PredictED: An Explainable ESI Level Classification and Length of Stay Prediction Using Machine Learning

Dhone Matthews Calibuyot[1] and Perlita Gasmen[2]

Department of Physical Sciences and Mathematics,
College of Arts and Sciences,
University of the Philippines Manila, Ermita, Manila, 1000 Philippines
dmcalibuyot@up.edu.ph[1], pegasmen@up.edu.ph[2]

Abstract. Even before the COVID-19 pandemic, many Emergency Departments (ED) were already dealing with overcrowding issues. This study investigates how machine learning (ML) can help ED triage patients and predict the length of their stay (LOS). Using the MIMIC-IV-ED dataset, this study's findings include that Histogram-based Gradient Boosting was the top-performing model for the ESI Level Classification, achieving an AUPRC of 69.48% and an F1-Score of 63.89%. Explainable AI tools, SHAP and LIME, were used to clarify how the models arrived at their predictions. Moreover, Histogram-based Gradient Boosting also excelled in predicting the LOS, with an MAE of 3.458, an MSE of 40.439, and an RMSE of 6.359. This use of ML streamlines the decision-making process and attempts to enhance the accuracy and efficiency of patient care in EDs.

Keywords: ED, emergency, ESI, LOS, overcrowding, triage, XAI

1 INTRODUCTION

The Emergency Department (ED) conducts several tasks such as triaging, diagnosing, and treating medical issues that can be either light or threatening [1]. This means that even without a pandemic, the ED is experiencing an influx of patients [2], which eventually causes overcrowding. Triage is part of the ED process to assess the patient's medical urgency. The correctness of the triage classification is important to maintain since most patients who died or were in critical condition are not classified with high acuity at triage [3]. Emergency Severity Index (ESI) is one of the many triaging tools being used to classify the medical urgency of the patient's case. Additionally, ESI has five levels that range from one, being the most urgent, to five, being the least priority [4].

Length of Stay (LOS) is another clinical metric that measures the patient's duration of stay, from the time of their arrival until their discharge from the ED. In the ED, LOS has been linked with clinical issues such as poor outcomes and overcrowding [6]. When it comes to overcrowding, longer LOS may contribute to the relative shortage of beds and other resources [5, 6].

J. Caro et al. (eds.), Proceedings of the Workshop on Computation: Theory and Practice (WCTP 2024), Atlantis Highlights in Computer Sciences 23,
https://doi.org/10.2991/978-94-6463-684-0_16

Machine Learning (ML) can now aid healthcare workers' decision-making in patients' diagnosis and prognosis. When compared with conventional care, the ML-driven approach exhibits superior predictive performance than the traditional approach in terms of assessing ED patients with diverse health conditions [8]. Explainable AI (XAI) is a powerful tool that aims to explain the decisions or predictions of the ML model [9]. In this manner, the model becomes more interpretable, thus, making the users understand the decision or prediction.

Medical Information Mart for Intensive Care IV (MIMIC-IV) is a large-scale Electronic Health Record (EHR) database. One of its databases is the MIMIC-IV-ED and it contains various de-identified records of patients of Beth Israel Deaconess Medical Center from 2011 to 2019 [7]. This study is designed to create a web-based system that will incorporate the best ML model for the ESI level classification with XAI and the model for the ED LOS prediction using the MIMIC-IV-ED datasets.

1.1 Related Works

Overcrowding is one of the challenges faced by the ED which happens globally. Increasing bed capacities and staff does not resolve the problem [10]. However, Machine Learning (ML) can be one of the solutions [11], specifically in the early detection of the hospitalization of a patient in the triaging stage.

A study by Xie et al. [12] utilized MIMIC-IV-ED tables to benchmark emergency department prediction models. The study provided relevant pre-processing techniques that can be used to handle such raw datasets. Moreover, thresholds were also provided by the study to handle outliers that may skew the results. Aside from the preprocessing techniques, the study has shown the use of Area Under the Precision-Recall Curve (AUPRC) as the main metric.

In the study of Mutegeki et al. [13], various ML techniques were used to classify the ESI levels of ED patients. They have used Logistic Regression, Decision Trees, Random Forest, Extreme Gradient Boosting, and Histogram-based Gradient Boosting. Among all of them, Histogram-based Gradient Boosting yields the highest value for metrics with accuracy, precision, and an F1-score of 0.70. Moreover, the dataset used vital signs and chief complaints as predictor variables.

A more complex approach is done by Larburu et al. [14] who have done systematic reviews of various research studies related to predicting hospital admissions from the Emergency Department. This study provided a good overview of predictor variables and ML techniques that can be useful for ESI Level classification. Moreover, it is stressed that there is a need for large datasets to train and validate the models to ensure that the models are generalizable to different patient populations and healthcare settings.

Length of Stay (LOS), as tackled in the study of Zeleke et al. [15], indicates the quality of service, specifically in patient outcomes and success of procedures. One of the study's components is the utilization of regression algorithms to predict the patient's LOS. The results show that their best-performing models have an overall prediction error that suggests a six to seven-day mean difference.

The study also has shown useful metrics such as Mean Absolute Deviation, Mean Square Error, and Root Mean Square Error.

As XAI explains the decisions of the ML model, it can help bring a high level of transparency [17]. In addition, the feature-based model explainability technique is discussed in the study of Dave et al. [16]. In the aforementioned study of Mutegeki et al. [13], the ML model was made interpretable using both SHAP and LIME. As LIME focuses on the individual contributions of each variable, SHAP, on the other hand, was used to see which variables have high impacts on the model.

2 METHODS

2.1 Dataset

The MIMIC-IV-ED Triage dataset is constructed by extracting and combining tables under the MIMIC-IV-ED database. All the triage data such as the Vital Signs, Pain Scale, Chief Complaints, and Acuity (target class) are obtained from the *triage* table. Demographic data such as Age and Gender are extracted from the *patients* table. The Mode of Transportation column is obtained from *edstays* table.

As for the ED LOS Prediction, the same dataset is used. However, **acuity** column is added as one of the predictors. The ED LOS is extracted by getting the time difference in hours between the *outtime* and *intime* from the *edstays* table.

In the configuration where Medication Reconciliation is included, the *gsn* column, which contains the code of the medicines taken by the patient before the ED visit, from the *medrecon* table is added as one of the features for both ESI Level Classification and LOS Prediction. Moreover, *n_meds* is feature-engineered by counting the number of unique medicines per patient record.

2.2 Data Cleaning and Preprocessing

The *pain* column is converted from string to float. The text inputs are set as *null*, and numeric inputs that are in range form are set as the range's average. In the *chiefcomplaint* column, Python's Regular Expression is used to access keywords and symbols and manipulate them to standardize the inputs. The top 20 complaints are retained and the others are marked as "others". Moreover, the *gsn* column has over 9000 unique values. The top 20 most occurring *gsn* are obtained. The upper and lower thresholds for the numeric columns, which are used to handle outliers, are based on recommendations from a domain expert as outlined in the study of Xie et al. [12].Values below the Lower Outlier or above the Upper Outlier are set to null. Additionally, values falling between the Lower Outlier and Valid Low are replaced with the Valid Low, while values between the Upper Outlier and Valid High are replaced with the Valid High.

The constructed MIMIC-IV-ED triage dataset is split into 80% for the training set and the remaining 20% for the testing set. This study removes rows with missing values in *acuity* and *age* since imputing them may not represent the actual truth. Additionally, rows with at least four missing values in numeric columns are also eliminated. As for data imputation, Median Imputation is employed to fill the missing values in numeric columns. Moreover, numeric columns are scaled using Scikit-learn's MinMaxScaler.

Feature	**Lower Outlier**	**Valid Low**	**Valid High**	**Upper Outlier**
Body Temperature	14.2	26.0	45.0	47.0
Heart Rate	0.0	0.0	350.0	390.0
Respiratory Rate	0.0	0.0	300.0	330.0
Oxygen Saturation	0.0	0.0	100.0	150.0
Systolic BP	0.0	0.0	375.0	375.0
Diastolic BP	0.0	0.0	375.0	375.0
Pain Scale	0.0	0.0	10.0	10.0
ESI Level	1	1	5	5

Table 1: Threshold for Numeric Features

2.3 Model Training and Evaluation

ESI Level Classification The dataset is trained using conventional ML algorithms such as Decision Trees (DT), Random Forest (RF), Logistic Regression - One vs. Rest (LOG-OVR), Logistic Regression - Multinomial (LOG-MULTI), Gradient Boosting (GB), and Histogram-based Gradient Boosting (HGB). The models are evaluated using the Area Under the Precision-Recall Curve (AUPRC), and F1-score.

For experimentation, various techniques for handling class imbalance are used:

1. SMOTE;
2. SMOTE-Tomek Links;
3. Random Undersampler; and
4. Class Weights.

ED LOS Prediction The dataset is trained using conventional regression algorithms such as Decision Trees (DT), Linear Regression (LR), Gradient Boosting (GB), and Histogram-based Gradient Boosting (HGB). The models are evaluated using the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

For research purposes, various experimental setups are done:

1. Log Transformation;
2. Cube Root Transformation;
3. Log Transformation with SelectKBest; and
4. Cube Root Transformation with SelectKBest.

3 RESULTS

3.1 ESI Level Classification

As mentioned, there are various experimental setups are done. Based on their performance, models that utilized the default balance of the target class are chosen. These models are represented in the first few rows of Table 3. Moreover, these base models are tuned, and among them, the tuned HistGradient Boosting performs best with a weighted AUPRC of 68.99% and a weighted F1-score of 63.08%. Therefore, this tuned model is chosen for training the dataset with added features. However, when the tuned model is used on the extended dataset, the performance has further improved with an AUPRC of 69.478% and an F1-score of 63.891%.

Table 2: Summary of Performance - ESI Level Classification

Model	**AUPRC**	**F1-Score**
Across different experimental setups		
Decision Tree (Default balance)	0.6266	0.5380
Random Forest (Default balance)	0.6676	0.6240
LOG-OVR (max_iter=3000) (Default balance)	0.6462	0.5851
LOG-MULTI (max_iter=3000) (Default balance)	0.6463	0.5895
Gradient Boosting (Default balance)	0.6812	0.6237
HistGradient Boosting (Default balance)	0.6883	0.6287
After Hyperparameter Tuning		
Decision Tree (max_depth = 100, min_samples_split = 9, min_samples_leaf = 150, max_leaf_nodes = 355)	0.6617	0.6185
Random Forest (max_depth = 50, max_leaf_nodes = 3000, n_estimators = 300)	0.6841	0.6241
LOG-OVR (solver = 'saga', max_iter = 1000, C = 200, penalty = 'l2')	0.6469	0.5859
LOG-MULTI (solver = 'saga', max_iter = 3000, C = 40, penalty = 'l2')	0.6469	0.5905
Gradient Boosting (n_estimators = 200, learning_rate = 0.27)	0.6888	0.6301

Continued on next page

Table 2 – continued from previous page

Model	AUPRC	F1-Score
HistGradient Boosting (learning_rate = 0.1, max_leaf_nodes = 60, max_depth = None, min_samples_leaf = 20, l2_regularization = 2)	0.6899	0.6308
Best tuned model on the extended dataset		
HistGradient Boosting (learning_rate = 0.1, max_leaf_nodes = 60, max_depth = None, min_samples_leaf = 20, l2_regularization = 2)	0.6948	0.6389

SHAP is used to explain the variable contribution for each ESI Level, as shown in Figure 1-2. In the figure, Wound Evaluation, Heart Rate, Abdominal Pain, and Abnormal Labs positively contribute to classifying a patient as ESI Level 1.

Suicidal Ideation, Dyspnea, and Syncope are some of the features that contribute to classifying a patient as ESI Level 2.

For ESI Level 3, Pain, Age, and Weakness are some of the positively contributing features. This makes sense as these variables are seen contributing positively to ESI Levels 1 and 2, which are higher priorities.

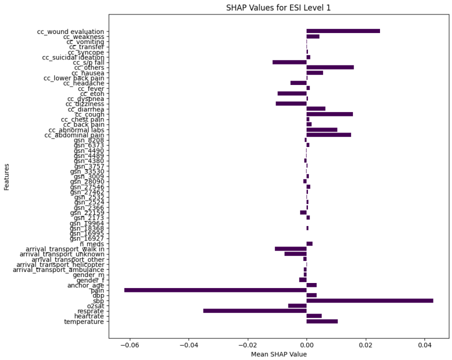
Moreover, Weakness, Nausea, and the Number of Medicines Taken Prior to the ED Visit are some of the features that have a positive contribution to ESI Level 4.

Finally, the ESI Level 5 is the most underrepresented class. Despite that, there are features that could still lead to the classification of a patient as ESI Level 5 namely Weakness, Transfer, Headache, and Number of Medicines Taken Prior to the ED Visit, among others.

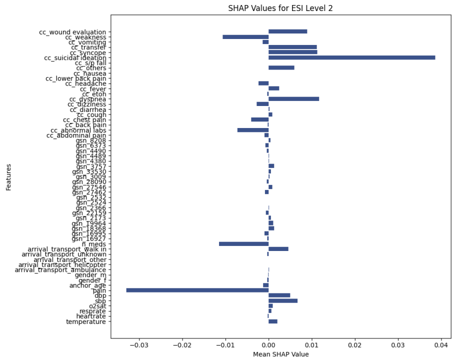
The sample LIME output from PredictED web application is shown in Figure 3. Aside from the LIME output, the predicted ESI Level and ED LOS are also shown. The LIME output explains the classification as it shows that the input Heart Rate, along with the absence of other features, positively contributed to the decision.

3.2 ED LOS Prediction

The ED LOS Prediction has various experiments. According to their performance, all the models that utilized Cube Root Transformation alone, except for Decision Trees, are chosen to undergo hyperparameter tuning. As for the Decision Trees, the setup that used Cube Root Transformation with SelectKBest yielded the highest performance. Moreover, as the models are tuned, HistGradient Boosting is the best-performing with an MAE of 3.458, MSE of 40.439, and RMSE of 6.359. When the best-performing model trained on the set of features suggested by the domain expert is tested against the extended feature, it still performs better.

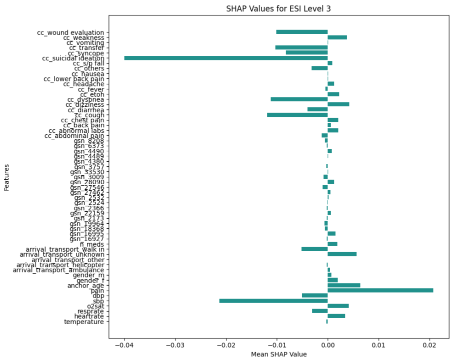


(a) ESI Level 1

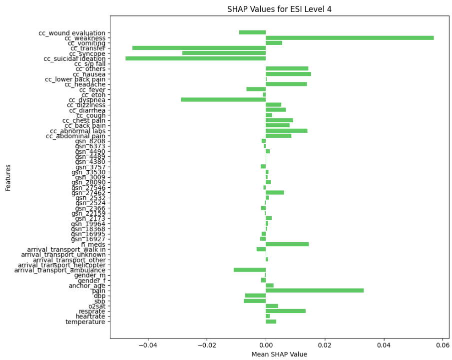


(b) ESI Level 2

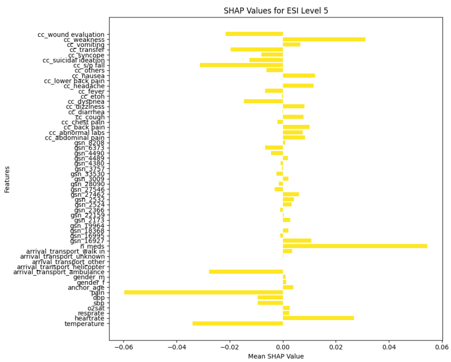
Fig. 1: Plots of SHAP Values for ESI Levels 1-2



(a) ESI Level 3



(b) ESI Level 4



(a) ESI Level 5

Fig. 2: Plots of SHAP Values for ESI Levels 3-5

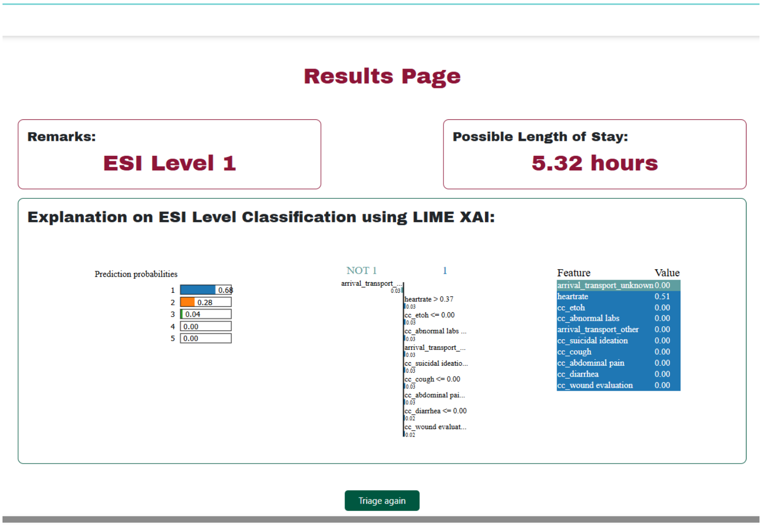


Fig. 3: Screenshot of Sample Output for ESI Level 1 with LIME

Table 3: Summary of Performance - LOS Prediction

Model	MAE	MSE	RMSE
Across different experimental setups			
Decision Tree (CRT w/ SelectKBest)	3.666	44.219	6.650
Linear Regression (CRT)	3.555	41.963	6.478
Gradient Boosting (CRT)	3.477	40.993	6.403
HistGradient Boosting (CRT)	3.459	40.511	6.364
After Hyperparameter Tuning			
Decision Tree (max_depth = 10, min_samples_split = 20, min_samples_leaf = 100)	3.515	41.078	6.409
Linear Regression (n_jobs = 21)	3.555	41.963	6.478
Gradient Boosting (learning_rate = 0.31 n_estimators = 500)	3.465	40.485	6.363
HistGradient Boosting (learning_rate = 0.13, max_iter = 100, max_leaf_nodes = 60, max_depth = 20, min_samples_leaf=100)	3.458	40.4395	6.359

Continued on next page

Table 3 – continued from previous page

Model	MAE	MSE	RMSE
Best tuned model on the extended dataset			
HistGradient Boosting (learning_rate = 0.13, max_iter = 100, max_leaf_nodes = 60, max_depth = 20, min_samples_leaf=100)	3.612	42.604	6.527

4 DISCUSSION

The ESI Level classification model had a performance that is average compared to the conventional high-accuracy prediction model. Handling class imbalance did not exhibit improvements, hence, the models that utilized the default balance are chosen to be tuned. Moreover, the hyperparameter tuning has shown slight improvements in the model's performance. However, when the tuned model trained using the features suggested by the expert is tested against the dataset with added features, the performance of the latter shows a slight improvement, which suggests that there may be other variables that could be predictive for ESI Level classification. Despite these constraints, the XAI, both SHAP and LIME were able to provide a better view of the model and the predictions by quantifying and visualizing the individual impact of each feature, offering a more granular understanding of how each variable contributes to the prediction of ESI levels. This contributes a novel layer of transparency and interpretability, aligning with current trends toward explainable AI in healthcare.

The ED LOS Prediction model had a performance that was below the typical high-value metrics. Moreover, the use of the dataset with added features did not show any improvement. Despite these, the implementation of this component remains beneficial to the target users, aspiring medical doctors and nurses, as they could have an estimate of the patient's possible LOS. Overall, this system would remain beneficial as the goal of this is not to replace the professionals but rather provide them with a tool to aid decision-making and resource allocation.

5 CONCLUSION

This study was able to come up with models for the classification of the ESI Level of patients, and the prediction of patient's ED LOS. This shows that Histogram-based Gradient Boosting performs best in classifying the ESI Level of a patient when using the dataset with additional features such as the number of medicines taken prior to the ED visit and the medicines taken prior to the ED visit. Moreover, the XAI, both SHAP and LIME, was able to provide explanations on the variable's contributions to the model in general and the individual predictions. Additionally, Histogram-based Gradient Boosting was also the best model for the prediction of a patient's ED LOS using the suggested features of the expert.

References

1. McKenna, P., Heslin, S. M., Viccellio, P., Mallon, W. K., Hernandez, C., Morley, E. J.: Emergency department and hospital crowding: causes, consequences, and cures (2019)
2. Kelen, G. D., Wolfe, R., D'Onofrio, G., Mills, A. M., Diercks, D., Stern, S. A., Wadman, M. C., Sokolove, P. E.: Emergency department crowding: the canary in the health care system (2021)
3. Hinson, J. S., Martinez, D. A., Cabral, S., George, K., Whalen, M., Hansoti, B., Levin, S.: Triage performance in emergency medicine: a systematic review (2019)
4. Sax, D. R., Warton, E. M., Mark, D. G., Vinson, D. R., Kene, M. V., Ballard, D. W., Vitale, T. J., McGaughey, K. R., Beardsley, A., Pines, J. M., and others: Evaluation of the emergency severity index in US emergency departments for the rate of mistriage (2023)
5. Farimani, R. M., Karim, H., Atashi, A., Tohidinezhad, F., Bahaadini, K., Abu-Hanna, A., Eslami, S.: Models to predict length of stay in the emergency department: a systematic literature review and appraisal (2024)
6. Andersson, J., Nordgren, L., Cheng, I., Nilsson, U., Kurland, L.: Long emergency department length of stay: A concept analysis (2020)
7. Bulgarelli, J. A., Pollard, L., Celi, T., Mark, L. A., Horng, R., Horng, S.: MIMIC-IV-ED (version 2.2) (2023)
8. Kareemi, H., Vaillancourt, C., Rosenberg, H., Fournier, K., Yadav, K.: Machine learning versus usual care for diagnostic and prognostic prediction in the emergency department: a systematic review (2021)
9. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G. Z.: XAI-Explainable artificial intelligence (2019)
10. Savioli, G., Ceresa, I. F., Gri, N., Bavestrello Piccini, G., Longhitano, Y., Zanza, C., Piccioni, A., Esposito, C., Ricevuti, G., Bressan, M. A.: Emergency department overcrowding: understanding the factors to find corresponding solutions (2022)
11. Sartini, M., Carbone, A., Demartini, A., Giribone, L., Oliva, M., Spagnolo, A. M., Cremonesi, P., Canale, F., Cristina, M. L.: Overcrowding in emergency department: causes, consequences, and solutions—a narrative review (2022)
12. Xie, F., Zhou, J., Lee, J. W., Tan, M., Li, S., Rajnthern, L., Chee, M. L., Chakraborty, B., Wong, A. I., Dagan, A., and others: Benchmarking emergency department prediction models with machine learning and public electronic health records (2022)
13. Mutegeki, H., Nahabwe, A., Nakatumba-Nabende, J., Marvin, G.: Interpretable Machine Learning-Based Triage For Decision Support in Emergency Care (2023)
14. Larburu, N., Azkue, L., Kerexeta, J.: Predicting Hospital Ward Admission from the Emergency Department: A Systematic Review (2023)
15. Zeleke, A. J., Palumbo, P., Tubertini, P., Miglio, R., Chiari, L.: Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis (2023)
16. Dave, D., Naik, H., Singhal, S., Patel, P.: Explainable ai meets healthcare: A study on heart disease dataset (2020)
17. Mohanty, A., Mishra, S.: A Comprehensive Study of Explainable Artificial Intelligence in Healthcare (n.d.)