals: A data-based queueing-science perspective, *Stochastic Systems, vol.* 5.1, pp. 146-194, 2015.

Moreno, Atilio, Lina A., Julián, F., Camilo, C., Sandra, T., and Oscar, M. M., Application of queuing theory to optimize the triage process in a tertiary emergency care (ER) department, Journal *of Emergencies, Trauma and Shock,* vol. 12, no. 4, pp. 268–273, 2019.

McMahan, B., and Streeter, M., Delay-tolerant algorithms for asynchronous distributed online learning. I*n Advances in Neural Information Processing Systems*, pp. 2915-2923, 2014.

Mahadevan, B, *Operations Management Theory and Practice,* 3rd Edition, Pearson Education, India, 2015.

Pak, A., Brenda, G., and Andrew, S., Predicting waiting time to treatment for emergency department patients, *International Journal of Medical Informatics*, vol. 145, pp. 104303, 2020.

Palmer, G. I., Vincent, A. K., Paul R. H., and Asyl, L. H., Ciw: an open-source discrete event simulation library, *Journal of Simulation,* vol. 13, no. 1, pp. 68–82, 2019.

Pargent, F., Bischl, B., and Thomas, J., A benchmark experiment on how to encode categorical features in predictive modeling, *Master Thesis*, 2019.

Peterson, M. D., Dimitris, J. B., and Amedeo, R. O., Models and algorithms for transient queueing congestion at airports, Management *Science,* vol. 41, no. 8, pp. 1279-1295, 1995.

Pianykh, O. S. and Daniel, I. R., Can we predict patient wait time? *Journal of the American College of Radiology,* vol. 12, no. 10, pp. 1058–1066, 2015.

Rasouli, H. R., Esfahani, A. A., and Mohsen, A. F., Challenges, consequences, and lessons for way-outs to emergencies at hospitals: a systematic review study, *BMC Emergency Medicine,* vol. 19, no. 1, pp. 1-10, 2019.

Ruder, S., An overview of gradient descent optimization algorithms, Available: https://arxiv.org/abs/1609.04747, 2016

Ruben, A., Billy, J. M., Ying, P. T., Mark, H. D., Christopher, A. C., Song, Z., Gary, R., Timothy, S. S., Ying, M., and Ethan, A. H., An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data, *Medical Care*, vol. 48, No. 11, pp. 981-988, 2010.

Sasanfar, S., Morteza, B., and Afrooz, M., Improving emergency departments: simulation-based optimization of patients waiting time and staff allocation in an Iranian hospital, *International Journal of Healthcare Management*. vol. 16, pp. 1-8, 2020.

Shafaf, N., and Hamed, M., Applications of machine learning approaches in emergency medicine; a review article, *Archives of Academic Emergency Medicine,* vol. 7, no. 1, pp. 34, 2019.

Srivastava, T., How to predict waiting time using queuing theory? Available: https://www.analyticsvidhya.com/blog/2016/04/predict-waiting-time-queuing-theory/, December 17, 2019.

Stagge, A., *A time series forecasting approach for queue wait-time prediction*, *Thesis,* Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1458832&dswid=9120, 2020.

Stintzing, J., and Fredrik, N., *Prediction of Queuing Behaviour through the Use of Artificial Neural Networks*, Thesis, Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1111289&dswid=9120, 2017.

Sun, B. C., Adams, J., Orav, E. J., Rucker, D. W., Brennan, T. A., and Burstin, H. R., Determinants of patient satisfaction and willingness to return with emergency care, *Annals of Emergency Medicine,* vol. 35, no. 5, pp. 426-434, 2000.

Ülkü, Sezer, Chris, H., and Shiliang, C., Making the wait worthwhile: experiments on the effect of queueing on consumption, *Management Science,* vol. 66, no. 3, pp.1149-171, 2020.

Ward, P. R., Philippa, R., Clinton, C., Mariastella, P., Nicola, D., Simon, A.C., and Samantha, M., Waiting for' and 'waiting in' public and private hospitals: a qualitative study of patient trust in south australia, *BMC Health Services Research,* vol. 17, no. 1, pp. 1-11, 2017.

Biography

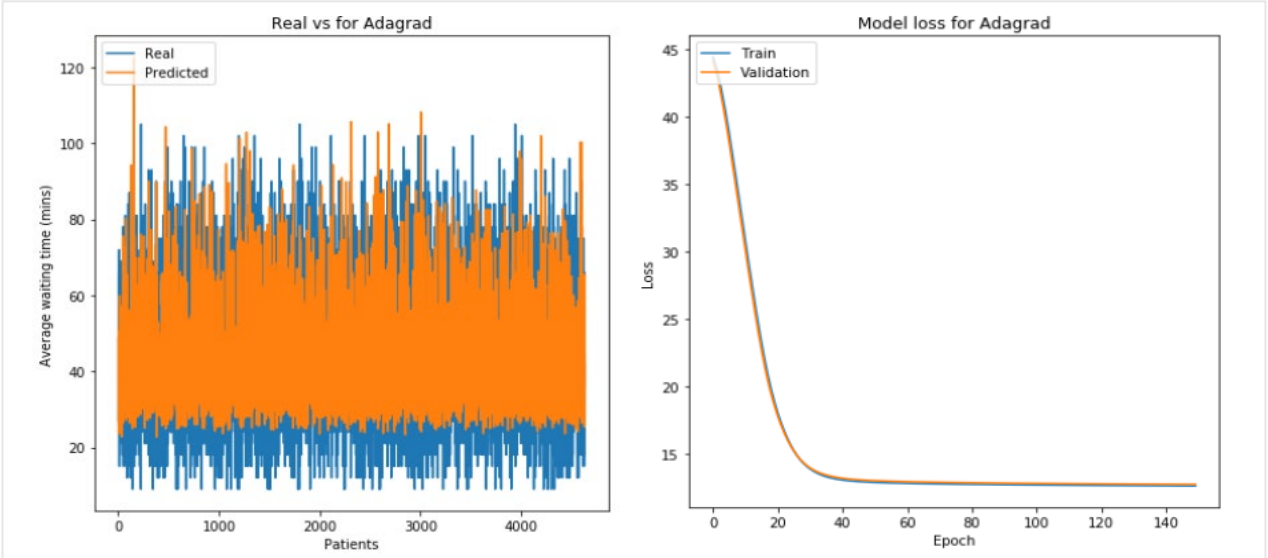
Hassan Hijry is lecturer of Industrial Engineering in department of Industrial Engineering at University of Tabuk, Tabuk, Saudi Arabia. Currently, He is a Ph.D. candidate at Oakland University (OU), USA, majoring in Systems Engineering. He earned his B.S. in Industrial Engineering from Jazan University, SA in 2012 and M.S. degree in Industrial Engineering from Lawrence Technological University (LTU), Detroit, MI, USA in 2017. And he worked in industry as a front-line manager at PEPSICO (Al-Riyadh, SA) in 2013. His research interests include artificial intelligence and machine learning applications, simulation, optimization, and operations research.

Richard Olawoyin is an Associate Professor of Industrial and Systems Engineering (ISE) at Oakland University (OU), Rochester, Michigan teaching Engineering Risk Analysis, Statistical Methods in Engineering, Safety Engineering, Industrial and Systems Engineering, Human Factors Engineering and Occupational Biomechanics. He holds a BS in Geology (University of Calabar, Nigeria), MS and PhD in Energy Engineering (Penn. State University). His research interests emphasize on Industry X.0 Systems in areas of; statistics and artificial intelligence and big data risk analytics, digital supply chain networks, blockchain and stochastic trend modelling. He is a book author and authored several book chapter and peer-reviewed journal publication (more than 35 as first author). He is the assessment coordinator for the ISE department at Oakland University, he is an advisory council member for the ABET Inclusion and Diversity and Equity Advisory (IDEA) Council.

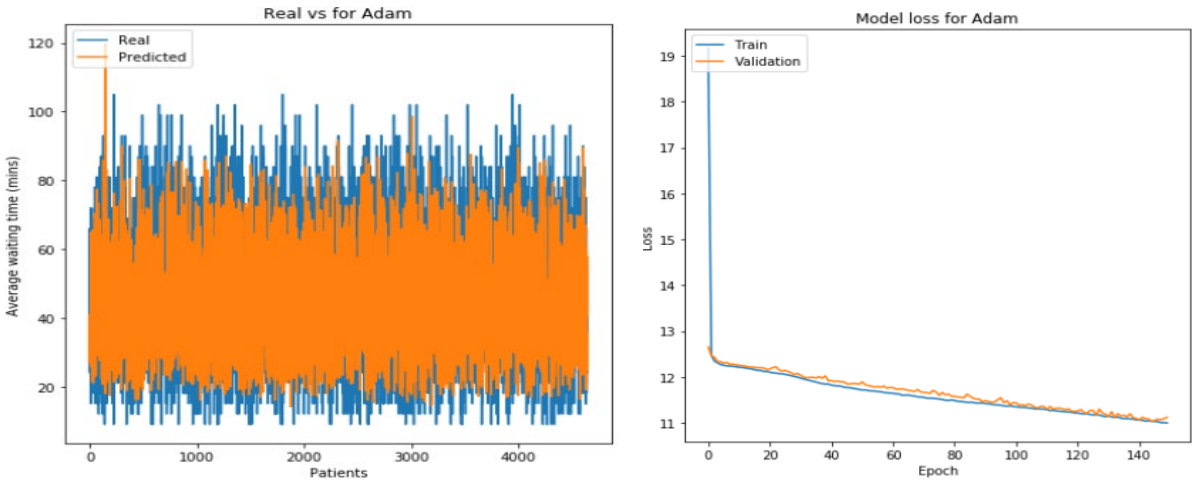
Appendix A. Supplements results

Table A. Patient arrival time data (one day)

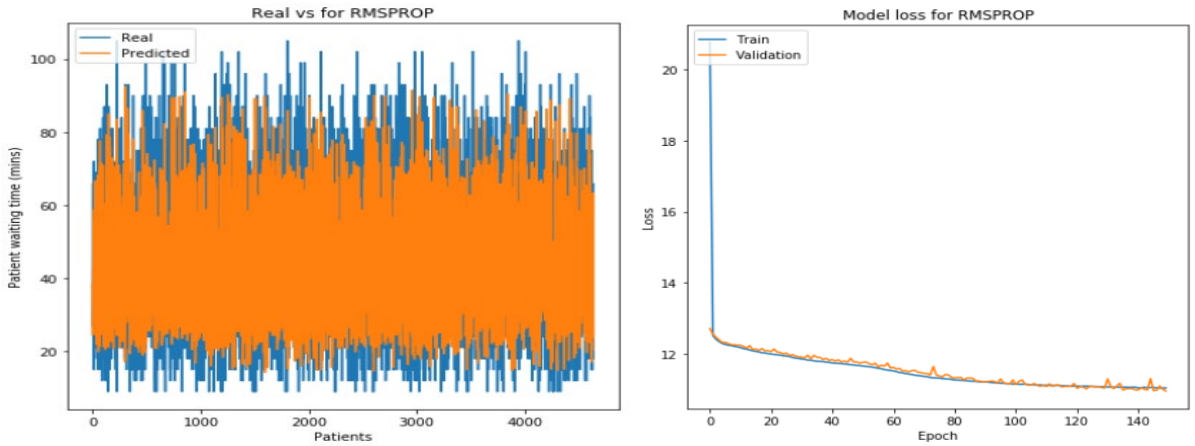
	Day 1		
	Arrival time	Waiting Time	Service Time
0	7:01	33	84
1	7:27	39	75
2	7:44	42	78
3	10:07	27	78
4	10:12	42	69
5	10:22	30	78
6	10:24	27	81
7	13:56	30	81
8	14:26	33	78
9	17:33	36	72
10	18:27	33	84
11	19:09	42	75
12	21:20	36	84
13	22:14	45	72
14	23:47	45	84
15	23:51	45	75



Adagrd aptimation algrithm resutls (Waiting time predicted vs actual and plot of model loss on training and validation dataset)



Adam aptimation algrithm resutls (Waiting time predicted vs actual and plot of model loss on training and validation dataset)



RMSprop aptimation algrithm resutls (Waiting time predicted vs actual and plot of model loss on training and validation dataset)