BMC Health Services Research

RESEARCH

Open Access

Interpretable machine learning models for prolonged Emergency Department wait time prediction

Hao Wang[1*], Nethra Sambamoorthi[2], Devin Sandlin[1] and Usha Sambamoorthi[3]

Abstract

Objective Prolonged Emergency Department (ED) wait times lead to diminished healthcare quality. Utilizing machine learning (ML) to predict patient wait times could aid in ED operational management. Our aim is to perform a comprehensive analysis of ML models for ED wait time prediction, identify key feature importance and associations with prolonged wait times, and interpret prediction model clinical relevance among ED patients.

Methods This is a single-centered retrospective study. We included ED patients assigned an Emergency Severity Index (ESI) level of 3 at triage. Patient wait times were categorized as <30 minutes and ≥30 minutes (prolonged wait time). We employed five ML algorithms - cross-validation logistic regression (CVLR), random forest (RF), extreme gradient boosting (XGBoost), artificial neural network (ANN), and support vector machine (SVM) - for predicting patient prolonged wait times. Performance assessment utilized accuracy, recall, precision, F1 score, false positive rate (FPR), and false negative rate (FNR). Furthermore, using XGBoost as an example, model key features and partial dependency plots (PDP) of these key features were illustrated. Shapley additive explanations (SHAP) were employed to interpret model outputs. Additionally, a top key feature interaction analysis was conducted.

Results Among total 177,665 patients, nearly half of them (48.20%, 85,632) experienced prolonged ED wait times. Though all five ML models exhibited similar performance, minimizing FNR is associated with the most clinical relevance for wait time predictions. The top features influencing patient wait times and gaining the top ranked interactions were ED crowding condition and patient mode of arrival.

Conclusions Nearly half of the patients experienced prolonged wait times in the ED. ML models demonstrated acceptable performance, particularly in minimizing FNR when predicting ED wait times. The prediction of prolonged wait times was influenced by multiple interacting factors. Proper application of ML models to clinical practice requires interpreting their predictions of prolonged wait times in the context of clinical significance.

Keywords Emergency department, Wait time, Machine learning, Performance

*Correspondence:
Hao Wang
hwang@ies.healthcare
[1]Department of Emergency Medicine, John Peter Smith Health Network, Integrative Emergency Services, 1500 S. Main St., Fort Worth, TX 76104, USA
[2]CRM Portals LLC, Fort Worth, TX 76126, USA
[3]College of Pharmacy, University of North Texas Health Science Center, 3500 Camp Bowie Blvd, Fort Worth, TX 76107, USA

Introduction

Emergency Department (ED) wait time, defined as the time interval between patients who completed the triage process, and the time that patients are placed to the examination room, is one of the ED flow metrics used for quality care measurement [1, 2]. However, ED wait time can be affected by multiple factors, including patient demographics, level of severity, transportation methods, and ED crowding statuses, etc. [3, 4]. Prolonged ED wait time leads to poor healthcare quality with an increased rate of patients left without being seen, poor patient satisfactions, and worsening of patient clinical outcomes [5–7]. Based upon the national ED survey, it is recommended that an ideal ED wait time should be less than 30 minutes [8].

Since ED wait time is an important quality metric, previous studies to predict ED wait time have been reported [9, 10]. Hemaya and Locker used a linear regression to predict ED wait time and found an average of 29 minute difference between the actual and predicted patient wait time [9]. Sun et al. reported good accuracies with the use of quantile regression models for ED wait time predictions. However, their predictions were grouped according to the patients' different triage acuity levels, which limited the models' broader applicability [10]. Previous attempts to predict ED wait times using the benchmark rolling average model demonstrated a lack of accuracy [11]. Researchers in Stanford university used a Q-Lasso regression model to predict patient wait time and found it more accurate than the rolling average prediction [12]. Overall, traditional statistical methods for predicting ED wait times have shown lower accuracy, limiting their clinical utility.

Recently, various ML algorithms have been employed to predict ED wait times, as reported in the literature [13, 14]. Cheng and Kuo developed a Long Short-Term Memory (LSTM) recurrent neural network model for predicting ED wait times, finding an average discrepancy of 17 minutes between actual and predicted wait times [13]. Similarly, Hijry and Olawoyin applied a deep learning stochastic gradient descent algorithm, achieving a lowest average difference of 10.8 minutes [14]. However, considering that an ideal ED wait time is generally less than 30 minutes, these predictions remain suboptimal. To enhance clinical impact on ED operational management, it may be more effective to predict wait times as a categorical indicator rather than as a continuous variable.

Additionally, before healthcare prediction models are applied in practice, a comprehensive assessment is essential. Typically, metrics such as accuracy, recall, and precision are used to evaluate the performance of ML model predictions [15, 16]. In recent years, there has been a growing emphasis on more thorough assessments by focusing on key parameters that directly relate to clinical applications [17]. This includes the use of techniques such as SHAP values, partial dependence plots (PDP), and analysis of key feature interactions [17]. SHAP values help explain how each feature contributes to the final model prediction, while PDPs illustrate whether a specific feature positively or negatively influences the prediction. Once key features are identified, interactions related to the target prediction can be ranked accordingly. These performance assessments not only provide a comprehensive evaluation of the prediction quality but also help establish standards and quality control measures for ML implementation in healthcare. A systematic review of predictive ML models for ED operational management, conducted by Porto BM, summarized that although numerous ML predictive models have been developed, feature engineering and explainable artificial intelligence remain underexplored in this field [18].

Only with a robust evaluation of ML model performance with appropriate model interpretation can such algorithms be effectively integrated into clinical practice. If ML models can be used to predict patient wait times, targeted interventions can be implemented to reduce delays for specific patients. Therefore, this study aims to: (1) develop various algorithmic models to predict prolonged ED wait times, (2) evaluate the predictive performance of these models, (3) identify key features and feature interactions associated with prolonged wait times, and (4) interpret the model predictions in terms of their clinical relevance for ED patients.

Methods

Study design and setting

We conducted a single-center observational study at an urban tertiary referral hospital. The ED at the study hospital has an annual volume of approximately 120,000 visits. The ED is divided into a main area, staffed by ED physicians and residents who manage high-acuity patients (i.e., ESI 1–3), and a fast-track area, staffed by advanced practice providers, who care for lower-acuity patients (i.e., ESI 4–5). The study was approved by the regional Institutional Review Board with a waiver of informed consent (IRB#1967558-1).

Inclusion and exclusion criteria

To assess ED wait times and the risks associated with prolonged wait times, we included all patients classified as ESI-3 who presented to the ED between January 1, 2019, and December 31, 2021. Patients with acuity levels other than ESI-3 were excluded. The primary objective of using ML models for wait time prediction is to provide alerts for patients experiencing prolonged wait times. Such prediction models are less applicable for ESI 1–2 patients, whose average wait times were typically less than 30 minutes. Conversely, the study's ED features

a dedicated fast-track area, where healthcare providers promptly attend to patients with lower acuity levels (ESI 4–5), minimizing their wait times. As a result, the practical benefit of wait time prediction alerts is most relevant for ESI-3 patients, who represent the majority of cases in our ED. Additionally, we excluded patients who: 1) did not have recorded wait times, 2) had missing sociodemographic data (e.g., age, sex, race, and ethnicity) or clinical information (e.g., comorbidities, mode of arrival, and vital signs at triage), and 3) left before triage completed.

Target variable (prolonged wait time)

Wait time is defined as the time from a patient who completed the triage process to the time that this patient was placed in an examination room. Wait time is recommended to be less than 30 minutes from patients entered to the hospital [8]. Therefore, patient wait time was further categorized into two group: 1) wait time<30minutes, and 2) wait time≥30minutes (i.e., prolonged wait time). We selected a threshold-based approach to guide decision-making, ensure consistent comparisons, and evaluate policy goals aimed at reducing health disparities. The selection of a 30-minute cutoff for prolonged wait times was informed by professional consensus, empirical evidence, and clinical operations management. National surveys and established standards support this benchmark. Academic emergency physicians recommend waiting times of less than 30 minutes, and *Emergency Physician Monthly* classifies this threshold as "excellent" [8, 19]. Empirical evidence further reinforces its validity, for example, the U.S. Government Accountability Office reported median wait times for ESI-3 patients ranging from 15 to 60 minutes, while the National Center for Health Statistics (2006) documented a national median of 30 minutes [20]. From an operational perspective, adopting this threshold aligns with emergency medicine targets, facilitating triage optimization, improving patient flow, and enhancing resource efficiency. Ultimately, this approach contributes to better patient outcomes and increased satisfaction.

Features

Features were divided into three domains: patient sociodemographic, clinical-related, and ED site-related. Patient sociodemographic domain included age, sex (male and female), marital status (single, married, and others), race and ethnicity [Non-Hispanic White (NHW), Non-Hispanic Black (NHB), Hispanic/Latino (Hispanic), Non-Hispanic Asian (NHA), and others], and language speaking (English, Spanish, and others). Patient clinical-related domain included insurance coverage (yes and no), primary care physician assignment (yes and no), patient chronic disease condition (no chronic condition, one chronic condition, and two or more chronic conditions), patient method of arrival at ED (private, ambulance, public transportation, and ambulatory), patient high blood pressure conditions at ED triage (yes and no), and abnormal vital signs (exclude high blood pressure) at ED triage (normal and abnormal). Abnormal vital signs were defined as abnormal if any one of the following criteria were met: heart rate>100 or <60, respiratory rate>20 or <12, systolic BP<90mmHg, diastolic BP<60mmHg, pulse oximetry <95%, and temperature >99.6°F or <95°F. ED site-related domain included ED crowding status (not crowded, crowded, and overly crowded), weekend (Saturday and Sunday) versus weekday (Monday through Friday) presentation, and clinic hours [on and off, clinic was open from 8am to 5pm Monday through Friday, therefore, we defined patients who arrived to ED from 8am to 5pm as having presented during clinic hours (i.e., on), whereas patients arriving to ED from 5pm to 8am Monday through Friday or during the weekends as having presented during non-clinic hours (i.e., off)]. We use SONET score to determine ED crowding status as reported previously [21].

Machine learning algorithms

We chose five ML algorithms to predict patient prolonged wait time including cross-validation logistic regression (CVLR), random forest (RF), extreme gradient boosting (XGBoost), artificial neural network (ANN), and support vector machine (SVM). Data was split into training (70% of data) and testing (30% of data) sections. During the data preprocessing, a feature correlation heatmap was included (Appendix Figure). Hyperparameter tuning and cross-validation were employed to optimize the ML models. Specifically, CVLR tuning focused on optimizing the regularization parameter and the type of regularization. RF tuning involved adjusting the number of trees, maximum tree depth, and minimum samples per split. XGBoost tuning included adjustments to the learning rate, maximum depth, and the number of boosting rounds. ANN tuning encompassed the number of hidden layers, neurons per layer, learning rate, and batch size. SVM tuning involved optimizing the regularization parameter and kernel coefficient. To ensure robust model selection, grid search with fivefold cross-validation was applied to all ML models and the best performance of each ML algorithm was chosen for model prediction. Due to the well-balanced nature of our dataset (e.g., the equivalence between patients who wait ≥30 minutes and those who wait <30 minutes), model under or oversampling investigations could be avoided.

Comprehensive performance assessment of model prediction

In terms of the model performance accuracy, we chose to report both training and testing model accuracy,

recall, precision, and F1 score. In addition, we also reported areas under the receiver operating characteristics (AUROC) of each ML algorithmic model predictions. Furthermore, we report the overall false positive rate (FPR) and false negative rate (FNR) of model prediction with the use of five different ML algorithms. Due to its clinical relevance, we placed greater emphasis on the FNR in our efforts to predict patients' wait times. A false negative occurs when a patient who waits longer than 30 minutes is misclassified by ML algorithms as waiting less than 30 minutes. Furthermore, to perform the global and local interpretation of feature importance, we utilized SHAP to interpret the output from the XGBoost model. We derived a SHAP value beeswarm summary plot including essential features predictive of patient wait time and illustrated their overall directionality. To illustrate the effects of features on predicting patient wait time, we utilized PDP. These plots reveal the patterns associated with the feature and the target, such as linear, monotonic, or complex patterns. Given that age was the only continuous feature in this study, we selected to display the PDP of age with other two top essential categorical features using the XGBoost model. Feature interaction constraints allow the users to decide which features are allowed to interact thus providing better predictive performance. Top key feature interactions were identified from the XGBoost model with the use of Xgbfir package [17]. Ranks of the Gain, FScore, and weighted FScore were reported. Gain refers to the total gain of each feature interaction. FScore is the feature importance score and is the number of splits taken on feature interactions. Weighted FScore is FScore weighted by the splits. A higher gain, FScore, or weighted FScore value all indicate such feature interactions are more influential in the model prediction.

Data analysis

Patients were divided into two groups (wait time<30min and ≥30min). Patients' sociodemographic, clinical-related, and hospital-related variables were compared between these two groups. Continuous data were compared either using the *Student-t* test for mean comparison or using the *Kruskal-Wallis'* test for median comparison. Categorical data were compared using the *Chi-square* test. STATA 14.2 was used for two group comparisons and Python 3.8 was used for different ML model predictions and performance assessments.

Results

A total of 177,665 patients were included in this study. Patient characteristics were compared between two groups based on wait times (i.e., <30 minutes vs. ≥30 minutes, see Table 1). The median wait time in the shorter wait time group (<30 minutes) was 5 minutes, compared to 95 minutes in the prolonged wait time group (≥30 minutes) (p<0.001). Patients in the prolonged wait time group were generally younger, had a higher proportion of Hispanic patients, a greater number of non-English speakers, and more patients without insurance coverage compared to those in the shorter wait time group. Additionally, fewer patients with prolonged wait times arrived via ambulance, while a larger proportion presented to the ED during periods of overcrowding (p<0.001, Table 1).

Five ML algorithms were utilized in this study, including CVLR, RF, XGBoost, ANN, and SVM. The performance accuracies of these models were comparable between the training and testing datasets, suggesting that the models are not overfitting. The final predictions from the testing data were also similar across the five ML algorithms (Table 2). Figure 1 illustrates the overall FPRs and FNRs for the different models. Slight variations in FPRs and FNRs were observed across different ML models. In general, FPR and FNR exhibited an inverse relationship, where an increase in FPR corresponded to a decrease in FNR, and vice versa. Therefore, selecting an appropriate balance between FPR and FNR is essential and should be guided by the specific clinical application.

Essential features that contribute to the XGBoost algorithmic model prediction are shown in the SHAP feature importance plot and beeswarm summary plot (Fig. 2). It is found that patient mode of arrival [e.g., much higher number of patients who arrived by ambulance (moa_ambulance) were predicted waiting<30min than ones who arrived by non-ambulance] and ED crowding status [e.g., much higher number of patients arrived at ED under not crowded status (crowd_notcrowded) were predicted waiting <30min than ones who arrived at ED under crowded status] played important roles on predicted patient wait time at ED. Apart from this, other domains such as patient demographics (e.g., sex), ED visit date (e.g., weekday), or clinical information (e.g., abnormal vital signs at triage: abnvs_yes) may also contribute to the final model prediction. The Beeswarm summary plot illustrates the relationship between key features and wait time predictions. Positive SHAP values indicate a predicted prolonged wait time, while negative SHAP values suggest a predicted wait time of less than 30 minutes. Categorical features are represented as either "0" or "1." For example, when individuals were transported by ambulance (i.e., "1"), the SHAP values are negative, indicating a lower likelihood of experiencing prolonged wait times. Regarding sex, "0" represents female and "1" represents male. The positive SHAP values associated with females (i.e., "0") suggest that female patients are more likely to experience prolonged wait times. We also presented the SHAP values for the first 25 samples (Fig. 3). Our results highlight the variability of SHAP values across individuals. Given the significant variability

Table 1 Characteristics of the study population

Comparisons of all variables by Wait-time among Emergency Department Visits Triaged as ESI-3 Level			
	30min or less Wait Time	**Wait time more than 30min**	***P* value**
Number of patient visits	92,033 (51.80)	85,632 (48.20%)	
Wait time --- min			
Median (IQR)	5 [3–14]	95 [55–161]	<0.001
Mean (SD)	9 (8)	124 (96)	<0.001
Age --- year			
Median (IQR)	45 [31–57]	42 [30–55]	<0.001
Mean (SD)	45 (17)	43 (15)	<0.001
Gender --- n (%)			<0.001
Male	48,259 (56.27)	37,501 (43.73)	
Female	43,774 (47.63)	48,131 (52.37)	
Marital status --- n (%)			<0.001
Single	50,510 (53.41)	44,060 (46.59)	
Married	21,915 (47.02)	24,688 (52.98)	
Others	19,608 (53.73)	16,884 (46.27)	
Race and ethnicity --- n (%)			<0.001
NHW	30,437 (56.46)	23,471 (43.54)	
NHB	31,280 (53.57)	27,115 (46.43)	
Hispanic/Latino	26,258 (46.39)	30,347 (53.61)	
NHA	1,263 (45.42)	1,518 (54.58)	
Others	2,795 (46.77)	3,181 (53.23)	
Language --- n (%)			<0.001
English	79,052 (53.32)	69,221 (46.68)	
Spanish	10,325 (44.28)	12,995 (55.72)	
Others	2,626 (43.74)	3,416 (56.26)	
Insurance --- n (%)			<0.001
No	32,548 (47.77)	35,585 (52.23)	
Yes	59,485 (54.31)	50,047 (45.69)	
Primary care physician --- n (%)			<0.001
Assigned	35,072 (49.12)	36,332 (50.88)	
Not assigned	56,961 (53.60)	49,300 (46.40)	
Comorbid --- n (%)			<0.001
No	40,114 (51.05)	38,460 (48.95)	
One	14,866 (50.15)	14,778 (49.85)	
Multimorbidity	37,053 (53.35)	32,394 (46.65)	
Crowding status --- n (%)			<0.001
Not-crowded	33,154 (79.33)	8,639 (20.67)	
Crowded	36,190 (53.12)	31,945 (46.88)	
Overly-crowded	22,689 (33.50)	45,048 (66.50)	
Mode of arrival --- n (%)			<0.001
Private car	42,047 (39.48)	64,444 (60.52)	
Ambulance	41,675 (83.02)	8,526 (16.98)	
Public transportation	492 (36.26)	865 (63.74)	
Ambulatory	7,819 (39.86)	11,797 (60.14)	
Clinical hours --- n (%)			<0.001
Within clinical hour	45,044 (51.36)	42,660 (48.64)	
Out of clinical hour	46,989 (52.23)	42,972 (47.77)	
Weekday vs. weekend --- n (%)			<0.001
Weekday	63,922 (48.71)	67,308 (51.29)	
Weekend	28,111 (60.54)	18,324 (39.46)	
Having High BP at Triage --- n (%)			<0.001
Yes	51,686 (50.90)	49,853 (49.10)	
No	40,347 (53.00)	35,779 (47.00)	

Table 1 (continued)

Comparisons of all variables by Wait-time among Emergency Department Visits Triaged as ESI-3 Level			
	30min or less Wait Time	**Wait time more than 30min**	***P* value**
Having abnormal vital signs at triage (exclude high BP) --- n (%)			<0.001
Yes	22,825 (57.61)	16,796 (42.39)	
No	69,208 (50.13)	68,836 (49.87)	

Based on 177,665 patient ED visits of two groups (wait time≤30min and wait time >30min) from January 1, 2019, to December 31, 2021. Categorical variables were compared using Chi-square test. Continuous variables were compared either using student-t test (mean) or using Kruskal-Wallis' test (median). $p<0.001$ among all variables when two groups were compared

Abbreviations: ESI Emergency Severity Index, *NHW* Non-Hispanic White, *NHB* Non-Hispanic Black, *NHA* Non-Hispanic Asian, *IQR* Interquartile range, *SD* Standard deviation, *BP* Blood pressure

Table 2 Performance accuracy comparison of using five different machine learning algorithms with all features to predict patient wait times

	CVLR		**Random Forest**		**XGBoost**		**ANN**		**SVM**	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**
Accuracy	0.75	0.74	0.76	0.74	0.75	0.75	0.75	0.74	0.75	0.75
Recall	0.80	0.80	0.80	0.79	0.79	0.79	0.80	0.79	0.80	0.80
Precision	0.71	0.71	0.73	0.71	0.71	0.71	0.72	0.71	0.71	0.71
F1 score	0.75	0.75	0.76	0.75	0.75	0.75	0.75	0.75	0.75	0.75
AUROC	0.81	0.81	0.81	0.81	0.81	0.81	0.82	0.81	0.79	0.78

The performance accuracy was reported with the use of training and testing data

Abbreviations: *XGBoost* eXtreme Gradient Boosting, *CVLR* Cross Validation Logistic Regression, *ANN* Artificial Neural Network, *SVM* Support Vector Machine, *AUROC* Areas Under the Receiver Operating Characteristics

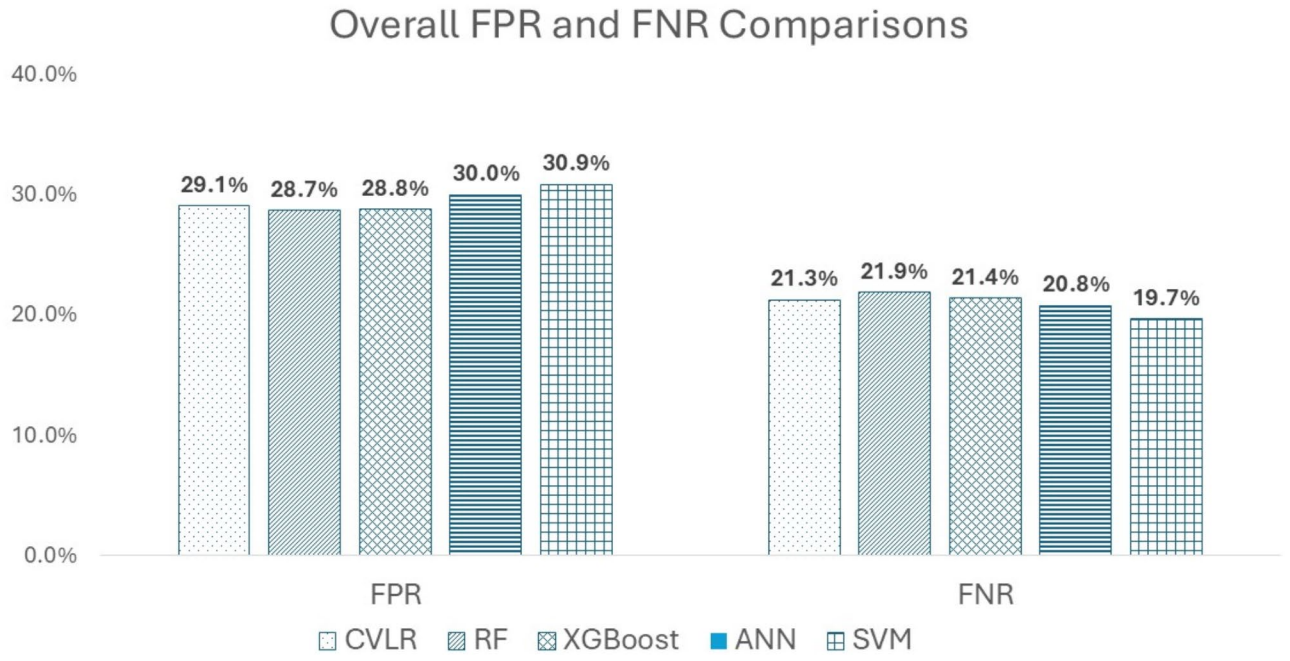

Fig. 1 Comparison of false negative rates and false positive rates when different machine learning algorithms utilized for wait-time prediction. Figure 1 depicts various false negative rates (FNRs) and false positive rates (FPRs) for predicting patient wait times using different ML algorithms. Our focus was primarily on the FNR in our attempts to predict patients' wait times. A false negative occurs when ML algorithms misclassify a patient who waits longer than 30 minutes as waiting less than 30 minutes. In Figure 1, the highest FNR was observed when using RF algorithm to predict patient wait times, while the lowest FNR was observed when SVM algorithm was utilized. Abbreviations: FNR, False Negative Rate; CVLR, Cross Validation Logistic Regression; RF, Random Forest; XGBoost, eXtreme Gradient Boosting; ANN, Artificial Neural Network; SVM, Select Vector Machine

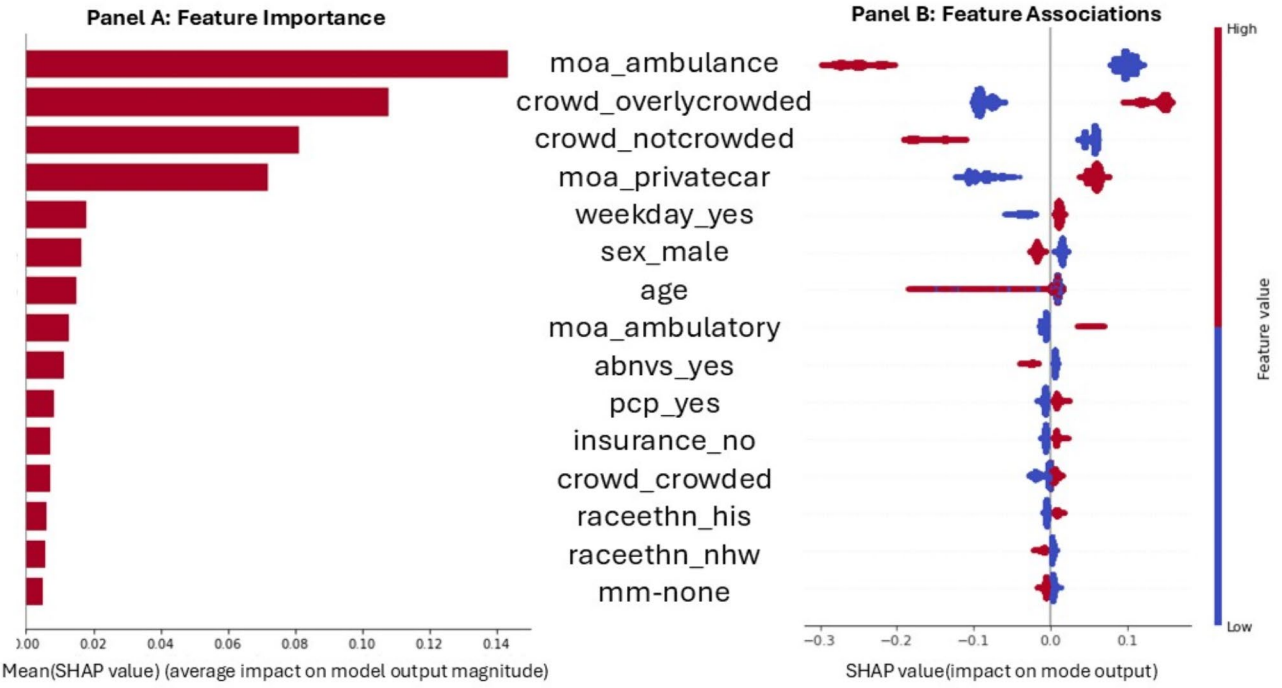

Fig. 2 Feature importance and associations from XGBoost classification model. Figure 2 illustrates the essential features contributing to wait time prediction using XGBoost algorithmic model. Panel **A** (Feature Importance): Feature importance of each feature contributing to the model prediction. The x-axis represents the marginal contribution of a feature to the change in the predicted probability of prolonged wait time (30min). Panel **B** (Feature Associations): The x-axis indicates the direction of each feature impact on the model output. SHAP values >0 indicates the prolonged wait time and <0 indicates patients wait time<30min. All features except age were dichotomous coded either 0 (no) or 1 (yes). For example, moa_ambulance (i.e., patients arrived by ambulance) had more negative values indicating the higher impact of predicting patient wait time<30min if patients arrived by ambulance

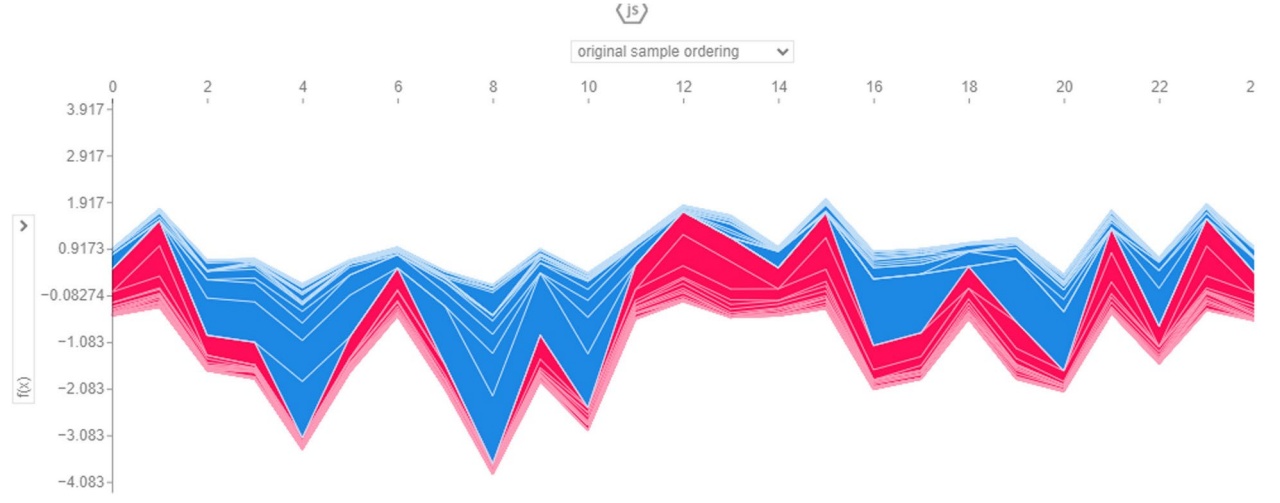

Fig. 3 Different SHAP values of the first 25 samples. Figure 3 shows the different SHAP values of the first 25 samples from this study. X-axis shows the number of samples and Y-axis shows the SHAP values of different features within the samples. It shows that certain features (i.e., red colored) can negatively contribute to the prolonged wait time predictions, whereas others (i.e., blue colored) may positively contribute to the wait time predictions. This figure shows the variability of each sample predicting prolonged ED wait times

observed, individualized wait time predictions are essential for improving ED operational management efficiency. Less crowded or arrival by ambulance, extremes of age, abnormal vital signs lead to shorter wait times.

PDPs of the key features are shown in Fig. 4, including two categorical features (mode of arrival by ambulance, and ED overly crowded) and one continuous feature (age). The PDPs indicate that patients who arrived by ambulance or during less crowded conditions experienced

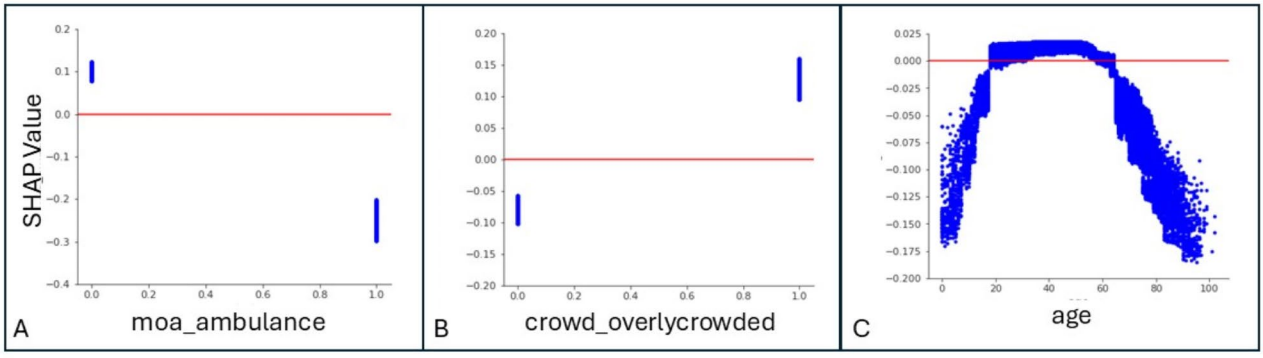

Fig. 4 Partial dependency plots of leading predictors from XGBoost. Figure 4 depicts Partial Dependency Plots (PDP) generated using the XGBoost algorithmic model to predict patient wait time. The categorical features include mode of arrival (moa) by ambulance (Panel **A**) and ED crowding status (overly crowded, Panel **B**), while age is presented as a continuous feature (Panel **C**). The categorical features demonstrate bidirectional effects on patient wait time prediction. Generally, patients who arrived at the ED under not overly crowded conditions, as well as those who arrived by ambulance, experienced shorter wait times, whereas patients arriving at an overly crowded ED, or arrived not by ambulance, experienced longer wait times. The PDP for age reveals a complex pattern; patients at extreme ages (i.e., very young or very old) tended to experience shorter wait times compared to others

Table 3 Top five feature interactions using XGBoost algorithm to predict patient wait times

	Gain rank	FScore rank	weighted FScore rank
Two feature interactions			
overlycrowded vs. moa_ambulance	1	16	2
not_crowded vs. moa_ambulance	2	58	4
abnvs_yes vs. moa_ambulance	3	51	6
overlycrowded vs. moa_private car	4	45	3
moa_ambulance vs. weekly_yes	5	52	12
Three feature interactions			
overlycrowded vs. moa_ambulance vs. abnvs_yes	1	6	1
overlycrowded vs. moa_ambulance vs. sex_male	2	10	3
moa_ambulance vs. abnvs_yes vs. weekday_yes	3	7	2
overlycrowded vs. moa_private car vs. age	4	11	4
overlycrowded vs. moa_private car vs. not_crowded	5	22	5

Based on 177,665 ED patients with the use of XGBoost model prediction

Abbreviations: moa mode of arrival, *abnvs* abnormal vital sign

shorter wait times compared to their counterparts. The PDP for age reveals patients at extreme ages (i.e., very young or very old) tend to experience shorter wait times compared to those with other age groups (Fig. 4).

Top feature interactions and their effect on wait time prediction are listed in Table 3. Gain, FScore, weighted FScore and their ranks are reported. ED overly crowded status and patient arrived by ambulance were the highest observed interactions in the model (Table 3). PDP interactions generated using the XGBoost to predict patient prolonged wait time are shown in Fig. 5.

Discussion

This study provides a comprehensive analysis of ML models for predicting prolonged patient wait times in the ED. Additionally, we offer both global and local interpretations of key features and feature interactions related to prolonged wait times. It is noteworthy that the chi-square tests indicated statistically significant relationships of all features with prolonged wait time. However, the differences between chi-square test results and ML predictions are expected due to differences in methodology, dependent features, and model structure. Thus, our paper reinforces ML approaches to assess prediction that goes beyond simple bivariate associations.

Our findings suggest that the performance accuracy of the ML models in predicting prolonged wait times is acceptable. However, for ML models to be effectively integrated into clinical practice, it is crucial to analyze and interpret their predictions in the context of clinical and operational significance. Our study emphasizes minimizing FNR to enhance patient safety and clinical outcomes. Only through such a thorough assessment can ML models be pragmatically applied in real-time emergency care settings.

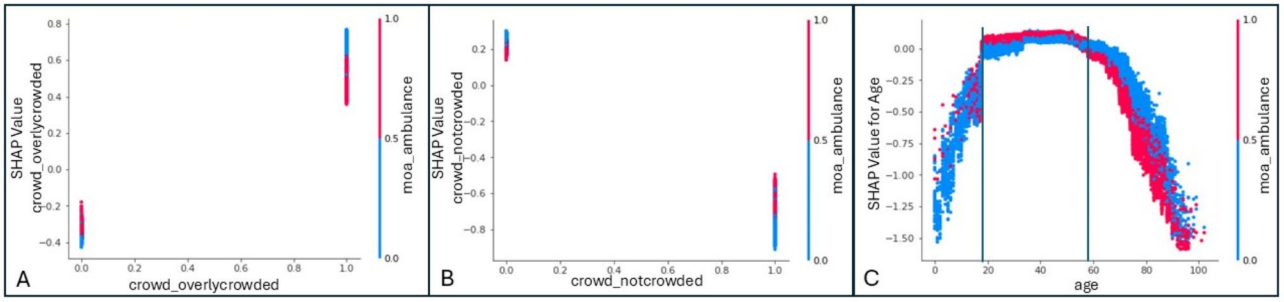

Fig. 5 Partial dependency plots of top feature interactions using XGBoost algorithm for prolonged wait time predictions. Figure 5 shows the PDP of top feature interactions. **A** shows the interactions between ED overly crowded conditions and ambulance transportation. When ED was overly crowded, more patients with ambulance transfer had less prolonged wait times than patients who arrived at ED without ambulance transfer. **B** shows when ED was not crowded, more patients who arrived without ambulance transfer had less prolonged wait times than patients who arrived via ambulance. **C** shows the interactions between age and ambulance transfer. Two perpendicular lines were drawn and indicated that patients' age ranges (20 to 60). Generally, less prolonged wait time occurred when patients aged either younger than 20 or older than 60. ED wait times were quite similar among patients aged ranging from 20- to 60-year-old. When patients older than 60, more patients had less prolonged ED wait times when arrived at ED via ambulance than ones without

In recent years, numerous ML models have emerged to predict operational metrics in the ED, including patient ED arrival, ESI, triage, disposition, and wait time predictions [11, 18, 22–25]. The performance accuracy of these model predictions range from 65% to over 90% [22–24]. Our model's performance falls within this range, affirming the feasibility and consistency of employing ML in predicting various ED operational metrics. Moreover, our study goes beyond conventional performance assessment by integrating interpretable evaluations to their clinical significance. For example, the selection of a specific metric is contingent upon the predicted outcome and its clinical significance. In our study, given that prolonged wait times may impact subsequent clinical outcomes, it is crucial to prioritize model accuracy by focusing on increasing sensitivity and reducing FNR. In this context, maximizing recall and minimizing FNR are more critical than maximizing specificity and minimizing FPR. Therefore, selecting the appropriate confusion matrix metrics for evaluating quality is essential to ensure the clinical effectiveness and relevance of the ML model in predicting patient wait times.

While all five models demonstrated similar accuracy performance, we selected the XGBoost algorithm for further comprehensive model assessment due to its relatively high performance and efficiency [26, 27]. In terms of global and local feature interpretations, SHAP has been widely used to analyze the associations between features and the output of predictive modeling in healthcare research, this includes the feature importance (identify individual feature contributing to the outcome prediction) and feature association (PDP, visualize the interaction effects of each feature associated with the model prediction) [28, 29]. Furthermore, feature interactions could capture complex relationships between different features, recognize hidden patterns that may not be apparent, and identify the combinations of features most predictive of the final outcomes, all leading to more accurate, interpretable, and generalizable predictions [30].

This study exhibits several notable strengths. We integrated electronic health records with ML for patient wait time predictions, focusing on their clinical and operational implications, rather than solely reporting model performance accuracy. Additionally, our study expanded upon traditional performance assessment by including a comprehensive evaluation of model predictions. These additional analyses allowed for global and local interpretation of key features, explored multi-feature interactions, and determined their associations with model prediction, thereby enhancing the overall robustness and applicability of our study findings.

However, this study is also subject to certain limitations. Firstly, being a retrospective study, the presence of missing/inaccurate data introduces potential bias. As a single-centered study, the generalizability of our findings is limited to EDs with similar patient populations. Secondly, the exploration was confined to five ML algorithms, and there may be other algorithms that could potentially yield more accurate wait time predictions. Thirdly, the inclusion of only a limited number of features for model predictions may be insufficient, and the inclusion of other features may enhance prediction accuracy. Fourth, we did not conduct a fairness evaluation of the model predictions, which is an important aspect of ML model quality assessment. Therefore, future studies should focus on a more comprehensive exploration of ML models, incorporate a broader range of features for model predictions, and provide robust fairness evaluations as part of model quality assessment.

Conclusion

ML models exhibit satisfactory performance in categorizing patient prolonged wait times at the ED. When assessing model performance accuracy of prolonged wait time predictions, maximizing recall and minimizing FNR are associated with better clinic significance. Top key features influencing prolonged patient wait times were found to be ED crowding status and patient mode of arrival. Implementing strategies to alleviate ED crowding and mitigate the impact of patient arrival modes could potentially improve ED operational management.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12913-025-12535-w.

Supplementary Material 1.

Acknowledgements
Not applicable.

Authors' contributions
Conceptualization: HW, NS, and US; Data curation: HW and DS; Formal analysis: HW, NS, and US; Investigation: HW and DS; Methodology: HW, NS, and US; Project administration: HW and DS; Supervisions: US; Validation: HW, and US; Visualization: HW and US; Writing -original draft: HW; Writing -review & editing: HW, NS, DS, and US.

Funding
The project described was supported by the National Institute on Minority Health and Health Disparities through the Texas Center for Health Disparities (NIMHD) 5S21MD012472-05 (Usha Sambamoorthi), and the National Institute of Health/Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity Grant # 1OT2OD032581-01 (Usha Sambamoorthi). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data availability
The study data includes patient information; therefore, data is available upon requesting to the corresponding author.

Declarations

Ethics approval and consent to participate
This study was approved by the University of North Texas Health Science Center Regional Institutional Review Board with a waiver of informed consent (IRB#1967558-1). The study was conducted in full compliance with the ethical principles outlined in the Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 18 November 2024 / Accepted: 6 March 2025
Published online: 18 March 2025

Reference

1. Ortiz-Barrios MA, Alfaro-Saiz JJ. Methodological Approaches to Support Process Improvement in Emergency Departments: A Systematic Review. Int J Environ Res Public Health. 2020;17(8):2664.
2. Horwitz LI, Green J, Bradley EH. US emergency department performance on wait time and length of visit. Ann Emerg Med. 2010;55(2):133–41.
3. Sonnenfeld N, Pitts SR, Schappert SM, Decker SL. Emergency department volume and racial and ethnic differences in waiting times in the United States. Med Care. 2012;50(4):335–41.
4. Savioli G, Ceresa IF, Gri N, Bavestrello Piccini G, Longhitano Y, Zanza C, Piccioni A, Esposito C, Ricevuti G, Bressan MA. Emergency Department Overcrowding: understanding the factors to find corresponding solutions. J Pers Med. 2022;12(2):279.
5. Plunkett PK, Byrne DG, Breslin T, Bennett K, Silke B. Increasing wait times predict increasing mortality for emergency medical admissions. Eur J Emerg Med. 2011;18(4):192–6.
6. Nyce A, Gandhi S, Freeze B, Bosire J, Ricca T, Kupersmith E, Mazzarelli A, Rachoin JS. Association of Emergency Department Waiting Times With Patient Experience in Admitted and Discharged Patients. J Patient Exp. 2021;8:23743735211011404.
7. Baker DW, Stevens CD, Brook RH. Patients who leave a public hospital emergency department without being seen by a physician. Causes and consequences JAMA. 1991;266(8):1085–90.
8. Benchmark survey by AAAEM (Academy of Administrators in Academic Emergency Medicine). https://www.saem.org/about-saem/academies-interest-group-affiliate/aaaem/benchmark-survey2. Accessed 1 Dec 2024.
9. Hemaya SA, Locker TE. How accurate are predicted waiting times, determined upon a patient's arrival in the Emergency Department? Emerg Med J. 2012;29(4):316–8.
10. Sun Y, Teow KL, Heng BH, Ooi CK, Tay SY. Real-time prediction of waiting time in the emergency department, using quantile regression. Ann Emerg Med. 2012;60(3):299–308.
11. Pak A, Gannon B, Staib A. Predicting waiting time to treatment for emergency department patients. Int J Med Inform. 2021;145:104303.
12. Ang E, Kwasnick S, Bayati M, Plambeck EL. Accurate Emergency Department wait time prediction. Manuf Serv Oper Manag. 2015;18(1):141–56.
13. Cheng N, Kuo A. Using Long Short-Term Memory (LSTM) Neural Networks to Predict Emergency Department Wait Time. Stud Health Technol Inform. 2020;270:1425–6.
14. Hijry H, Olawoyin R. Predicting patient waiting time in the queque system using deep learning algorithms in the Emergency Room. Int J Ind Eng Oper Manag. 2021;3(1):33–45.
15. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, Aardoom JJ, Debray TPA, Schuit E, van Smeden M, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med. 2022;5(1):2.
16. Rashidi HH, Albahra S, Robertson S, Tran NK, Hu B. Common statistical concepts in the supervised Machine Learning arena. Front Oncol. 2023;13:1130229.
17. Shaikh NF, Shen C, LeMasters T, Dwibedi N, Ladani A, Sambamoorthi U. Prescription Non-Steroidal Anti-Inflammatory Drugs (NSAIDs) and Incidence of Depression Among Older Cancer Survivors With Osteoarthritis: A Machine Learning Analysis. Cancer Inform. 2023;22:11769351231165160.
18. Porto BM. Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review. BMC Emerg Med. 2024;24(1):219.
19. Emergency Physicians Monthly: 11 Benchmarks that should matter to EPs. https://epmonthly.com/article/11-benchmarks-that-should-matters-to-eps. Accessed 1 Dec 2024.
20. Hospital Emergency Department: crowding continues to occur, and some patients wait longer than recommended time frames. https://www.gao.gov/assets/gao-09-347.pdf. Accessed 1 Dec 2024.
21. Wang H, Robinson RD, Garrett JS, Bunch K, Huggins CA, Watson K, Daniels J, Banks B, D'Etienne JP, Zenarosa NR. Use of the SONET Score to Evaluate High Volume Emergency Department Overcrowding: A Prospective Derivation and Validation Study. Emerg Med Int. 2015;2015:401757.
22. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA Jr, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care. 2019;23(1):64.
23. Ivanov O, Wolf L, Brecher D, Lewis E, Masek K, Montgomery K, Andrieiev Y, McLaughlin M, Liu S, Dunne R, et al. Improving ED Emergency Severity Index Acuity Assignment Using Machine Learning and Clinical Natural Language Processing. J Emerg Nurs. 2021;47(2):265-278 e267.
24. Chen CH, Hsieh JG, Cheng SL, Lin YL, Lin PH, Jeng JH. Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data. Int J Med Inform. 2020;139:104146.

25. Porto BM, Fogliatto FS. Enhanced forecasting of emergency department patient arrivals using feature engineering approach and machine learning. BMC Med Inform Decis Mak. 2024;24(1):377.
26. Yuan KC, Tsai LW, Lee KH, Cheng YW, Hsu SC, Lo YS, Chen RJ. The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. Int J Med Inform. 2020;141:104176.
27. XGBoost: a comprehensive guide, model overview, analysis, and code demo using paperspace GPUs. https://blog.paperspace.com/xgboost-a-comprehensive-guide-to-model-overview-analysis-and-code-demo-using/. Accessed 1 Dec 2024.
28. Choi JH, Choi Y, Lee KS, Ahn KH, Jang WY. Explainable model using Shapley additive explanations approach on wound infection after wide soft tissue Sarcoma resection: "Big Data" analysis based on health insurance review and assessment service hub. Medicina (Kaunas). 2024;60(2):327.
29. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak. 2019;19(1):146.
30. Oh S. Feature interaction in terms of prediction performance. Appl Sci. 2019;9(23):5191.

Publisher's Note