JAMIA Open, 2025, **8(2)**, ooae138
https://doi.org/10.1093/jamiaopen/ooae138
Research and Applications

Research and Applications

Artificial intelligence-driven forecasting and shift optimization for pediatric emergency department crowding

Izzet Turkalp Akbasli, MD[1,*], Ahmet Ziya Birbilen, MD[1], Ozlem Teksam, MD, PhD[1]

[1]Division of Pediatric Emergency, Department of Pediatrics, Faculty of Medicine, Hacettepe University, Ankara 06270, Turkey

*Corresponding author: Izzet Turkalp Akbasli, MD, Division of Pediatric Emergency, Department of Pediatrics, Faculty of Medicine, Hacettepe University, Gevher Nesibe Avenue, Altindag, Ankara 06230, Turkiye (izzetakbasli@gmail.com)

Abstract

Objective: This study aimed to develop and evaluate an artificial intelligence (AI)-driven system for forecasting Pediatric Emergency Department (PED) overcrowding and optimizing physician shift schedules using machine learning operations (MLOps).

Materials and Methods: Data from 352 843 PED admissions between January 2018 and May 2023 were analyzed. Twenty time-series forecasting models—including classical methods and advanced deep learning architectures like Temporal Convolutional Network, Time-series Dense Encoder and Reversible Instance Normalization, Neural High-order Time Series model, and Neural Basis Expansion Analysis—were developed and compared using Python 3.8. Starting in January 2023, an MLOps simulation automated data updates and model retraining. Shift schedules were optimized based on forecasted patient volumes using integer linear programming.

Results: Advanced deep learning models outperformed traditional models, achieving initial R^2 scores up to 75%. Throughout the simulation, the median R^2 score for all models was 44% after MLOps-based model selection, the median R^2 improved to 60%. The MLOps architecture facilitated continuous model updates, enhancing forecast accuracy. Shift optimization adjusted staffing in 69 out of 84 shifts, increasing physician allocation by up to 30.4% during peak hours. This adjustment reduced the patient-to-physician ratio by an average of 4.32 patients during the 8-16 shift and 4.40 patients during the 16-24 shift.

Discussion: The integration of advanced deep learning models with MLOps architecture allowed for continuous model updates, enhancing the accuracy of PED overcrowding forecasts and outperforming traditional methods. The AI-driven system demonstrated resilience against data drift caused by events like the COVID-19 pandemic, adapting to changing conditions. Optimizing physician shifts based on these forecasts improved workforce distribution without increasing staff numbers, reducing patient load per physician during peak hours. However, limitations include the single-center design and a fixed staffing model, indicating the need for multicenter validation and implementation in settings with dynamic staffing practices. Future research should focus on expanding datasets through multicenter collaborations and developing forecasting models that provide longer lead times without compromising accuracy.

Conclusions: The AI-driven forecasting and shift optimization system demonstrated the efficacy of integrating AI and MLOps in predicting PED overcrowding and optimizing physician shifts. This approach outperformed traditional methods, highlighting its potential for managing overcrowding in emergency departments. Future research should focus on multicenter validation and real-world implementation to fully leverage the benefits of this innovative system.

Lay Summary

Overcrowding in Pediatric Emergency Departments (PEDs) can lead to long wait times and reduced quality of care for children. This study developed an artificial intelligence (AI) system to predict when overcrowding might occur and adjust doctors' work schedules accordingly. By analyzing data from over 350 000 past patient visits, the AI learned patterns of when more children are likely to come to the emergency department. The system used advanced machine learning techniques and a process called machine learning operations to continuously update and improve its predictions. With these forecasts, the hospital could plan ahead, assigning more doctors during busy times and fewer during slower periods. This adjustment increased doctor availability during peak hours by up to 30%, reducing the number of patients each doctor had to see. This approach helped balance the workload among doctors and improved the patient-to-doctor ratio during busy times. The study found that the AI system was effective in forecasting busy periods and optimizing staffing, outperforming traditional methods. Implementing such an AI-driven system could enhance patient care, reduce wait times, and improve overall efficiency in PEDs. Future steps include testing this system in multiple hospitals and integrating it into real-world settings to fully realize its benefits.

Key words: Pediatric Emergency Department; forecasting; artificial intelligence; machine learning operations; time-series analysis; shift optimization; overcrowding.

Introduction

Artificial intelligence (AI) usage for assessing electronic health records (EHRs) has been increasing significantly in recent years. The latest advancements in AI hold significant potential and are primarily centered around clinical-focused solutions, such as prediction of prognosis of critically ill children,

Received: September 30, 2024; **Revised:** November 4, 2024; **Editorial Decision:** November 18, 2024; **Accepted:** November 19, 2024

forecasting overcrowding, predicting revisits, pediatric early warning, and clinical decision support systems.[1–6] Although these solutions demonstrate promising outcomes, useful implementation into clinical workflows are challenging tasks that currently hinder their real-world application.[7]

The Pediatric Emergency Department (PED) is one of the busiest units in hospitals. It accommodates a wide variety of patients, including those with high-risk medical conditions. Overcrowding is a significant issue in emergency departments, and various studies have shown that it negatively affects patient care in pediatric emergency services.[8,9] Many factors including overcrowding decrease staff satisfaction and increase burnout.[10]

The causes of hospital overcrowding can be summarized into 3 main factors: Input factors refer to external elements beyond the control of emergency departments, such as emerging healthcare needs, an aging population, and seasonal diseases, all of which increase patient admissions. Throughput factors are internal processes that prolong patient length of stay, including specialist consultations, additional diagnostic tests, and inadequate staff capacity. Output factors involve bottlenecks caused by bed shortages or delays in transferring patients, leading to prolonged stays in the emergency department.[11] This supply and demand imbalance results in prolonged waiting times, delayed treatments, and increased complications, all of which negatively impact patient care and staff well-being.[12]

Various strategies, including fast track processes, telephonic triage, and telemedicine, are being explored internationally to mitigate overcrowding in emergency departments.[12] Tools for predicting real-time overcrowding, such as The Emergency Department Work Index (EDWIN), International Crowding Measure in Emergency Departments (ICMED), Skåne Emergency Department Assessment of Patient Load (SEAL), and the National ED Overcrowding Scale (NEDOCS), have been implemented in diverse emergency department.[13–16] Both models have undergone external validation with varied outcomes. Additionally, these models are oriented towards identifying crowding post-occurrence. Effective crowding models should enable ongoing assessment of crowding severity and its immediate causes. The ability to predict congestion before it materializes allows for preemptive measures to be taken.

In the context of PEDs facing overcrowding and resource limitations, the forecasting of patient demand on an hourly or daily scale emerges may be a critical strategy to improve this predicament.[17] With reliable forecasting of PED demands, hospital managers are empowered to select the most effective strategies to manage expected high PED demands, thus ensuring optimal allocation of resources. Furthermore, these data can assist hospital management in establishing a proactive patient flow strategy, optimizing the use of available resources and preventing overcrowding that can lead to stressful situations.[4,18,19]

This study presents SO-SAFED (Shift Optimization and System for Anticipating and Forecasting Emergency Department Crowding), an AI-driven architecture utilizing the as-yet infrequently published dynamic machine learning process known as machine learning operations (MLOps). Trained with dynamic EHR data, SO-SAFED aims to develop dynamic and highly consistent forecasting models for emergency department management. These models are intended to forecast the number of visits and optimize shift schedules for emergency department staff based on these results.

Thus, this study is the first to address overcrowding by using dynamic time-series forecasting models in an MLOps architecture to forecast patient volumes and optimize workflows in a PED.

Methods

Data acquisition

Admissions of patients to the emergency department between January 2018 and May 2023 were analyzed in the study. Given the study's nature as a time-series analysis, where the number of admissions was of significant interest, no patient was excluded from the study. Admission data were collected via structured query language (SQL) queries from the hospital's EHR system, resulting in datasets comprising visit counts across hourly, daily, weekly, monthly, and yearly time periods. Exogenous data were not included in the analysis. Following the acquisition of approval (GO 23/508) from the Institutional Review Board to enhance hospital service quality, this study was conducted as a retrospective analysis of the hourly clinical activity at a single PED at our reference academic medical center in Turkey is a high-volume unit equipped with 10 examination rooms and 20 observation beds.

Data pre-processing and analysis

Time series often consist of single-feature linear data and aggregated features from specific time intervals. A dataset was constructed by aggregating patient application numbers according to hourly, daily, weekly, monthly, and yearly time periods. The analysis was performed univariately, and univariate descriptive analyses were conducted. Datasets were examined for trends, seasonality, and stationarity characteristics using time-series analysis tests. During the analysis, the augmented Dickey-Fuller (ADF) test was employed to evaluate stationarity. The ADF test was used to determine if the time series was stationary, ensuring that the data did not contain a unit root which could affect the reliability of the model. It was observed that hourly data exhibited a higher degree of stationarity, and thus, the data were aggregated into hourly intervals. Analyses and models were developed using Python 3.8 version. Data flows were facilitated through an automated MLOps data pipeline, which transferred data from weekly EHR to object storage, converted it into hourly data, and then prepared it for model training.

Forecasting models

In this study, models such as AutoRegressive Integrated Moving Average (ARIMA), Seasonal AutoRegressive Integrated Moving Average (SARIMA), Theta, Exponential Smoothing, Croston's Method, Prophet, Greykite, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), CatBoost, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), Deep Linear Models (D-Linear), Temporal Fusion Transformer (TFT), Temporal Convolutional Network (TCN), Neural High-order Time Series(N-HiTS) model, Neural Basis Expansion Analysis (N-BEATS), and Time-series Dense Encoder and Reversible Instance Normalization (TiDE RIN) were utilized. Both classical time-series analysis methods and advanced deep learning architectures from the open-source Darts library (version 0.8.1) were employed.[20,21]

During significant events such as pandemics or natural disasters, widespread changes in human behavior can induce concept drift, which can severely impact the performance of real-time forecasting models. Consequently, monitoring data patterns in models that employ the MLOps architecture becomes essential. In this study, the Kolmogorov-Smirnov (KS) test was employed to address this issue. The KS test was conducted on identically sized windows, comparing the empirical cumulative distribution functions to detect concept drift without making any assumptions about the data distribution. This approach ensures that any deviations in data patterns are promptly identified and addressed, thereby maintaining the accuracy and reliability of the forecasting models.[22]

Model performance metrics

In this study, a comprehensive set of metrics was used to evaluate the precision and robustness of the time-series forecasting models. These metrics included Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Squared Error (RMSE), Root Mean Squared Logarithmic Error (RMSLE), and R-squared (R^2), each selected for a thorough assessment of model performance.

Particular emphasis was placed on the R^2 metric, which indicates the proportion of variance in the predicted values explained by the independent variables. R^2 ranges from 0 to 1, with higher values indicating better goodness-of-fit. In this study, 20 models were developed for each of the 4 architectures, totaling 80 models. While hundreds of metrics were evaluated, the results section focused on R^2 to demonstrate the models' fit and to facilitate comparisons with existing literature. The prominence of R^2 in time-series MLOps frameworks further highlights its importance in evaluating model performance.

Machine learning operating system architecture

In this study, due to its retrospective nature, the direct use of a live MLOps architecture was precluded. Instead, an MLOps architecture was simulated starting from January 2023. In this simulation, it was assumed that data from the hospital information system were continuously and securely transferred to object storage weekly and accumulated there. It was also presumed that the cumulatively updated data underwent specified data preprocessing operations and were used to train models weekly.

In our MLOps simulation, we established an architecture demonstrated in Figure 1 to evaluate the post-training test area and subsequent real-world performance for 4 selected models. The models were developed on Sundays, starting with the first model constructed using data from January 1, 2018, to December 25, 2022, up to 1 week prior to the start of the simulation. Data from December 25, 2022, to January 1, 2023, were predicted, and based on the highest R^2 score, the best-performing model was selected to produce forecasts for the week of January 1 to January 7, 2023. This process was repeated weekly until May 31, 2023, amounting to a total of 20 iterations. Each week, the models were retrained with a cumulative training series that incorporated data from the previous week. Every Sunday, the forecasting results were compiled and presented to the emergency department manager for approval. This process allowed for prompt adjustments to the staffing schedule based on the most recent data, ensuring that staffing levels were aligned with the predicted patient demand for the upcoming week. Additionally, model deviation was monitored using the Kolmogorov-Smirnov test, and models showing a significant difference in performance metrics were disregarded.

Shift optimization by forecasted data

Shift scheduling involved 3 patterns: 08-16 (08:00-16:00), 16-24 (16:00-24:00), and 24-08 (24:00-08:00), with 4

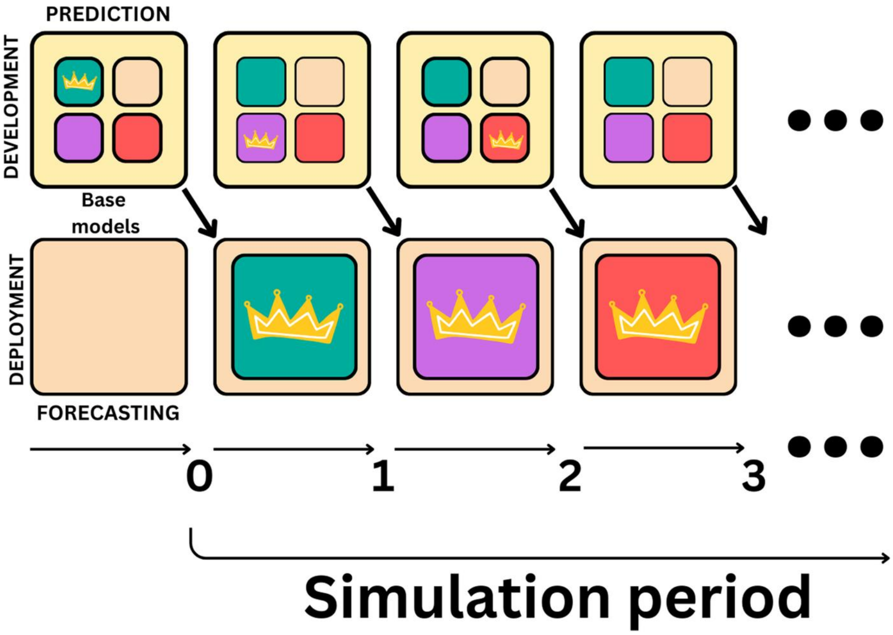

Figure 1. Overview of model training and deployment in the MLOps simulation. In the MLOps simulation, 2 primary areas are identified: the development area, where models are trained and tested, and the deployment area, where models are utilized in real-world applications. In this simulation, models trained on weekly cumulative data in the development phase were evaluated for prediction performance using real data. The model achieving the highest R^2 score was then deployed for generating forecasting results for the following week. Initially, base models were developed 1 week prior to the start of the simulation. Actual models were determined at the commencement of the simulation and were subsequently used every week until the simulation concluded.

standard physicians per shift. Weekly predictions were generated using hourly forecasting models, with patient admissions as the workload metric. April 2023 served as an example. The linear programming method from Python's Pulp library (version 2.9.0) was used to optimize shifts based on forecasted hourly admissions. The optimization ensured a minimum of 2 and a maximum of 8 physicians per shift, without requiring external labor. This process was repeated for 4 weeks in April 2023, and the impact was assessed by comparing patient-to-physician ratios across the period.

Results

Among the 352 843 total PED visits, the median age of the patients was 75 months [interquartile range (IQR): 24-117], 54.3% were male, and 45.6% were female. The most frequent complaint was fever at 18.31% (n = 58 937), followed by cough at 12.49% (n = 44 070), nausea and vomiting at 5.51% (n = 19 464), general weakness at 5.43% (n = 19 159), and abdominal pain at 4.71% (n = 16 620). The most common hours for patient visits were 21:00 (n = 28 002), 20:00 (n = 27 925), and 22:00 (n = 23 676). The busiest days of the week were Sunday (n = 58 021) and Saturday (n = 53 536), and the busiest months were January (n = 41 028), December (n = 36 526), and November (n = 32 040). Monthly evaluations of visit numbers up to 2023, as shown in Figure 2, revealed a sudden drop in 2020 due to the COVID-19 pandemic and the resulting quarantine, which dramatically reduced the number of visits. By June 2021, after the pandemic measures were fully lifted, the visit patterns returned to previous levels starting in August. In August 2022, a dramatic increase in visits was observed due to a regional infection outbreak, reaching unprecedentedly high numbers.

The ADF test conducted on the hourly aggregated data resulted in a test statistic of −7.11 with a *P*-value of less than 3e−10. Therefore, the ADF test rejects the null hypothesis ($P < .05$) of non-stationarity, indicating that the residuals are stationary. Consequently, the data were subjected to aggregation based on the number of hourly applications. In the development phase of the experimental model, the adequacy of baseline models was evaluated using time-series analysis techniques. It was observed that advanced deep learning architectures such as TCN (R^2: 73%), TiDE-RIN (R^2: 75%), N-BEATS (R^2: 72%), and N-HiTS (R^2: 57%) outperformed traditional time-series models like ARIMA (R^2: 16%), LSTM (R^2: 32%), and RNN (R^2: 40%). Many of these models were significantly impacted by concept drift during the COVID-19 period and struggled to improve their performance in the post-COVID era. Consequently, the models were retrained using data exclusively from the post-COVID period, but this approach resulted in lower performance compared to training with the entire dataset. The 4 selected models demonstrated superior forecasting performance when trained on data encompassing both pre- and post-COVID periods, offering better generalization, scalability, and adaptability to the problem at hand. Therefore, model training continued using the complete dataset covering the entire study period. As a result, these models were chosen for further development in the project, with a focus on enhancing time-series forecasting accuracy.

Per the simulation protocol (Figure 3), the models were trained weekly on Sundays using a dataset that also incorporated data from the preceding week. A 1-month forecasting was conducted with these trained models, and the predictions for next week were stored in the database for subsequent model drift analysis. The performance of the models was assessed by comparing the weekly forecasts to the actual data, with the weekly R^2 scores presented in Figure 4. Throughout the simulation, the performance metrics of the 20 developed models are displayed.

Throughout the simulation, it was observed that the median R^2 score for all models was 44% (IQR: −2.5% to 60%). On the other hand, when only the first model was

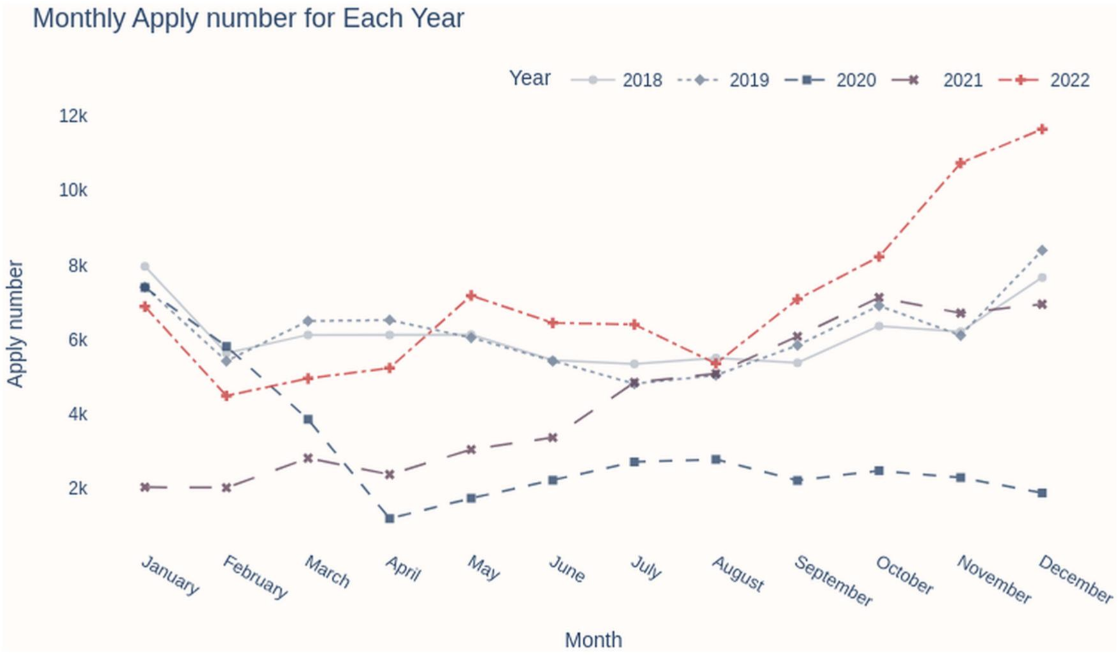

Figure 2. Monthly patient visit numbers by year. Monthly patient visit trends from 2018 to 2022 are shown. The graph displays the fluctuations in the number of patient visits each month, highlighting significant variations across different years.

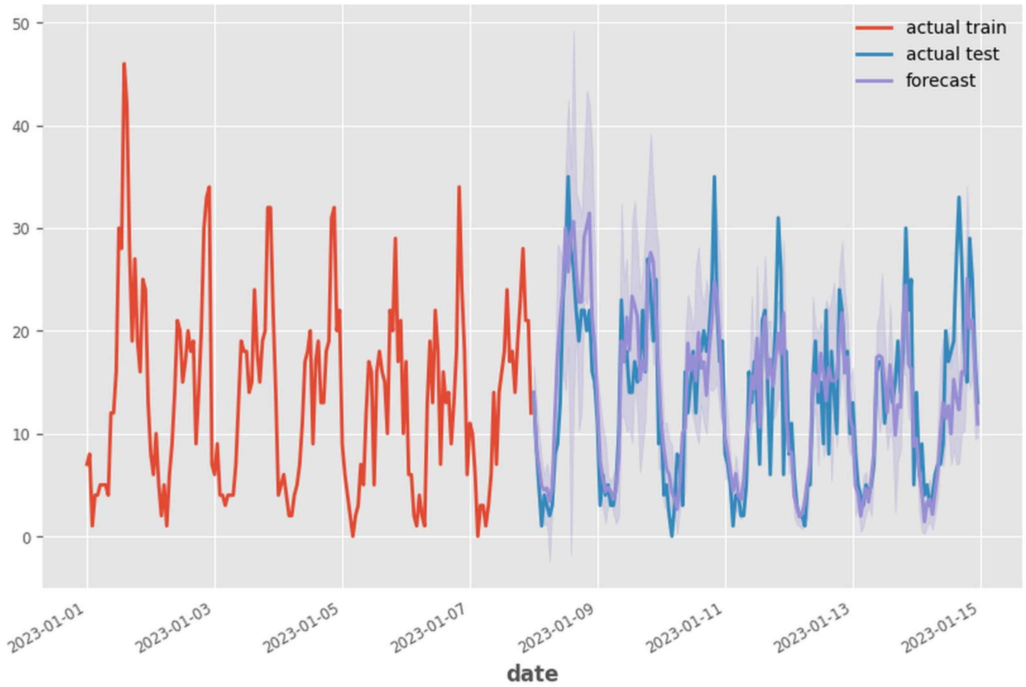

Figure 3. Comparison of TCN model forecasts with actual data for a 1-week simulation cycle. Monthly patient visit trends from 2018 to 2022 are shown. The graph displays the fluctuations in the number of patient visits each month, highlighting significant variations across different years.

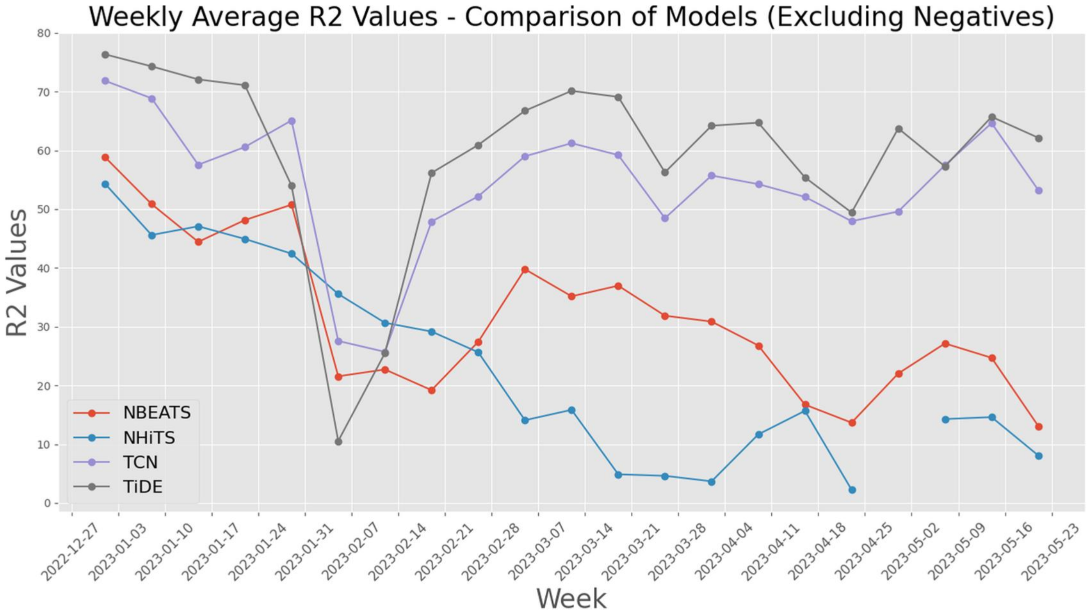

Figure 4. Weekly average R^2 values for forecasting models. After developing the TCN and TiDE-RIN models, 50 different forecasts were made, and the average values of these results were used to create the final forecasting results. In the line graph above, a cycle of the first week of the simulation for the TCN model is shown. The red line represents the training data, the purple shaded area indicates the distribution range of the fifty different forecasting results, and the central line within this area shows their average values. The blue line represents the actual data of the forecasting horizon. N-BEATS = Neural Basis Expansion Analysis; N-HiTS = Neural High-order Time Series model; TCN = Temporal Convolutional Network; TIDE = Time-series Dense Encoder.

used throughout the entire simulation, the median value for the 4 models was calculated as 12% (IQR: −12.3% to 36.3%). During the prediction phase, the median R^2 score for models selected based on the highest R^2 score monitoring was 63% (IQR: 58% to 69.5%). For the forecasting phase, the median R^2 score of the selected models was 60% (IQR: 57% to 67.2%). Detailed weekly performances of the model R^2 scores are presented in a heatmap in Figure 5.

Before the optimization, it was standard for 4 doctors to work each shift, with each doctor expected to attend to an average of 16 patients per shift. After the optimization,

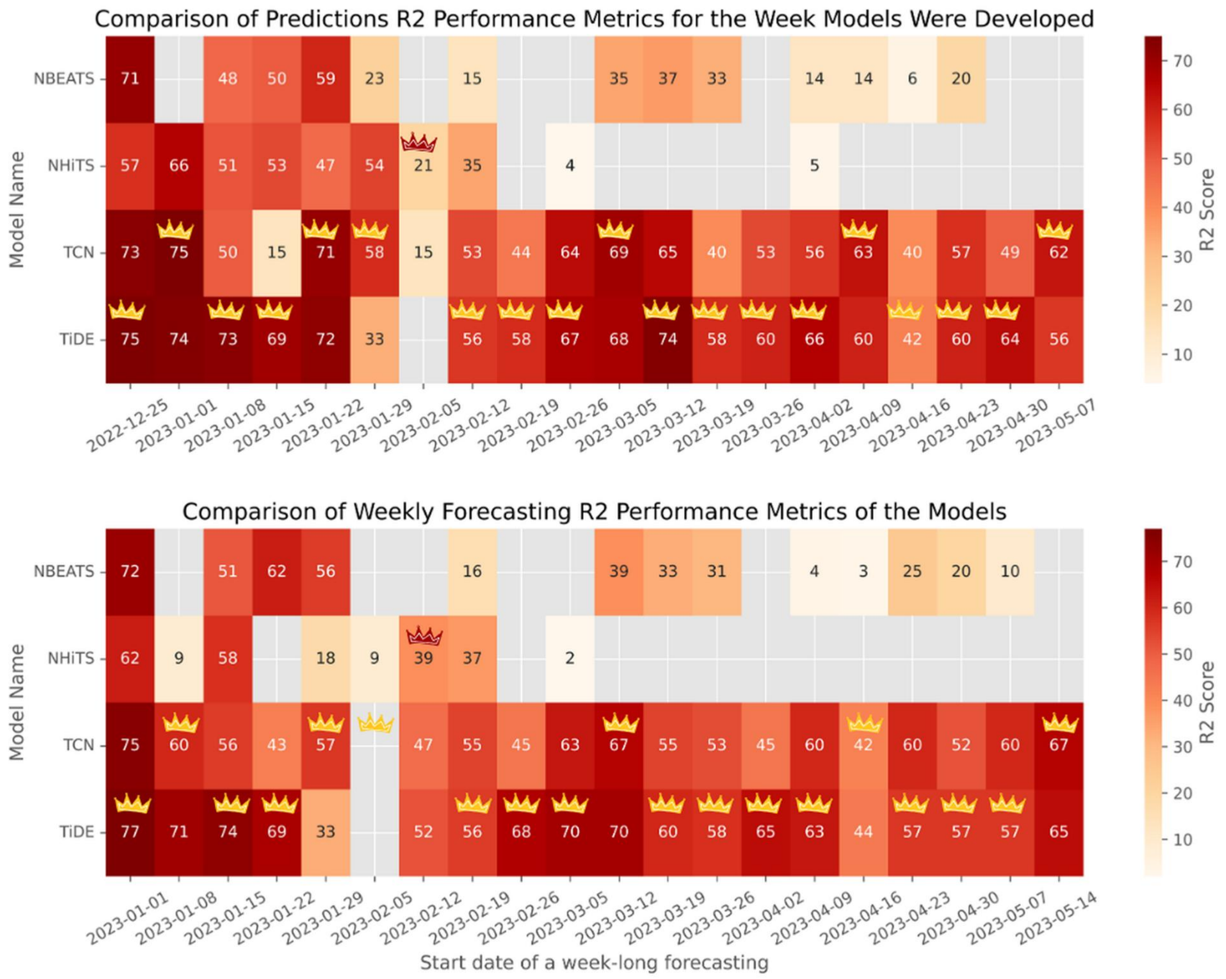

Figure 5. Weekly R^2 performance monitoring of selected models during development and deployment phases. The heatmap presented illustrates the performance monitoring of 4 selected models throughout the simulation period. Weekly R^2 scores are depicted, with values less than zero omitted to enhance the clarity of the color scale on the right. The upper heatmap details the performance of models during the development phase, highlighting those with the highest R^2 scores marked with a crown. Conversely, the lower heatmap demonstrates the forecasting outcomes when these developed models were deployed and tested against actual data 1 week later. This layout facilitates a direct comparison between the models' performances during their development and after their deployment, enhancing the understanding of their predictive accuracy in real-world scenarios. During the week of February 5, a significant disaster led to concept drift, affecting the forecasting accuracy, which was evaluated as underfitted. N-BEATS = Neural Basis Expansion Analysis; N-HiTS = Neural High-order Time Series model; TCN = Temporal Convolutional Network; TIDE = Time-series Dense Encoder.

changes were made in 69 out of 84 shifts. Following the shift optimization, it was observed that 30.4% (n = 21) of shifts were assigned 2 physicians, 10.1% (n = 7) of shifts were assigned 3 physicians, 47.8% (n = 33) of shifts were assigned 5 physicians, and 11.5% (n = 8) of shifts were assigned 6 physicians, with 85% of shifts modified accordingly. In the optimized shifts, an analysis of the number of patients seen per physician in shifts where 4 doctors were not working revealed that, on average, each physician saw 3.75 (IQR: -4.25 to 4.55) fewer patients per shift. When examined across 3 different shifts, the median number of patients seen per physician decreased by 4.32 (IQR: 4.01 to 5.13) during the 8-16 shift, by 4.40 (IQR: 3.88 to 4.75) during the 16-24 shift, while it increased by 5.37 (IQR: 3.23 to 6.06) during the 24-08 shift. These results are visualized in Figure 6.

Discussion

In this study, time-series forecasting models, automated via MLOps architecture, were used to forecast hourly patient admissions on a monthly basis in a simulation environment with a retrospective dataset, enabling the use of the SO-SAFED technique. The forecast results were then utilized to optimize PED physician shifts, ensuring appropriate workforce distribution.

The Pediatrics Optimizing Pediatric Patient Safety in the Emergency Care Setting policy emphasizes data science and AI for improving patient safety and outcomes.[23] AI-based clinical decision support is increasingly used for diagnosing patients, predicting admissions, and estimating hospitalization needs.[24–27] In this study, optimizing physician shifts based on patient volume demonstrated the potential to improve hospital staff management. Similarly, machine learning methods help manage operating room demand by predicting surgery durations, reducing unnecessary cancellations and optimizing surgery schedules.[28–30] These studies show the benefits of AI in optimizing hospital workflows, but more validated studies are needed to demonstrate their real-world benefits.[7]

The potential for unexpected surges in emergency department crowding has always been a critical area of research.

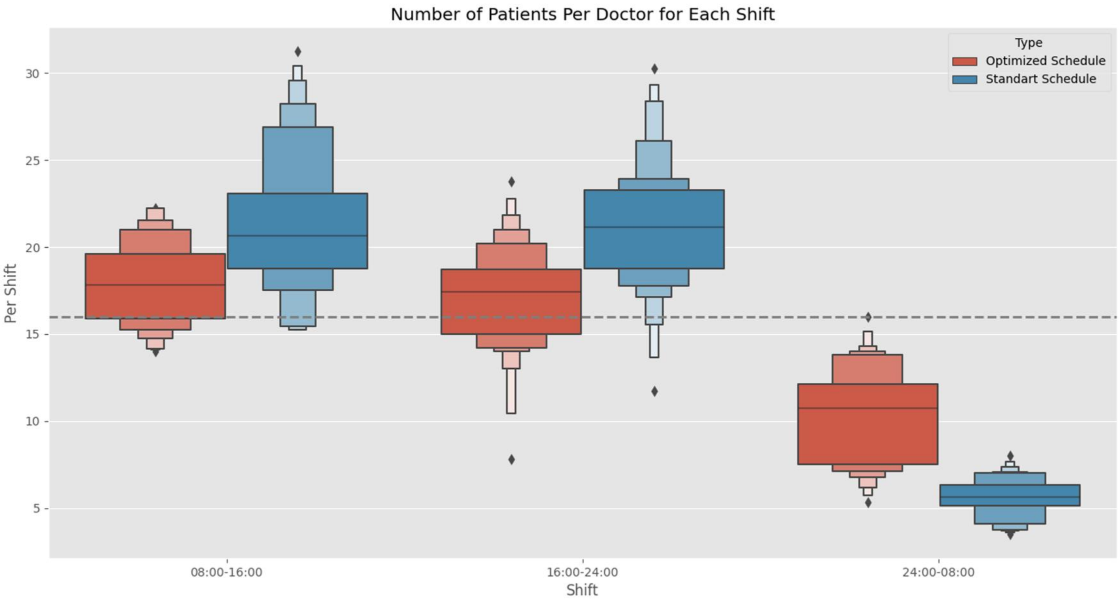

Figure 6. Comparison of patient distribution per physician across different shift schedules. The illustrated boxen plot shows the distributions in deciles, with a horizontal dashed line indicating an average of approximately 16 patients per physician when 4 physicians are assigned per shift as standard. The blue and red boxen plots represent the patient distribution per physician for 3 daily shifts according to the shift schedule, with the red indicating the optimized shifts and the blue indicating the standard schedule. According to the plot, the number of patients per physician decreased during the 8-16 and 16-24 shifts, while it increased during the 24-08 shift. However, the increased number of patients during this shift is still shown to be below the average number of patients.

Unlike previous studies, this research not only utilizes retrospective data but also integrates continuous data streams through the MLOps architecture, enabling dynamic model training processes and providing seamless weekly and monthly forecasting capabilities. Previous studies, such as EDWIN, ICMED, SEAL, and NEDOCS, predicted crowding severity based on real-time data or provided short-term forecasts (4-5 hours) with diminishing performance over time.[4,31] Whitt et al. demonstrated that in classical time forecasting models such as SARIMA and ARIMA, forecasting horizons longer than 4 hours decreased model performance. They stated that a 4-hour forecasting horizon was sufficient, as patient processes would be completed by then.[30] However, it was observed that although crowding could be predicted in the PED with this approach, the advantage of this prediction could not be realized in real life. In contrast, SO-SAFE aims to optimize the workload based on the forecasted crowding. In the studies by Harrou et al., the R^2 metric for forecasting performance exceeded 90% for 1-hour predictions but dropped to around 50% for 4-hour forecasts. In the SO-SAFE study, among the 4 most consistently accurate models developed experimentally, the one with the highest weekly forecasting R^2 performance was automatically selected, consistently providing a relatively high accuracy of around 70%.

In this study, R^2 was chosen as the primary performance metric for its consistency and interpretability across datasets. Unlike MAE, MSE, MAPE, SMAPE, and RMSE, which are influenced by data scale and distribution and can be difficult to interpret, R^2 provides a standardized measure of a model's explanatory power, ranging from 0 to 1 for good fits. A value of 1 indicates a perfect fit, while negative values highlight poor performance, offering a clearer assessment of regression models. Chicco et al. reported that R^2 is more informative and realistic than SMAPE and avoids the interpretability limitations of MSE, RMSE, MAE, and MAPE, recommending it as the standard metric for evaluating regression analyses in any scientific domain.[32]

Advanced deep learning models like TCN, TiDE-RIN, N-BEATS, and N-HiTS outperform traditional time-series models (ARIMA, LSTM, RNN) by effectively capturing complex temporal dependencies and patterns. Traditional models often struggle with long-term dependencies and non-linear relationships, especially in dynamic environments like post-COVID-19 healthcare settings. TCN models long sequences without information loss, and TiDE-RIN captures intricate temporal interactions, providing robust forecasting frameworks. N-BEATS and N-HiTS enhance predictive accuracy by decomposing time-series data into meaningful components and modeling multi-scale temporal patterns. The decline in model performance with post-pandemic data underscores the challenges of data shifts, highlighting the need for models that are both accurate and adaptable to changing conditions. Therefore, these advanced models are more suitable for continued development, offering a strong foundation for improving time-series forecasts in pediatric emergency room settings.

Although machine learning (ML) has existed since the 1980s, early failures led to the "AI winter."[33] Recent advancements, however, have successfully integrated AI into various business processes, driving significant transformations across sectors.[34–36] In healthcare, ML has revolutionized clinical decision-making, telemedicine, and patient management.[37,38] However, many projects fail to progress from development to real-world application due to manual workflow management that hinders full automation.[39,40] This underscores the need for standardized, automated ML processes, positioning MLOps as critical.[41] MLOps merges software engineering with ML development to ensure efficient pipelines, training, and deployment, maintaining security and efficiency through continuous monitoring.[42,43] By integrating automation, Continuous Integration (CI), and Continuous Delivery (CD), MLOps facilitates efficient deployment and sustained accuracy.[44] Despite its importance, MLOps research is nascent; understanding factors affecting its implementation is crucial as ML adoption grows.[39,43,45]

In time-series models, AI learns patterns from longitudinal data for accurate forecasts. Retrospective data enhances performance more than prospective data. Streaming data models need continuous updates; manual updates cause inefficiency and reduced performance. By incorporating prospective data post-deployment, MLOps enhances robustness against data or concept drift, preventing performance degradation from infrequent retraining.[22] This study demonstrates how MLOps automates data ingestion, model training, and deployment, enabling continuous updates and improving prediction accuracy, making the transition from theory to practice more reliable.

In simulations, automatic model selection in response to data drift improved median performance by approximately 16%. Without MLOps, median R^2 values were around 12%; with MLOps, performance improved by 38%, due to robust monitoring, automatic retraining, and seamless data integration. MLOps also enhances collaboration, governance, and resource optimization, leading to better decision-making and operational efficiency.

Data drift can significantly impact the performance of ML-based software in medical applications, as abrupt shifts in data distribution can challenge models trained on historical data patterns.[46,47] In our dataset, 2 major instances of data and concept drift were observed: the COVID-19 pandemic and the February 2023 earthquake that affected a significant portion of Turkey.

In many time-series analyses, the COVID-19 period is excluded due to the concept drift it introduces, with modeling typically starting from the post-pandemic normalization phase. However, this approach substantially reduces the dataset length, limiting the model's ability to capture historical trends. In our study, we found that including the entire dataset from 2018 to 2023, covering both pre- and post-COVID periods, resulted in better model performance compared to using only post-COVID data. The extended time series allowed the models to capture a wider range of patterns and seasonal trends, thereby enhancing predictive accuracy.

Following the February 2023 earthquake in Turkey, a significant drop in patient admissions introduced a sudden data drift that impacted model accuracy. For instance, this disruption led to decreased predictive performance in previously effective models, such as N-HiTS and N-BEATS, which struggled to adapt. In contrast, models like TiDE-RIN and TCN, which utilize recurrent example normalization, demonstrated greater resilience and accuracy under these conditions, with predictive performance improving as more post-earthquake data became available (Figure 4).[21] This experience underscores the importance of accounting for data and concept drift, as these factors critically influence model adaptability and robustness in the face of unforeseen events.

Without adding physicians, the algorithm-optimized shifts showed notable improvements. Initially, both the busiest (16-24) and least busy (24-08) shifts had the same number of physicians due to equal distribution. Post-optimization, patient load per physician increased during the 24-08 shift but stayed below the average of 16 patients. Conversely, during the 16-24 shift, each physician saw nearly 5 fewer patients, just above the overall average. Without optimization, physicians during peak hours would have managed 5 more patients each. The optimization reduced per-physician patient load during the busiest hours, enhancing resource utilization efficiency. However, it did not account for unmeasured variables; it was based on hourly aggregated patient arrivals, while patients have diverse medical needs. Consequently, fewer arrivals during the 24-08 shift may still involve ongoing cases from the busier 16-24 shift.

It is important to note that our hospital traditionally employed a steady-state staffing model, assigning 4 physicians per shift regardless of temporal fluctuations in patient volume. This practice, intended to maintain fairness by equalizing shift assignments among physicians, led to significant discrepancies between staffing levels and actual patient demand. Our optimization intervention highlighted this inefficiency and demonstrated the potential benefits of adjusting staffing based on predictive models. However, we acknowledge that this baseline staffing strategy is not representative of the dynamic scheduling practices used in many other healthcare settings.

The adoption of AI in healthcare encounters significant challenges due to the complex nature of health data, hindering its widespread use. Effective implementation requires strategies that support behavioral changes and clear results for various roles.[7] Technical challenges related to model validation and integration into clinical workflows must be addressed with consideration of ethical, legal, and moral issues. The perception of AI by healthcare providers and users is crucial for acceptance. We encountered difficulties in real-world application, necessitating further studies to ensure program acceptance and adaptation. Transparent model explanations and careful validation are vital for accuracy and reliability, contributing positively to patient care and potentially reducing clinician burnout.[48] Organizing working hours based on patient arrivals may improve care quality and decrease burnout, though more research is needed.[49]

The limitations of this study are its focus on data from a single PED, which limits the generalizability of the findings. To enhance applicability, future studies must collaborate with multiple departments to collect larger and more diverse datasets. Federated learning (FL) represents a significant advancement in this context, allowing data processing from different centers without centralization, thereby preserving privacy while enabling extensive data collection.[50] Reviews indicate that FL facilitates multicenter collaborations without sharing patient data, which is crucial for evaluating model performance across diverse demographics and pathologies.[51] Employing techniques like FL in future research will help overcome current limitations by continuously evaluating and improving models with data from various centers, ensuring the findings apply to a broader patient population. Focusing on multicenter validation is essential to fully realize the benefits of this innovative approach in enhancing healthcare efficiency and patient outcomes, although additional efforts are needed for real-world implementation.

Additionally, our staffing model and prediction timing present limitations. We optimized a staffing strategy that uniformly assigned 4 physicians per shift, not accounting for fluctuations in patient volume, which may not reflect settings with dynamic staffing practices. Also, generating predictions on Sundays for the upcoming week provides limited time to adjust staff schedules, affecting practicality. Future research should implement the optimization model in environments with variable staffing and develop forecasting models that provide longer lead times without compromising accuracy. We are currently applying these adjustments in a real-world setting to evaluate their effectiveness.

Conclusion

In conclusion, our study successfully applied a novel MLOps architecture to forecast PED overcrowding and optimize physician shift schedules. This approach demonstrated superior performance compared to traditional forecasting methods by utilizing dynamic and automated machine learning processes, ensuring continuous data updates and adjustments for data drift. Tested in a simulated environment, the SO-SAFED system has demonstrated that it optimizes workforce allocation, has the potential to improve resource allocation, and may improve patient care by anticipating periods of high demand. This study is the first to utilize MLOps architecture for PED overcrowding forecasting, contributing significantly to the literature.

Author contributions

Izzet Turkalp Akbasli conceptualized the study, developed the methodology, handled the software, validated the results, performed the formal analysis, and contributed to data curation, writing the original draft, and visualization. Ahmet Ziya Birbilen contributed to the conceptualization, validation, resources, data curation, and drafting the original manuscript. Ozlem Teksam was responsible for the investigation, provided resources, reviewed and edited the manuscript, supervised the project, managed project administration, and acquired funding. All authors have read and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicts of interest

All authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data produced in the present study are available upon reasonable request to the authors.

References

1. Malhotra A, Molloy EJ, Bearer CF, et al. Emerging role of artificial intelligence, big data analysis and precision medicine in pediatrics. *Pediatr Res*. 2023;93:281-283. https://doi.org/10.1038/s41390-022-02422-z
2. Nijman J, Zoodsma RS, Koomen E. A strategy for artificial intelligence with clinical impact-eyes on the prize. *JAMA Pediatr*. 2024;178:219-220. https://doi.org/10.1001/jamapediatrics.2023.6259
3. Ramgopal S, Sanchez-Pinto LN, Horvat CM, et al. Artificial intelligence-based clinical decision support in pediatrics. *Pediatr Res*. 2023;93:334-341. https://doi.org/10.1038/s41390-022-02226-1
4. Harrou F, Dairi A, Kadri F, et al. Forecasting emergency department overcrowding: a deep learning framework. *Chaos Solitons Fractals*. 2020;139:110247. https://doi.org/10.1016/j.chaos.2020.110247
5. Lee Y-C, Ng C-J, Hsu C-C, et al. Machine learning models for predicting unscheduled return visits to an emergency department: a scoping review. *BMC Emerg Med*. 2024;24:20. https://doi.org/10.1186/s12873-024-00939-6
6. Kwon J-M, Jeon K-H, Lee M, et al. Deep learning algorithm to predict need for critical care in pediatric emergency departments. *Pediatr Emerg Care*. 2021;37:e988-e994. https://doi.org/10.1097/pec.0000000000001858
7. Sahni NR, Carrus B. Artificial intelligence in U.S. Health Care Delivery. *N Engl J Med*. 2023;389:348-358. https://doi.org/10.1056/NEJMra2204673
8. Moylan A, Maconochie I. Demand, overcrowding and the pediatric emergency department. *CMAJ*. 2019;191:E625-E626. https://doi.org/10.1503/cmaj.190610
9. Doan Q, Wong H, Meckler G, et al.; for Pediatric Emergency Research Canada (PERC). The impact of pediatric emergency department crowding on patient and health care system outcomes: a multicentre cohort study. *CMAJ Can Med Assoc J*. 2019;191:E627-E635. https://doi.org/10.1503/cmaj.181426
10. Yazıcı MU, Teksam O, Agın H, et al. The burden of burnout syndrome in pediatric intensive care unit and pediatric emergency department: a multicenter evaluation. *Pediatr Emerg Care*. 2021;37:e955-e961. https://doi.org/10.1097/PEC.0000000000001839
11. Sartini M, Carbone A, Demartini A, et al. Overcrowding in emergency department: causes, consequences, and solutions—a narrative review. *Healthcare*. 2022;10:1625. https://doi.org/10.3390/healthcare10091625
12. Pines JM, Hilton JA, Weber EJ, et al. International perspectives on emergency department crowding. *Acad Emerg Med*. 2011;18:1358-1370. https://doi.org/10.1111/j.1553-2712.2011.01235.x
13. Bernstein SL, Verghese V, Leung W, et al. Development and validation of a new index to measure emergency department crowding. *Acad Emerg Med*. 2003;10:938-942. https://doi.org/10.1111/j.1553-2712.2003.tb00647.x
14. Boyle A, Coleman J, Sultan Y, et al. Initial validation of the International Crowding Measure in Emergency Departments (ICMED) to measure emergency department crowding. *Emerg Med J*. 2015;32:105-108. https://doi.org/10.1136/emermed-2013-202849
15. Wretborn J, Khoshnood A, Wieloch M, et al. Skåne Emergency Department Assessment of Patient Load (SEAL)—a model to estimate crowding based on workload in Swedish emergency departments. *PLoS One*. 2015;10:e0130020. https://doi.org/10.1371/journal.pone.0130020
16. Weiss SJ, Derlet R, Arndahl J, et al. Estimating the degree of emergency department overcrowding in academic medical centers: results of the National ED Overcrowding Study (NEDOCS). *Acad Emerg Med*. 2004;11:38-50. https://doi.org/10.1197/j.aem.2003.07.017
17. Kadri F, Harrou F, Chaabane S, et al. Time series modelling and forecasting of emergency department overcrowding. *J Med Syst*. 2014;38:107. https://doi.org/10.1007/s10916-014-0107-0
18. Aregbesola A, Abou-Setta AM, Okoli GN, et al. Implementation strategies in emergency management of children: a scoping review. *PLoS One*. 2021;16:e0248826. https://doi.org/10.1371/journal.pone.0248826
19. Hirner S, Dhakal J, Broccoli MC, et al. Defining measures of emergency care access in low-income and middle-income countries: a scoping review. *BMJ Open*. 2023;13:e067884. https://doi.org/10.1136/bmjopen-2022-067884

20. Das A, Kong W, Leach A, et al. Long-term forecasting with TiDE: time-series dense encoder. 2023. Accessed April 4, 2024. https://arxiv.org/abs/2304.08424
21. Kim T, Kim J, Tae Y, et al. Reversible instance normalization for accurate time-series forecasting against distribution shift. 2021. Accessed February 13, 2023. https://openreview.net/forum?id=cGDAkQo1C0p
22. Tembhekar P, Malaiyappan JNA, Shanmugam L. Cross-domain applications of MLOps: from healthcare to finance. *JKLST*. 2023;2:581-598. https://doi.org/10.60087/jklst.vol2.n2.p598
23. Joseph MM, Mahajan P, Snow SK, American Academy of Pediatrics Committee on Pediatric Emergency Medicine, American College of Emergency Physicians Pediatric Emergency Medicine Committee, and Emergency Nurses Association Pediatric Committee, et al. Optimizing pediatric patient safety in the emergency care setting. *Pediatrics*. 2022;150:e2022059673. https://doi.org/10.1542/peds.2022-059673
24. Silva E, Pereira MF, Vieira JT, et al. Predicting hospital emergency department visits accurately: a systematic review. *Int J Health Plann Manage*. 2023;38:904-917. https://doi.org/10.1002/hpm.3629
25. Kao H-J, Chien T-W, Wang W-C, et al. Assessing ChatGPT's capacity for clinical decision support in pediatrics: a comparative study with pediatricians using KIDMAP of Rasch analysis. *Medicine (Baltimore)*. 2023;102:e34068. https://doi.org/10.1097/MD.0000000000034068
26. Adebayo O, Bhuiyan ZA, Ahmed Z. Exploring the effectiveness of artificial intelligence, machine learning and deep learning in trauma triage: a systematic review and meta-analysis. *Digit Health*. 2023;9:20552076231205736. https://doi.org/10.1177/20552076231205736
27. Leonard F, O'Sullivan D, Gilligan J, et al. Supporting clinical decision making in the emergency department for paediatric patients using machine learning: a scoping review protocol. *PLoS One*. 2023;18:e0294231. https://doi.org/10.1371/journal.pone.0294231
28. Bellini V, Guzzon M, Bigliardi B, et al. Artificial intelligence: a new tool in operating room management. Role of machine learning models in operating room optimization. *J Med Syst*. 2019;44:20. https://doi.org/10.1007/s10916-019-1512-1
29. Ozen A, Marmor Y, Rohleder T, et al. Optimization and simulation of orthopedic spine surgery cases at Mayo Clinic. *M&SOM*. 2016;18:157-175.
30. Whitt W, Zhang X. Forecasting arrivals and occupancy levels in an emergency department. *Oper Res Health Care*. 2019;21:1-18. https://doi.org/10.1016/j.orhc.2019.01.002
31. Cheng Q, Argon NT, Evans CS, et al. Forecasting emergency department hourly occupancy using time series analysis. *Am J Emerg Med*. 2021;48:177-182. https://doi.org/10.1016/j.ajem.2021.04.075
32. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput Sci*. 2021;7:e623. https://doi.org/10.7717/peerj-cs.623
33. Lloyd JW. Surviving the AI winter. In: *Logic Programming: The 1995 International Symposium*. MIT Press; 1995:33-47.
34. Mariani MM, Perez-Vega R, Wirtz J. AI in marketing, consumer research and psychology: a systematic literature review and research agenda. *Psychol Market*. 2022;39:755-776. https://doi.org/10.1002/mar.21619
35. Hildebrand C, Bergner A. AI-driven sales automation: using chatbots to boost sales. *NIM Mark Intell Rev*. 2019;11:36-41. https://doi.org/10.2478/nimmir-2019-0014
36. Li RC, Tee ML. Developing an implementation framework for automated customer support service in collaborative customer relationship management systems. In: *2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Singapore, Singapore. IEEE; 2021:1092-1096. https://doi.org/10.1109/IEEM50564.2021.9672894
37. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis*. 2018;66:149-153. https://doi.org/10.1093/cid/cix731
38. Das SD, Bala PK. Artificial intelligence in Telemedicine: a brief survey. In: Mishra S, Tripathy HK, Mallick P, et al., eds. *Augmented Intelligence in Healthcare: A Pragmatic and Integrated Analysis*. Springer Nature; 2022:453-461.
39. Kreuzberger D, Kühl N, Hirschl S. Machine Learning Operations (MLOps): overview, definition, and architecture. *IEEE Access*. 2023;11:31866-31879. https://doi.org/10.1109/ACCESS.2023.3262138
40. Lwakatare LE, Crnkovic I, Rånge E, et al. From a data science driven process to a continuous delivery process for machine learning systems. In: Morisio M, Torchiano M, Jedlitschka A, eds. *Product-Focused Software Process Improvement*. Springer International Publishing; 2020:185-201.
41. John MM, Olsson HH, Bosch J. Towards MLOps: a framework and maturity model. In: *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Palermo, Italy. IEEE; 2021:1-8. https://doi.org/10.1109/SEAA53835.2021.00050
42. Li B, Qi P, Liu B, et al. Trustworthy AI: from principles to practices. *ACM Comput Surv*. 2023;55:1-46.. https://doi.org/10.1145/3555803
43. Das SD, Bala PK. What drives MLOps adoption? An analysis using the TOE framework. *J Decis Syst*. 2024;33:376-412. https://doi.org/10.1080/12460125.2023.2214306
44. Granlund T, Stirbu V, Mikkonen T. Towards regulatory-compliant MLOps: Oravizio's journey from a machine learning experiment to a deployed certified medical product. *SN Comput Sci*. 2021;2:342. https://doi.org/10.1007/s42979-021-00726-1
45. Recupito G, Pecorelli F, Catolino G, et al. A multivocal literature review of MLOps tools and features. In: *2022 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Gran Canaria, Spain. IEEE; 2022:84-91.
46. Sahiner B, Chen W, Samala RK, et al. Data drift in medical machine learning: implications and potential remedies. *Br J Radiol*. 2023;96:20220878. https://doi.org/10.1259/bjr.20220878
47. Rahmani K, Thapa R, Tsou P, et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. medRxiv, https://doi.org/10.1101/2022.06.06.22276062, 2022, preprint: not peer reviewed.
48. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019;25:1337-1340. https://doi.org/10.1038/s41591-019-0548-6
49. Rotenstein LS, Torre M, Ramos MA, et al. Prevalence of burnout among physicians: a systematic review. *JAMA*. 2018;320:1131-1150. https://doi.org/10.1001/jama.2018.12777
50. Dhade P, Shirke P. Federated learning for healthcare: a comprehensive review. *Eng Proc*. 2023;59:230. https://doi.org/10.3390/engproc2023059230
51. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep*. 2020;10:12598. https://doi.org/10.1038/s41598-020-69250-1

JAMIA Open, 2025, 8, 1–10
https://doi.org/10.1093/jamiaopen/ooae138
Research and Applications