ORIGINAL ARTICLE

https://doi.org/10.1590/1806-9282.20260287

Artificial intelligence-driven triage in emergency departments: a comparative study of machine learning approaches

Ayşe Şule Akan[1]*, Sefa Kucuk[2], Şule Yakar[3], Mahmut Tutam[4], Hatice Esen[2], İbrahim Toker[3]

SUMMARY

OBJECTIVE: In emergency departments, accurate and timely triage plays a critical role in prioritizing patient care. Traditional triage often suffers from subjectivity and inconsistency, especially under time pressure and overcrowded conditions. The aim of this study was to evaluate the predictive potential of multiple supervised machine learning models for three-class emergency department triage classification and compare their performance using routinely available triage variables.

METHODS: We conducted a comparative analysis of eight supervised machine learning models using a retrospective dataset comprising 510 patient records from the emergency department triage. The dataset included demographic features, vital signs, and chief complaints. A 10-fold cross-validation strategy was employed to evaluate performance in a three-class triage outcome scenario based on urgency levels using accuracy, precision, recall, and F1-score.

RESULTS: Among the evaluated models, logistic regression, random forest, support vector machine, and majority voting achieved the highest predictive performance, with overall accuracies exceeding 89%. The confusion matrix analysis revealed that low-urgency cases were consistently classified with the highest accuracy across all models, while misclassifications occurred primarily between high-urgency and medium-urgency cases.

CONCLUSION: The results show that machine learning models appear well-suited to support the emergency departments triage decision-making process. Integrating such models into emergency departments workflows can improve patient prioritization, streamline clinical operations, and ultimately contribute to better quality of care.

KEYWORDS: Emergency departments. Machine learning. Triage.

INTRODUCTION

The recent global pandemic, coupled with an aging population, has led to a significant increase in demand for emergency medical assistance worldwide. In 2021, member countries of the Organization for Economic Cooperation and Development (OECD) with available data reported an average of 27 emergency department (ED) visits per 100 people annually[1]. Examples include Canada with 36.6 visits, England with 43.1, Australia with 34.2, and the United States of America with 43 visits per 100 people[2]. In Türkiye, the rate is significantly elevated, approximately 5.56 times the average of OECD countries[3]. When demand exceeds the capacity of EDs, several challenges in patient care can emerge. These include prolonged waiting times for medical evaluation, insufficient physical infrastructure, overburdened healthcare personnel, and delays in both treatment initiation and patient discharge[4]. In response to these challenges, emergency care in EDs typically begins with a triage process, which enables the prioritization of patients based on the severity and urgency of their medical needs.

Most widely adopted emergency department triage systems use five-level classification models. The most recognized are the Emergency Severity Index (ESI), Manchester Triage System (MTS), Canadian Triage and Acuity Scale (CTAS), and Australian Triage Scale (ATS), which form the basis of many global triage models. While most systems follow five levels, China uses a four-level system, the Chinese Emergency Triage Scale (CETS), whereas Türkiye uses a three-level, color-coded system called the Türkiye Triage System (TTS), established by a directive of the Turkish Ministry of Health in 2009[5]. TTS classifies patients into three levels: Red, Yellow, and Green, with Red and Yellow further divided into two subcategories[6]. Red represents the most critical cases: Subcategory 1 requires immediate intervention for life-threatening conditions (e.g., cardiac arrest, major trauma), while Subcategory 2 includes

[1]Erzurum City Hospital, Department of Emergency Medicine – Erzurum, Türkiye.
[2]Erzurum Technical University, College of Engineering and Architecture, Department of Electrical and Electronics Engineering – Erzurum, Türkiye.
[3]Kayseri City Hospital, Department of Emergency Medicine – Kayseri, Türkiye.
[4]Erzurum Technical University, College of Engineering and Architecture, Department of Industrial Engineering – Erzurum, Türkiye.
*Corresponding author: dr.aysesuleakan@gmail.com
Conflicts of interest: the authors declare there is no conflicts of interest. Funding: none.
Received on March 01, 2026. Accepted on July 04, 2026.
Scientific Editor: José Maria Soares Júnior

high-risk cases needing treatment within 10 min (e.g., severe chest pain, acute neurological symptoms). Yellow covers serious but less urgent conditions: Subcategory 1 involves risk of limb loss or significant morbidity (e.g., severe hypertension, moderate bleeding), and Subcategory 2 includes prolonged yet non-critical issues such as minor trauma or gastrointestinal complaints. Green refers to stable patients with mild symptoms suitable for outpatient management.

Effective triage implementation ensures appropriate patient prioritization while optimizing ED resource allocation[7]. Reliable triage systems should be based on validated scoring methods[8]; however, variations in data collection, cultural context, and individual or environmental factors may introduce assessment errors[9]. Integrating artificial intelligence (AI) can mitigate these issues by enhancing the precision, consistency, and overall effectiveness of traditional triage systems[10]. Machine learning (ML) identifies actionable patterns within complex datasets through computational algorithms. For instance, a Gradient Boosted Decision Tree model implemented with the CatBoost library was developed to support ED triage officers in identifying potentially life-threatening mis-triage cases based on CETS[11]. This approach significantly reduced mis-triage among critically ill patients and outperformed logistic regression (LR).

Several ML models, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF), were evaluated for triage in the TTS[12], with RF achieving the best performance. Beyond outcome prediction, an XGBoost model was developed to recommend critical clinical interventions, improving timely decision-making[13]. An XGBoost-based model optimized with Hyperopt replicated clinicians' decisions, reducing over-triage and under-triage while improving efficiency[14]. Another study compared multiple ML algorithms for predicting discharge, hospitalization, ICU admission, and mortality, using oversampling and genetic algorithms for optimization; RF emerged as the top model[15]. Additionally, six ML models integrating structured and unstructured triage data via NLP were evaluated against clinicians' predictions, accurately forecasting ICU admission, mortality, and hospitalization, demonstrating their value in patient prioritization[16].

This study aims to compare the performance of several supervised ML algorithms for three-class ED triage prediction within the Turkish triage system using routinely available triage variables. By evaluating these models on a retrospective, single-center dataset, we seek to assess their predictive performance and identify methodological considerations for future externally validated triage decision-support studies.

METHODS

Dataset

This study used anonymized patient data collected from the ED of Kayseri City Hospital. Ethical approval for data collection was obtained from the Non-Interventional Clinical Research Ethics Committee. The data includes structured triage information for adult patients (aged 18 and above), including demographic characteristics (e.g., age, gender), vital signs, and clinical complaints. Only patients with complete triage data and approved urgency classifications were included in the analysis. The dataset comprises 510 patients equally represented in three triage categories: red, yellow, and green. The features were selected based on their routine availability during ED triage and relevance to urgency classification. The dataset includes structured information on patient demographics, vital signs, and chief complaints recorded at triage. Demographic variables comprise age and gender. Vital signs include systolic and diastolic blood pressure, heart rate, respiratory rate, body temperature, and oxygen saturation. Chief complaints represent the primary condition reported at admission, with a total of 58 distinct complaints recorded in the dataset.

Data preparation

We applied several preprocessing steps to the dataset to ensure machine readability and consistency. Initially, categorical values such as gender and urgency labels were encoded into numerical representations using label encoding, allowing for direct processing by algorithms. To ensure comparability between variables, the six vital sign features were normalized between 0 and 1. The preprocessed data was exported in the attribute-relationship file format, which allows for seamless integration with the WEKA ML environment.

Machine learning approaches

This study applied a supervised ML approach to train eight models on the ED triage dataset: BayesNet, DT, KNN, LR, Multilayer Perceptron (MLP), RF, SVM, and Majority Voting (MV). We employed model evaluation using WEKA. Given the relatively small size of the dataset, extensive hyperparameter search was avoided; instead, limited tuning was performed for the main tunable classifiers using WEKA's CVParameterSelection within the 10-fold cross-validation framework. BayesNet was kept with the standard WEKA configuration, and MV was reconstructed using optimized base classifiers. Table 1 summarizes the tuning strategy. Model performances were evaluated using accuracy, precision, recall, F1-score, specificity, and AUROC, and 95%CI for accuracy were calculated using the Wilson method.

Implementation details

We performed data analysis and visualization in Python and classification experiments in WEKA due to its robust collection of ML algorithms and filters. Specifically, we applied the Partition Membership filter to transform the data into a propositionalized representation, as in[17]. This filter partitions the instance space using a partition generator based on DT models such as J48, RF, or REPTree and encodes membership values as new attributes. The resulting transformed instances are combined with the original class labels to form a sparse representation suitable for classification. This transformation allows traditional classifiers to be used on the data while preserving the essential structural information obtained from the partitioning process.

RESULTS

In this section, we first explore the correlations between the features to understand their mutual relationships and then present the comparative performance of the ML models on the triage classification task.

Figure 1 presents the feature correlation matrix for the vital signs and TTS class. A positive correlation exists between diastolic and systolic blood pressure (r=0.69), aligning with physiological expectations. Heart rate shows a moderate positive correlation with respiratory rate (r=0.40), while oxygen saturation shows a moderate negative correlation with respiratory rate (r=-0.48). Correlations between individual vital signs and TTS class are generally weak to moderate, with the highest association observed for oxygen saturation (r=0.36), followed by respiratory rate (r=-0.20). The positive association between oxygen saturation and TTS suggests that higher saturation values are linked to less urgent triage categories, whereas the negative association of respiratory rate with TTS indicates higher respiratory rates in more urgent cases. Overall, no single vital sign strongly predicts triage class independently.

We evaluated the performance of several ML algorithms to classify patients into urgency levels. All models were evaluated using a 10-fold cross-validation strategy to obtain an internal estimate of predictive performance. In each iteration, nine folds were used for training, and the remaining fold was used for testing. This process was repeated until each fold served once as the test set, and the final performance scores were averaged across all folds.

The results in Table 2 show that LR and RF yield the highest overall classification performance in most metrics. LR achieves the highest accuracy (0.902; 95%CI 0.873–0.925), precision (0.902), and specificity (0.951), while RF closely follows and

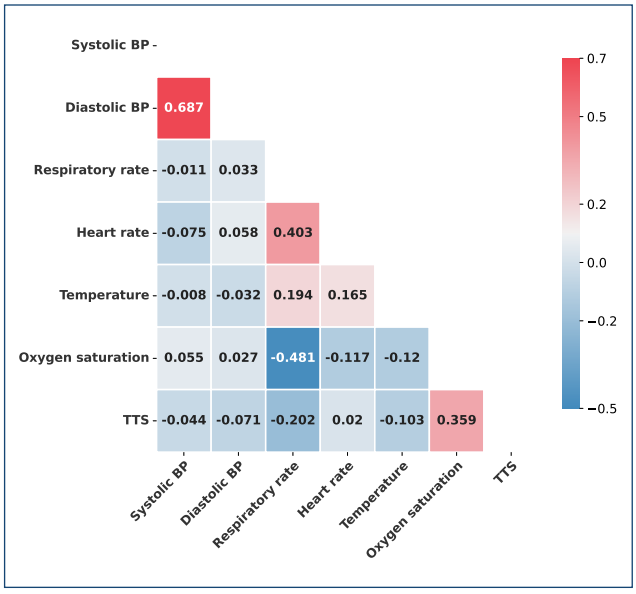

Figure 1. Correlation matrix of the features and Türkiye Triage System class.

Table 1. Hyperparameters used for the machine learning models.

Models	Tuned parameter(s)	Search range	Selected value
BayesNet	Not tuned	Standard configuration	Default
DT	Confidence factor	0.1–0.5	0.1
KNN	Number of neighbors	1–15	1
LR	Ridge parameter	1.0E-8–1.0E-4	1.0E-8
MLP	Learning rate	0.1–0.5	0.1
RF	Number of trees	50–300	50
SVM	Complexity parameter	0.1–10	2.575
MV	Not independently tuned	Optimized base classifiers	Majority voting

DT: Decision Tree; KNN: k-Nearest Neighbors; LR: logistic regression; MLP: Multilayer Perceptron; RF: Random Forest; SVM: Support Vector Machine; MV: Majority Voting.

achieves the highest AUROC. KNN, SVM, and MV also demonstrate competitive performance, with accuracy values close to 0.89. However, overlapping CI indicates small numerical differences among the best-performing models. In contrast, BayesNet underperforms across all metrics, with an accuracy of only 0.578 and an F1-score of 0.539, suggesting limited discriminative capacity on this dataset.

Figure 2 presents the confusion matrices, offering a detailed view of class-wise performance and highlighting prediction trends across urgency levels. LR, RF, SVM, and the MV ensemble demonstrate the most consistent and accurate results, with high true positive rates across all three TTS classes. In contrast, BayesNet performs substantially worse, frequently misclassifying TTS-2 and TTS-3, indicating limited ability to capture class overlap. DT and KNN show moderate performance. DT achieves reasonable accuracy, particularly for TTS-3, but often misclassifies TTS-1 as TTS-2, while KNN occasionally confuses TTS-1 and TTS-2, suggesting insufficient separation in local feature neighborhoods. MLP maintains a good balance across classes, though minor confusion between TTS-1 and TTS-2 persists, likely due to overlapping feature distributions.

The dataset is class-balanced, with 170 instances per urgency level. Among all classifiers, TTS-3 is the most accurately detected class: LR, RF, SVM, and MV correctly classify 162 of 170 TTS-3 cases. In contrast, TTS-2 is the most challenging class, with BayesNet misclassifying nearly half of its instances, often as TTS-3. Performance on TTS-1 is generally stable, though DT and MLP occasionally confuse it with TTS-2.

Table 2. Classification metrics comparison.

Methods	Accuracy (95%CI)	Precision	Recall	F1-score	Specificity	AUROC
BayesNet	0.578 (0.535–0.621)	0.621	0.578	0.539	0.789	0.786
DT	0.845 (0.811–0.874)	0.845	0.845	0.845	0.923	0.909
KNN	0.894 (0.864–0.918)	0.895	0.894	0.894	0.947	0.935
LR	0.902 (0.873–0.925)	0.902	0.902	0.902	0.951	0.931
MLP	0.880 (0.849–0.906)	0.880	0.880	0.880	0.940	0.935
RF	0.898 (0.869–0.921)	0.898	0.898	0.897	0.949	0.945
SVM	0.890 (0.860–0.914)	0.890	0.890	0.890	0.945	0.929
MV	0.892 (0.862–0.916)	0.892	0.892	0.892	0.946	0.919

DT: Decision Tree; KNN: k-Nearest Neighbors; LR: logistic regression; MLP: Multilayer Perceptron; RF: Random Forest; SVM: Support Vector Machine; MV: Majority Voting; AUROC: area under the receiver operating characteristic curve.

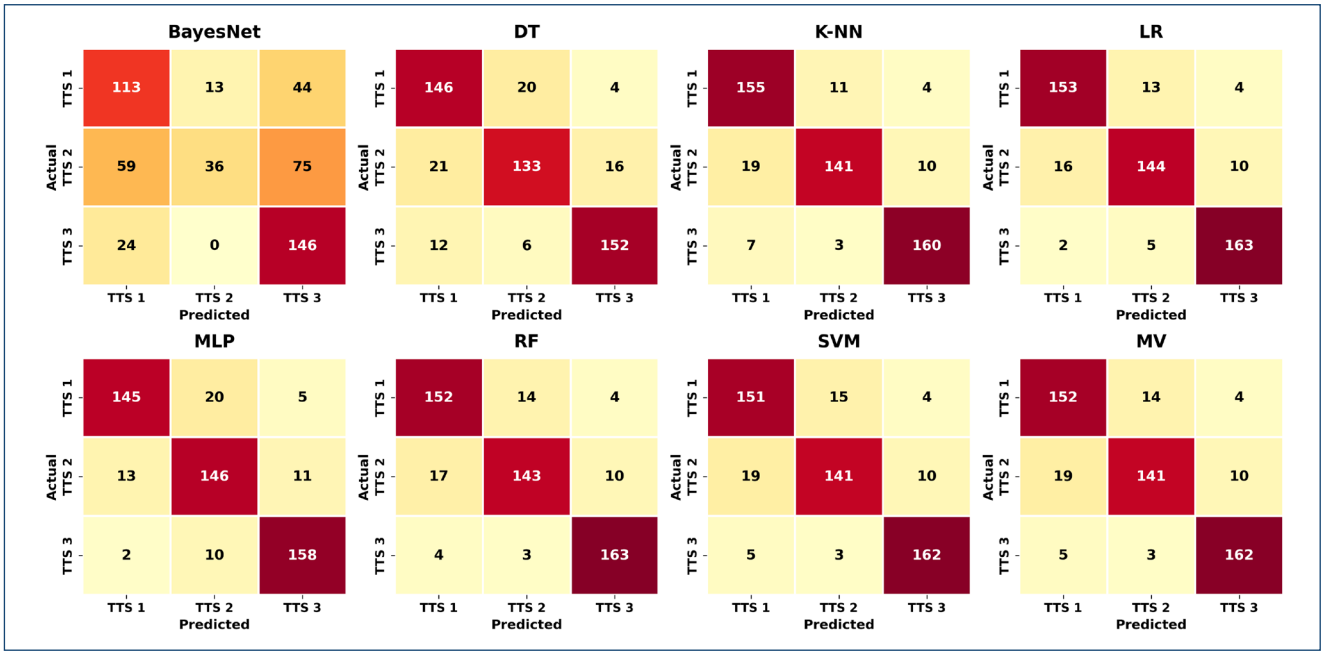

Figure 2. Confusion matrices obtained from a 10-fold cross-validation.

DISCUSSION AND CONCLUSION

This study evaluated several ML algorithms for three-class ED triage prediction using a balanced single-center dataset. LR achieved the highest numerical accuracy, while RF showed the highest AUROC. However, overlapping Wilson CIs indicate that differences among the best-performing models should be interpreted with caution. Additionally, correlation analysis showed that routinely measured vital signs, particularly oxygen saturation, contained relevant information for TTS classification; however, no single variable could independently explain triage severity. These findings suggest that machine learning–based models can support triage decision-making processes but should be considered as complementary tools rather than replacements for clinical judgment.

Several limitations must be considered. The relatively small sample size, single-center retrospective design, and lack of external validation may restrict generalizability. While the study aimed to compare ML algorithms within a controlled and balanced dataset, further validation using larger, prospective, and multi-institutional cohorts is necessary to confirm robustness, fairness, and clinical applicability in real-world emergency care.

AUTHORS' CONTRIBUTIONS

AŞA: Conceptualization, Writing – original draft. **SK:** Conceptualization, Methodology, Software, Visualization, Writing – review & editing. **ŞY:** Data curation. **MT:** Conceptualization, Methodology, Software, Writing – review & editing. **HE:** Software, Visualization. **İT:** Data curation.

DATA AVAILABILITY STATEMENT

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

REFERENCES

1. OECD. Health at a glance 2023: OECD indicators. Paris: OECD Publishing; 2023. https://doi.org/10.1787/7a7afb35-en
2. National Center for Health Statistics. National Hospital Ambulatory Medical Care Survey. 2022. Available from: https://www.cdc.gov/nchs/nhamcs/about/index.html
3. Birinci S, Ülgü MM, Gözükara MG. Critical insights based on the ministry of health's 6-year data analysis: an epidemiological study of patient visits trends of emergency departments in Türkiye. Haydarpasa Numune Train Res Hosp Med J. 2023;63(3):334-9. https://doi.org/10.14744/hnhj.2023.17048
4. Gottlieb M, Farcy DA, Moreno LA, Vilke GM, Guittard JA. Triage nurse-ordered testing in the emergency department setting: a review of the literature for the clinician. J Emerg Med. 2021;60(4):570-5. https://doi.org/10.1016/j.jemermed.2020.11.004
5. N. 27378 of the Republic of Turkey. Notification on the rules and principles of emergency services in in-patient health care facilities. 2009. Available from: https://www.resmigazete.gov.tr/eskiler/2009/10/20091016-16.htm
6. Aslanoğlu A, Ayyıldız M. Triage in the world and in Turkey and the triage knowledge levels of emergency ward nurses. Sağlık Bilim Derg. 2023;32(2) :316-22. https://doi.org/10.34108/eujhs.1159047
7. Andersson J, Nordgren L, Cheng I, Nilsson U, Kurland L. Long emergency department length of stay: a concept analysis. Int Emerg Nurs. 2020;53:100930. https://doi.org/10.1016/j.ienj.2020.100930
8. Cairós-Ventura LM, Mercedes Novo-Muñoz M, Rodríguez-Gómez JÁ, Ortega-Benítez ÁM, Ortega-Barreda EM, Aguirre-Jaime A. Validity and reliability of the emergency severity index in a Spanish hospital. Int J Environ Res Public Health. 2019;16(22):4567. https://doi.org/10.3390/ijerph16224567
9. Souza CC, Toledo AD, Tadeu LF, Chianca TC. Risk classification in an emergency room: agreement level between a Brazilian institutional and the Manchester Protocol. Rev Lat Am Enfermagem. 2011;19(1):26-33. https://doi.org/10.1590/s0104-11692011000100005
10. Stewart J, Sprivulis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. Emerg Med Australas. 2018;30(6):870-4. https://doi.org/10.1111/1742-6723.13145
11. Liu Y, Gao J, Liu J, Walline JH, Liu X, Zhang T, et al. Development and validation of a practical machine-learning triage algorithm for the detection of patients in need of critical care in the emergency department. Sci Rep. 2021;11(1):24044. https://doi.org/10.1038/s41598-021-03104-2
12. Bora Ş, Kantarci A, Erdogan A, Beynek B, Kheibari B, Evren V, et al. Machine learning for E-triage. Int J Multidiscip Stud Innov Technol. 2022;6(1) :86. https://doi.org/10.36287/ijmsit.6.1.86
13. Chang H, Yu JY, Yoon S, Kim T, Cha WC. Machine learning-based suggestion for critical interventions in the management of potentially severe conditioned patients in emergency department triage. Sci Rep. 2022;12(1):10537. https://doi.org/10.1038/s41598-022-14422-4
14. Gao Z, Qi X, Zhang X, Gao X, He X, Guo S, et al. Developing and validating an emergency triage model using machine learning algorithms with medical big data. Risk Manag Healthc Policy. 2022;15:1545-51. https://doi.org/10.2147/RMHP.S355176
15. Elhaj H, Achour N, Tania MH, Aciksari K. A comparative study of supervised machine learning approaches to predict patient triage outcomes in hospital emergency departments. Array. 2023;17:100281. https://doi.org/10.1016/j.array.2023.100281
16. Chang YH, Lin YC, Huang FW, Chen DM, Chung YT, Chen WK, et al. Using machine learning and natural language processing in triage for prediction of clinical disposition in the emergency department. BMC Emerg Med. 2024;24(1):237. https://doi.org/10.1186/s12873-024-01152-1
17. Daşdemir Y, Özakar R. Affective states classification performance of audio-visual stimuli from EEG signals with multiple-instance learning. Turkish J Electr Eng Comput Sci. 2022;30(7):2707-24.