Array 17 (2023) 100281

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

A comparative study of supervised machine learning approaches to predict patient triage outcomes in hospital emergency departments

Hamza Elhaj [a,*], Nebil Achour [b], Marzia Hoque Tania [c], Kurtulus Aciksari [d]

[a] PhD Researcher @ Anglia Ruskin University. Chelmsford, UK
[b] Associate Professor in Disaster Mitigation @ Anglia Ruskin University. Cambridge, UK
[c] Postdoctoral Researcher @ Department of Engineering Science of Oxford University. Oxford, UK
[d] Emergency Medicine Professor @ Istanbul Medeniyet University. Istanbul, Turkey

ARTICLE INFO

Keywords:
Triage
Emergency departments
Machine learning
Ensemble learning
Critical outcomes
Patient outcomes prediction

ABSTRACT

Background: The inconsistency in triage evaluation in emergency departments (EDs) and the limitations in practice within the standard triage tools among triage nurses have led researchers to seek more accurate and robust triage evaluation that provides better patient prioritization based on their medical conditions. This study aspires to establish the best methodological practices for applying machine learning (ML) techniques to build an automated triage model for more accurate evaluation.

Methods: A comparative study of selected supervised ML models was conducted to determine the best-performing approach to evaluate patient triage outcomes in hospital emergency departments. A retrospective dataset of 2688 patients who visited the ED between April 1, 2020 and June 9, 2020 was collected. Data included patient demographics (age and gender), Vital signs (body temperature, respiratory rate, heart rate, blood pressure and oxygen saturation), chief complaints, and chronic illness. Nine supervised ML techniques were investigated in this study. Models were trained based on patient disposition outcomes and then validated to evaluate their performance.

Findings: ML models show high capabilities in predicting patient disposition outcomes in ED settings. Four models (KNN, GBDT, XGBoost, and RF) performed better than the rest. RF was selected as the optimal model as it demonstrated a slight advantage over the other models with 89.1% micro accuracy, 89.0% precision, 89.1% recall, and 89.0% F1-score, exhibiting outstanding performance in differentiation between patients with critical outcomes (e.g., Mortality and ICU admission) from those patients with less critical outcomes (e.g., discharged and hospitalized) in ED settings.

Conclusion: Machine learning techniques demonstrate high promise in improving predictive abilities in emergency medicine and providing robust decision-making tools that can enhance the patient triage process, assist triage personnel in their decision and thus reduce the effects of ED overcrowding and enhance patient outcomes.

1. Introduction

Emergency departments are considered one of the most stressful healthcare areas due to budget constraints [1]. They are encountering a considerable rise in attendance, causing overcrowding and delays internationally [2]. EDs worldwide suffer from overcrowding, mainly risking lives and affecting the quality of care [3]. Overcrowding is described as *"the greatest obstacle to providing timely and appropriate emergency care"* worldwide [4]. Over the last decade, hospitals struggled to maintain overcrowding problems with limited resources and rising patient demand [5]. This massive increase in demand can adversely impact healthcare service provision and lead to many negative consequences. Long waiting times, length of stay (LOS), patient revisits, leave without being seen (LWBS) [6,7], and patient dissatisfaction [8] are some of the negative consequences that have been identified in the literature. Overcrowding caused more admissions to inpatient departments, poorer care provision [9], and delays in diagnosis and treatment [10], affecting those in more need [11] in a way that increases the rate of morbidity and mortality [12]. It is also estimated that 40% of patients attending EDs have non-urgent problems [13], raising the question of *how to effectively identify those in urgent need from those in less urgent need*. It indicates that succeeding in reducing these numbers will

* Corresponding author.
E-mail address: htme100@pgr.aru.ac.uk (H. Elhaj).

https://doi.org/10.1016/j.array.2023.100281
Received 31 October 2022; Received in revised form 14 January 2023; Accepted 20 January 2023
Available online 30 January 2023
2590-0056/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

reduce some of the problems of EDs and the wider healthcare system.

Medical triage is one of the techniques applied to minimize the effect of ED overcrowding by identifying patients with urgent medical situations and determining the care level and resources required by patients before delivering care [14]. It is implemented to ensure patients with the most acute conditions are prioritized for assessment when the waiting room is loaded with new patients [15]. Hence, hospital administrators need accurate identification and prioritization of patient conditions by implementing accurate and precise triage systems. However, due to the time pressure, various medical complaints, limited information and resources, and the subjectivity of triage evaluations make the practice of triage more challenging [16]. Since standard triage tools rely heavily on human judgments, they are subject to high variation and individual bias and may affect the accuracy and precision of triage evaluation [16]. Raising the need for a more robust assessment of patient conditions.

The rapid growth of technology-based systems such as Electronic Health Records (EHRs) generates enormous amounts of data for more effective decision-making. Artificial Intelligence (AI) enabled a better understanding of this data. It helped predict individual patient outcomes [17]. Machine learning, as an element of AI, has been adopted in healthcare applications due to its advanced ability to analyze complex and large-scale data [18]. ML is a set of computational methods that learn patterns in data without being explicitly programmed [16]. They offer benefits for predictive clinical applications as they can be designed to yield stable predictions and recognize patterns within a dataset to predict patient outcomes and assist in diagnosis and treatment in many clinical scenarios [19]. Hence, this study will explore the advantages of applying ML techniques to automate the triage process and whether the application of ML will improve the decision-making of patient triage within hospital ED settings compared to the current triage system. This study will also provide a comparative approach of selected supervised machine learning techniques in the case of predicting triage outcomes within the hospital emergency department to establish the best machine learning technique to be applied for such a problem.

2. Literature review

Triage assessment in emergency care systems has become a challenge due to the rapid increase of patients with different acuity levels [20]. Despite traditional triage systems showing a good ability to tackle ED overcrowding, it lacks better patient sorting in busy ED settings. Overcrowding also forces nursing staff to rush the triage process to see other patients, resulting in human errors that can endanger patient lives. The recent advancement of ML technologies has emerged as a solid candidate to automate the triage decision-making process to obtain an accurate and rapid examination of patients, ensure timely treatment according to the severity of their conditions and reduce human error and judgmental bias evaluations [21]. That has led many researchers to take advantage of ML technologies to develop models to help predict patient needs for hospitalization and prioritize them according to the intensity of care provided by a registered nurse at the time of triage in hospital EDs [13,16,22–24]. Many researchers have proven the ability of ML to offer better performance in predicting hospitalization and critical-care outcomes over the reference triage models through the evaluation of nursing triage [16,22,25], with the potential to tackle overcrowding and provide better health services for patients and reduce morbidity and mortality rate among patients [26].

Despite the high potential of predictive abilities of these intelligent triage models, there is still considerable scope for improvement for these models in the reported studies. We have investigated multiple recent academic publications that attempted to employ ML techniques to predict triage outcomes, as mentioned in Table 1. Many of these studies lack clarity in some areas. For example, Raita et al. [23] and Jiang et al. [22] lack clarity on how datasets were collected and organized. Levin et al. [16] and Gao et al. [28] neither explained how algorithms were chosen nor compared with other approaches with sufficient clarity. These

Table 1
The recent publications related to ML-based Triage models.

Authors (year)	Triage System in Use	ML technique	Prediction outcomes	Performance measures
Dugas et al. (2016) [25],	Emergency Severity Index (ESI).	LR	Triage score & Patient outcomes.	AUC-ROC.
Levin et al. (2018) [16],	Pediatric ESI.	RF	Triage score & Patient outcomes.	AUC-ROC.
Raita et al. (2019) [23],	ESI.	Lasso, RF, GBDT, DL.	Triage score & Patient outcomes.	Sensitivity, Specificity, AUC-ROC, Accuracy, PPV, NPV.
Goto et al. (2019) [24],	Pediatric ESI.	Lasso, RF, GBDT, DL.	Patient outcomes.	Sensitivity, Specificity, AUC-ROC, Accuracy, PPV, NPV.
Choi et al. (2019) [27],	Korean Triage and Acuity Scale (KTAS).	LR, RF, XGBoost.	Triage score.	Precision, Recall, F1-score, AUC-ROC.
Wolff, Ríos and Graña (2019) [13],	Rule-based expert system.	DL, RF, NB, SVM.	Patient outcomes.	Sensitivity, Specificity, AUC-ROC, Accuracy, PPV, NPV.
Jiang et al. (2021) [22],	4-Level Chinese Emergency Triage Scale (CETS)	LR, RF, XGBoost, GBDT.	Triage score.	Accuracy, AUC-ROC, macro-F1.
Gao et al. (2022) [28],	A four-level triage standard	XGBoost	Triage score	AUC-ROC
Chang et al. (2022) [29],	KTAS	XGBoost	Critical interventions	AUC-ROC, Sensitivity, Specificity, PPV, NPV.
This article	Manchester Triage System (MTS)	LR, DT, KNN, SVM, MLP, GBDT, XGBoost, AdaBoost, RF	Patient outcomes.	Accuracy, Precision, Recall, F1-score.

Table 2
Patients triage information.

General (Input)	
ED visits - number	4540
Age - median ± SD	53 ± 20.00
Gender - Female %	1637 [36.05%]
Vital Signs (Input)	
Temperature (∘C) - median ± SD	36.7 ± 0.77
Heart rate (Pulse/min) - median ± SD	93 ± 15.47
Respiratory Rate (Breath/min) - median ± SD	19 ± 6.23
Oxygen saturation (%) - median ± SD	96 ± 6.55
Systolic BP (mmHg) - median ± SD	131 ± 15.63
Diastolic BP (mmHg) - median ± SD	76 ± 7.75
Chronic Illness indicator (Input)	
Has chronic illnesses	1389
No chronic illnesses	3151
Chief Complaints (CCs) (Input)	9 different CCs
Patient Outcomes (Output)	
Discharge	2278
Hospitalization to COVID-19 Services	1040
Hospitalization	862
ICU	214
Dead	146

clarity issues are essential as they help future researchers whether to adopt or avert these issues for similar problems. Dugas et al. [25] indicated that some ED data used in their study might not be based on the same triage system evaluations included in the analysis. This selection of various data sources results in systematic differences between different EDs, where the data collected might have been based on further clinical evaluations of other triage systems or liable to human bias. This can impact predictor outcomes (e.g., accuracy), leading to a significant difference in the statistical performance evaluation. Most studies tried to predict the triage nurses' actual decisions. These triage labels may not be the most accurate or liable to human error; therefore, researchers like Goto et al. [24] used patient outcomes instead of triage class to train the model. Levin et al. [16] could not attain good recall (sensitivity) for high-severity levels, which means that an algorithm returns more irrelevant results than relevant ones, which is required for an effective triage practice. None of the reviewed studies considered fixing the class imbalance within their datasets, except Wolff et al. [13]; hence, they have not applied any correction strategy to enhance the model building. Class imbalance is considered an issue as most ML algorithms presume that data is fairly distributed. When there is a class imbalance in the training dataset, the ML classifier tends to be more biased towards the majority classes, causing inaccurate classification of the minority class. Most studies failed to clarify how they developed their models and tunned model hyperparameters in the manner of best methodological practices for applying ML techniques to similar problems. Wolff et al. [13] followed a manual selection for model hyperparameter tuning. However, these parameters might not be the best to apply to minimize the total error and obtain optimal performance. Moreover, applying random or manual hyperparameter tunning takes time away from essential steps of the ML pipeline, like feature engineering and interpreting results. This study will address the reported issues in the aforementioned studies by implementing the best methodological practices to achieve the best possible outcomes.

3. Methods

3.1. Study design, settings, and data source

A comparative study of selected supervised ML models was conducted to determine the best-performing approach to evaluate triage outcomes in hospital ED settings. The collected dataset in this study was extracted from a single-center ED of a university hospital in Istanbul with an annual visit of 200,000 patients. A retrospective dataset of 4540 patients admitted to the ED between April 1, 2020 and June 30, 2020 has been collected. No personally identifiable data is part of the dataset. Data processing and analysis were done using Python 3.9 programming language and some related scientific libraries, including Pandas, NumPy, and scikit-learn [30,31]. Ethical approvals have been granted from all relevant parties.

3.2. Data analysis approach

This study followed the experimental setup shown in Figure (1) to construct the triage models. We evaluated nine ML models ranging from linear models (e.g., logistic regression) to non-linear models (e.g., ensemble learning). We guided the learning process of the triage model by patient outcomes instead of the actual triage labels as we believe that actual triage labels might be liable to human bias and data entry errors.

3.3. Data preparation

Data preparation is an essential step in the ML pipeline and is one of

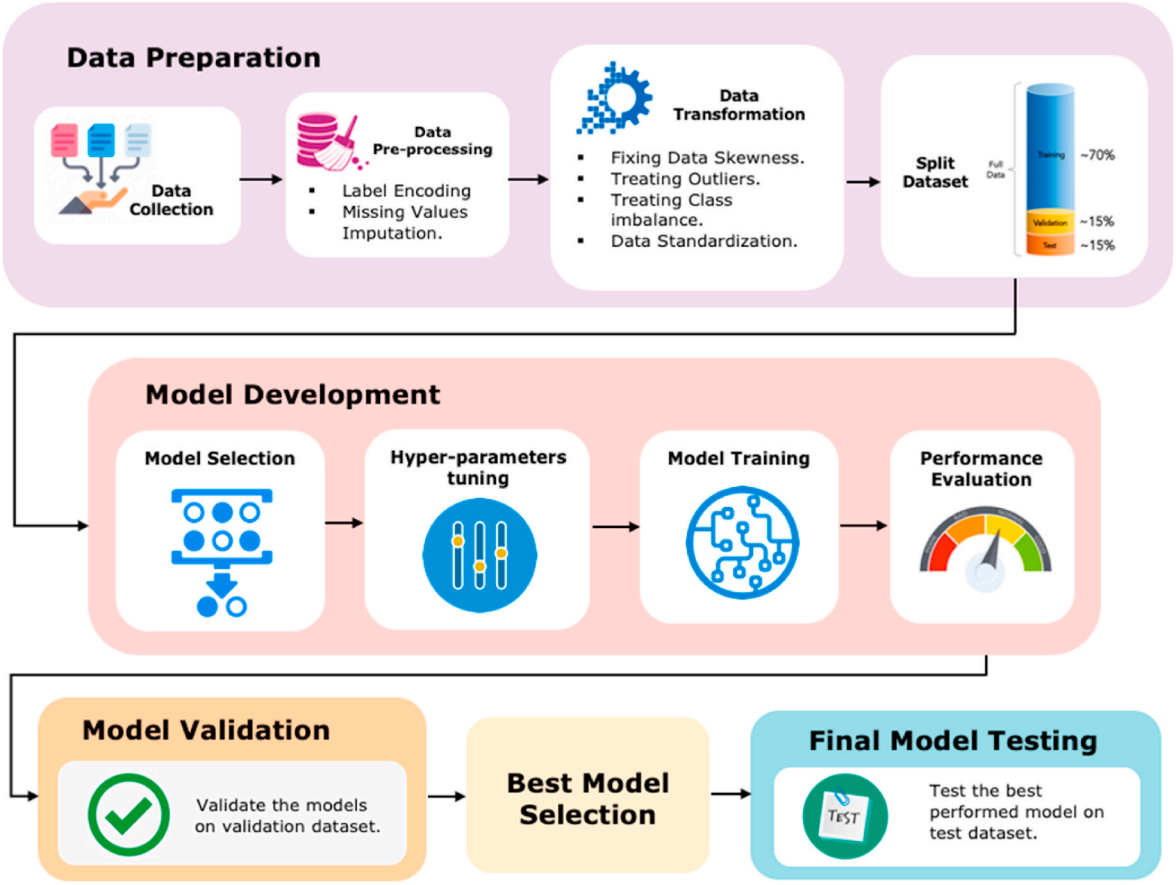

Fig. 1. Study experimental setup.

the most time-consuming tasks in building ML models [32]. After collecting the data, we carried out the task of manual inspection of the data. We performed data translation to some features within the dataset from Turkish into English. Due to the small data sample size, all syntax errors were manually checked and corrected. Data pre-processing, including label encoding and missing data imputation, was conducted. Labels of categorical features were encoded into a numeric form to convert them into machine-readable form. Missing values were imputed using K-NN imputer. K-NN imputer is an approach designed to search for k-nearest neighbors for a missing datum from all complete instances in a given dataset and then fill in the missing datum with the most frequent one occurring in the neighbors in the target feature [33].

3.4. Variables

The data used in this study was extracted from the patient information system during their visit to the ED at the time of triage. The clinical variables (features) investigated in this study are shown in Tables (2) and (3).

Data include patient demographics (age and gender), vital signs (respiratory rate, body temperature, oxygen saturation, blood pressure and pulse rate), chief complaints (CCs), and chronic illness. Since we aim to create a simple dataset with less complexity during the model training stage, the chronic illness variable was analyzed based on whether the patient has past or current chronic illness instead of creating multiple columns for each illness. Patient outcomes evaluation was identified based on Manchester Triage System (MTS) as the current triage system used within the hospital. Most of the CCs identified during data analysis were limited to patient complaints related to patients with Coronavirus (COVID-19) symptoms, including (cough, chest pain, fever, etc.). The reason behind that is there was a national lockdown during the data collection stage. Most patients were shielded and avoided visiting hospitals due to the fear of contracting the virus. The target variable was patient disposition outcome during their visit to the ED, whether they were discharged, hospitalized (admitted into hospital wards), hospitalized to isolated wards for COVID-19 suspicion, transferred to the Intensive care unit (ICU), or died.

3.5. Exploratory data analysis

We have conducted an exploratory data analysis (EDA) to identify the relationship between variables and understand how each variable contributes to the learning process. EDA helps to identify which variable contains the most relevant information that can provide accurate predictions [34]. For this purpose, we analyzed the correlation between each feature before elaborating on the predictive model. We used the Pearson correlation coefficient to determine which features have the most significant impact on others. The Pearson correlation coefficient measures the linear dependence between two random variables (real--valued vectors) [34]. We used Python Matplotlib and Seaborn data visualization libraries to plot the matrix.

3.6. Data transformation

After completing the data pre-processing task, we applied multiple transformation techniques to our dataset. We started by inspecting the skewness in each feature to improve the normality of the data. Fixing the skewness in data is a vital task that must be carried out before building the model. The tail region of the data distribution in skewed data may act as an outlier for the statistical model that can affect the training process. We applied the Box-cox transformation technique for this task to transform the skewed data into normal distribution [35]. We then inspected the outliers within the dataset. We used Interquartile Range (IQR) method to eliminate the outliers from the dataset, which are the patterns that are not in the range of normal behavior [36].

Our dataset contains features that are highly varying in magnitudes, units, and ranges. Since most ML algorithms use *Euclidean Distance* between two data points in their computations, it can generate wrong computations and, therefore, affect the model evaluation. To solve this problem, we applied standardization as a feature scaling technique. Standardization is the transformation of features by subtracting from the mean and dividing by standard deviation. Standardization rescales a dataset with a mean of zero and a standard deviation of one. This is often called Z-score normalization [37]. We also merged the features of chief complaints into one feature to reduce the training time and the model's overfitting. We applied a hold-out scheme for splitting the dataset. The data was split into training, validation, and testing samples (70%–3178 samples for model training, 15%–681 samples for model validation, and 15%–681 samples for testing the best-achieved model on new data to see how robust its performance and detect overfitting within the final model).

In the final data preparation stage, a significant challenge encountered in designing ML triage models is the high-class imbalance of the collected data. As indicated in Table 3, the sample of patients who were discharged from ED (2278–50.37%) is much higher than the sample of patients who were hospitalized in ICU (214–4.71%) or died in the hospital (146–3.21%). Many methods are applied to handle the class imbalance in ML to achieve robust models. This study applied Synthetic Minority Over-sampling Technique (SMOTE) to handle the imbalance between classes. It is an approach in which the minority class is over-sampled by creating *"synthetic"* examples rather than over-sampling with replacement, which creates extra training data samples by performing certain operations on actual data [38]. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors, depending on the amount of over-sampling required, where neighbors from the k nearest neighbors are randomly chosen [38]. After applying SMOTE technique to the training dataset, the final sample size included 7950 data samples, with 1590 samples for each target class.

3.7. Machine learning models

This study applied a traditional (supervised) ML algorithm to train several models on a real dataset collected during hospital ED triage. Various algorithms were initially selected, including Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Multi-layer Perception Neural Network (MLP), and selected ensemble learning techniques, including Gradient Boosting Decision Trees (GBDT), XGradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), and Random Forests (RF). These algorithms were initially selected as they show higher predictive ability in triage classification problems within the literature [13,16,22–25,27]. LR has been selected as it is a widely used classification modeling technique in clinical research primarily suitable for predicting a binary dependent variable and handling linear problems [39]. DT identifies ways to split a dataset based on different conditions [40]. This technique has a hierarchical structure like that in the tree, where each node has exactly two outgoing edges. DT learning algorithm automatically learns an optimal decision tree structure given a dataset. KNN is a conventional nonparametric classifier that provides good performance for optimal values of k, where examples are classified based on the class of their nearest neighbors [41]. Support Vector Machine (SVM) is a modern learning system that delivers state-of-the-art performance in real-world data mining applications. It makes some highly complex data transformations and then generates the results by maximizing the distance between the defined labels within the dataset, usually called the optimal boundaries [42]. The Multilayer Perceptron (MLP) is a type of artificial neural network that mimics the human brain's functioning. MLP consists of a system of simple interconnected neurons, or nodes, which is a model representing a nonlinear mapping between an input and output vector [43]. It can have multiple hidden layers between the input and output

Table 3
Analysis of patient triage data in regard to disposition outcomes.

General (Input)	Discharged	Hospitalization	COVID-19 Services	ICU	Died
ED visits - number	2278 [50.37%]	862 [18.98%]	1040 [22.90%]	214 [4.71%]	146 [3.21%]
Male	1231	675	683	114	100
Female	1047	187	357	100	46
Age (Year) – (number)					
17–25	409	1	11	0	0
26–35	530	38	61	0	0
36–45	455	103	90	0	0
46–55	459	175	176	0	0
56–65	230	200	244	53	0
66–85	176	258	380	110	87
≥86	19	87	78	51	59
Vital Signs (Average)					
Temperature (∘C) NR (36.5 °C and 37.3 °C)	36.6	37.1	37.1	37.7	38.5
Heart rate (Pulse/min) NR (60–100)	88	95	91	96	146
Respiratory Rate (Breath/min) NR (12–18)	18	24	23	35	25
Oxygen Saturation (%) NR (95%–100%)	97.21	89.96	93.98	86.93	90.54
Systolic BP (mmHg) NR (90–120)	128.89	128.59	129.12	117.68	157.10
Diastolic BP (mmHg) NR (60–80)	76.73	74.71	74.61	72.04	86.76

layers.

Ensemble learning has proven spectacular capacities to enhance the prediction accuracy of base learning algorithms [44]. Ensemble learning is a supervised ML technique where multiple weak (base) models are trained on data and then combined to create a single strong model [45]. A weak model is referred to as one that is not strongly correlated with the target variable. It is only slightly better than random guessing; however, a strong model is referred to as one that is very strongly correlated with the true outcome or the target [45]. Adaptive Boosting (AdaBoost) is an example of ensemble learning where a sequence of classifiers is trained. In AdaBoost, each training sample is assigned a weight, and a higher weight is set for a training sample that has not been trained by the previous classifier [46]. Gradient boosting is another boosting ensemble learning that their base model is usually decision trees (DT); each DT base model learns *sequentially* rather than in parallel because often, each weak learner depends on the previous one. XGBoost is an improved model of GBDT that trains many decision trees in a sequential and additive manner [47]. XGBoost makes GBDT's training scalable by distributed learning with several approximation methods to parallelize training [27]. Since XGBoost was proposed, it has outperformed other algorithms in many studies and competitions [27]. Random forest is a trapping ensemble learning technique. Each weak model in RF is learned *independently* and often in parallel and then uses the high probability value to generate a classification decision [48]. Although both GBDT and RF are based on decision trees, they train trees differently. In GBDT, the model trains small trees, which may underfit the training set. In RF, each tree is trained independently and becomes a so-called fully-grown tree, which overfits a subset of the training dataset. However, trees in GBDT are trained sequentially, and each tree is trained in consideration of the errors of other tree models in the previous steps. In this way, as more trees are trained, the error of GBDT is reduced.

3.8. Model selection and development

The selection of ML models mainly depends on those algorithms that initially achieved better prediction performance. Multiple algorithms were preselected to train the model before tuning the hyperparameters to select the best-performed algorithms for the primary model selection. Nine models were chosen for the principal analysis used in this study, including (LR, DT, KNN, SVM, MLP, GBDT, XGBoost, AdaBoost, and RF). After selecting the best-performed models, it is essential to tune the models' hyperparameters to achieve the optimal performance of each trained model. Hyperparameters are a set of configurations that are external to the data and manually tuned by a researcher [49]. Finding a set of hyperparameters that delivers optimal performance in a reasonable time is part of the hyperparameter optimization problem that needs to be done before the final model evaluation. It is a task that is usually done to reduce the total error of the function by finding the trade-off between bias-variance errors. However, testing all the possible sets of hyperparameters can be computationally expensive [50]. For example, random and grid searches are time-consuming techniques as they require a long run time evaluating unpromising areas of the search space. That led researchers to seek an automated and structured approach for model hyperparameters optimization. For this purpose, we chose an automated approach to optimize the models using Genetic Algorithms (GA). Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on biologically inspired operators such as mutation, crossover, and selection [51]. GA can be applied even when there is no information about the gradient function at evaluated points. At the same time, it can still achieve good results when there are several local minima or maxima [52]. GA has advantages over local methods as they do not remain trapped in suboptimal local maximum or minimum [52]. We applied GA using the Tree-based Pipeline Optimization Tool (TPOT) auto-machine learning library in Python. A set of hyperparameters are initially selected for each model to find the hyperparameter that can lead to optimal performance for each model, as indicated in Table 4. To achieve more reliable results and avoid models' overfitting, cross-validation was applied during model training. Since our dataset is relatively small and some ML models (e.g., GBDT and SVM) require a long time to train, we have selected 5-fold cross-validation to ensure a good trade-off of low bias and low computational cost in an estimate of model performance.

Performance evaluation metrics in multi-class classification

Table 4
The set of hyperparameters for each selected model.

Model	Hyperparameter Selected for Training
KNN	- Leaf size: range (1–50). - Number of neighbors: range (1–30). - The parameter metric p: [1,2].
GBDT	- Number of estimators: [50, 250, 500]. - Maximum depth of the tree: [3,5,7,9]. - Learning rate: [0.1, 1, 10, 100].
XGBoost	- Number of estimators: [5, 50, 250, 500]. - Maximum depth of the tree: [1,3,5,7,9]. - Learning rate: [0.01, 0.1, 1, 10, 100].
RF	- Criterion: Gini, Entropy. - Maximum depth of the tree: [1,3,5,7,9]. - Maximum features: range (5, 13). - Minimum Sample Leaf: range (2, 12). - Minimum Samples Split: range (2,10). - Number of estimators: [5, 50, 250, 500].

problems are fundamental in evaluating the quality of learning methods and comparing the performance of different models. It is critical to select the right evaluation metrics to report the performance of predictive models, especially in the context of the healthcare evaluation models [13]. Since triage is a high-class imbalanced problem, reporting only one evaluation metric, such as accuracy or precession, will not be enough to evaluate model performance, such as Levin et al. [16] evaluation. However, there are no gold-standard performance metrics that can be applied to all types of multi-class problems. Since this study is a multi-class classification problem, only a limited set of performance metrics is available (typically accuracy, recall, precision, and F1-score) [53]. Fawcett [54] also indicated that Receiver operating characteristics (ROC) could be applied to evaluate multiclass problems; however, the complexity of analysis increases with the number of classes. Therefore, this study will only consider the predetermined evaluation metrics to evaluate the performance of the selected models. We first generated the confusion matrix of each model, which is a machine learning concept that provides information about predicted and actual classification values generated by a prediction system [55]. It consists of two dimensions, the first is the actual classes of the data, and the second is the predicted classes of the dataset. The evaluation metrics defined for binary classification do not apply to the full extent in multi-class classification problems due to the difference in the matrix dimension between both matrices [53]. Instead, the analysis of the confusion matrix in multi-class classification focuses on a specific class based on the characterization provided in Figure (2).

A number of performance evaluation measures can be defined based on the confusion matrix as indicated in Table 5.

We considered selecting the micro weighted average of each Precision, Recall, and F1 score for each class as a final evaluation.

4. Results

4.1. Features analysis

After visualizing the correlation matrix in Figure (3), we observed some correlation between features as follow.

- Patient age has a high impact on the respiratory rate, as confirmed by the Pearson R-value (0.47). This positive correlation indicates that older people have a higher respiratory rate than young people due to the decrease in the oxygen saturation in their blood.
- Oxygen saturation has a high impact on respiratory rate and age, as confirmed by Pearson R-values (−0.58) and (−0.38), respectively. The negative correlation value indicates that the higher the patient's age, the lower the oxygen saturation will get; thus, the need for more oxygen in the body will increase, which is substituted by increasing the respiration rate.

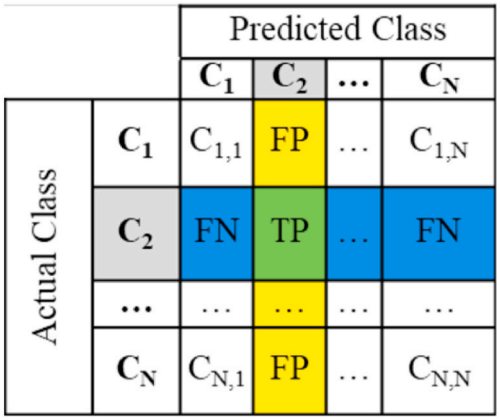

Fig. 2. A confusion matrix for multi-classification problem [53].

Table 5
Performance Evaluation Metrics used in this study.

Formula	Performance Measure	Description	Formula
(1)	Accuracy	Evaluates the degree of correct predictions of a classification model.	Accuracy $= \sum_{i=1}^{n} TP(C_i) / \sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij}$
(2)	Precision of class C_i	Evaluates the ratio of the correct predictions of a specific class to the total predicted for the same class.	$Precision_i = \frac{TP(C_i)}{TP(C_i) + FP(C_i)}$
(3)	Recall (Sensitivity) of class C_i	Evaluates the ratio of correct predictions of a specific class to the total number of the actual class.	$Recall_i = \frac{TP(C_i)}{TP(C_i) + FN(C_i)}$
(4)	F1 score of class C_i	Evaluates the harmonic mean of precision and recall.	F1 score $= \frac{2 * Precision\ (C_i) * Recall\ (C_i)}{Precision\ (C_i) + Recall\ (C_i)}$

- Systolic and diastolic blood pressure have a high correlation with a Pearson R-value of (0.60). It is expected to have a linear correlation since both features are extracted from the same feature (blood pressure).
- Age and chronic illness have a higher correlation, confirmed by the Pearson R-value (0.43). It concludes that elderly patients are highly possible to encounter chronic illness during their lifetime.

Among the selected variables, oxygen saturation, age, and respiratory rate are shown to have a higher correlation with the other variables, either positive or negative, delivering relevant information that helps generate accurate predictions. It is worth mentioning that the Pearson correlation matrix investigates the correlation between two features, not the causation. Determining correlation is much simpler than determining causation since causation may require independent experimentation.

4.2. Model outcomes and comparison

We tested several ML models with various parameter configurations for triage patient outcomes prediction on a class-balanced dataset, as shown in Table 4. Nine models were trained and tested on the generated dataset with a range of prediction accuracy (66.1%–89.1%, 95% CI), as shown in Table 6. Models training was repeated five times to examine the consistency of our results. Findings show that out of nine models, four models (KNN, GBDT, XGBoost, and RF) achieved better performance than the rest of the models. RF and KNN models outperformed the other models with the best performance metrics of prediction accuracy (89.1%, 95%CI = 88.8 to 89.5; 88.7%, 95%CI = 88.5 to 88.9), micro average precision (89.0%, 95%CI = 88.9 to 89.1; 88.7%, 95%CI = 88.4 to 89.0), average micro recall (89.1%, 95% CI = 89.0 to 89.3; 88.7%, 95%CI = 88.4 to 89.0) and average micro F1-score (89.0%, 95% CI = 88.9 to 89.1; 88.6%, 95% CI = 88.4 to 88.9) respectively. XGBoost model achieved the third-highest results in terms of accuracy (88.4%, 95% CI = 88.2 to 88.6). On the other hand, the AdaBoost technique seems to be the lowest performing model among the other models, with a prediction accuracy of (66.1%, 95% CI = 65.9 to 66.3) to be excluded from the comparison and model selection.

Multiple hyperparameters were initially selected for model tuning. Table 7 provides the hyperparameters that deliver the optimal prediction performance for the best four performed models.

Models were also evaluated regarding their precision, recall, and F1

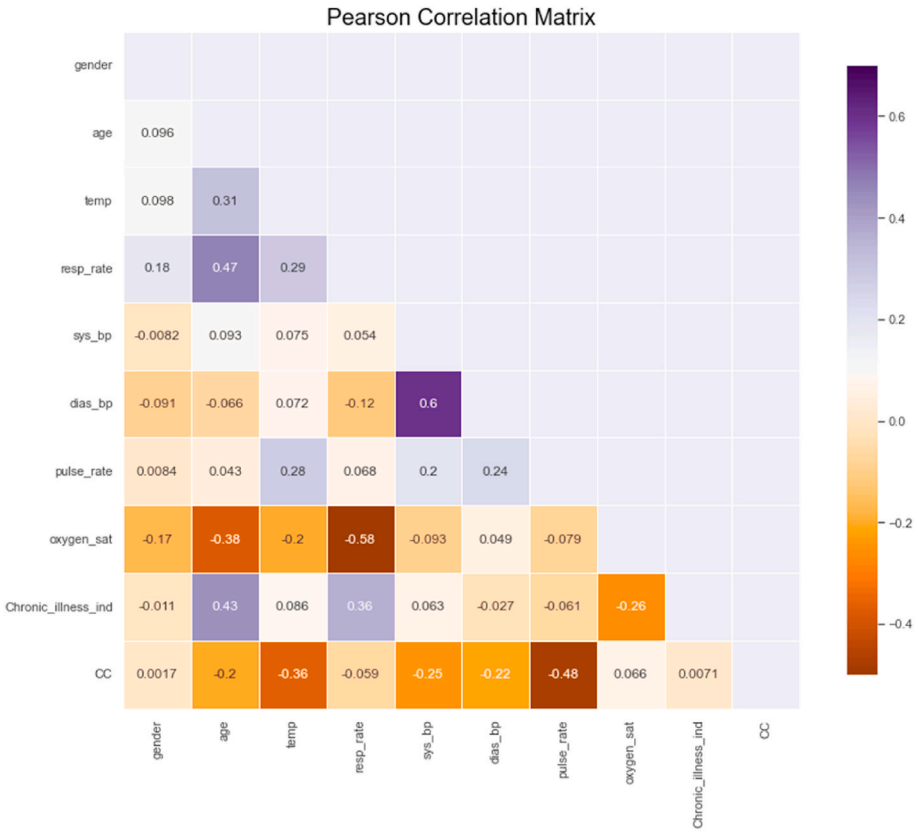

Fig. 3. Pearson correlation matrix.

Table 6
Training and validation performance of all models with 95% confidence intervals.

Model	Training Accuracy %	Validation Accuracy %	Precision %	Recall %	F1-score %
LR	80.5 (79.9, 80.8)	72.4 (71.3, 73.1)	71.4 (71.1, 71.6)	72.4 (71.9, 72.8)	71.6 (70.9, 72.2)
DT	98.0 (97.5, 98.5)	79.7 (78.8, 80.7)	79.1 (78.5, 79.6)	79.7 (77.6, 81.2)	79.1 (78.4, 79.6)
KNN	97.1 (96.3, 97.8)	88.7 (88.5,88.9)	88.7 (88.4, 89.0)	88.7 (88.4, 89.0)	88.6 (88.4, 88.9)
SVM	91.4 (90.9, 91.9)	82.1 (81.7, 82.5)	81.7 (81.4, 82.0)	82.1 (81.7, 82.5)	81.7 (81.4, 82.0)
MLP	97.4 (97.0, 97.8)	86.6 (86.2, 86.9)	86.6 (84.8, 88.3)	86.6 (86.0, 87.1)	86.6 (86.0, 87.1)
GBDT	99.9 (99.8, 100)	87.8 (87.5, 88.0)	87.7 (87.4, 88.0)	87.8 (87.4, 88.2)	87.8 (87.4, 88.2)
XGBoost	99.9 (99.8, 100)	88.4 (88.2, 88.6)	88.4 (88.2, 88.6)	88.4 (88.1, 88.7)	88.4 (88.2, 88.6)
AdaBoost	68.3 (67.6, 70.0)	66.1 (65.9, 66.3)	82.3 (82.0, 82.6)	66.1 (65.7, 66.7)	73.2 (72.3, 74.1)
RF	99.7 (99.5, 99.9)	89.1 (88.8, 89.5)	89.0 (88.9, 89.1)	89.1 (89.0, 89.3)	89.0 (88.9, 89.1)

score. Figure (4) presents the confusion matrix of the four best-performed models in this study. As a multiclass classification problem, it is essential to understand the model performance regarding each class

Table 7
Models' hyperparameters that delivers the optimal prediction performance.

Model	Hyperparameters with the Highest Performance
KNN	- Leaf size: 18. - Number of neighbors: 4. - The parameter metric p: 2.
GBDT	- Number of estimators: 250. - Maximum depth of the tree: 9. - Learning rate: 0.1.
XGBoost	- Number of estimators: 500. - Maximum depth of the tree: 9. - Learning rate: 0.1.
RF	- Criterion: entropy. - Maximum depth of the tree: None. - Maximum features: 10. - Minimum Sample Leaf: range 2. - Minimum Samples Split: 4. - Number of estimators: 50

within the dataset. Each class in the confusion matrix in Figure (4) is represented numerically and described as follows (0: Discharge, 1: Hospitalization to COVID Service, 2: Hospitalization, 3: Intensive Care Unit (ICU), 4: Died). When evaluating the process of predicting triage within ED settings, it is essential to consider the performance of each model for each class. For example, patients with high acuity conditions need higher prioritizations and fast, accurate and precise prediction of their conditions; any misclassification within these categories may cause delays that result in severe injuries and mortality, which come up against the actual purpose of triage. Therefore, we must tune our models for higher precision and recall for labels 3 and 4. Results indicated that the two best-selected models show increased sensitivity in predicting high acuity outcomes (ICU and dead patients), as indicated in Table 8. Furthermore, RF and KNN models not only achieve the most heightened sensitivity in predicting high acuity but also achieves the best prediction

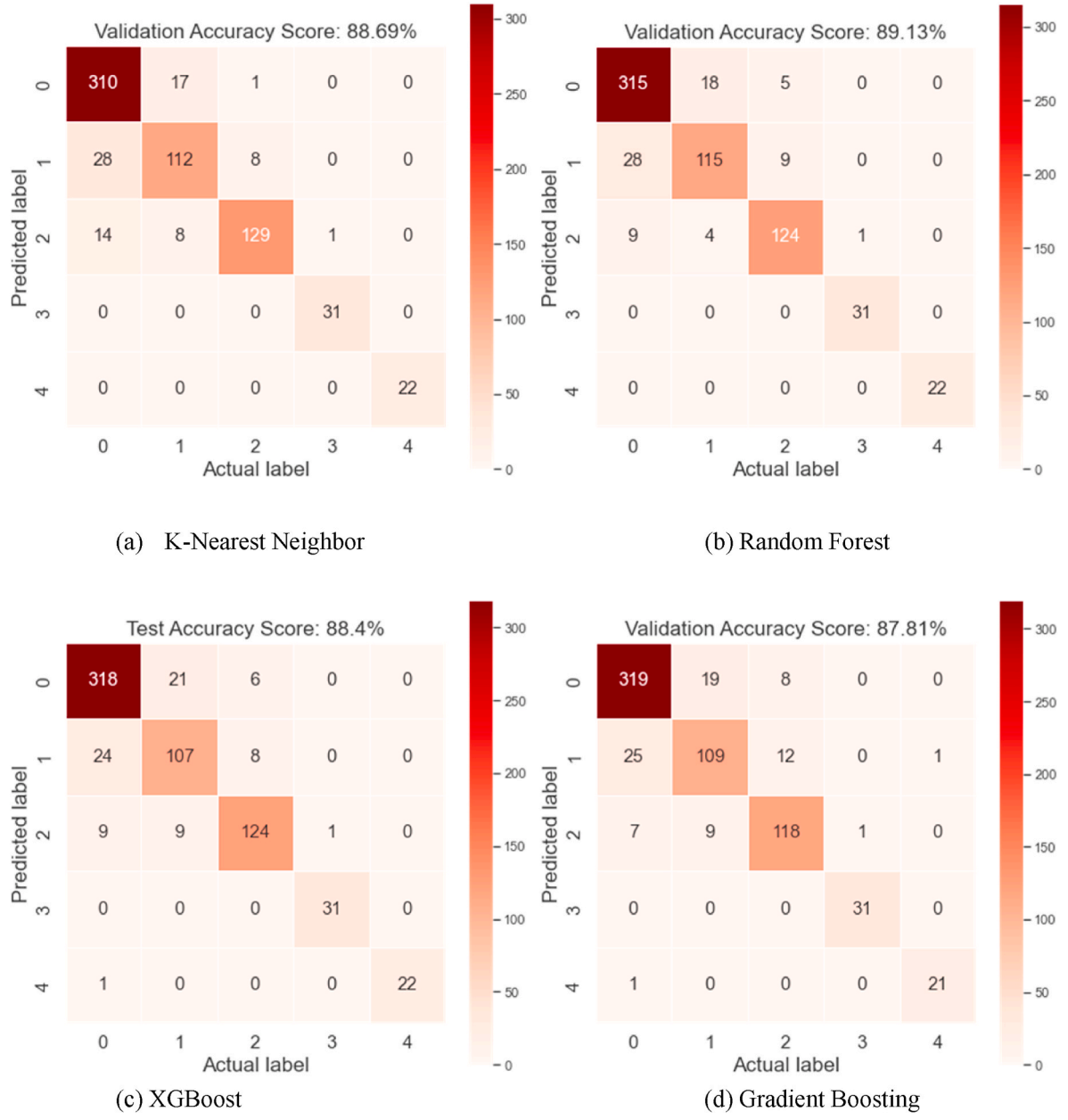

Fig. 4. The confusion matrix of the models with the highest prediction ability.

Table 8
The class wise precision and recall for the best two performed models with 95% Confidence Intervals.

Model	Class	Precision	Recall
RF	0	0.895 ± 0.009	0.932 ± 0.009
	1	0.839 ± 0.012	0.757 ± 0.011
	2	0.899 ± 0.003	0.899 ± 0.008
	3	0.969 ± 0.001	1.000 ± 0.004
	4	1.000 ± 0.001	1.000 ± 0.001
KNN	0	0.881 ± 0.012	0.945 ± 0.011
	1	0.818 ± 0.009	0.757 ± 0.009
	2	0.935 ± 0.004	0.849 ± 0.003
	3	0.969 ± 0.001	1.000 ± 0.002
	4	1.000 ± 0.002	1.000 ± 0.003

of low acuity outcomes (Discharge and hospitalization than other models, making them the best-performed models.

After selecting the best two-performed models, we tested the model using the testing sample to verify both models' performance, as indicated in Table 9. Both models also have almost similar performances, with slightly better performance of the RF model of 88.9% accuracy, 88.9% precision, and 88.9% recall over the KNN model with 87.8% accuracy, 87.8% precision, and 87.8% recall, with huge advantage of RF over KNN in latency (prediction time).

Table 9
Testing performance of the best two performed models with 95% confidence intervals.

Model	Test Accuracy %	Test Precision %	Test Recall %	Test Latency (ms)
KNN	87.8 (86.9, 88.7)	87.8 (87.4, 88.2)	87.8 (87.4, 88.2)	442.9
RF	88.9 (87.5, 89.3)	88.9 (87.4, 89.4)	88.9 (87.4, 89.4)	44.4

4.3. Training and prediction time

Results in Table 10 show a considerable difference in the training time between models (48 s–3.5 h). The Decision Tree model required the least time (48 s) to fit the model, while on the other hand, Gradient boosting required the longest time (3hrs, 26mins, 12sec) to fit the model. Since the XGBoost model is an improved GBDT algorithm, it is noticeably that XGBoost performed better than GBDT and took much less time (17min, 25sec) to fit. Regarding the best two performed models on the validation dataset, although the KNN model takes less time to fit compared to the RF model, it took nearly 26-times longer to predict outcomes. After testing both models on the test dataset, similar outcomes related to prediction time were acquired with a significant advantage of the RF over the KNN model, with almost 10-times faster to predict. It is necessary to emphasize that the faster the decision of triage evaluations for patients with high acuate conditions, the most rapid patients will receive care and save lives. Hence, the RF model seems to be the best-performed model, with better performance in accuracy and latency than the other models.

4.4. Feature importance

This study investigated feature importance correlation and their contribution toward the model training using Random Forest model. Figure (5) presents a descending order of the feature importance and their contribution towards the model building. Respiratory rate, age, and oxygen saturation appear to dominate the significant contribution toward triage outcomes prediction. It is noticeably that these outcomes came in line with the earlier analysis of feature correlation using the Person correlation matrix since these three features show a high correlation with each other and the other variables. Oxygen saturation and age also seemed to significantly contribute to predicting triage class within the literature [16,22]. Despite the fact that CCs have a considerable role in determining the actual triage implementation, our analysis shows less contribution of chief complaints toward the model analysis. This can be attributed to the scarcity of complaints among the data sample used in this study. Other features such as chronic illness indicators and gender did not show a significant contribution to the model classification, indicating that they had less correlation with patient outcomes. The findings emphasize the need for triage nurses to give more attention when examining these demographics and vitals during their triage screening.

For that purpose, we applied Principal Component Analysis (PCA) as a dimensionality reduction technique to eliminate the number of features that do not add much value to the learning process. Seven features were selected to train the model as they were accountable for 93% of the information within the dataset, as shown in Figure (6). The analysis findings show that the average accuracy of the selected models was dropped by 0.9%, with an improvement in the training time for each model.

Table 10
Training and Prediction Time of each selected models.

Model	Training Time (Minutes)	Prediction Time (m seconds)
LR	1.55	4.6
DT	0.48	2.8
KNN	4.20	626.4
SVM	10.16	1061.5
MLP	42.00	19.8
GBDT	206.75	89.2
XGBoost	17.52	23.8
AdaBoost	18.60	152.3
RF	5.75	23.7

5. Discussion

This study provides a comparative approach between a selected supervised machine learning techniques in the case of predicting triage outcomes within hospital ED. The study presented the best methodological practices for the application of machine learning in predicting patient triage in hospital EDs settings with in-depth clarification and explanation of each step taken within this study. We developed multiple ML-based triage models on patient data collected at the time of ED triage. We evaluated multiple triage models and compared their performance using standard evaluation metrics. The novelty of this study was not in offering a new machine-learning technique to build the model; instead, this study contributed to tackling some of the shortcomings reported in the literature and revealed some important points about the choice of the models as follows: (1) This study applied an automated hyperparameters optimization using genetic algorithm to reduce model bias-variance errors and obtain the optimal performance. To the best of the authors' knowledge, none of the studies in Table 1 has revealed the use of automated hyperparameter tuning to optimize the classification models except [13] with the use of manual parameters inspection. Hence, this study highlighted the significance of using automated hyperparameter tuning techniques in obtaining the best parameters that provide optimal prediction performance. The use of these techniques can be reflected in the performance of each model, as shown in Table 5. Table 2 We applied SMOTE techniques to handle the high-class imbalance within the dataset, which none of the work referred in Table 1 took into consideration the big class imbalance within the dataset except [28] with little description provided, thus they did not apply any correction strategy to improve model construction. It is known that imbalanced datasets degrade the performance of data mining and machine learning techniques as the overall accuracy and decision-making be biased toward the majority class, which leads to misclassifying the minority class samples or, furthermore, treating them as noise [56]. The use of SMOTE in this study has helped to reduce the high bias found within the dataset and rebalanced the internal distributions among the classes with an equal distribution of each class. We have trained the model on both data (class imbalanced data & class balanced data), where results show an average improvement in model prediction accuracy of RF about 0.8% (89.1 after SMOTE vs 88.3 before SMOTE) & KNN of 2.8% (88.7 after SMOTE vs 85.9 before SMOTE). Some models such as GBDT show a significant improvement after applying SMOTE with an improvement of 9.39% (87.8 after SMOTE vs 78.41 before SMOTE). Prior research also shows that models built via SMOTE are shown to attain higher prediction performance (in terms of recall and F1-score) than other over-sampling techniques [57]. (3) We found a substantial lack of clarity in the model development process in the literature, with little reporting of the techniques used to process the data and train the models. This study has provided a clear experimental setup with a clear guidance on data preparation, model selection, and model development and evaluation, which what many prior studies failed to clarify. (4) This study highlighted the power of ensemble learning techniques (XGBoost, GBDT, and RF) when used to predict the model performance on noisy biomedical data, which seem to achieve outstanding performance comparing to the other techniques in this study, with a prediction accuracy of (88.4%, 87.8%, and 89.1%) respectively. Making them as favorable choice for similar problems, since RF and XGBoost has been shown to be widely used in sub-signal recognition contests due to its high prediction accuracy and outstanding efficiency [58]. (5) In regard to comparing our model performance to those in the existing literature. Many studies failed to show the class wise outcomes (e.g., recall/sensitivity and precision) as it is very essential for the triage model to provide an accurate and precise evaluation of patients who need urgent attention than patients who are in stable health conditions and are able to wait, where the error in classifying patients with critical conditions can result in health complications and death. Our models show outstanding performance in

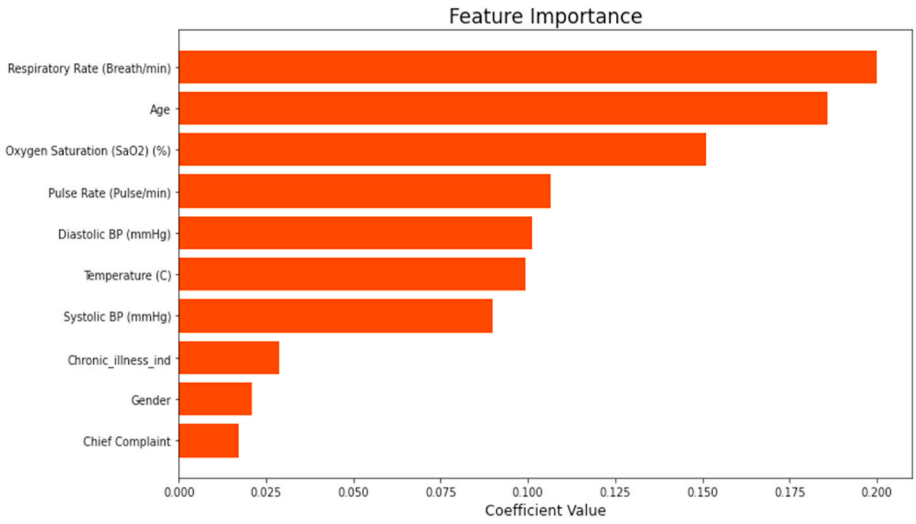

Fig. 5. Feature importance toward model training.

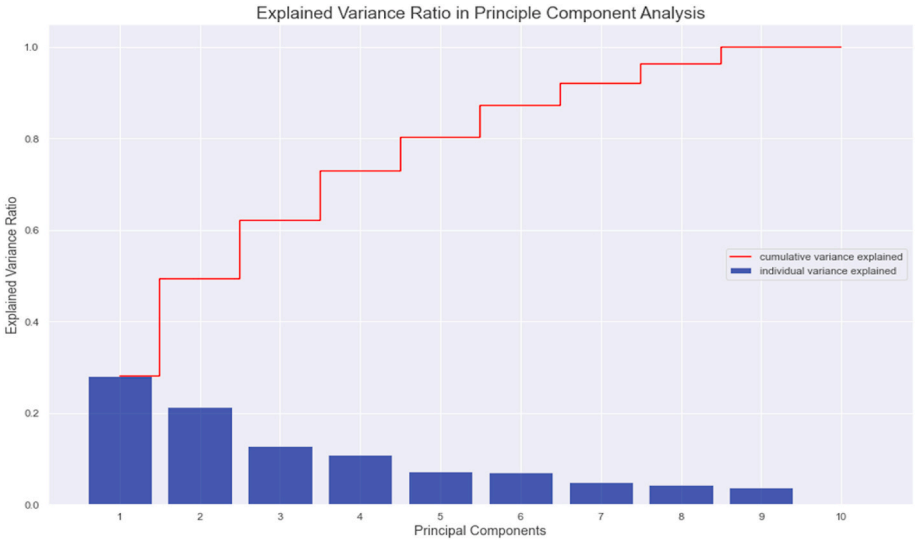

Fig. 6. Explained variance ratio in principle component analysis.

predicting patient critical outcomes (i.e., ICU or Died) as presented in Table 8 comparing to the work in Refs. [13,22].

With the increased level of ED overcrowding, waiting times and LOS, the shortcomings of the current triage practice and the massive availability of healthcare data emphasize the need for more robust and intelligent triage evaluation in emergency settings. Literature shows that machine-learning techniques enables the automation of processes that can help analyze complex data to extract a meaningful relationship from datasets to assist in prediction, such as triage evaluations [59]. They can identify patient symptoms, access patients' medical results and history, and predict patient needs based on the available data through EHRs [60]. It can also capture complex interactions that are likely to be present when predicting less specific outcomes [21]. ML techniques have shown promise to enhance predictive triage abilities in many conditions (e.g., Congestive Heart Failure, sepsis) [23] and better analyze patient conditions across a wide range of medical conditions and illness severity relevant to triage [16]. The patterns, recognized by ML, can assist physicians' ability in decision-making, personalize the care tailored to individual patient needs and reduce dependence on subjective human judgment and relative experience [61]. ML-based triage models have proven superior abilities in predicting critical-care outcomes and hospitalization over the reference standard triage models [22]. These models provided accurate patient distribution through the five severity levels and enhanced triage nurses' decision-making [61]. These superior predictive abilities would contribute to improve patient pathways, better manage hospital ED resources, reduce costs for both the hospital and patients and reduce patients' waiting times and LOS. It, therefore, tackles overcrowding, provides better healthcare services for patients, and reduces morbidity and mortality rates. Our developed ML triage models show a high capability in predicting patient disposition outcomes in ED settings for both high acuity and low acuity patients, indicating that machine learning techniques demonstrate high promise in improving predictive abilities in emergency medicine, especially in triage evaluation in ED settings. The implementation of this developed model within the hospital settings, therefore, will translate into a better Manchester Triage evaluation tool that will assist triage nurses in their judgement and provide faster and more accurate evaluation for many complicated cases.

Although this study achieves outstanding prediction outcomes, there is still room for improvement. First, the analysis in this study was based solely on a retrospective data sample collected from patients during their visit to the ED. However, these retrospective data might be liable to

human bias or data input error. Moreover, the models were validated on the same retrospective data. Future research will consider evaluating the model on prospective data. The data sample used for the model development was relatively small. Finally, this research was conducted on a small data sample compared to the one in the literature; however, the main aim of this research was mainly to compare different ML approaches rather than to implement the model within a hospital environment. Future research will consider collecting multi-sources large data samples for model development and validation to enhance the generalization and the performance of the model in a broader sample of data.

6. Conclusion

Machine learning techniques have proven good ability in supporting triage decision-making within hospital ED settings. These models demonstrated a considerable performance in differentiation between patients with critical outcomes (Mortality and ICU admission) from patients with less critical outcomes (e.g., discharged and hospitalized) in ED settings. Implementing these models can enhance patient outcomes, reduce waiting times and length of stay, manage hospital resources more effectively and thus reduce overcrowding. Moreover, the outcomes of this study will benefit future research whether to adapt or ignore certain criteria when designing triage prediction models and guide them toward the best methodological practices for such a problem.

Credit author statement

Hamza Elhaj: Methodology, Data curation, Coding, Writing – original draft preparation, Visualization, Writing- Reviewing and Editing. Nebil Achour: Supervision and Revision. **Marzia Hoque Tania**: Supervision and Revision. **Kurtulus Aciksari:** Data Collection

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

References

[1] Zlotnik A, Alfaro MC, Pérez MCP, Gallardo-Antolín A, Martínez JMM. Building a decision support system for inpatient admission prediction with the manchester triage system and administrative check-in variables. Comput Inf Nurs 2016;34:224–30. https://doi.org/10.1097/CIN.0000000000000230.

[2] Cameron A, Rodgers K, Ireland A, Jamdar R, McKay GA. A simple tool to predict admission at the time of triage. Emerg Med J 2015;32:174–9. https://doi.org/10.1136/emermed-2013-203200.

[3] Derlet RW, Richards JR. Overcrowding in the nation's emergency departments: complex causes and disturbing effects. Ann Emerg Med 2000;35:63–8. https://doi.org/10.1016/S0196-0644(00)70105-3.

[4] Parker CA, Liu N, Wu SX, Shen Y, Lam SSW, Ong MEH. Predicting hospital admission at the emergency department triage: a novel prediction model. Am J Emerg Med 2019;37:1498–504. https://doi.org/10.1016/j.ajem.2018.10.060.

[5] Pines JM, Hilton JA, Weber EJ, Alkemade AJ, Al Shabanah H, Anderson PD, Bernhard M, Bertini A, Gries A, Ferrandiz S, Kumar VA, Harjola V-P, Hogan B, Madsen B, Mason S, Öhlén G, Rainer T, Rathlev N, Revue E, Richardson D, Sattarian M, Schull MJ. International perspectives on emergency department crowding. Acad Emerg Med 2011;18:1358–70. https://doi.org/10.1111/j.1553-2712.2011.01235.x.

[6] Pines JM, Hollander JE, Localio AR, Metlay JP. The association between emergency department crowding and hospital performance on antibiotic timing for pneumonia and percutaneous intervention for myocardial infarction. Acad Emerg Med 2006;13:873–8. https://doi.org/10.1197/j.aem.2006.03.568.

[7] Carter EJ, Pouch SM, Larson EL. The relationship between emergency department crowding and patient outcomes: a systematic review. J Nurs Scholarsh 2014;46:106–15. https://doi.org/10.1111/jnu.12055.

[8] Pines JM, Iyer S, Disbot M, Hollander JE, Shofer FS, Datner EM. The effect of emergency department crowding on patient satisfaction for admitted patients. Acad Emerg Med 2008;15:825–31. https://doi.org/10.1111/j.1553-2712.2008.00200.x.

[9] Ro YS, Do Shin S, Song KJ, Cha WC, Cho JS. Triage-based resource allocation and clinical treatment protocol on outcome and length of stay in the emergency department. Emerg Med Australasia (EMA) 2015;27:328–35. https://doi.org/10.1111/1742-6723.12426.

[10] ACEP Crowding. https://www.acep.org/patient-care/policy-statements/crowding/. [Accessed 3 August 2021]. accessed.

[11] van der Linden MC, Meester BEAM, van der Linden N. Emergency department crowding affects triage processes. Int. Emerg. Nurs. 2016;29:27–31. https://doi.org/10.1016/j.ienj.2016.02.003.

[12] Sprivulis PC, Da Silva J, Jacobs IG, Jelinek GA, Frazer ARL. The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. Med J Aust 2006;184:208–12. https://doi.org/10.5694/j.1326-5377.2006.tb00203.x.

[13] Wolff P, Ríos SA, Graña M. Setting up standards: a methodological proposal for pediatric Triage machine learning model construction based on clinical outcomes. Expert Syst Appl 2019;138:112788. https://doi.org/10.1016/j.eswa.2019.07.005.

[14] Jayaraman PP, Gunasekera K, Burstein F, Haghighi PD, Soetikno HS, Zaslavsky A. An ontology-based framework for real-time collection and visualization of mobile field triage data in mass gatherings. Hawaii Int. Conf. Syst. Sci., IEEE, 2013; 2013. p. 146–55. https://doi.org/10.1109/HICSS.2013.92. 46th.

[15] Fernandes M, Vieira SM, Leite F, Palos C, Finkelstein S, Sousa JMC. Clinical decision support systems for triage in the emergency department using intelligent systems: a review. Artif Intell Med 2020;102:101762. https://doi.org/10.1016/j.artmed.2019.101762.

[16] Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, Dugas A, Linton B, Kirsch T, Kelen G. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. Ann Emerg Med 2018;71:565–574.e2. https://doi.org/10.1016/j.annemergmed.2017.08.005.

[17] Liu N, Zhang Z, Wah Ho AF, Ong MEH. Artificial intelligence in emergency medicine. J. Emerg. Crit. Care Med. 2018;2:82. https://doi.org/10.21037/jeccm.2018.10.08. 82.

[18] Georgopoulos VC, Stylios CD. Introducing fuzzy cognitive maps for developing decision support system for triage at emergency room admissions for the elderly. IFAC Proc 2012;45:484–9. https://doi.org/10.3182/20120829-3-HU-2029.00107.

[19] Ramesh A, Kambhampati C, Monson J, Drew P. Artificial intelligence in medicine. Ann R Coll Surg Engl 2004;86:334–8. https://doi.org/10.1308/147870804290.

[20] Weber EJ. Triage: making the simple complex? Emerg Med J 2018;36. https://doi.org/10.1136/emermed-2018-207659. emermed-2018-207659.

[21] Rahimian F, Salimi-Khorshidi G, Payberah AH, Tran J, Ayala Solares R, Raimondi F, Nazarzadeh M, Canoy D, Rahimi K. Predicting the risk of emergency admission with machine learning: development and validation using linked electronic health records. PLoS Med 2018;15:e1002695. https://doi.org/10.1371/journal.pmed.1002695.

[22] Jiang H, Mao H, Lu H, Lin P, Garry W, Lu H, Yang G, Rainer TH, Chen X. Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. Int J Med Inf 2021;145:104326. https://doi.org/10.1016/j.ijmedinf.2020.104326.

[23] Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019;23:64. https://doi.org/10.1186/s13054-019-2351-7.

[24] Goto T, Camargo CA, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning–based prediction of clinical outcomes for children during emergency department triage. JAMA Netw Open 2019;2:e186937. https://doi.org/10.1001/jamanetworkopen.2018.6937.

[25] Dugas AF, Kirsch TD, Toerper M, Korley F, Yenokyan G, France D, Hager D, Levin S. An electronic emergency triage system to improve patient distribution by critical outcomes. J Emerg Med 2016;50:910–8. https://doi.org/10.1016/j.jemermed.2016.02.026.

[26] Farahmand S, Shabestari O, Pakrah M, Hossein-Nejad H, Arbab M, Bagheri-Hariri S. Artificial intelligence-based triage for patients with acute abdominal pain in emergency department; a diagnostic accuracy study. Adv. J. Emerg. Med. 2017;1:e5. https://doi.org/10.22114/AJEM.v1i1.11.

[27] Choi SW, Ko T, Hong KJ, Kim KH. Machine learning-based prediction of Korean triage and acuity scale level in emergency department patients. Healthc. Inform. Res. 2019;25:305. https://doi.org/10.4258/hir.2019.25.4.305.

[28] Gao Z, Qi X, Zhang X, Gao X, He X, Guo S, Li P. Developing and validating an emergency triage model using machine learning algorithms with medical big data. Risk Manag Healthc Pol 2022;15:1545–51. https://doi.org/10.2147/RMHP.S355176.

[29] 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;12:10537. https://doi.org/10.1038/s41598-022-14422-4.

[30] Müller A, Guido S. Introduction to machine learning with Python: a guide for data scientists. first ed. California: O'Reilly Media, Inc; 2016.

[31] Brownlee J, Duchesnay E, Löfstedt T, Baloch QB, Jason B, Brownlee J. Statistics and Machine Learning in Python. Mach. Learn. Mastery. 2017;91:399–404. https://hal.science/hal-03038776v3.

[32] Munson MA. A study on the importance of and time spent on different modeling steps. ACM SIGKDD Explor. Newsl. 2012;13:65–71. https://doi.org/10.1145/2207243.2207253.

[33] Zhang S. Nearest neighbor selection for iteratively kNN imputation. J Syst Software 2012;85:2541–52. https://doi.org/10.1016/j.jss.2012.05.073.
[34] Jebli I, Belouadha F-Z, Kabbaj MI, Tilioua A. Prediction of solar energy guided by pearson correlation using machine learning. Energy 2021;224:120109. https://doi.org/10.1016/j.energy.2021.120109.
[35] Osborne JW. Improving your data transformations: applying the Box-Cox transformation. Practical Assess Res Eval 2010;15:12. https://doi.org/10.7275/qbpc-gk17.
[36] Vinutha HP, Poornima B, Sagar BM. Detection of outliers using interquartile range technique from intrusion dataset. In: Adv. Intell. Syst. Comput. Singapore: Springer; 2018. p. 511–8. https://doi.org/10.1007/978-981-10-7563-6_53.
[37] GeeksforGeeks. Normalization vs standardization. 2021. n.d.), https://www.geeksforgeeks.org/normalization-vs-standardization/. [Accessed 1 May 2022]. accessed.
[38] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.
[39] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inf 2002;35:352–9. https://doi.org/10.1016/S1532-0464(03)00034-0.
[40] Panje CM, Glatzer M, von Rappard J, Rothermundt C, Hundsberger T, Zumstein V, Plasswilm L, Putora PM. Applied Swarm-based medicine: collecting decision trees for patterns of algorithms analysis. BMC Med Res Methodol 2017;17:123. https://doi.org/10.1186/s12874-017-0400-y.
[41] Kataria A, Singh MD. A review of data classification using K-nearest neighbour algorithm. Int. J. Emerg. Technol. Adv. Eng. 2013;3:354–60. www.ijetae.com. [Accessed 8 May 2022]. accessed.
[42] Huang Xiaolin, Shi Lei, Suykens JAK. Support vector machine classifier with pinball loss. IEEE Trans Pattern Anal Mach Intell 2014;36:984–97. https://doi.org/10.1109/TPAMI.2013.178.
[43] Gardner M, Dorling S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ 1998;32:2627–36. https://doi.org/10.1016/S1352-2310(97)00447-0.
[44] Webb GI, Zheng Z. Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. IEEE Trans Knowl Data Eng 2004;16:980–91. https://doi.org/10.1109/TKDE.2004.29.
[45] Zhou Z-H. Ensemble learning. In: Mach. Learn. Springer; 2021. p. 181–210.
[46] Taherkhani A, Cosma G, McGinnity TM, AdaBoost-CNN. An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. Neurocomputing 2020;404:351–66. https://doi.org/10.1016/j.neucom.2020.03.064.
[47] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proc. 22nd ACM SIGKDD int. Conf. Knowl. Discov. Data min. New York, NY, USA: ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
[48] Jackins V, Vimal S, Kaliappan M, Lee MY. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. J Supercomput 2021;77:5198–219. https://doi.org/10.1007/s11227-020-03481-x.
[49] Gaspar A, Oliva D, Cuevas E, Zaldívar D, Pérez M, Pajares G. Hyperparameter optimization in a convolutional neural network using metaheuristic algorithms. Stud. Comput. Intell. 2021:37–59. https://doi.org/10.1007/978-3-030-70542-8_2.
[50] Bochinski E, Senst T, Sikora T. In: Hyper-parameter optimization for convolutional neural network committees based on evolutionary algorithms. 2017 IEEE Int. Conf. Image Process., IEEE; 2017. p. 3924–8. https://doi.org/10.1109/ICIP.2017.8297018.
[51] Aszemi NM, Dominic PDD. Hyperparameter optimization in convolutional neural network using genetic algorithms. IJACSA) Int. J. Adv. Comput. Sci. Appl. 2019;10. May 5, 2022, www.ijacsa.thesai.org.
[52] Rojas R. Neural networks. Berlin, Heidelberg: Springer Berlin Heidelberg; 1996. https://doi.org/10.1007/978-3-642-61068-4.
[53] Markoulidakis I, Rallis I, Georgoulas I, Kopsiaftis G, Doulamis A, Doulamis N. Multiclass confusion matrix reduction method and its application on net promoter score classification problem. Technologies 2021;9:81. https://doi.org/10.3390/technologies9040081.
[54] Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006;27:861–74. https://doi.org/10.1016/j.patrec.2005.10.010.
[55] Deng X, Liu Q, Deng Y, Mahadevan S. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Inf. Sci. (Ny). 340–341 2016:250–61. https://doi.org/10.1016/j.ins.2016.01.033.
[56] Abd Elrahman SM, Abraham A. A review of class imbalance problem. May 15, 2022, www.mirlabs.net/jnic/index.html; 2013.
[57] Chakraborty J, Majumder S, Menzies T. Bias in machine learning software: why? how? what to do?. In: Proc. 29th ACM jt. Meet. Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng. New York, NY, USA: ACM; 2021. p. 429–40. https://doi.org/10.1145/3468264.3468537.
[58] Song K, Yan F, Ding T, Gao L, Lu S. A steel property optimization model based on the XGBoost algorithm and improved PSO. Comput Mater Sci 2020;174:109472. https://doi.org/10.1016/j.commatsci.2019.109472.
[59] Ong MEH, Lee Ng CH, Goh K, Liu N, Koh Z, Shahidah N, Zhang T, Fook-Chong S, Lin Z. Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. Crit Care 2012;16:R108. https://doi.org/10.1186/cc11396.
[60] Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. Acad Emerg Med 2016;23:269–78. https://doi.org/10.1111/acem.12876.
[61] LaMantia MA, Platts-Mills TF, Biese K, Khandelwal C, Forbach C, Cairns CB, Busby-Whitehead J, Kizer JS. Predicting hospital admission and returns to the emergency department for elderly patients. Acad Emerg Med 2010;17:252–9. https://doi.org/10.1111/j.1553-2712.2009.00675.x.