Analysis and Prediction of Emergency Department Length of Stay Categories Based on Triage Acuity

Salwa Nur JB, Muhammad Syahputra Novelan, Solly Aryza

Abstract

This research used a machine learning model to classify the length of stay in the *Emergency Department Length of Stay* (ED LOS) RSUD Langsa. Source workbook has 16,739 non-empty visit records from May 21 to November 6, 2026. Visits with incomplete timestamps, negative duration, ED length of stay greater than 24 hours, or invalid triage category were excluded, leaving 15,415 visits from 12,596 patient group visits. Dummy Classifier, Logistic Regression, Random Forest and XGBoost are tested on 5 random groups for cross validation. ED LOS is the number of exercise data and categorized Short, Medium and Long. XGBoost achieves 43.95% out-of-fold macro-F1, 44.70% holdout random grouped and 39.04% holdout temporal. Currently SHAP logs the month, hour and day of arrival of each.Triase Hijau menunjukkan bias seleksi dengan nilai timestamp tidak lengkap tertinggi (67.91%). The results do not support use as an automatic clinical decision tool because it functions only.

Keywords: Emergency Department Length of Stay, Triage Acuity, Arrival-Time Features, Machine Learning, XGBoost, Patient-Grouped Validation, SHAP.

Salwa Nur JB[1]
[1]Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: salwanur2015@gmail.com [1]

Muhammad Syahputra Novelan[2], Solly Aryza[3]
[2,3]Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: putranovelan@dosen.pancabudi.ac.id[2], sollyaryzalubis@pancabudi.ac.id [3]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

Patient congestion (ED overcrowding) in the Emergency Department (IGD) burdens the operational efficiency of hospitals and emergency medical handling. When the surge in visits exceeds the facility's capacity, the Emergency Department Length of Stay (ED LOS) increases, the risk of mortality rises, patient safety decreases, and medical staff become fatigued more quickly [1], [2], [3]. Delays in the emergency department (ED) flow generally stem from fluctuations in patient arrival times, the complexity of diagnostic procedures, delays in specialist consultations, and the difficulty in obtaining inpatient beds (boarding time) [2], [4].

Hospital triage codes the severity level of patients upon arrival to determine the priority of treatment. In Indonesia, patients are categorized as Red (resuscitation/emergent), Yellow (urgent), Green (non-urgent), and Black (deceased). Operational practitioners often assume that the higher the triage urgency, the longer the patient stays in the ER. Data from regional healthcare facilities show a more complex relationship. Red triage patients often require resuscitation, stabilization of vital signs, and extensive laboratory and radiology examinations, resulting in a very long service duration [5]. On the other hand, Green triage patients generally finish quickly, while the duration for Yellow triage patients varies greatly due to the queue for supporting examinations and the availability of beds [6]. The distribution of triage color durations overlaps, so triage alone is not sufficient to predict a patient's service duration.

Previous research on predicting emergency department service duration using machine learning mostly utilized large-scale electronic health records (EHR) data from tertiary hospitals in developed countries, complete with laboratory variables and detailed diagnoses [1], [6], [7]. Implementing it in regional hospitals in developing countries like Indonesia is often hindered by transferability issues, due to limited real-time clinical attributes, different local operational patterns, and varying information technology infrastructure [5], [8]. It is still unclear to what extent local administrative operational data can predict the duration of emergency department services, and how the arrival time feature interacts with the triage level in affecting prediction accuracy.

To address this issue, this study aims to develop and evaluate a machine learning-based predictive model to classify the duration of emergency department services for patients at RSUD Langsa, Aceh, Indonesia. First, ED LOS is operationally defined from local SIMRS administrative data and includes all discharge patterns (discharge, hospitalization, referral, and death) up to a 24-hour limit. Second, the target is created with tercile quantiles calculated only from the training set to prevent data leakage. Third, validation uses patient-grouped data splitting, so repeated visits from the same patient do not leak information between sets. Fourth, the model's resilience to data distribution shifts (concept drift) is tested by comparing random grouped holdout evaluation and temporal holdout evaluation. The contribution of predictors is explained through feature ablation studies and SHapley Additive exPlanations (SHAP), and the potential selection bias due to missing data is also analyzed.

Literature Review

Triage is a clinical risk management instrument specifically designed to prioritize medical treatment based on the urgency of the patient's condition, rather than on a first-come, first-served basis [8], [9]. International standard triage systems such as the Emergency Severity Index (ESI) and the Australasian Triage Scale (ATS) have been proven to reduce delays in handling critical cases [1]. However, a systematic study by Farimani et al. [8] stating that the triage scale is designed to control the time to the initial doctor examination (time-to-physician), not to indicate the total duration of service or the time a patient leaves the ER. Service flow barriers (access block) and delays in supporting examinations often extend the stay duration of moderate to severe triage patients [2], [8].

Machine learning is increasingly used to predict length of stay and disposition decision in the emergency department. Kishore et al. [1] utilizing EMR data from 599,015 visits in Australia to predict decisions with an AUROC value of 0.93 within the first 30 minutes of patient arrival. Canellas et al. [2] distinguishing the duration of the emergency department into three operational

intervals (arrival-to-room, room-to-disposition, and admit disposition-to-departure) and finding that waiting room congestion and bed capacity are dominant predictors in the delay phase.

Sezik et al. [3] comparing seven machine learning algorithms on 75,803 patients and reporting that Random Forest and XGBoost produced the best AUC value (0.958) in predicting disposition Etu et al. [4] applying Gradient Boosting on 3,301 COVID-19 patients to predict an ED LOS threshold of 4 hours with 85% accuracy. Sulaiman et al. [5] applying rule extraction techniques (TE2Rules) and data balancing (SMOTETomek) on the MIMIC-IV ED dataset (>400,000 medical records) to identify long-duration predictive rules (>4.5 hours). Next, Moreno-Sánchez et al. [6] and Akhlaghi et al. [7] proving that the use of explainable AI (XAI) methods can provide visual transparency for clinicians in understanding the weight of operational predictor variables. Meanwhile, the Porto study [9] emphasizing the importance of integrating natural language processing (NLP) in triage notes to improve the precision of patient flow classification.

Research on predicting service duration and patient flow in the Emergency Department has applied various machine learning approaches with diverse data characteristics and prediction goals. Kishore et al. [1] developing a model to predict the likelihood of patient hospitalization based on initial arrival data. Canellas et al. [2] analyzing ED Length of Stay (ED LOS) through several stages of the service process and demonstrating the importance of operational variables in understanding the duration of patients in the ED. Meanwhile, Sezik et al. [3] focusing on predicting the final decision on patient management (disposition), while Etu et al. [4] applying a machine learning approach to classify the duration of COVID-19 patient care. Sulaiman et al. [5] developing ED LOS prediction modeling equipped with rule extraction to enhance the transparency of classification results. Research by Moreno-Sánchez et al. [6] and Akhlaghi et al. [7] applying explainable AI (XAI) so that clinicians can see the weights of operational predictor variables. Porto [9] examining the use of natural language processing (NLP) in triage notes to improve the precision of patient flow classification. Systematic review by Farimani et al. [8] finding that the approach and quality of methodology in research on predicting length of stay in the emergency department are very diverse.

Those studies have not yet tested whether simple administrative data, namely triage urgency levels and arrival times, are sufficient to classify ED LOS categories in regional hospitals. It is also unclear how significant the relative contributions of triage features and arrival times are to classification performance, and whether the results are consistent across evaluations based on different patient groups and time periods.

This study evaluates Logistic Regression, Random Forest, and XGBoost on the operational data of the Langsa Regional General Hospital's Emergency Department. Feature ablation compares the contribution of each group of features, SHAP analysis shows the contribution of predictors to the model output, and temporal testing assesses the model's generalization ability over different time periods.

Research Methodology

This research uses an experimental quantitative approach oriented toward predictive analytics. Data was collected from the Emergency Department (IGD) of RSUD Langsa, Aceh, Indonesia, for the period from May 21, 2026, to November 6, 2026.

The outcome variable is the Emergency Department Length of Stay (ED LOS), which is the difference in time between the patient's arrival and the time the patient leaves the ED unit:

$$\boldsymbol{ED\ LOS} = Time\ patient\ leaves\ ED - Time\ of\ arrival \quad (1)$$

The "Service Completion Time" attribute in the RSUD Langsa information system records the timestamp when the patient leaves the physical area of the ED. This definition includes all valid discharges: patients discharged home, patients admitted to the hospital, and patients transferred to other healthcare facilities.

The initial dataset obtained from transactions includes 16,752 patient visit records. The data cleaning stages were carried out meticulously through the following steps: The first stage is anonymization and the removal of non-predictive features. Patient names are removed and not used in modeling to protect privacy. The patient code was changed to *patient_group_id*, which is

only used to group data during data splitting and bootstrapping, not as a predictor, so the model does not memorize specific patient patterns. The *record_id* code is only used for auditing. The "Unit" attribute was removed because it did not provide any variation in the analyzed data.

The second stage is the verification of timestamp completeness. As many as 131 records do not have Arrival Time and 1,017 records do not have Service Completion Time. Six records lost both, so the unique count that lost at least one timestamp is 1,142 (6.82%). This record was excluded because ED LOS cannot be calculated without both timestamps.

The third stage is the exclusion of negative durations. From the second recording with complete timestamps, 125 recordings (0.75%) had negative durations, meaning the recorded end time was earlier than the arrival time. These recordings were removed, leaving 15,472 recordings with positive durations.

The fourth stage is the exclusion of durations longer than 24 hours. A total of 56 records (0.33%) were excluded based on the operational limits of the study, leaving 15,416 records. The 24-hour limit is the selection criterion. Longer recordings are not automatically considered misrecorded, as they may reflect certain service conditions. Therefore, the results of this study only apply to visits with a positive ED LOS and no more than 24 hours.

The fifth stage is the exclusion of invalid triage categories and the formation of the final cohort. One record was excluded because its triage category was invalid or unavailable. After this stage, the final cohort consisted of 15,415 visits, which is 92.09% of the raw data. Table 1 summarizes the selection process.

Table 1. Cohort Dataset Filtering and Cleaning Process

Data Filtering Stages	Number of Data Records	Percentage (%)
Non-empty raw data	16.739	100,00%
At least one timestamp unavailable	1.142	6,82%
Negative duration	125	0,75%
Countable positive duration	15.472	92,43%
Duration over 24 hours	56	0,33%
Positive duration ≤24 hours	15.416	92,09%
Invalid or unavailable triage for duration ≤24 hours	1	0,01%
Final research cohort	15.415	92,09%

In the raw data with triage labels (16,737 records), there are 1,020 Red, 14,550 Yellow, 1,125 Green, and 42 Black records. Two other records do not have triage labels. After selection, the final cohort consisted of Yellow 14,120 visits (91.60%), Red 996 (6.46%), Green 268 (1.74%), and Black 31 (0.20%). This composition is unbalanced, as the number of Green and Black is much smaller than that of Yellow.

The completeness of timestamps in the source data also varies sharply between triage categories. Records that lost at least one timestamp numbered 764 out of 1,125 in the Green triage (67.91%) and 8 out of 42 in the Black triage (19.05%). In the Yellow triage, the number was 350 out of 14,550 (2.41%), and in the Red triage, 19 out of 1,020 (1.86%). Because the likelihood of a visit entering the cohort is not the same across all categories, the model's results are potentially subject to selection bias. This model primarily represents visits that have valid timestamps, durations, and triage categories, so it cannot be directly applied to all ER visits, especially Green triage.

The target variable is divided into three categories of statistical duration based on the 33rd percentile (Q_{33}) and the 67th percentile (Q_{67}) of ED LOS in the training data. The Short category includes visits with ED LOS less than or equal to Q33, which is a relatively low service duration. The Medium category includes visits with ED LOS above Q_{33} up to Q_{67}, which is a relatively medium service duration. The Long category includes visits with ED LOS above Q_{67}, which is a relatively high service duration.

To prevent data leakage, both thresholds are calculated only from the training data and then applied to the corresponding validation or test data. In cross-validation, the thresholds are recalculated from the training data in each fold. In random grouped holdout, the thresholds from the training data are Q_{33} = 38.8167 minutes and Q_{67} = 79.5167 minutes. In temporal holdout, the thresholds are Q_{33} = 39.3833 minutes and Q_{67} = 81.0167 minutes. Because both scenarios use different thresholds, the comparison of metrics between scenarios should be interpreted as a comparison of performance under each scenario's target definition. The Short, Medium, and Long categories are statistical categories relative to the training data distribution, not clinical thresholds, quality service standards, or official hospital deadlines.

This research compares four classification algorithms. Dummy Classifier with the *most_frequent* strategy serves as a baseline: this model always predicts the majority class, making its results the lower bound that other models must exceed. Logistic Regression is a linear model used to observe the basic relationship between features and the target category. Random Forest is a bagging-based ensemble that combines many decision trees. XGBoost is a gradient boosting that builds trees sequentially, with each tree correcting the errors of the previous tree.

Predictor features are selected from the information available at or around the time of arrival, namely triage category, arrival time, day of the week, and arrival month. The triage category is encoded with one-hot encoding and imputed with the mode if empty. Numeric features are imputed with the median, and StandardScaler is used on numeric features only for Logistic Regression, while Random Forest and XGBoost do not require it. Guaranty variables, pain status, and fall risk are not included in the main model because their availability at the prediction point cannot be consistently ensured. The arrival date is used only for temporal separation, and ED LOS is used only to form the target.

XGBoost was run with 350 estimators, *max_depth* 4, *learning_rate* 0.05, subsample 0.85, *colsample_bytree* 0.85, *min_child_weight* 3, and *reg_lambda* 1. The objective is multi: softprob with three classes, internal evaluation metric mlogloss, and *tree_method* hist. Model selection is based on Macro-F1 group cross-validation, and holdout data is used for evaluation after the model is selected. The testing was conducted with three schemes. In the random grouped holdout, the data is divided approximately 80:20 based on *patient_group_id*, so that all visits from the same patient are placed in only one part, either training or testing. From 12,596 unique patient groups, the training data consists of 12,341 visits from 10,076 groups, and the testing data consists of 3,074 visits from 2,520 groups. There are no identical patient groups in both parts.

Five-fold group cross-validation (StratifiedGroupKFold) was performed on the training data resulting from the above division, which consists of 12,341 visits. *Patient_group_id* becomes the group so that visits from the same patient are not separated between the training and validation parts within one fold, while the triage category is used for stratification to maintain its composition in each fold. The reported metric is the pooled out-of-fold, which is the combined validation predictions from all folds (12,341 predictions), so the number is not the same as the average score per fold.

In the temporal holdout, the cutoff point is October 7, 2026. The training data includes May 21 to October 6, 2026 (12,328 visits from 10,225 patient groups), and the test candidates include October 7 to November 6, 2026 (3,087 visits). A total of 622 visits during the testing period were removed through purging because the patients had already appeared in the training data. The final test data consists of 2,465 visits from 2,371 patient groups, with no group overlap. This scenario tests the model's performance over a more recent time period.

The main evaluation metric is Macro-F1, which is the average F1-score of each class with equal weight. The supporting metrics are accuracy, balanced accuracy (the average recall between classes), and Macro ROC-AUC one-vs-rest, which is the average AUC of one class against the others for the three target categories. Additional experiments include feature ablation to separate the influence of time and triage features, as well as global SHAP (SHapley Additive exPlanations) analysis on 500 random samples from the random holdout test data. SHAP values indicate the contribution of features to the model's output and do not show causal relationships.

Results

Exploratory analysis of 15,454 visits in the modeling cohort confirmed the dominance of Yellow triage patients (91.54%).

Table 2. Patient Distribution and Service Duration Based on Triage Color

Triage Color	Severity Level	Amount (n)	Cohort Proportion	Median ED LOS (minutes)	Duration Unavailable in Raw Data
Red	Resuscitation / Emergent	996	6,45%	82,97	23 / 1.020 (2,25%)
Yellow	Urgent	14.146	91,54%	54,98	390 / 14.559 (2,68%)
Green	Non-urgent	281	1,82%	39,60	842 / 1.129 (74,58%)
Black	Deceased	31	0,20%	15,60	11 / 42 (26,19%)
Total	**Overall**	**15.454**	**100,00%**	**56,29 (overall median)**	**1.266 / 16.750 with triage labels (7,56%)**

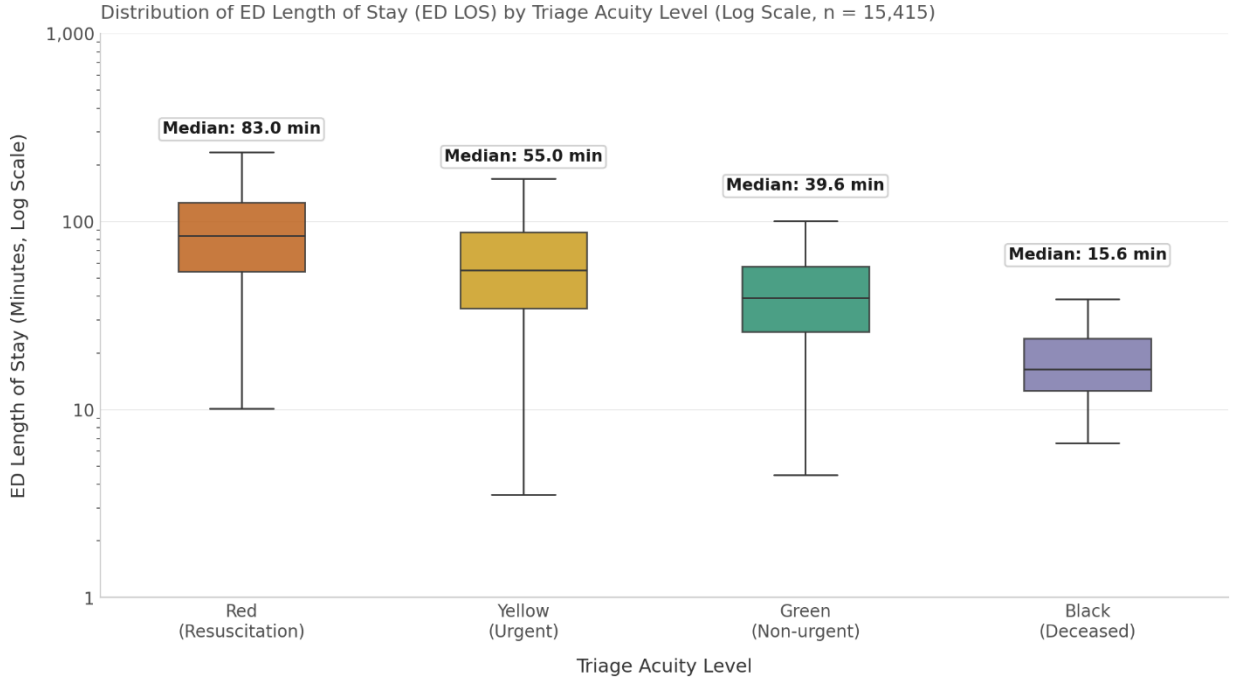


Figure 1. Distribution of ED Service Duration at RSUD Langsa Based on Triage Category.

As presented in Table 2 and Figure 1, red triage patients recorded the highest median ED LOS (82.97 minutes). This is due to the complexity of medical resuscitation, stabilization of vital conditions, and the need for further diagnostic examinations. Yellow triage patients have a median ED LOS of 54.98 minutes with a wide interquartile range variability. Green triage patients recorded a median ED LOS of 39.60 minutes, while black triage has a median of 15.60 minutes. Although the median for green triage is relatively low, the average duration reaches 170.43 minutes due to a right-skewed distribution with the presence of extreme values. This phenomenon emphasizes that the triage level cannot be used as a single linear indicator to determine service duration.

The evaluation results of the classification model performance in the 5-fold grouped cross-validation scenario (5-fold grouped CV), random grouped holdout, and temporal holdout are presented in Table 3 and illustrated in Figure 2.

Table 3. Evaluation Results of the Service Duration Category Prediction Model Performance

Model Algorithm	Accuracy	Macro-F1	Macro-AUC	Evaluation Scheme
Majority Classifier (*Dummy*)	33,33%	16,66%	0,500	Grouped CV 5-fold
Logistic Regression	40,28%	39,47%	0,581	Grouped CV 5-fold
Random Forest	42,42%	42,34%	0,606	Grouped CV 5-fold
XGBoost (CV 5-*Fold*)	44,43%	44,17%	0,632	Grouped CV 5-fold (Highest)
XGBoost (*Random Holdout*)	44,41%	44,10%	0,631	Random Grouped Holdout Test Set
XGBoost – Month Ordinal	38,92%	38,79%	0,570	Temporal Holdout
XGBoost – Cyclic Month	37,74%	37,70%	0,571	Temporal Holdout

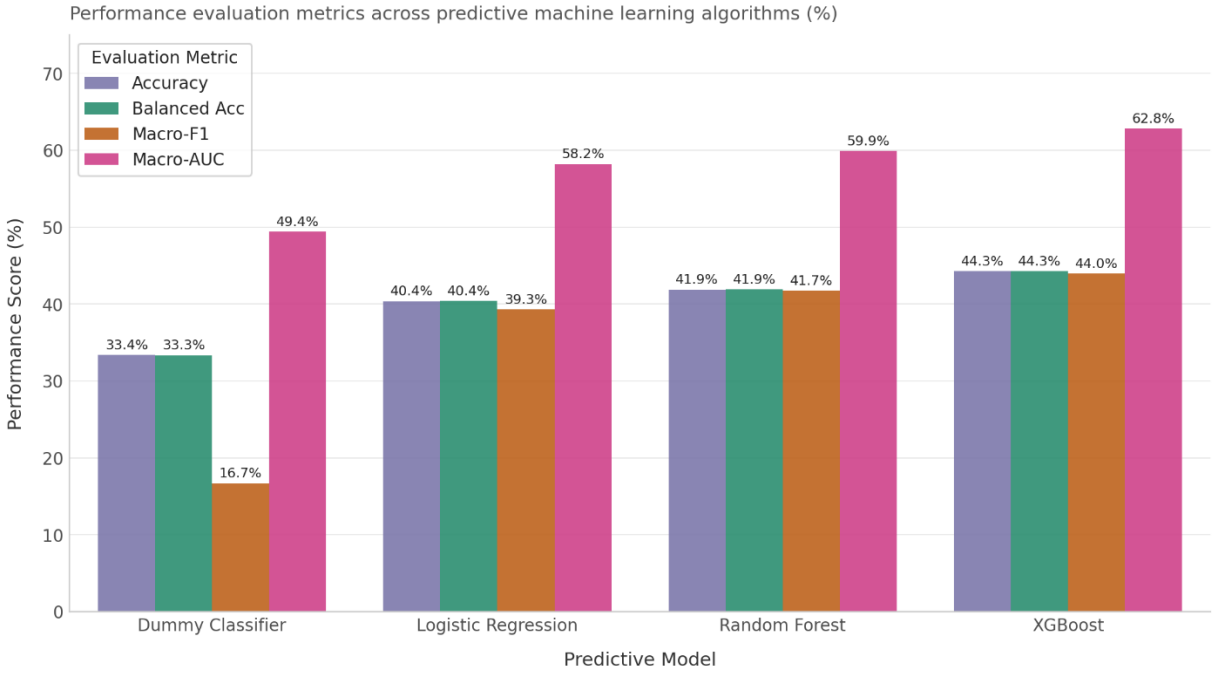


Figure 2. Comparison of ED LOS Category Prediction Algorithm Performance (CV 5-Fold).

As shown in Figure 2 and Table 3, the ensemble learning-based models (XGBoost and Random Forest) consistently outperform the linear models Logistic Regression and Majority Classifier. The XGBoost algorithm achieved the highest classification performance with a Macro F1-score of 44.17% on 5-fold CV and 44.10% on random grouped holdout. The F1-score evaluation per class on the XGBoost model resulted in scores of 45.86% for the Short class, 33.74% for the Medium class, and 52.69% for the Long class. Figure 3 displays the XGBoost confusion matrix to map the dynamics of misclassification between classes.

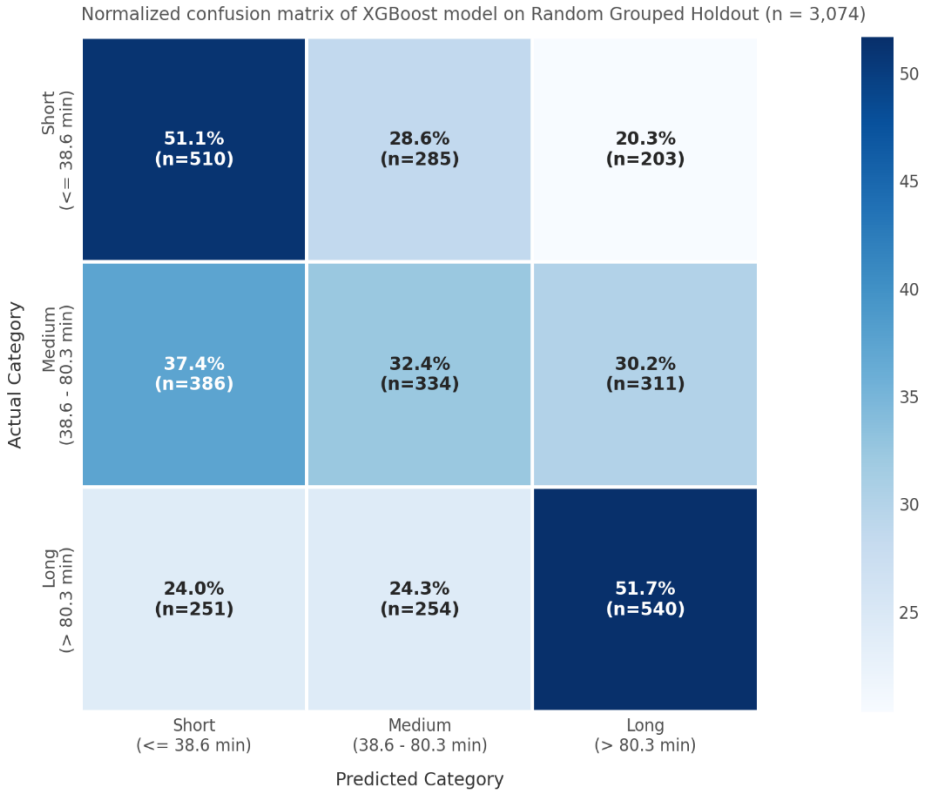


Figure 3. XGBoost Confusion Matrix (%)

Figure 3 shows that the Medium class is the hardest to predict. In the random grouped holdout, this class has a recall of 32.40% and an F1-score of 35.08%, the lowest among the three classes. Of the 1,031 actual Medium observations, only 334 were classified correctly. A total of 386 observations (about 37.44%) were predicted as Short and 311 observations (about 30.16%) were predicted as Long. Thus, the errors in the Medium class were spread to the two surrounding classes.

The cause may be related to the way the categories are formed: the Q_{33} and Q_{67} thresholds cut through the continuous duration distribution, so visits near the thresholds can switch classes with a small difference in duration. However, the confusion matrix alone has not proven this. The explanation requires an analysis of the ED LOS distance of each observation to Q_{33} and Q_{67}, which has not been conducted in this study.

This study performs global SHAP analysis on the XGBoost model and feature ablation on out-of-fold predictions to determine the relative contribution of each feature group.

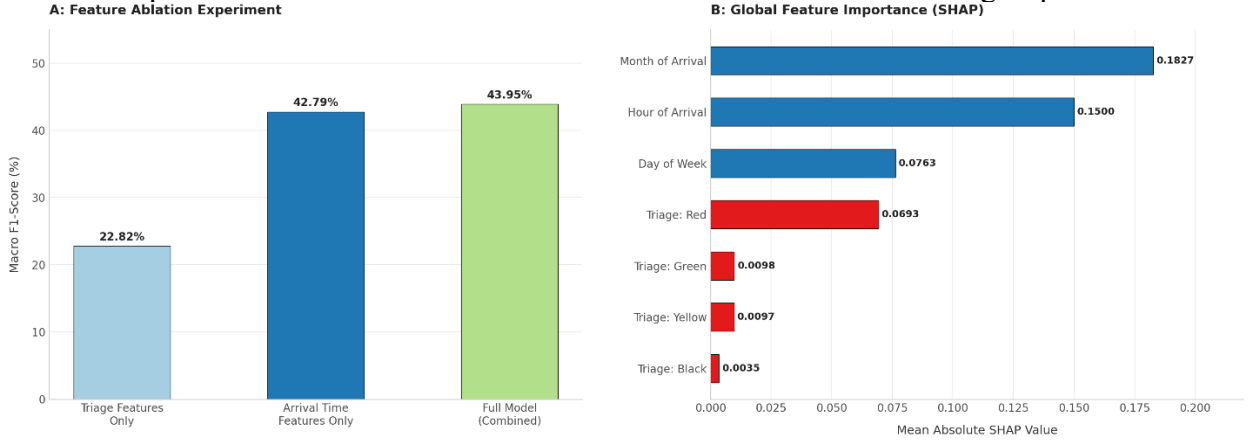


Figure 4. (A) Feature Ablation Study and (B) SHAP Global Feature Importance.

In feature ablation, the model that only uses triage features produces a Macro-F1 of 22.82%. The model that only uses arrival time features (hour, day, and month) achieves a Macro-

F1 of 42.79%. The combination of triage features and arrival time features results in 43.95%, an increase of about 1.16 percentage points compared to using only time features. In the tested configuration and data, the time features have a much greater predictive power than triage alone, while the addition of triage only provides a limited additional improvement.

Global SHAP analysis was conducted on 500 random samples from the random holdout test data, and the results are displayed in Figure 4. The *arrival_month* feature has the highest mean absolute SHAP (0.18275), followed by *arrival_hour* (0.14997) and *arrival_dayofweek* (0.07629). Among the triage features, *triage_Red* recorded a value of 0.06934, while *triage_Green*, *triage_Yellow*, and *triage_Black* had values of 0.00984, 0.00969, and 0.00349, respectively. These values indicate the extent of the feature's contribution to the model's output for the analyzed sample. The value does not indicate the direction of influence and does not prove a cause-and-effect relationship.

The interpretation of the month feature needs to be limited. The data only covers May to November 2026, and the months are represented as ordinal values, so this feature can serve as a proxy for the observation period or unmeasured operational changes. These findings do not mean that triage is not clinically important. Triage serves to prioritize emergencies, while ED LOS is also influenced by examinations, consultations, disposition decisions, and the availability of resources that have not yet been covered in the model.

In the temporal holdout, the training data includes visits before October 7, 2026, and the test data includes October 7 to November 6, 2026. After purging overlapping patient groups, the test data consisted of 2,465 visits from 2,371 patient groups. In this data, XGBoost produced an accuracy of 39.07%, balanced accuracy of 39.56%, Macro-F1 of 39.04%, and Macro-AUC of 0.569. The Macro-F1 decreased from 44.70% in the random grouped holdout to 39.04%, or about 5.66 percentage points. This decline indicates a challenge in generalizing to newer periods. The Medium class again had the lowest F1-score (34.47%), compared to Short (40.61%) and Long (42.03%). Temporal metrics for the comparison model were not reported.

This decline cannot yet be interpreted as a pure estimate of concept drift. The Q_{33} and Q_{67} thresholds were recalculated from the training data in each scenario, namely 38.8167 and 79.5167 minutes in the random holdout and 39.3833 and 81.0167 minutes in the temporal holdout, so the target class definitions in both scenarios are not identical. One temporal separation point is also not enough to conclude the occurrence of concept drift or identify the cause of performance decline. Differences in patient composition, arrival patterns, or operational dynamics may play a role, but this has not yet been tested. Rolling-origin evaluation at multiple time points is necessary to assess the stability of the model's performance.

Conclusions

This study evaluates four machine learning models to classify ED LOS categories in the emergency department of RSUD Langsa. From 16,739 non-empty records during the period from May 21 to November 6, 2026, 15,415 visits met the research criteria, which were a positive ED LOS and not more than 24 hours, as well as a valid triage category. ED LOS is defined as the time difference between the patient's arrival and the moment the patient physically leaves the ED unit, which includes patients discharged, admitted, referred, and deceased.

XGBoost produced a Macro-F1 of 43.95% in pooled out-of-fold grouped cross-validation, 44.70% in random grouped holdout, and 39.04% in temporal holdout. In pooled out-of-fold, these figures were higher than Random Forest (41.74%) and Logistic Regression (39.32%), and in random holdout, XGBoost also outperformed Random Forest (41.30%) and Logistic Regression (37.59%). Its predictive ability is still moderate and not sufficient for individual clinical use.

Feature ablation shows that the arrival time feature alone produces a Macro-F1 of 42.79%, far above the triage feature alone (22.82%), and the combination of both reaches 43.95%. SHAP places the month, hour, and day of arrival as the largest global contributors. These findings illustrate predictive patterns in the dataset and do not prove a cause-and-effect relationship. The decline in performance in the temporal test indicates a challenge in generalizing to newer periods, but a single temporal split has not proven the existence of concept drift.

The main limitation of this study is the sharp difference in timestamp completeness across triage categories. In the source data, 67.91% of Green triage records are missing at least one timestamp, which can introduce selection bias. The 24-hour limit also restricts the results to the selected cohort. Therefore, this model is more appropriately viewed as exploratory analytics and is not yet suitable for use as a clinical decision-making tool or operational prediction tool without additional validation.

References

[1] K. Kishore, G. Braitberg, N. E. Holmes, and R. Bellomo, "Early prediction of hospital admission of emergency department patients," *EMA - Emerg. Med. Australas.*, vol. 35, no. 4, pp. 572–588, 2023, doi: 10.1111/1742-6723.14169.

[2] M. M. Canellas *et al.*, "A Granular View of Emergency Department Length of Stay: Improving Predictive Power and Extracting Real-Time, Actionable Insights," *Ann. Emerg. Med.*, vol. 84, no. 4, pp. 386–398, 2024, doi: 10.1016/j.annemergmed.2024.02.004.

[3] S. Sezik, M. Ö. Cingiz, and E. İbiş, "Machine Learning-Based Model for Emergency Department Disposition at a Public Hospital," *Appl. Sci.*, vol. 15, no. 3, pp. 1–14, 2025, doi: 10.3390/app15031628.

[4] E. E. Etu *et al.*, "Prediction of Length of Stay in the Emergency Department for COVID-19 Patients: A Machine Learning Approach," *IEEE Access*, vol. 10, pp. 42229–42237, 2022, doi: 10.1109/ACCESS.2022.3168045.

[5] W. A. Sulaiman *et al.*, "Leveraging machine learning and rule extraction for enhanced transparency in emergency department length of stay prediction," *Front. Digit. Heal.*, vol. 6, no. February, pp. 1–19, 2024, doi: 10.3389/fdgth.2024.1498939.

[6] P. A. Moreno-Sánchez, M. Aalto, and M. van Gils, "Prediction of patient flow in the emergency department using explainable artificial intelligence," *Digit. Heal.*, vol. 10, 2024, doi: 10.1177/20552076241264194.

[7] L. Song *et al.*, "Developing interpretable machine learning models to predict length of stay and disposition decision for adult patients in emergency departments," *BMJ Heal. Care Informatics*, vol. 32, no. 1, pp. 1–8, 2025, doi: 10.1136/bmjhci-2024-101152.

[8] R. M. Farimani *et al.*, "Models to predict length of stay in the emergency department: a systematic literature review and appraisal," *BMC Emerg. Med.*, vol. 24, no. 1, pp. 1–39, 2024, doi: 10.1186/s12873-024-00965-4.

[9] B. M. Porto, "Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review," *BMC Emerg. Med.*, vol. 24, no. 1, 2024, doi: 10.1186/s12873-024-01135-2.