INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Predicting Patient Arrival Rates in Outpatient Department to Optimize Patient Flow

Jagriti Gupta[1], **Naresh Sharma**[1]*

1 School of Engineering and Sciences, GD Goenka University, Gurugram, Haryana, India

OPEN ACCESS

Received: 12/12/2024

Accepted: 22/05/2025

Published: 19/09/2025

Citation: Gupta J, Sharma N (2025) Predicting Patient Arrival Rates in Outpatient Department to Optimize Patient Flow. Indian Journal of Science and Technology 18(35): 2845-2852. https://doi.org/10.17485/IJST/v18i35.gupta

* **Corresponding author**.

naresh.sharma2006@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This real-life study explores the application of Machine Learning (ML) models to predict hourly patient arrivals in multi-speciality outpatient departments. Also to find the features which affect the arrival rate of patients in different departments. **Methods**: The patient arrival data, 21298 patients in four specialities (Dental, General medicine, Orthopaedic, ENT) of the Outpatient Department (OPD) of a Multi-Speciality hospital in Gurugram, during the Nov – Dec 2023 have been extracted and pre-processed. In this study, ML models, including linear regression, K nearest neighbors, random forests, decision tree and gradient boosting machine are applied to identify the most effective approach for predicting queuing behaviour using historical arrival data. Afterwards, random forest feature selection algorithm was also applied to select the best features for each speciality. **Findings**: Random Forest algorithm produced lowest mean absolute error of 2.70 and 3.29 for orthopaedic and dental department respectively. However, Gradient Boosting Machine is found to be the best predictive ML models with least mean absolute error of 2.53 and 3.25 for the general medicine and ENT department respectively. Best subset for dental and general medicine speciality contains day of the month, lagged arrivals, revisit number and hour of the day. For ENT and orthopaedic speciality, day of the month, lagged arrivals, hour and morning shift appear to be the best predictors. **Novelty**: This work is one of the first to use ML to predict the number of patients in a multi-speciality OPD, emphasising feature selection and model comparison. By better predicting and controlling patient flow, the results give hospital managers important insights to maximise resource allocation and increase patient satisfaction. With accurate prediction of arrival rate, hospitals can better manage resources and staffing, thereby reducing wait times and improving patient care.

Keywords: Arrival Rate; Outpatient Department; Hospital; Multi-speciality; Machine Learning Algorithm

1 Introduction

The arrival process of patients is one of most critical processes in outpatient departments (OPDs) of hospital. In multi-specialist hospital, a random arrival process increases the crowd in the hospital, due to which planning about staff and resources is materializing. Arrival of the unscheduled patients follows the Poisson process[1]. Arrival processes for unscheduled outpatient presentations seem to be different from the arrival processes generated by appointment schedules. As a result, the medical services provided within these departments are even more pervaded by their environment and remain highly affected by the uncertainty due to patient demand variability and the randomness of their arrival. Generally, the appointment schedules determine the structure of appointment-driven arrival processes. Nevertheless, there are certain factors, such as patient no shows and resource availability, that can disturb the schedules of real arrival process[2].

Patient tardiness poses significant challenges for healthcare administrators. When patients arrive ahead of schedule instead of being late, it results in overcrowding[3], while arriving late can lead to idle costs and make scheduling systems less effective[4]. The probability of missing an outpatient visit varies with the speciality of the departments, such as 80.3% in general practice department and 11.5% in multi-speciality orthopaedic department[5]. The walk-in patients who visit the OPD without an appointment represent a different pattern of arrival associated with criticality, speciality and location; similar to the variance observed with no-shows[6]. Though the arrival rates of walk-ins and emergency patients fluctuate dramatically during the day, demand for patients in OPD varies throughout the days and months. Due to issues posed by patients who come late or drop in without an appointment, predicting the arrival of patients in hospitals is crucial for improving the decision-making process of managers. Supervised learning algorithms such as machine learning (ML) employed in different domains for predicting values of a target variable[7] can also be used by healthcare organisations to make better decisions by acquiring insights into their future problems[8].

Available research in public domain focuses majorly on daily and hourly arrival predictions in emergency departments of hospitals, as most of the EDs operate on daily staffing and resource planning. According to Vollmer *et al.*[9], there hasn't been much research done in the literature on techniques for forecasting admission rates while comprehending the underlying dynamics. Zhang *et al.*[10] used six linear and nonlinear models to predict the flow of emergency radiology patients each day to enhance the efficiency of the appointment scheduling system. The researchers found that the Lasso model was the most effective for predicting radiology visits, with an average absolute percentage error of 7.06 %. Asheim *et al.*[11] develop a continuous monitoring system to forecast patient arrival in ED department using a poison time series regression model. The results show that 20% of arrivals have been updated more than 1 hour prior to arrival. This approach to predicting arrivals can be an effective tool for decision making and ED resource management. But this study had been done in single centre, thus it cannot be generalised to other healthcare facilities. ML methods specifically predict patient arrival in emergencies in the healthcare sector[12].

Porto[13] predicted the patient's arrival in ED using 11 ED datasets from three different countries. They investigated that metrological and calendar variable, which contribute more to the prediction of arrival rate. The XGBoost algorithm performed better among other machine learning algorithms. The literature stated that simpler methods may outperform complex ones in ED patient arrival forecasting[3]. Forecasting the patient arrival rate in ED can be done using single and hybrid methods such as LR, ARIMA, ANN, ES, ARIMA + ANN and ARIMA + LR. Yousefi *et al.*[14] looked into the factors that affect how many patients visit the emergency room each day and used DNN to build a reliable model that predicts patient arrivals for the next seven days.. The authors argued that applying ML algorithms' skills can lead to more accurate forecasts. The suggested approach performs better than statistical alternatives such as seasonal ARIMA, ARIMA paired with linear regression, generalized linear models, Autoregressive Integrated Moving Average (ARIMA), and multiple linear regressions, according to the numerical results.

The Outpatient Department (OPD) is the hospital's public face to the outside world. The external service, OPD in the hospital, is under increasing pressure year to year due to rising patient numbers, growing difficulty of health issues and so on. The task of analysing and predicting daily outpatient visits in hospitals is limited to the domain of time series prediction. In this field, the most significant factors that impact the accuracy of the forecasts are the random nature of the data and the cyclic variations seen. Nyoni and Nyoni[15] used monthly time series dataset to predict outpatient visits using artificial neural networks. A literature review of publicly available resources shows that there are very few studies which have used ML algorithms to forecast the hourly arrival rate of patients in an outpatient department that serves multiple specialities. Random forest[16] and vector regression are found to be supporting ML algorithms in predicting hourly arrival rate of patients[17]. Similarly, the RF and XGBoost algorithms can reduce service time in OPD[18]. Also, RF model outperformed in predicting waiting time in Internal Medicine department with an MAE of 5.03 minutes[19].

The present research applies more than one ML method to predict the patient arrival in different outpatient departments of a multi-speciality hospital. The work facilitates the comparison of different algorithms on a single dataset and offers a cross-comparison of different datasets obtained for different departments simultaneously. As far as we know, this study uses more than one ML method to predict how many patients will arrive at a multi-speciality hospital. This study is one of the novel studies applying the ML in predicting arrival time in OPD compared to other studies predicting waiting time, service time in OPD, and arrival time in emergency.

The scientific allocation of healthcare resources, including medical supplies, surgical equipment, and hospital beds, can be facilitated by an accurate and dependable prediction of the volume of various types of outpatients. Predicting daily outpatient visits accurately is important because it can assist hospital administrators in making timely and effective decisions to satisfy anticipated healthcare demand, especially in public hospitals.

2 Methodology

2.1 Data collection

In this study, the data was collected from the hospitals information system of the Multi specialist hospital, Gurugram, for predicting arrival rate. Patient arrival data is extracted for four specialities between November 1, 2023, and December 31, 2023. During this period, a total of 21298 patients visited the hospital. This study's target variable is the hourly patient arrival rate. The explanatory variables, such as arrival hour, weekdays, and day shifts, were investigated from the arrival day.

The obtained dataset has been pre-processed, and ML algorithms were trained, and performance of ML models was evaluated.

2.2 Data cleaning

This is the critical step that involves preparing dataset by correcting dataset that is incomplete, inaccurate, and inconsistent due to several reasons. Data was pre-processed and cleaned for the missing values to obtain reliable and meaningful results in machine learning project. The final distribution of data includes 41% patients for dental, 26% patients for general medicine, 18% patients for orthopaedic, 15% patients for ENT.

2.3 Data transformation

Feature encoding is a fundamental process that involves transforming data into a suitable input format, enabling the computer to extract meaningful information without modifying the data's original meaning. Mathematical equations play a major role in both machine learning-based algorithms and standard prediction models. In the pre-processed dataset, there are four categorical variables (shift of the day, month, weekday and weekday) that are encoded in numerical variables using one-hot encoding technique[(20)].

Furthermore, feature scaling is another step in data preprocessing, which normalises the variables in a dataset to a specific range. This helps in avoiding the construction of incorrect predictions during training and testing of the models.

2.4 ML algorithms implementation:

With the use of ML algorithms, one can learn a target function f, such as $f(X) = Y$, that best connects input variables X to an output variable Y. Before implementing ML algorithms, dataset is split into training and testing dataset to avoid leakage of data. The training dataset, also known as model training, typically contains 80% of the data while testing will use the remaining 20%. It is used to determine the parameters of a specific model.

Furthermore, in order to mitigate the possibility of overfitting, one can implement k fold cross validation during the training phase. This work employed several ML techniques to improve the accuracy of predicting the patient arrival rate. Specifically, our focus is on the following five machine learning techniques:

2.4.1. Linear regression (LR)

Linear regression (LR) is a supervised machine learning algorithm which is capable of handling the variety of data patterns found in dataset[(21)]. The algorithm follows the mathematical approach to predict the dependent variable. In this study, LR models were used in OPD patient arrival forecasting as shown in **Equation (1)**:

$$y_i = a_i + b_1 x_{i1} + \ldots + b_p x_{ip} + \varepsilon_i (i = 1, \ldots, n) \tag{1}$$

where y_i is the hourly predicted arrival rate of patients, a_i is a constant, x_{i1}, x_{i2}, …, x_{ip} are independent variables such as day of the week, month of the year, time of day, b is the weighting coefficient of each of these variables and ε_i is called the error term.

2.4.2. K-Nearest Neighbor (KNN)

KNN is generally recognised as a basic model for making predictions; it learns from the training dataset and uses that information to generate predictions when the test object's properties match those of an example in the training dataset. For each new predicted instance, our method can produce local and distinct estimations of complicated target concepts. While KNN has demonstrated strong performance in various domains, its performance degrades when dealing with giant datasets due to the necessity to store and browse all the training data. As the amount of noise and irrelevant features in the training dataset increases, its accuracy also falls.

2.4.3. Decision Tree (DT)

This involves splitting a given dataset into smaller subsets based on feature values and then continually splitting each subset in a manner that minimizes the variance in the data [22]. The resultant model builds a tree with leaf and decision nodes. Two or more branches of a decision node stand for the values for the features during testing. A decision on the target value is represented by the leaf node. The most common criterion for making splits is the mean squared error (MSE) which chooses the split that minimises the MSE across the two resulting nodes from a split. The MSE for each node is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{2}$$

2.4.4. Random Forest Algorithm (RF)

Random Forest combines multiple decision trees to improve the prediction accuracy and control over-fitting. Each tree in a random forest is built from a bootstrap sample, which is a sample taken with replacement from the original dataset, typically the same size as the original dataset.

2.4.5. Gradient Boosting Algorithm (GBM)

The gradient boosting algorithm is a supervised machine learning algorithm used for both classification and regression problems. It is based on the concept of boosting, which involves training individual weak models and assessing their performance [23]. Inputs that were not accurately predicted by earlier models are given higher weights, whereas inputs that were successfully predicted are given less weight. Furthermore, GBM decreases overfitting by explicitly specifying the number of trees and the learning rate.

2.5 ML algorithms Evaluation

After training the ML algorithms, the performance of models was evaluated using two metrics: R^2 score and mean absolute error (MAE). R^2 score is the coefficient of determination, which indicates the correlation between the target variable and the value predicted by the ML model [24]. Its value, which ranges from 0 to 1, indicates how effectively the model is expected to predict samples that have not yet been seen.

MAE is the useful measure in ML model evaluation. It measures the average magnitude of the errors in a set of predictions [25]. It is the average of the absolute differences, giving each individual difference equal weight, between the actual observation and the predicted outcome for the entire test sample.

3 Results and Discussion

It is critical to emphasize that the patient arrival rate at the OPD is marked by considerable variation throughout the year and hour, in addition to a high degree of uncertainty.

Figure 1 clearly demonstrates that most patients in the four specialties tend to arrive during the hours of 10 am and 1 pm, which is the busiest period. Moreover, the period with the fewest people is after 2 pm. Figure 2 illustrates the patient arrival rate in days. Saturdays seem to be the busiest day for patients visiting the OPD. Additionally, the number of arrivals of dental department is higher compared to other.

After a thorough examination of the demand variations across the four OPDs is completed, the patient arrival rate is predicted using the four steps of the technique as discussed above. We used Python for all the processing.

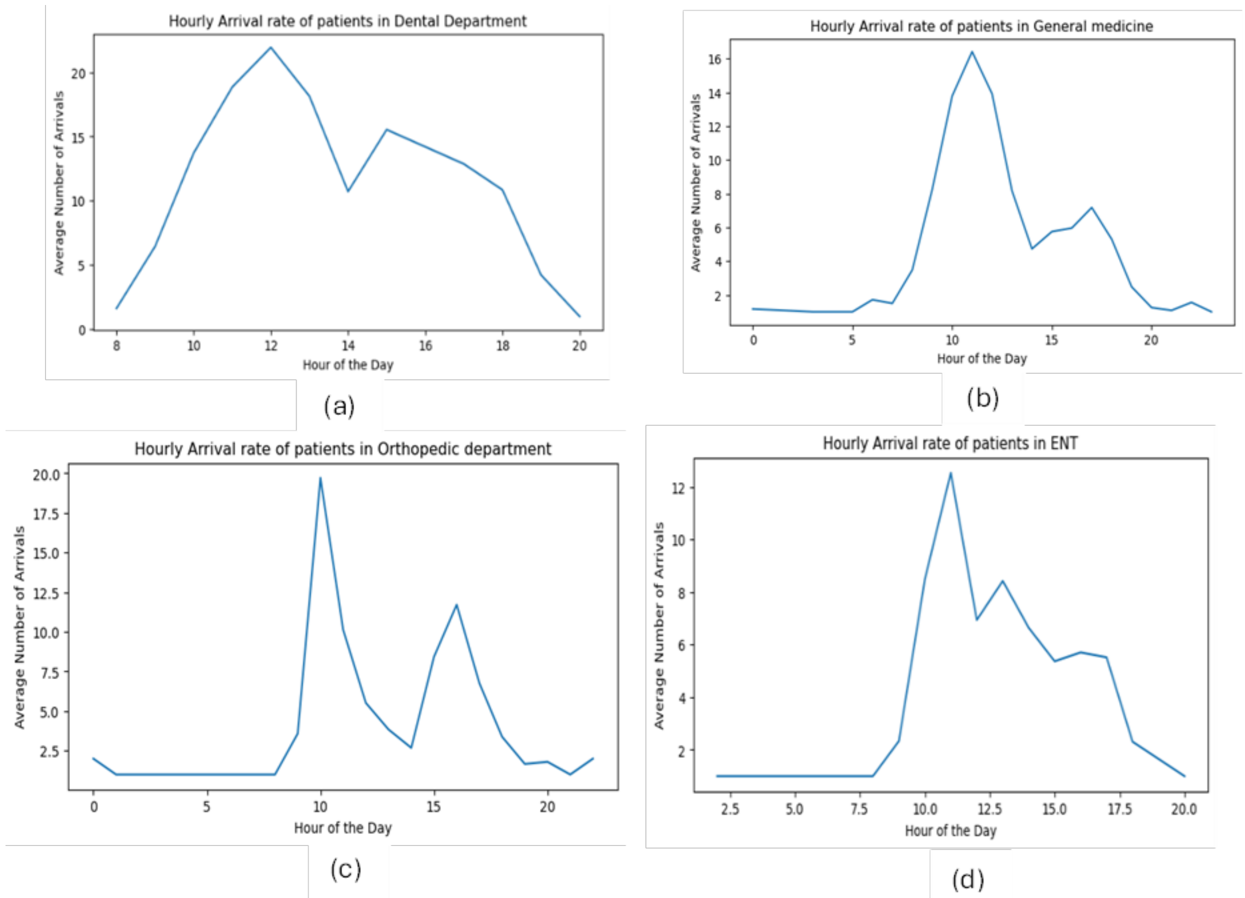

Fig 1. Hourly patient arrival rate for (a) General medicine (b) Dental (c) Orthopaedic (d) ENT OPD's

3.1 ML algorithms implementation.

Initially, the pre-processed dataset is randomly split into two datasets: a training dataset (80%) and testing dataset (20%). Afterwards, min max scaling [26] technique was used to scale the dataset values within the range of 0 to 1. To implement all five ML algorithms as described in above section, initially the algorithms were trained using the training data. The ML algorithms with the fivefold cross validation technique were applied. The patient arrival rates are predicted using the model that performs the best in relation to the MAE.

Table 1 demonstrates that the most effective predictive ML models are the RF for dental and orthopaedic conditions, and the GBM for general medicine and ENT specialities. For dental patients, RF model performed with least MAE 3.29 minutes, while for General Medicine, GBM model performed with least MAE 2.53 minutes. However, the LR model showed relatively poor predictive performance for all departments. These findings indicate that no single ML algorithm outperforms others across all hospital specialities, emphasising the need of testing several algorithms. The result are in line with prior studies such as [16] and [17], where RF was able to forecast hourly arrival rate in allergology and cardiology specialities compared to vector regression predicting the arrival rate with lowest mean absolute in pulmonary department. Similarly, the results of [19] validate the effectiveness of RF model, though GBDT performed better in other cases. These variations demonstrate how the dataset size, feature interactions, and the inherent patterns in the data, such as linearity or nonlinearity affect model performance.

However, some previous studies present contradictory results. In a study about predicting patient wait times and appointment delays in radiology facilities, the elastic net model did better than other algorithms, such as RF and GBM, in a number of different ways. Similarly, research on predicting next day patient flows using calendar and metrological variables found that SVR (Support Vector Regression) outperformed all other algorithms, including k-Nearest Neighbors (kNN), SVR, XGBoost, RF and Gradient Boosting Machine (GBM), as well as traditional regression techniques like Ordinary Least Squares (OLS) and Ridge Regression. This indicates that SVR may be more effective in capturing patterns in patient arrival data

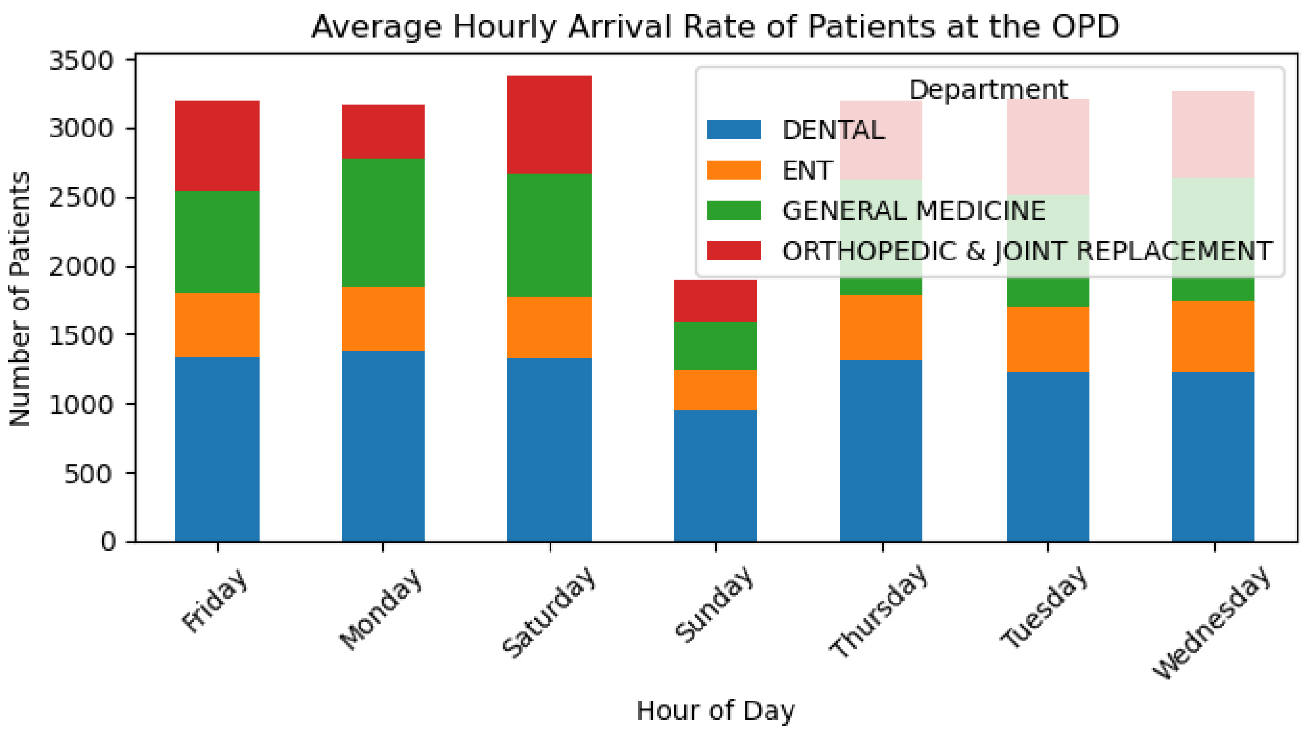

Fig 2. Number of hourly arrivals at the four specialities per day

compared to ensemble models like RF and GBM [27]. This indicates that the effectiveness of a predictive model can vary based on factors such as the nature of the medical speciality, the specific workflow, and the features incorporated into the model.

In this study, the Random Forest feature selection method was used to find the best predictive features of each speciality, which affect arrival rate of patients. The mean absolute error was used as the loss function in this procedure to analyse the most significant features. It was found that the best subset for dental and general medicine speciality contains day of the month, lagged arrivals, revisit number and hour of the day. For ENT and orthopaedic speciality, day of the month, lagged arrivals, hour and morning shift appear to be the best predictors. This analysis verifies the significance of speciality-specific model selection by incorporating data from recent literature and discussing both supporting and conflicting results. The findings show that no one model is universally relevant across all specialities and that predictive performance should be assessed independently if patient flow optimisation is to be maximised.

Table 1. Mean absolute score of five ML algorithms for four different specialities.

ML	**OPD Speciality**			
Algorithm	**Dental**	**General Medicine**	**ENT**	**Orthopaedic**
LR	4.46	3.27	3.58	4.44
KNN	4.92	3.38	3.80	4.47
DT	4.42	3.54	4.61	2.87
RF	**3.29**	2.73	3.38	**2.70**
GBM	3.49	**2.53**	**3.25**	2.78

4 Conclusion

This study predicts the hourly arrival rate of patients using ML algorithms. A two-month dataset of patient arrivals was collected and then pre-processed for four specialities of a Multi specialist hospital of Gurugram in India. A dataset containing information on patients' visits to the relevant outpatient department (OPD) was retrieved from the hospital's database in order to ensure the reliability of the results. Five ML regression algorithms were applied with five-fold cross validation to predict patient arrival rates. The most effective predictor for the dental and orthopaedic speciality was the Random Forest (RF) model with the lowest MAE value, while the best predictor for the general medicine and ENT specialities was gradient boosting (GB). Furthermore, the best features influencing the hourly arrival rate were identified using the random forest feature selection method. The features such as day of the month, lagged arrivals and patients revisit number are the best explanatory predictors for all speciality which gives the most reliable predictions. Thus, it can be concluded that not only emergency departments, but also OPD are also impacted by hourly effects, weekend effects. However, different departments have different arrival rates and for each department different machine algorithm work differently. Results obtained in hourly arrival rate obtained using the ML models can help manager in resource optimization and delivering services. Large dataset from different hospitals with similar specialities can be used in future to establish the cross validation.

5 Acknowledgement

The article was presented in "International Conference on Transforming Lives Through Adoption of SDGs: Role of Higher Education Institutions" (TLASH 2024) during 30th May-31st May, 2024, organized by GD Goenka University, India, in alliance with the Association of Indian Universities (AIU). The organizers claim the review responsibility.

References

1) Grot M, Kugai S, Degen L, Wiemer I, Werners B, Weltermann BM. Small Changes in Patient Arrival and Consultation Times Have Large Effects on Patients' Waiting Times: Simulation Analyses for Primary Care. *International Journal of Environmental Research and Public Health*. 2023;20(3). https://doi.org/10.3390/ijerph20031767.
2) Li J, Zhu G, Luo L, Shen W. Big Data-Enabled Analysis of Factors Affecting Patient Waiting Time in the Nephrology Department of a Large Tertiary Hospital. *Journal of Healthcare Engineering*. 2021;2021. https://doi.org/10.1155/2021/5555029.
3) Yucesan M, Gul M, Celik E. A multi-method patient arrival forecasting outline for hospital emergency departments. *International Journal of Healthcare Management*. 2020;13(S1):283–295. https://doi.org/10.1080/20479700.2018.1531608.
4) Incze E, Holborn P, Higgs G, Ware A. Using machine learning tools to investigate factors associated with trends in 'no-shows' in outpatient appointments. *Health and Place*. 2021;67. https://doi.org/10.1016/j.healthplace.2020.102496.
5) Suk MY, Kim B, Lee SG, You CH, Kim TH. Evaluation of Patient No-Shows in a Tertiary Hospital : Focusing on Modes of Appointment-Making and Type of Appointment. *International Journal of Environmental Research and Public Health*. 2021;18:3288. https://doi.org/10.3390/ijerph18063288.
6) Munavalli JR, Boersma HJ, Rao SV, van Merode GG. Real-Time Capacity Management and Patient Flow Optimization in Hospitals Using AI Methods. Springer International Publishing. 2021. https://doi.org/10.1007/978-3-030-45240-7_3.
7) Ouerghi A, Hasni M, Jaidi Z, Layeb S. A new combined linear-artificial neural network-based model for accurate inflation forecasting in Tunisia. *International Journal of Decision Sciences Risk and Management*. 2019;8(4):220. https://doi.org/10.1504/IJDSRM.2019.10028769.
8) Pianykh OS, et al. Improving healthcare operations management with machine learning. *Nature Machine Intelligence*. 2020;2(5):266–273. https://doi.org/10.1038/s42256-020-0176-3.
9) Vollmer MAC, et al. A unified machine learning approach to time series forecasting applied to demand at emergency departments. *BMC Emergency Medicine*. 2021;21(1). https://doi.org/10.1186/s12873-020-00395-y.
10) Zhang Y, et al. Emergency patient flow forecasting in the radiology department. *Health Informatics Journal*. 2020;26(4):2362–2374. https://doi.org/10.1177/1460458220901889.
11) Asheim A, Bjørnsen LPBW, Næss-Pleym LE, Uleberg O, Dale J, Nilsen SM. Real-time forecasting of emergency department arrivals using prehospital data. *BMC Emergency Medicine*. 2019;19(1):1–6. https://doi.org/10.1186/s12873-019-0256-z.
12) Nas S, Koyuncu M. Emergency Department Capacity Planning : A Recurrent Neural Network and Simulation Approach. *Computational and Mathematical Methods in Medicine*. 2019;2019. https://doi.org/10.1155/2019/4359719.
13) Porto BM. Enhancing emergency department patient arrival forecasting : a study using feature engineering and advanced machine learning algorithms. *Research Square*. 2024. https://doi.org/10.21203/rs.3.rs-3891200/v1.
14) Yousefi M, Yousefi M, Fathi M, Fogliatto F. Patient visit forecasting in an emergency department using a deep neural network approach. *Kybernetes*. 2019;49(9):2335–2338. https://doi.org/10.1108/K-10-2018-0520.
15) Nyoni SP, Nyoni T. Forecasting the Number of Outpatient Visits at Silobela District Hospital in Zimbabwe Using Artificial Neural Networks. *EPRA International Journal of Multidisciplinary Research (IJMR)*. 2020;6(2):198–210. https://doi.org/10.36713/epra2013.
16) Aldhoayan MD, Alobaidi RM. The Use of Machine Learning to Predict Late Arrivals at the Adult Outpatient Department. *Cureus*. 2023;15(6). https://doi.org/10.7759/cureus.39886.
17) Mbarek HB, Layeb SB, Aissaoui NO, Jaoua A. Predicting Patient Arrival Rates in a Multi-Specialty Outpatient Department. *Proceedings of the 3rd Asia Pacific International Conference on Industrial Engineering and Operations Management*. 2023;p. 2134–2142. https://doi.org/10.46254/ap03.20220358.
18) Joseph J, Senith S, Kirubaraj AA, Ramson JSR. Machine Learning for Prediction of Wait Times in Outpatient Clinic. *Procedia Computer Science*. 2023;215:230–239. https://doi.org/10.1016/j.procs.2022.12.026.
19) Li X, et al. Prediction of outpatient waiting time : using machine learning in a tertiary children ' s hospital. *Translational Pediatrics*. 2023;12(11):2030–2043. https://doi.org/10.21037/tp-23-58.
20) Brownlee J. Why one-hot encode data in machine learning? *Machine Learning Mastery*. Available from: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/.
21) Maulud D, Abdulazeez AM. A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*. 2020;1(4):140–147. https://doi.org/10.38094/jastt1457.

22) Vadapalli P. Random Forest Vs Decision Tree: Difference Between Random Forest and Decision Tree. 2020. Available from: https://www.upgrad.com/blog/random-forest-vs-decision-tree/.
23) Aziz N, Akhir EAP, Aziz IA, Jaafar J, Hasan MH, Abas ANC. A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems. *2020 International Conference on Computational Intelligence, ICCI 2020*. 2020;p. 11–16. https://doi.org/10.1109/ICCI51257.2020.9247843.
24) Valbuena R, et al. Evaluating observed versus predicted forest biomass: R-squared, index of agreement or maximal information coefficient? *European Journal of Remote Sensing*. 2019;52(1):1–14. https://doi.org/10.1080/22797254.2019.1605624.
25) Vinicius T. Comparing Robustness of MAE, MSE and RMSE. *Towards Data Science*. Available from: https://towardsdatascience.com/comparing-robustness-of-mae-mse-and-rmse-6d69da870828.
26) Chong J. What is Feature Scaling & Why is it Important in Machine Learning? *Towards Data Science*. 2020. Available from: https://towardsdatascience.com/what-is-feature-scaling-why-is-it-important-in-machine-learning-2854ae877048.
27) Susnjak T. Forecasting patient flows with pandemic induced concept drift using explainable machine learning. *EPJ Data Science*. 2023;12(1). https://doi.org/10.1140/epjds/s13688-023-00387-5.