BMC Emergency Medicine

SYSTEMATIC REVIEW

Open Access

Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review

Bruno Matos Porto[1*]

Abstract

Background In Emergency Departments (EDs), triage is crucial for determining patient severity and prioritizing care, typically using the Manchester Triage Scale (MTS). Traditional triage systems, reliant on human judgment, are prone to under-triage and over-triage, resulting in variability, bias, and incorrect patient classification. Studies suggest that Machine Learning (ML) and Natural Language Processing (NLP) could enhance triage accuracy and consistency. This review analyzes studies on ML and/or NLP algorithms for ED patient triage.

Methods Following Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) guidelines, we conducted a systematic review across five databases: Web of Science, PubMed, Scopus, IEEE Xplore, and ACM Digital Library, from their inception of each database to October 2023. The risk of bias was assessed using the Prediction model Risk of Bias Assessment Tool (PROBAST). Only articles employing at least one ML and/or NLP method for patient triage classification were included.

Results Sixty studies covering 57 ML algorithms were included. Logistic Regression (LR) was the most used model, while eXtreme Gradient Boosting (XGBoost), decision tree-based algorithms with Gradient Boosting (GB), and Deep Neural Networks (DNNs) showed superior performance. Frequent predictive variables included demographics and vital signs, with oxygen saturation, chief complaints, systolic blood pressure, age, and mode of arrival being the most retained. The ML algorithms showed significant bias risk due to critical bias assessment in classification models.

Conclusion NLP methods improved ML algorithms' classification capability using triage nursing and medical notes and structured clinical data compared to algorithms using only structured data. Feature engineering (FE) and class imbalance correction methods enhanced ML workflows' performance, but FE and eXplainable Artificial Intelligence (XAI) were underexplored in this field.

Registration and funding.

This systematic review has been registered (registration number: CRD42024604529) in the International Prospective Register of Systematic Reviews (PROSPERO) and can be accessed online at the following URL: https://www.crd.york.

*Correspondence:
Bruno Matos Porto
bmatosporto@gmail.com
Full list of author information is available at the end of the article

ac.uk/prospero/display_record.php?RecordID=604529. Funding for this work was provided by the National Council for Scientific and Technological Development (CNPq), Brazil.

Keywords Triage of patients, Triage systems, Artificial intelligence, Feature engineering

Introduction

In the Emergency Department (ED), patient care begins with triage, which is a preliminary clinical assessment performed to identify the severity of the patient's health condition before diagnostic and therapeutic evaluation [1, 2]. Triage is essential for identifying patients who require urgent care and must be attended to immediately [1, 3]. The most commonly used emergency triage systems employ five-level priority scales associated with the patient's condition, such as the Korean Triage and Acuity Scale (KTAS), Emergency Severity Index (ESI), and Manchester Triage Scale (MTS) [4].

Traditional triage systems are widely adopted in EDs to prioritize patients and efficiently allocate available resources [3, 5–9]. However, some studies report under-triage (which occurs when a patient is not classified at the severity level corresponding to their condition, resulting in increased morbidity, mortality, and costs) [10–12] and over-triage (when less urgent patients are designated for urgent care, diverting resources from patients who genuinely require such care) [11–13] as frequently occurring phenomena [9, 11, 14].

The use of triage scales depends on human judgment, which can result in high variability and individual bias, affecting the accuracy of the assessment [3, 15, 16]. Studies report that some triage systems, such as the ESI, exhibit suboptimal predictive capacity for identifying severely ill patients, as well as low inter-rater agreement, high variability within the same triage level [5, 11, 14, 17], and a predominance of classifying patients at the medium acuity level [11, 18].

Incorrect patient classifications are common in traditional triage systems [4, 9, 12] and result in issues in EDs such as: (i) overcrowding [19]; (ii) under-triage; (iii) over-triage; (iv) failures to identify patients with cardiac events [4]; (v) increased safety risks, patient wait times, and deterioration in the quality of care [2, 19]; and (vi) a high degree of variability in triage assignment by nurses within the same region [16]. This context creates the need for more accurate classification of patient conditions at the time of triage, which can be achieved through Machine Learning (ML) and Natural Language Processing (NLP) [11, 14, 15]. An effective strategy to improve triage systems and support nurses' decision-making in patient stratification is the use of ML models [4].

Various ML models have been used in patient triage classification, including both multiclass classification [8, 20–22] and binary classification [23–25]. ML has demonstrated high performance in predicting various clinical outcomes, such as hospital admissions [26–29], critical care in patients with chest pain [30], patients with sepsis [31–33] and patient no-shows for medical appointments [34].

Predominant models in the literature include Logistic Regression (LR) [6, 24] and Random Forest (RF) [17, 35]. High-performance models include eXtreme Gradient Boosting (XGBoost) [8, 12] and Deep Neural Networks (DNNs) [36, 37]. Studies have shown high performance in triage prediction [9, 15] using ML with structured patient triage variables. Recently, incorporating triage clinical notes has demonstrated improved ML classification performance based on NLP [6, 10, 13, 37–41], highlighting the advantage of combining both approaches to achieve superior performance.

NLP uses computational models to analyze human language, its structure, and meaning [6, 42]. Initially, in patient triage, NLP methods were simple (e.g., Bag-of-Words [43, 44]), considering the relative frequencies of words in triage notes, ignoring word order and context [43]. More recent NLP methods based on DNNs process triage notes through layers of neural networks, providing more complex representations of the data [13, 37, 39]. Another advanced method used in the field is Bidirectional Encoder Representations from Transformers (BERT), a model pre-trained on large text datasets, applied in patient triage prediction [45–47]. Recent studies have employed the Chat Generative Pre-trained Transformer (ChatGPT) model [48, 49] for patient triage. This study describes the methods and evaluates the performance of NLP applications on unstructured free-text triage notes.

Three systematic reviews [1, 18, 50] and one literature review [19] on ML models for triage highlight the performance of XGBoost and Gradient Boosting (GB), with LR showing inferior performance. Previous reviews focused on various algorithms for disease predictions, hospital admissions, and triage, but only one concentrated on ED triage, limited to a two-year period. The current review addresses multiclass patient classification, crucial for the efficient allocation of human and material resources in EDs [19], and explores less-discussed aspects: feature selection, feature engineering (FE), eXplainable Artificial Intelligence (XAI), class imbalance correction, and particularly the use of NLP, which are essential for improving

the performance of ML algorithms. FE involves creating new features based on domain knowledge or exploratory analysis of available data [51], while XAI refers to systems that provide transparent and understandable explanations of their processes, making it easier for users to comprehend how the ML model functions and why specific outcomes are produced [52].

In this work, 60 original studies were comprehensively analyzed, with the objective of systematically reviewing studies that employed ML and/or NLP methods for classifying the triage of adult and pediatric patients in EDs. The studies were organized based on their main characteristics: (i) quality assessment and risk of bias, (ii) ML and NLP methods used, (iii) variable selection, FE, and resampling techniques, (iv) predictors tested, and (v) performance metrics. The main findings include: (i) ML models exhibited a high risk of bias when evaluated using the Prediction Model Risk of Bias Assessment Tool (PROBAST) [53], (ii) NLP enhanced the performance of ML algorithms in predicting patient triage, (iii) FE and XAI approaches were underutilized in this field, (iv) the most commonly used and retained predictor variables were identified, (v) the classification performance was assessed using key metrics, and (vi) an overview of the methods employed in the ML workflow was provided. This review underscores the need for more robust and explainable approaches in the development and evaluation of predictive models for patient triage.

Materials and methods

In the literature, ML algorithms have been widely used in recent years, being the subject of three systematic reviews [1, 18, 50] and one literature review [19]. However, a comprehensive understanding of the five aspects is lacking: feature selection, FE, XAI, class imbalance correction, and particularly the use of NLP. The previous reviews were summarized in Table 1 to identify the unexplored aspects. These aspects are fundamental, as they all impact the performance of ML models, and the quality of the predictions made in the studies.

The methodological steps used to conduct the systematic review included: (i) registering the systematic review protocol in PROSPERO, (ii) applying the predefined inclusion and exclusion criteria, (iii) following the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) 2020 guidelines to ensure a comprehensive and transparent reporting of the review process, (iv) formulating research questions using the Participants, Intervention, Comparison, Outcome, Study Design (PICO-SD) format [54], followed by a search strategy conducted across five databases, covering the period from the inception of each database until October 2023, (v) extracting and synthesizing data from the included studies to answer the research questions, (vi) analyzing the risk of bias using the PROBAST tool to ensure the validity and applicability of the predictive models, and (vii) performing a sensitivity and ROC-AUC analysis to assess the robustness of the findings.

Inclusion and exclusion criteria

The primary outcome was defined as the triage-levels of patients (multiclass classification), while the secondary outcome was the triage of critical patients, including mortality or admission to the intensive care unit (binary classification). This review covers ML and/or NLP methods for the classification of patient triage in EDs. The following criteria presented in Table 2 were used to select the studies.

Research questions

We analyzed studies that applied ML and/or NLP methods for patient triage classification to answer the research questions outlined in Table 3. These research questions were developed following the PICO-SD format [54]. Specifically, the patient/population/problem refers to patients in the ED. The intervention refers to the use of ML algorithms or NLP methods for patient triage in the ED. The comparison involves at least one ML or NLP method for patient triage, either with or without a comparison to conventional triage systems, such as the ESI or MTS. The outcomes focus on triage of severely ill patients, including mortality, admission to the intensive care unit (ICU), and triage levels. Finally, the study design includes both retrospective and prospective studies.

Search strategy and registration

This study followed the PRISMA 2020 protocol [55] to select studies based on the defined criteria. PRISMA is used to address the three research questions in Table 3. The steps of the PRISMA checklist are detailed in Table A1, in Appendix A. This systematic review is registered with the International Prospective Register of Systematic Reviews (PROSPERO) and is available for access online at the following URL: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=604529. The protocol has been registered on PROSPERO (registration number CRD42024604529).

First, Medical Subject Headings were used to define keywords related to the topic. Next, the databases Web of Science, PubMed, Scopus, IEEE Xplore, and ACM Digital Library were selected to search for studies published up to October 2023. Searches in the five databases were initially conducted on July 20, 2023, and the last search was on October 13, 2023.

The databases were selected based on four main criteria: (i) Web of Science and IEEE Xplore were excluded

Table 1 Systematic review research on ML for patient triage in EDs

Ref. / Year	Study aims	Period analyzed	Articles analyzed	Data base	Main results
Shafaf and Malek [19]	Evaluate ML models for predicting diseases, hospital admission, mortality, and triage in EDs	2008–2018	30 studies for all outcomes 5 articles on patient triage	Not reported	The most important ML models were LR, SVM, RF, GB, and DL, describing the potential of these algorithms in clinical outcomes. It is concluded that ML can be used primarily in emergency medicine to predict clinical outcomes
Miles et al. [18]	Assess the accuracy of machine learning models in their application of triaging the acuity of patients	2010–12/2019	25	MEDLINE, CINAHL, PubMed and the grey literature	For predictions of hospitalization, ICU admission, and severe illnesses, neural networks and tree-based algorithms achieved the best performances. However, for triage, no ML approach was superior in terms of performance. Although some approaches showed better AUCs than others, when comparing different ML approaches, no statistically significant differences were found between the models
Gao et al. [50]	Assess the performance of machine learning models for patient triage in emergency departments, and to identify future challenges	2018–10/2021	21	ScienceDirect, PubMed, Google Scholar and Springer Link	ML algorithms were analyzed for triage classification, hospital admission, and critical illnesses. Key predictor variables included demographic data, clinical variables, and ED arrival information. The most used algorithms were DNN, LR, RF, and GB. Most studies evaluated the models through cross-validation, using metrics such as F1 score, Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), sensitivity, specificity, and accuracy. Variable importance was determined using specific methods for each model, such as LR coefficients or permutation-based variable importance. It was identified that GB showed better performance in predicting critical illnesses, while DNN excelled in predicting hospital admission
Sánchez-Salmerón et al. [1]	Analyze the effectiveness of ML systems in triage for making predictions at the ED	2018–11/2020	11	CINAHL, Cochrane, Cuiden, Medline and Scopus	ML algorithms surpass the discriminatory capacity of triage scales used in EDs, such as ESI. XGBoost and DNN are superior performing models compared to other ML models

Table 2 Inclusion and exclusion criteria for prediction models in patient triage in EDs

Inclusion criteria	Exclusion criteria
I—Population: Patients undergoing the triage process to receive emergency care (adults and pediatrics) in EDs	**VII**—Studies that performed patient triage with coronavirus disease 2019 (COVID-19), prehospital triage in emergency medical dispatch, triage of call center, triage of sepsis, triage of stroke, and ophthalmology triage
II—Prediction outcomes: Triage of patients in EDs	**VIII**—Other outcome prediction: Studies on hospital admission, hospital readmission, length of stay, ED admission, fast-track section of EDs, detection of sepsis, and other illnesses
III—Study design: All retrospective and prospective studies	**IX**—Studies that conducted patient triage (e.g., ESI, MTS) without using ML or NLP methods for patient classification
IV—Studies that used at least one ML model (including LR) or NLP for patient triage in EDs	**X**—Studies that were not accessible in full text
V—Studies published in English, from any publication date, and peer-reviewed	**XI**—Studies published in languages other than English
VI—Studies that used variables collected during triage, such as structured data (e.g., demographics, vital signs) and unstructured data (e.g., nursing or medical triage notes)	**XII**—Gray literature and conference papers

Table 3 Research questions

#	Research Question	Rationale
RQ1	What ML and/or NLP methods were used for patient triage classification in EDs?	Identifying which ML models were used for patient triage classification
RQ2	Do ML and/or NLP methods show high performance in predicting patient triage in EDs?	Evaluate the discriminative ability of methods in patient triage in EDs
RQ3	Does using nursing or clinical notes through NLP improve prediction performance compared to studies using only structured data?	Evaluate whether the incorporation of free-text improves the performance of ML algorithms compared to algorithms using only structured data. Previous systematic reviews did not address this question

Since the first research question (RQ1) is broad and involves the application of other methods, it has been divided into three sub-questions: RQ1.1 Classification algorithms, RQ1.2 Variable selection techniques, and RQ1.3 Class imbalance correction methods

from previous reviews; (ii) Scopus was used in only one review [1] covering 2 years (2018–11/2020); (iii) PubMed was chosen for its extensive coverage in medicine and health informatics, including access to MEDLINE; and (iv) IEEE Xplore and ACM Digital Library are leading databases in the field of NLP and ML.

The screenings and full-text evaluations were performed independently by the researchers, following the criteria and extracting data using a standardized form (Table A2). Conflicts were resolved by discussion between the researchers. It was not necessary to contact the authors of the included studies. The search strings used are listed in Table 4. Additional references were retrieved from the reference lists of key reviews [1, 18, 19, 50] for analysis.

Study selection

For the review, we used the participants and interventions from the PICO-SD framework as search criteria. Initially, 2,292 results were identified across five databases. After excluding articles not published in journals, not in English, and 241 duplicates, 668 remained for screening. Ninety-three were excluded due to their titles being out of scope, leaving 575 articles. Of these, 352 did not report the clinical outcomes of interest. The main focus of these excluded articles is shown in Fig. 1. Two hundred twenty-three articles were retained for full-text reading, of which 167 were removed for dealing with ML or NLP related to hospital admission, chest X-rays, among other outcomes. We included 4 articles from the references of the reviewed articles. Sixty studies met all criteria. The screening process is detailed in Fig. 1.

The Rayyan.ai software (https://new.rayyan.ai/reviews/806188/screening) was used in the screening and selection process for the study. We utilized Rayyan.ai to facilitate the systematic review process, starting with the import of RIS files from five databases. Rayyan's features enabled us to efficiently manage and organize the large volume of articles retrieved from these sources. The software automatically detected and removed duplicate records across the five databases, streamlining the initial phase of data cleaning.

Following duplicate removal, Rayyan was used to screen the remaining articles against the predefined inclusion and exclusion criteria. The tool allowed for quick and flexible tagging of articles based on specific

Table 4 Literature search strategy strings

Database	Search term	Query box/search within	Search in	Results in 11/10/2023
Web of Science	(triage OR "clinical triage" OR "emergency calls" OR "telephone triage" OR "classification of patient" OR "Patient severity" OR "Patient acuity" OR "patient prioritization") AND ("machine learning" OR "deep learning" OR "natural language processing" OR NLP OR "artificial intelligence" OR "artificial neural network" OR "text mining") AND ("emergency department" OR "emergency room" OR "emergency medical system" OR "emergency medical service" OR "emergency unit" OR "emergency medicine" OR "emergency service" OR "emergency care" OR "urgent care" OR "accident and emergency" OR "accident & emergency" OR A&E)	All Fields	Core Collection	302
PubMed Central	(triage OR "clinical triage" OR "emergency calls" OR "telephone triage" OR "classification of patient" OR "Patient severity" OR "Patient acuity" OR "patient prioritization") AND ("machine learning" OR "deep learning" OR "natural language processing" OR NLP OR "artificial intelligence" OR "artificial neural network" OR "text mining") AND ("emergency department" OR "emergency room" OR "emergency medical system" OR "emergency medical service" OR "emergency unit" OR "emergency medicine" OR "emergency service" OR "emergency care" OR "urgent care" OR "accident and emergency" OR "accident & emergency" OR A&E)	All Fields	MEDLINE	91
Scopus	(triage OR "clinical triage" OR "emergency calls" OR "telephone triage" OR "classification of patient" OR "Patient severity" OR "Patient acuity" OR "patient prioritization") AND ("machine learning" OR "deep learning" OR "natural language processing" OR NLP OR "artificial intelligence" OR "artificial neural network" OR "text mining") AND ("emergency department" OR "emergency room" OR "emergency medical system" OR "emergency medical service" OR "emergency unit" OR "emergency medicine" OR "emergency service" OR "emergency care" OR "urgent care" OR "accident and emergency" OR "accident & emergency" OR A&E)	Article title, Abstract, Keywords	SCOPUS	609
IEEE Xplore	(triage OR "clinical triage" OR "emergency calls" OR "telephone triage" OR "classification of patient" OR "Patient severity" OR "Patient acuity" OR "patient prioritization") AND ("machine learning" OR "deep learning" OR "natural language processing" OR NLP OR "artificial intelligence" OR "artificial neural network" OR "text mining") AND ("emergency department" OR "emergency room" OR "emergency medical system" OR "emergency medical service" OR "emergency unit" OR "emergency medicine" OR "emergency service" OR "emergency care" OR "urgent care" OR "accident and emergency" OR "accident & emergency" OR A&E)	All Fields	IEEE	344
ACM Digital Library	(triage OR "clinical triage" OR "emergency calls" OR "telephone triage" OR "classification of patient" OR "Patient severity" OR "Patient acuity" OR "patient prioritization") AND ("machine learning" OR "deep learning" OR "natural language processing" OR NLP OR "artificial intelligence" OR "artificial neural network" OR "text mining") AND ("emergency department" OR "emergency room" OR "emergency medical system" OR "emergency medical service" OR "emergency unit" OR "emergency medicine" OR "emergency service" OR "emergency care" OR "urgent care" OR "accident and emergency" OR "accident & emergency" OR A&E)	All Fields	ACM Full-Text collection	946

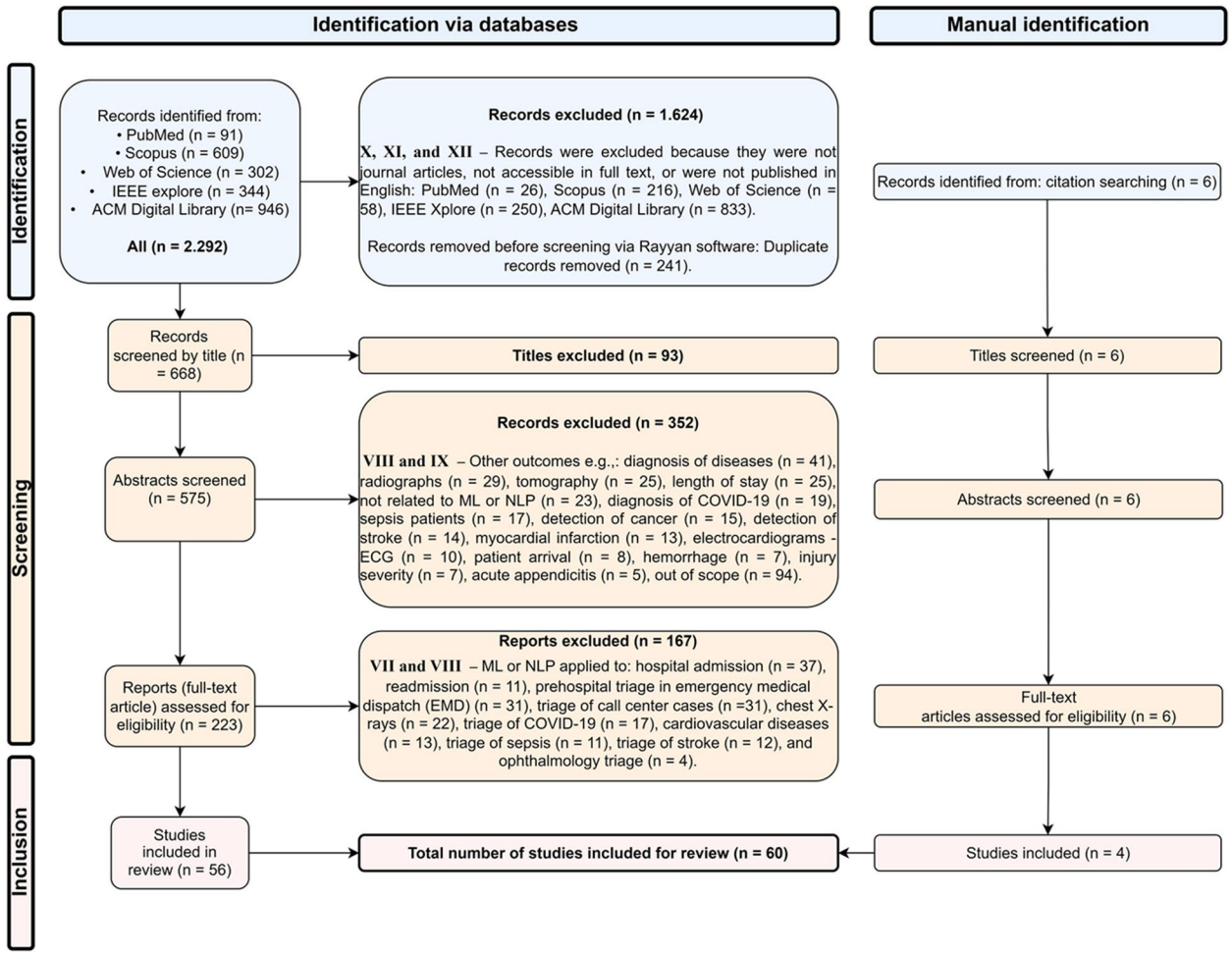

Fig. 1 PRISMA [55] flow chart adapted for the selection of articles using ML or NLP in triage systems

criteria, such as population, and use of ML or NLP methods. Articles that matched the inclusion criteria were retained, while those that fell under the exclusion criteria—like studies involving COVID-19, hospital admissions, or non-peer-reviewed sources—were flagged and excluded.

The platform also facilitated collaborative screening, allowing two reviewers to categorize articles independently while tracking conflicts for later resolution. Overall, Rayyan.ai improved the efficiency and accuracy of the study selection process by automating duplicate detection and enhancing the collaboration between reviewers.

Risk of bias assessment

PROBAST [53] was used to assess the risk of bias and reproducibility of ML models in the studies. PROBAST is a tool with twenty signaling questions, designed to evaluate four distinct domains, providing an overall assessment of bias risk and model applicability. The Excel template from Fernandez-Felix et al. [56] was used, a well-established and highly standardized tool, widely accepted for evaluating the risk of bias and model applicability in healthcare predictions. Two researchers independently conducted the risk of bias assessment using PROBAST, and any disagreements were resolved through discussion between them. The results of the PROBAST assessment are available in Tables A3 and A4 (Appendix A).

Assessment

The main information from the selected articles was summarized in Table A2 in Appendix A. For each article, the following were assessed: (1) author, year, and country of the ED; (2) triage system used; (3) classification outcome; (4) predictors tested; (5) ML algorithms; (6) best-performing predictors; (7) dataset partitioning and type of validation; (8) best-performing algorithms; (9) performance metrics; and (10) main results.

Classification tasks were categorized into triage of critically ill patients (binary) and triage levels using systems

such as ESI and MTS (multiclass). The most common ML algorithms (e.g., LR and RF) and the best-performing ones (e.g., XGBoost and DNNs) were identified based on the Receiver Operating Characteristic-Area Under the Curve (ROC-AUC). Table 5 summarizes the best ML algorithms, variable selection, FE, NLP methods, and class imbalance correction methods used in the studies.

Results

Summary of characteristics of included studies

Figure 2 shows a quantitative summary of the selected corpus. The search resulted in 60 articles meeting the criteria described above, with 88% of the included studies being retrospective and 12% prospective, conducted in various countries. The United States, South Korea, China, and Taiwan were responsible for 69% of the total scientific output. Most of the studied locations, 71.4%, were tertiary hospitals or academic medical centers located in urban environments. Of the analyzed studies, 63.3% were conducted in single centers and 36.7% in multiple emergency centers. Twenty-five percent of the studies used large national samples, with more than 20 EDs. The total number of patients included ranged from 15,000 to 600,000 for 57.3% of the analyzed articles.

Figure 2b shows the evolution of the number of articles and their annual percentages, with about 78% published in the last 5 years. The first study using ML for patient triage was published in 2008 in the Journal of Biomedical Informatics (JCR = 4.5). Since 2018, there has been an increase in the use of ML for patient triage prediction in the literature. PLoS One published the highest number of articles (Fig. 2c). The literature is not concentrated in a few journals, as about 80% of the total studies were published only two times or less in each journal.

The most commonly used traditional triage systems (human-based) were ESI, KTAS, and MTS (Fig. 2d). The top four triage systems accounted for 67% of the total studies. Regarding clinical outcomes analyzed, 21.6% of the articles studied more than one outcome. Among the studies, 58.3% analyzed the prediction of mortality or ICU admission, while 50% evaluated the prediction of triage-levels. The most commonly used predictor variables in the studies were demographic data, vital signs, and unstructured data such as nursing or medical triage notes. About 78.6% of the studies used fewer than 30 predictors, while only 16.3% used more than 50. Of the total, 32.7% implemented variable selection, most frequently retaining vital signs, age, mode of arrival, triage notes, chief complaint, laboratory tests, previous visits, pain score, and visit reasons.

ML involves the use of algorithms that enable the inference of patterns from training data, providing the capability for generalization—making predictions on the test set [3, 42]. A total of 57 ML algorithms were tested in 60 articles, ranging from one to twelve algorithms per study. The five most frequently used ML models were: LR at 53.1%, RF at 46.6%, XGBoost at 36.6%, DNN, and Multilayer Perceptron Neural Network at 20%. Fig. 3 presents the most commonly used ML models and those with the best performance in each study. LR is the benchmark model, while XGBoost performed best in 52.6% of comparisons with other algorithms, aligning with the review by Sánchez-Salmerón [1].

LR analyzes the linear relationships between independent and dependent variables to estimate the probabilities of outcomes occurring [1]. In LR, feature importance is determined by the estimated coefficients, which indicate the impact of each feature on the predicted outcome [3, 25]. Each coefficient represents the change in the log-odds of the outcome for a one-unit increase in the feature, with larger absolute values signifying greater influence [87]. Positive coefficients increase the likelihood of the outcome, while negative coefficients decrease it [87]. This analysis helps identify the most significant features, enhancing the model's interpretability in triage levels. XGBoost is a ML algorithm utilized for regression and classification tasks, creating a robust prediction model by aggregating a collection of weak prediction models [1, 43, 71]. Unlike other boosting models, GB trains new models directly on the errors of its predecessors. XGBoost, an extension of GB, enhances this approach by incorporating processing optimization techniques, resulting in improved outcomes while requiring fewer computational resources and less time [1, 43, 71].

DNN outperformed other algorithms in 66.6% of comparisons. DNN is a type of neural network that consists of multiple hidden layers and is capable of learning complex patterns by applying hierarchical, nonlinear transformations across sequential layers [45]. In DNNs, assessing feature importance is more complex than in LR due to their non-linear nature. Rather than using coefficients, DNNs evaluate feature significance based on their influence on predictions across multiple layers [13, 37, 39]. Techniques like permutation importance, SHapley Additive exPlanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME) are employed to determine feature contributions in triage levels [45, 46]. DNN and decision tree-based GB algorithms, such as categorical boosting (CatBoost), had better performance in 82.7% of comparisons. Only 30% of the studies compared traditional triage systems with the performance of ML algorithms, showing that ML models are consistently superior.

The combination of NLP methods and ML algorithms occurred in only 26.6% of the studies. The most commonly used NLP methods were: BERT [45, 46],

Table 5 Articles that used ML or NLP methods in triage systems ($n = 60$)

Ref.	Year/country	Prediction outcomes	Best algorithms	Feature engineering	Feature selection/ XAI	NLP	Class imbalance correction	Accuracy	Sensitivity	ROC-AUC	CI 95%
[57]	2008/Spain	Triage-level	NB	-	Naive Bayes	-	-	0.879	-	-	-
[58]	2012/Israel	Triage-level	NB	-	-	-	-	0.567	-	-	-
[23]	2012/Singapore	Triage critically ill patients with cardiac arrest	SVM	-	-	-	Oversampling	-	0.814	0.781	not reported
[59]	2013/USA	Triage-level	MLP	-	LR (Odds Ratio)	-	-	0.671	-	-	-
[60]	2013/Malaysia	Triage-level	MLF	-	-	-	-	0.967	0.85	-	-
[20]	2013/Taiwan	Triage-level	SVM	PCA	Genetic Algorithm (wrapper-based)	-	Undersampling	0.892	-	-	-
[24]	2016/USA	Triage-level and ICU admission	LR	-	LR (Odds Ratio)	-	-	-	-	0.83 0.83	not reported
[21]	2017/Ecuador	Triage-levels	MLP	-	-	-	-	0.9417	-	-	not reported
[3]	2018/South Korea	Triage critically ill patients	MLP	-	-	-	-	-	-	0.935 0.894	0.935–0.936 0.894–0.895
[41]	2018/USA	Triage critically ill patients	BiLSTM	one-hot encoding	-	WE, skip-gram	Undersampling	-	0.883	0.949	± 0.003 std
[25]	2018/USA	Prehospital triage	DNN	-	RF (relative importance)	-	Stratified sampling	-	-	0.89	0.882–0.890
[61]	2018/ USA	Triage critically ill patients	GBDT	-	RF (relative importance)	-	-	-	0.79	0.80	± 0.05 std
[17]	2018/USA	Triage-level and ICU admission	RF	-	RF (relative importance)	-	Bootstrapping	0.766	-	0.92	not reported
[5]	2019/USA	Triage critically ill patients	DNN	-	LR with Lasso regularization and RF (permutation-based)	-	Dropout	-	0.80	0.86	0.85–0.87
[62]	2019/Chile	Triage critically ill children	RF	-	RF (information gain)	-	SMOTE, Bootstrapping	0.853	0.667	0.86	±0.062
[22]	2019/Chile	Triage-levels	DT	-	-	-	-	0.969	-	-	not reported
[63]	2019/USA	Triage critically ill children	DNN	-	LR with Lasso regularization, importance feature in the RF and GBDT	-	Dropout	-	0.78	0.85	0,78–0,92

Table 5 (continued)

Ref.	Year/country	Prediction outcomes	Best algorithms	Feature engineering	Feature selection/ XAI	NLP	Class imbalance correction	Accuracy	Sensitivity	ROC-AUC	CI 95%
[6]	2019/South Korea	Triage-level	XGBoost	-	-	SoynIp, BoW	-	-	0.753	0.922	0.919–0.925
[64]	2019/South Korea	Triage critically ill children	MLP	one-hot encoding	RF (Gini index) LR (Deviance Difference)	-	Adjusted the data ratio in the training	-	-	0.908	0.903–0.910
[65]	2019/UK	Triage-level	RODDPSO	-	-	-	-	-	-	-	-
[7]	2020/South Korea	Triage critically ill patients	INA-ML	-	RF (information gain)	-	SMOTE	-	-	0.876	0.863–0.889
[66]	2020/South Korea	Triage critically ill patients	AI–ESI	-	-	-	Bootstrapping	-	0.799	0.923	0.920–0.926
[13]	2020/USA	Triage critically ill patients	DNN	-	LIME	WE	-	-	0.845	0.857	0.856–0.858
[10]	2020/Portuguese and USA	Triage critically ill patients	LR	-	LR (absolute values of the coefficients)	TF-IDF	Bootstrapping	-	0.73 0.81	0.86 in HBA, 0.91 in BIDMC	0.85–0.87 0.90–0.92
[67]	2020/Portuguese	Triage critically ill patients	XGBoost	-	XGBoost (relative importance)	TF-IDF	Stratified sampling	-	0.84	0.96	0.95–0.97
[40]	2020/Israel	Triage critically ill patients	XGBoost	target encoding and one-hot encoding	XGBoost (information gain)	WE	Bootstrapping	-	0.919	0.962	0.956–0.968
[37]	2020/Taiwan	Triage-level and ICU admission	DNN	-	-	WE, paragraph vectors, skip-gram	-	-	0.805 0.733	0.863 0.875	0.858–0.868 0,871–0,878
[9]	2021/China	Triage-level	CatBoost	✓	SHapley Additive exPlanation (SHAP)	-	-	-	-	0.875	± 0.006
[8]	2021/China	Triage-level	XGBoost	one-hot encoding	XGBoost (relative importance)	-	-	0.785	0.982	0.937	Not reported
[68]	2021/China	Triage critically ill patients	LightGBM	Time Series Feature Extraction	LightGBM	-	Bootstrapping	0.936	0.971	0.976	± 0.003
[69]	2021/Singapore	Triage critically ill patients	AutoScore	-	AutoScore (Parsimony plot)	-	-	-	0.763	0.823	0.814–0.832
[30]	2021/China	Triage critically ill patients	LAR	-	LR with Lasso regularization	-	-	0.890	0.864	0.953	0.922–0.984
[70]	2021/USA	Triage-level	XGBoost	Various approaches to FE	-	C-NLP	-	0.785	0.695	0.849	0.828–0.867
[43]	2021/USA	Triage critically ill patients	XGBoost	-	-	BoW	Bootstrapping, Class weights	-	0.80	0.93	0.92–0.95

Table 5 (continued)

Ref.	Year/country	Prediction outcomes	Best algorithms	Feature engineering	Feature selection/ XAI	NLP	Class imbalance correction	Accuracy	Sensitivity	ROC-AUC	CI 95%
[71]	2021/USA	Triage-levels	XGBoost	-	XGBoost (relative importance) and SHAP	-	-	-	0.970 0.922	0.982 0.968	0.980–0.983 0.967–0.969
[72]	2021/USA	Triage critically ill patients	GBM	-	Ablation study	-	Bootstrapping	-	-	0.88 0.86	0.87–0.89 0.85–0.87
[44]	2021/USA	Triage critically ill patients	Ensemble	-	-	BoW, TF-IDF	Bootstrapping	0.95	0.95	0.99	0.98–0.99
[35]	2021/USA	Triage-level	DNN	-	-	-	Undersampling	0.956	0.385	0.883	0,881–0,884
[73]	2021/USA	Trauma patient triage	OCT	-	-	-	-	-	-	0.941	not reported
[74]	2021/Thailand	Triage-levels	SVM	-	-	-	-	0.990	0.990	-	not reported
[36]	2021/Netherlands	Triage-level	XGBoost	one-hot encoding	SHAP	-	Bootstrapping	-	0.78	0.84	0.77–0.88
[45]	2021/South Korea	Triage-level	SVM	-	SHAP	BERT	Bootstrapping	-	0.78	0.90	0.86–0.94
[75]	2021/South Korea	Triage critically ill patients	XGBoost	-	XGBoost (permutation-based)	-	-	-	0.85	0.861	0.848–0.874
[12]	2022/China	Triage-level	XGBoost	-	LIME	-	-	0.825	-	0.9629	not reported
[76]	2022/South Korea	Triage critically ill children	RF	one-hot encoding	Random forest	-	Undersampling	-	-	0.991	0,991–0,992
[77]	2022/South Korea	Trauma patient triage	DNN	-	AdaBoost (Gini index)	-	SMOTE	0.8834	0.8599	0.9513	± 0.0023
[78]	2022/USA	Triage critically ill patients	GB	-	Random forest	-	Bootstrapping	-	0.809	0.880	0.876–0.884
[79]	2022/Germany	Triage-levels	BN	-	-	-	-	0.91	-	-	not reported
[15]	2023/Turkey	Triage-level	RF	PCA and label encoding	Random forest	-	SMOTE	0.889	0.889	-	-
[48]	2023/Turkey	Triage-levels	ChatGPT	-	-	ChatGPT	-	-	0.889	0.846	0.724–0.969
[46]	2023/UK	Triage-level	LightGBM	one-hot encoding	LIME	BERT	No sampling, oversampling, undersampling	0.77	0.77	0.81	not reported
[80]	2023/Romanian	Triage-level	NN-Sequentail	one-hot encoding	-	-	SMOTE and ADASYN	-	0.71	0.71	not reported
[81]	2023/ USA	Triage-levels	ASA-CaB	integer encoding	SFM, RFE, and SKB	-	SMOTE	83.3	83.3	-	not reported
[82]	2023/South Korea	Trauma patient triage	AdaBoost	-	AdaBoost (Gini index)	-	SMOTE	0.9887	0.9739	0.9974	± 0.0005

Table 5 (continued)

Ref.	Year/country	Prediction outcomes	Best algorithms	Feature engineering	Feature selection/ XAI	NLP	Class imbalance correction	Accuracy	Sensitivity	ROC-AUC	CI 95%
[83]	2023/South Korea	Triage critically ill patients	XGBoost	-	XGBoost (information gain)	-	Undersampling	-	0.61	0.961	not reported
[84]	2023/South Korea	Triage-levels	XGBoost	-	LR (Odds Ratio)	-	-	-	-	0.777	0.776–0.777
[85]	2023/Taiwan	Triage critically ill patients	CNN	-	XGBoost (information gain)	-	-	-	0.875	0.8915	not reported
[47]	2023/China	Triage-levels	TransNet	one-hot encoding	-	BERT and Tokenizer encoding	-	0.853	0.840	0.906-0.947	not reported
[39]	2023/Taiwan	Triage critically ill patients	DNN	one-hot encoding	-	CBoW	SMOTE	0.90	0.499	0.874	0.873–0.882
[86]	2023/Canada	Triage-levels	GB	✓	RF (permutation-based)	-	-	0.63	0.63	-	-

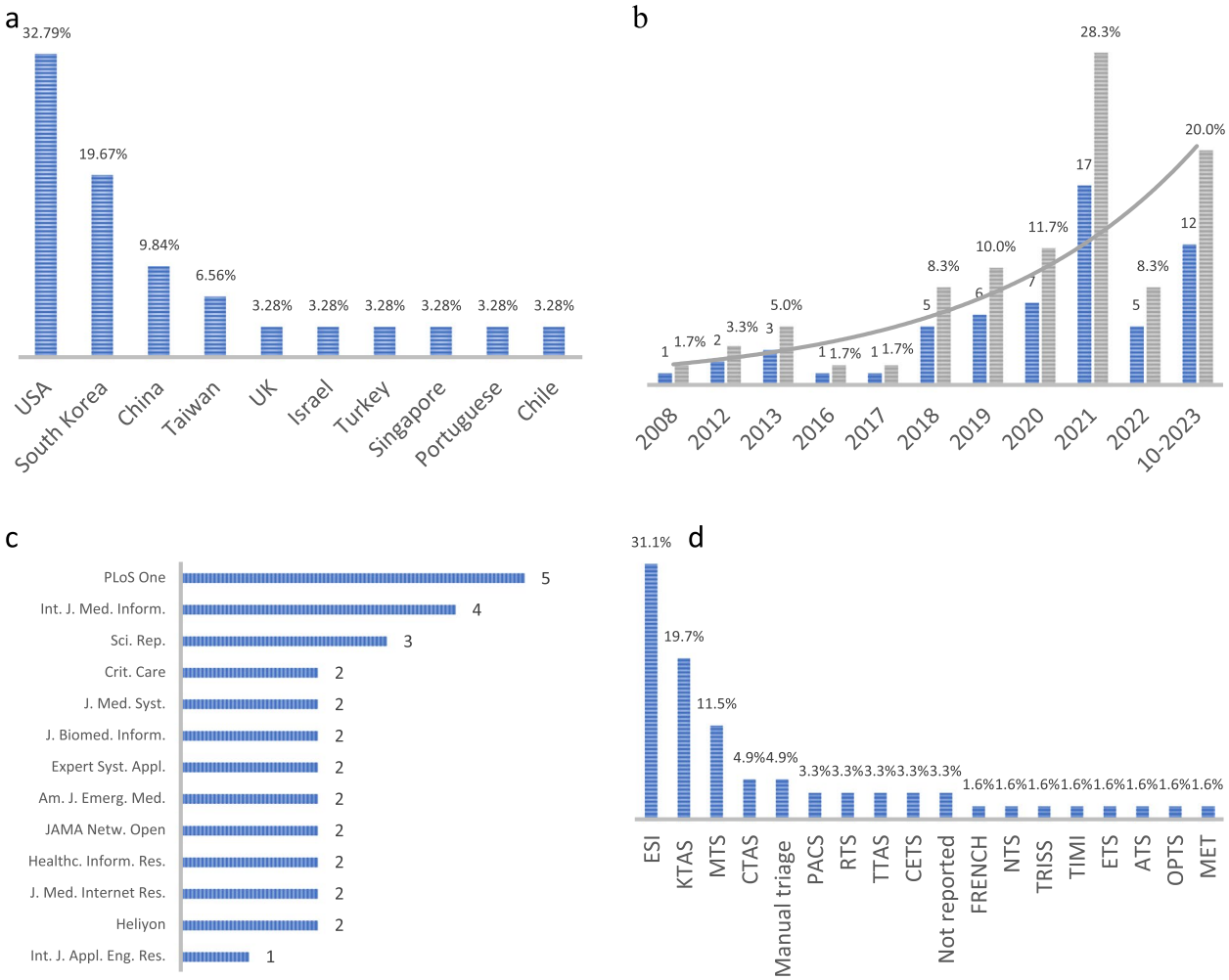

Fig. 2 Quantitative summary of the selected corpus. **a** Articles stratified by country. **b** Number of Articles and percentages per year. **c** Rankings of journals. **d** Articles stratified by triage system

Bag-of-Words (BoW) [43, 44], word embedding (WE) [40, 41], Skip-gram [37, 41], and Term Frequency—Inverse Document Frequency (TF-IDF) [44, 67]. Using unstructured data (triage notes) improved the performance of ML algorithms in all studies [6, 10, 13, 39, 70]. Additionally, algorithms that used both structured data and triage notes performed better than those that used only one of the two types of data [6, 13, 37, 41, 43, 67, 70].

Due to class imbalance in medical datasets [39, 62], it is important to assess whether the studies used correction methods. Fifty-three point three percent of the studies applied techniques such as bootstrapping [10, 78], synthetic minority oversampling technique (SMOTE) [15, 81], undersampling [35, 76], and oversampling [23, 46]. This is important because most classification models assume balanced classes [62], and imbalanced data tend to bias the model towards the majority class, impairing performance.

In model validation, 81% of the studies divided the data into training and testing sets, with common proportions of 80%/20% (14 articles), 90%/10% (10 articles), and 70%/30% (8 articles). Cross-validation was used in 56.6% of the studies, with K-fold being the most common (tenfold in 50% and fivefold in 41% of cases). Eighteen percent of the studies did not report data splitting. Internal–external validation was used in 10% of the articles. The main metrics for evaluating the performance of different models included ROC-AUC (C-statistic) in 78.3%, Se – Sensitivity (Recall) in 71.6%, positive predictive value (Precision) in 58.3%, Sp – Specificity in 56.6%, Accuracy in 41.6%, negative predictive value and F1-score in 35%, and Area Under the Precision and Recall Curve in 13.3%.

Quality assessment and risk of bias

The risk of bias in the predictive models was assessed using the PROBAST tool [53], following

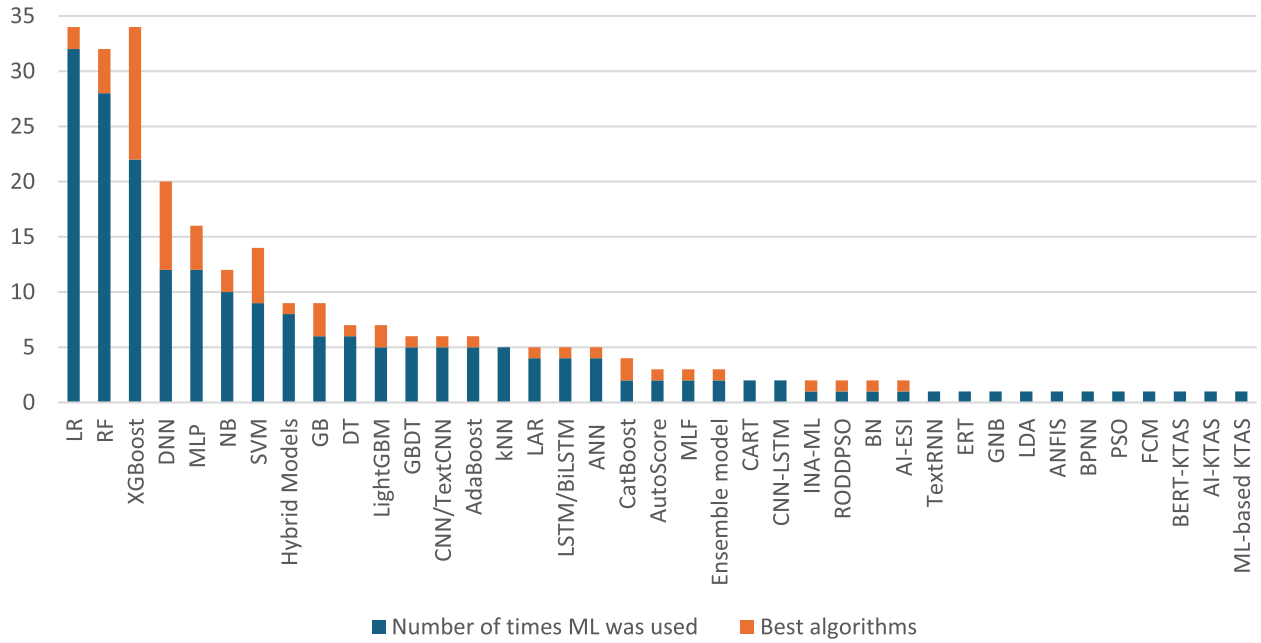

Fig. 3 Frequency of ML algorithms used and best models in included studies

the standardized approach developed by Fernandez-Felix et al. [56]. This assessment examined four key domains: participants, predictors, outcomes, and analysis, with guidelines for implementation and interpretation provided in [88]. The results of this evaluation are presented in Fig. 4 and detailed in Tables A3 and A4.

Of the 60 studies reviewed, 12 were found to have a low risk of bias overall. Specifically, as shown in Fig. 4 (a) and (b), only 10% of the studies exhibited a low risk of bias in patient triage, while a significant 73% presented a high risk of bias in the model analysis domain. Conversely, 80% of the evaluated models showed low risk of bias in the domains of predictors and participants. These findings underscore the presence of a high risk of bias in most prediction studies, emphasizing the need to mitigate bias to improve the reliability of these models in clinical practice.

When examining studies that focused on outcomes such as mortality and ICU admission, 30% demonstrated low risk of bias, and 33% showed low risk in model analysis. Furthermore, in the assessment of the domains related to participants, predictors, and outcomes, 90% of the studies indicated low risk of bias. In terms of model applicability, 87% of the studies were found to have low risk. Overall, these results suggest that models designed to predict mortality and ICU admission were less prone to bias and displayed better applicability compared to those used in patient triage.

Natural language processing

Text processing of clinical notes uses NLP, a subfield of Artificial intelligence (AI) that analyzes human language, including its structure and meaning [6, 42]. Initially, text representation was based on word frequencies in each text fragment, but now algorithms like DNN capture the meaning of words and phrases. They transform words into numerical vectors (embeddings), allowing models to process context and meaning.

Fifteen studies used NLP to process patients' free-text chief complaints. Six studies used complaints recorded by nursing professionals [6, 10, 13, 40, 43, 67], and six used triage medical notes [37, 39, 44, 46, 47, 70]. One study applied NLP to simulated triage dialogues [45], another to diagnostic medical history [41], and one study did not report the unstructured data used in the ChatGPT model [48].

The NLP methods used to transform clinical notes into numerical variables were BoW [6, 43], CBoW – Continuous Bag-of-Words [39], BERT [45–47], ChatGPT [48], Paragraph Vectors [37], Skip-gram [37, 41], TF-IDF [10, 44, 67], and WE [13, 40, 41]. Tang et al. [41] were the first to use WE and Skip-gram to classify patient mortality, comparing eight ML algorithms. Bidirectional Long Short-Term Memory achieved the highest area under the receiver-operating-characteristics curve (AUC) of 0.949 using medical history text transformed into numerical vectors by the Skip-gram model. The Skip-gram model is a technique for learning word embeddings by predicting

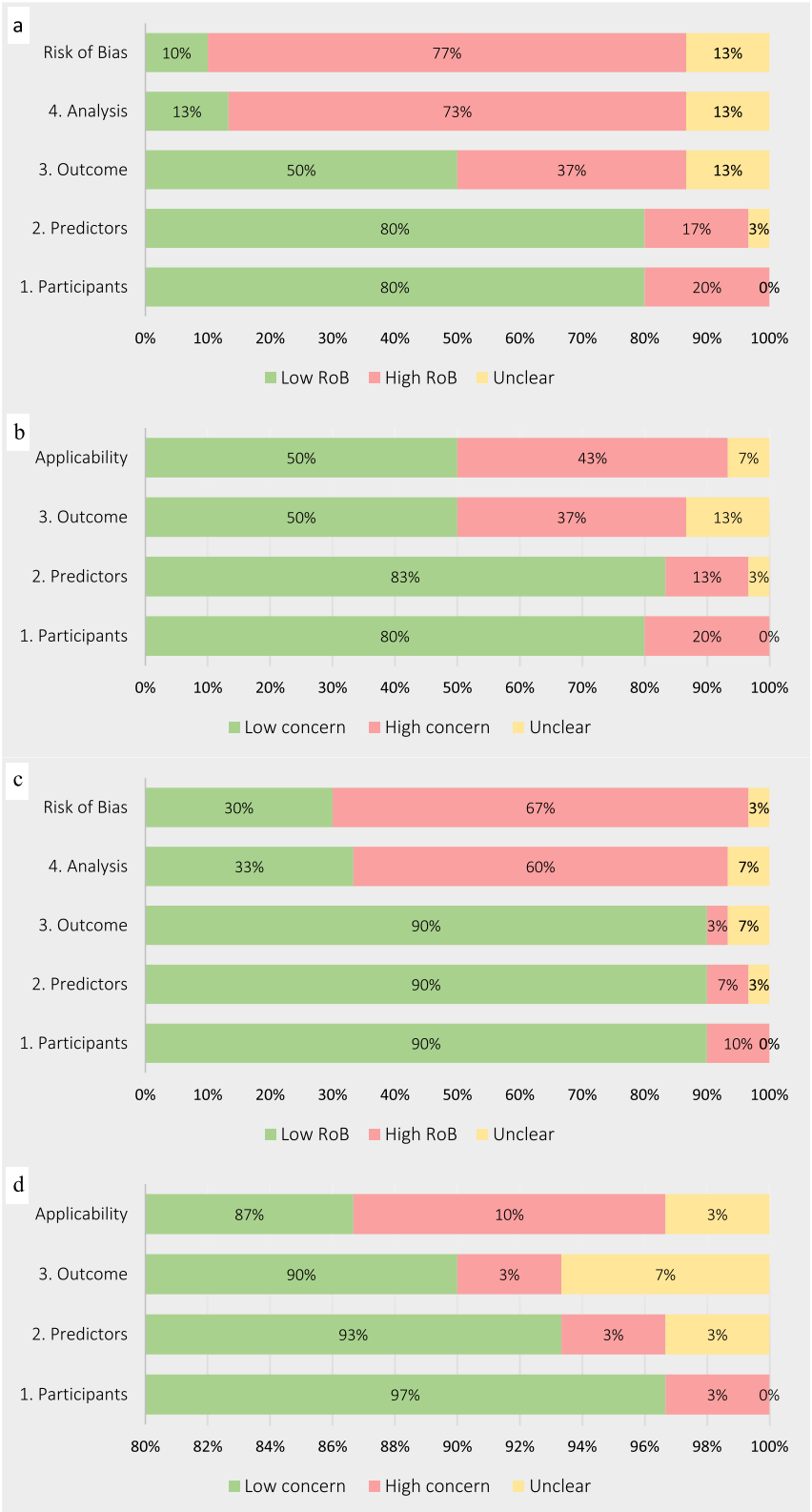

Fig. 4 Summary of PROBAST tool using data from the review of prognostic models for triage-levels, mortality and ICU admission. **a** Summary of risk of bias assessment for triage-levels. **b** Summary of applicability assessment for triage-levels. **c** Summary of risk of bias assessment for mortality and ICU admission. **d** Summary of applicability assessment for mortality and ICU admission

context words surrounding a target word, capturing semantic relationships in a high-dimensional vector space [37, 41].

Choi et al. [6], Klang et al. [43], and Klang et al. [44] used the BoW model to convert triage notes into numerical vectors, along with demographic data and vital signs as predictors. XGBoost performed best in the first two studies, with AUCs of 0.922 for predicting KTAS and 0.93 for ICU admission. In study [44], the Ensemble model achieved an AUC of 0.99 in predicting mortality using clinical data and medical triage notes. The BoW model is a simple technique that represents text as an unordered set of words, capturing word frequency while ignoring grammatical structure and word order, making it useful for tasks like text classification and sentiment analysis [6, 43].

Joseph et al. [13] and Chen et al. [37] used DNNs and WE to classify mortality or ICU admission and Taiwan Triage and Acuity Scale (TTAS) levels, relying on structured data and nursing and medical notes. The notes were transformed with WE and used as predictors in the DNNs, which achieved AUCs of 0.857 for mortality or ICU admission and 0.863 for TTAS. WE represented each word with a single embedding, ignoring contextual information and treating the same word identically, regardless of its surrounding words in the text [40, 41].

Chen et al. [39] used the CBoW model on medical notes, encoded as numerical vectors, in a DNN to predict ICU admission. The DNN with clinical narrative-aware achieved an AUC of 0.874. The CBoW model extends the BoW approach by capturing the meaning of words based on their surrounding context, predicting a target word from its context words to enhance understanding of word meanings in different contexts [39].

Fernandes et al. [10, 67] used LR, XGBoost, and TF-IDF to vectorize patients' chief complaints with the aim of predicting ICU admission and mortality. The ML algorithms achieved high performance by combining structured and unstructured variables, with AUCs of 0.91 for ICU admission and 0.96 for mortality. The TF-IDF is a numerical statistic that measures the importance of a word in a document relative to a corpus by combining term frequency, which counts word occurrences, with inverse document frequency, which assesses how informative a word is across documents [10, 67].

Kim et al. [45] and Wang et al. [46] applied BERT to pre-process texts from dialogues between doctors and patients and medical triage notes of patients' physical conditions. Support Vector Machine and LightGBM were the best classifiers for KTAS and MTS triage systems, with AUCs of 0.90 and 0.81, respectively. BERT is a model pre-trained on large text datasets, to generate unique word vectors based on the surrounding context of each word in a sentence, using attention layers to capture relationships and improve performance in various NLP tasks [45, 46].

Sarbay et al. [48] employed ChatGPT for ESI triage classification using a sample of 50 patients. ChatGPT achieved an AUC of 0.846 for triage-levels 1–2 (high acuity). The ChatGPT model, developed by OpenAI, utilizes the Transformer architecture to analyze patient data and predict triage levels in the ESI system, leveraging extensive pre-training on unstructured text to understand language patterns and generate severity assessments [49, 89].

Xiao et al. [47] employed TransNet with Tokenizer encoding to predict the triage levels of the Chinese Emergency Triage Scale using medical triage notes. The TransNet model vectorized the chief complaints and achieved superior performance, with AUCs of 0.947, 0.906, 0.910, and 0.922 for levels 1 to 4, respectively. NLP methods improved the classification capability of ML algorithms using triage notes text and structured clinical data compared to algorithms that used only structured data. The Tokenizer encoder converts words into unique integer codes to create a sequence of encoded tokens, while also establishing a reverse vocabulary for efficient word-to-code lookup, effectively capturing the semantic and structural information in medical triage notes [47]. Table 6 provides an overview of the NLP methods used in the field.

Feature selection, XAI, feature engineering and resampling techniques

Feature selection or XAI techniques were employed in 63.33% of the studies (Table 5). The most commonly used feature selection methods included RF in 12 studies, XGBoost in 7 (utilizing relative importance, information gain, Gini index, and permutation-based methods), and LR in 8 studies (using lasso regularization, odds ratio, and deviance difference). In 36.67% of the studies, all predictors were included without feature selection.

For XAI, the most frequently used methods were SHAP in 4 studies and LIME in 3. The importance of these model explanation techniques is directly linked to the need for medical specialists to understand and justify ML predictions [52], particularly in patient triage. Given that such predictions can significantly impact diagnostic and treatment decisions, it is crucial for healthcare professionals to both trust the results and comprehend the factors behind patient classification.

In this context, XAI plays a vital role in providing transparency and interpretability to ML models used in triage. It is crucial that ML algorithms used for triage classification are both explainable and transparent, as this helps physicians identify underlying conditions that may

Table 6 Natural language processing models used in patient triage

Ref	Unstructured data	Overview of NLP methods	Tools (Python, R packages)
[41]	Medical history (MH) – diagnostic history	The **Skip-gram** model is a technique for learning word embeddings by forecasting the context words surrounding a target word. Each word in a text corpus is represented as a high-dimensional vector [41]. Skip-gram aims to capture semantic relationships between words by forecasting adjacent context words given a specific target word [41]. During training, the model adjusts word vectors to improve the accuracy of predicting context words. For example, in the sentence "The patient presents chest pain," if "patient" is the target word, Skip-gram predicts "The," "presents," "chest," and "pain" as context words. By iteratively training on a large corpus of text, the model embeds words in a continuous vector space where similar words are positioned closer together	• Python (SciKit-Learn)
[6]	Nursing triage notes (chief complaints)	The **BoW** model is a simple technique that represents a text as an unordered set of words, disregarding grammatical structure and word order. To create a BoW representation, the text is divided into individual words, and then a vector is generated. Each element of this vector corresponds to a unique word in the text, and its value represents the frequency or count of occurrences of that word in the text. BoW captures the presence and frequency of words but loses contextual and word order information. This technique is more useful in NLP tasks such as text classification and sentiment analysis, where grammatical structure may be less important, and the focus is on the keywords present in the text	• Python (Pandas, SciKit-Learn, soynlp libraries)
[10]	Nursing triage notes (chief complaints)	The Term Frequency—Inverse Document Frequency (**TF-IDF**) is a numerical statistic that reflects how important a word is to a document in a collection or corpus [10, 67]. Term frequency (TF) measures how many times each token appears in each observation [10, 67]. Inverse document frequency (IDF) is a measure of how informative a word is [10, 67], e.g., how common or rare the word is across all the observations. If a word appears in all the observations, it might not give that much insight, but if it only appears in some it might help differentiate between observations. TF-IDF is the product of two statistics: the TF and the IDF	• Python (NumPy and Pandas)
[67]	Nursing triage notes (chief complaints)	**TF-IDF**—The model was explained briefly in the previous line	• Python (NumPy and Pandas)
[13]	Nursing triage notes (chief complaints)	Word embedding (**WE**)—The first word embeddings did not take the context of a word into account, i.e. the same word is represented in the same way (i.e. by one embedding) independently on where it appears in the text, i.e. independently on its surrounding words	• Python (SciPy and SciKit-Learn)
[40]	Nursing triage notes (chief complaints)	Word embedding—The model was explained briefly in the previous line	• Python (SciKit-Learn)

Table 6 (continued)

Ref	Unstructured data	Overview of NLP methods	Tools (Python, R packages)
[37]	Text of medical triage notes (chief complaints)	Paragraph vectors (**PV**) is NLP technique aimed at representing paragraphs of text as continuous numerical vectors in a high-dimensional space. They extend the WE algorithm by not only mapping individual words to vectors but also capturing the overall context and meaning of entire paragraphs. During training, the model learns to predict words based on the context of the entire paragraph, adjusting the paragraph vector to maximize prediction accuracy. As a result, similar paragraphs tend to have close vector representations, allowing NLP models to capture semantic and contextual relationships between different parts of the text. This facilitates tasks such as text classification, clustering, and sentiment analysis	• Python (Tensorflow and SciKit-Learn)
[70]	Text of medical records (clinical terms from patient record free text)	In the Clinical Natural Language Processing (C-NLP) method, they process raw text following these steps: sentence tokenization, word tokenization, text normalization, part-of-speech tagging (POS tagging), chunking, and extraction of clinical terms. Steps 1 to 5 are performed using the OpenNLP library from the Apache Software Foundation. The extraction of clinical terms in step 6 involves the following substeps: 1) extract noun phrases from the chunker (step 5); 2) permute the text in each noun phrase, generating all possible word combinations; 3) match the text combinations with a Unified Medical Language System (UMLS) dictionary to extract corresponding clinical terms; and 4) extract the unique UMLS code (concept unique identifier) for each medical term, which is then used as a feature	• Java 8.0, OpenNLP Java library, • Python (Sklearn and SciPy)
[43]	Nursing triage notes (chief complaints)	The Bag-of-Words (**BoW**) was applied to represent the nursing triage notes. In this method, each triage note is viewed as a "bag" containing its constituent words, with no regard to their order within the text [43]. These words are then structured into a table based on their frequency and count [43]. Subsequently, a statistical classifier is trained to categorize each note by analyzing its word frequency and count [43]. BoW approach does not capture word order	• not reported
[44]	Text of medical triage notes (chief complaints)	BOW—The model was explained briefly in the previous line	• Python (libraries not reported)
[45]	Texts from the dialogue between doctors and patients (recorded conversations)	Contextual models learn a different vector for a word based on the current complete sentence or a note context in which the word occurs. These vectors are often referred to as contextual word representations. **BERT** is based on attention layers; the model updates a learned word representation and will pay different attention (i.e. vector) to each word around simultaneously. The model learns general correlation patterns of words in context. The contextual representations produced by BERT improve the overall performance in most downstream applications becoming the dominant approach in the NLP community	• Python (NumPy, SciKit-Learn, and PyTorch)
[46]	Text of medical triage notes of the patient's physical conditions	**BERT**—The model was explained briefly in the previous line	• not reported

Table 6 (continued)

Ref	Unstructured data	Overview of NLP methods	Tools (Python, R packages)
[48]	Not reported	The **ChatGPT** model, developed by OpenAI, represents a specialized application of the Transformer neural network architecture and was used to predicting patient triage levels in the ESI system. ChatGPT is a pre-trained model with a vast amount of unstructured data from OpenAI used to learn language patterns through unsupervised learning techniques. By analyzing textual data from patient records, symptoms, and historical triage assessments, ChatGPT can generate predictions regarding the severity level of a patient's condition	• OpenAI (ChatGPT)
[47]	Text of medical triage notes (chief complaints)	The Tokenizer encoder functions by incorporating each word into a vocabulary and assigning it a unique integer code, thus producing a sequence of encoded tokens [47]. Furthermore, this encoder establishes a reverse vocabulary to facilitate word-to-code lookup [47]. Such an approach effectively captures semantic and structural information inherent in the text of medical triage notes [47]	• Python (Keras)
[39]	Text of medical triage notes (chief complaints)	The Continuous Bag-of-Words (**CBoW**) is an extension of BoW designed to understand the meaning of words within a given context. Unlike BoW models that do not analyze the exact word order in a sentence, CBoW focuses on capturing the overall meaning of words based on their surrounding context. CBoW considers the words surrounding a specific word and attempts to forecasting the central word from that context. For example, if we have the phrase "the patient presents dizziness and shortness of breath," CBoW would try to predict "patient" using the words "presents," "dizziness," "and," and "shortness of breath." This approach is useful for understanding word meanings in different contexts	• Python (Keras)

Abbreviation: Artificial intelligence (*AI*), Bag-of-words (*BoW*), Bidirectional Encoder Representations from Transformers (*BERT*), Chat Generative Pre-trained Transformer (*ChatGPT*), Continuous Bag of Words (*CBoW*), Natural Language Processing (*NLP*), Term Frequency-Inverse Document Frequency (*TF-IDF*), Word embedding (*WE*)

influence the severity assigned to patients [46]. Providing clear explanations enhances trust in the model's predictions, enabling healthcare professionals to make more informed decisions. To achieve this, two XAI (SHAP [9, 36, 45, 71] and LIME [12, 13, 46]) methods were explored in studies, specifically aimed at providing transparency and interpretability in ML models for patient triage.

FE was applied in 28.33% of the studies (Table 5), with one-hot encoding being the most common approach, used in 10 studies. Other techniques, such as principal component analysis, were less frequently employed. However, FE strategies were rarely explored to improve the quality of input data or create new predictors based on clinical domain knowledge, which could enhance ML performance in patient triage.

Class imbalance correction methods were used in 51.66% of the studies. Approximately half of the articles did not apply class balancing techniques, despite the fact that all studies exhibited class imbalance in outcomes related to mortality, ICU admission, and triage levels. The most commonly used methods for correcting class imbalance were bootstrapping (12 studies), SMOTE (8 studies), undersampling (6 studies), and oversampling, stratified sampling, and dropout (each used in 2 studies). Notably, no study combined oversampling of the minority class with undersampling of the majority class, even though previous research [90] suggests that combining these resampling techniques can improve classification performance in ML models. Fig. 5 provides an overview of the methods employed in the ML workflow across the 60 studies.

Modeling variables

Figure 6 provides an overview of the 30 most frequently used and retained predictors across the studies. Among the most common predictors were demographics (such

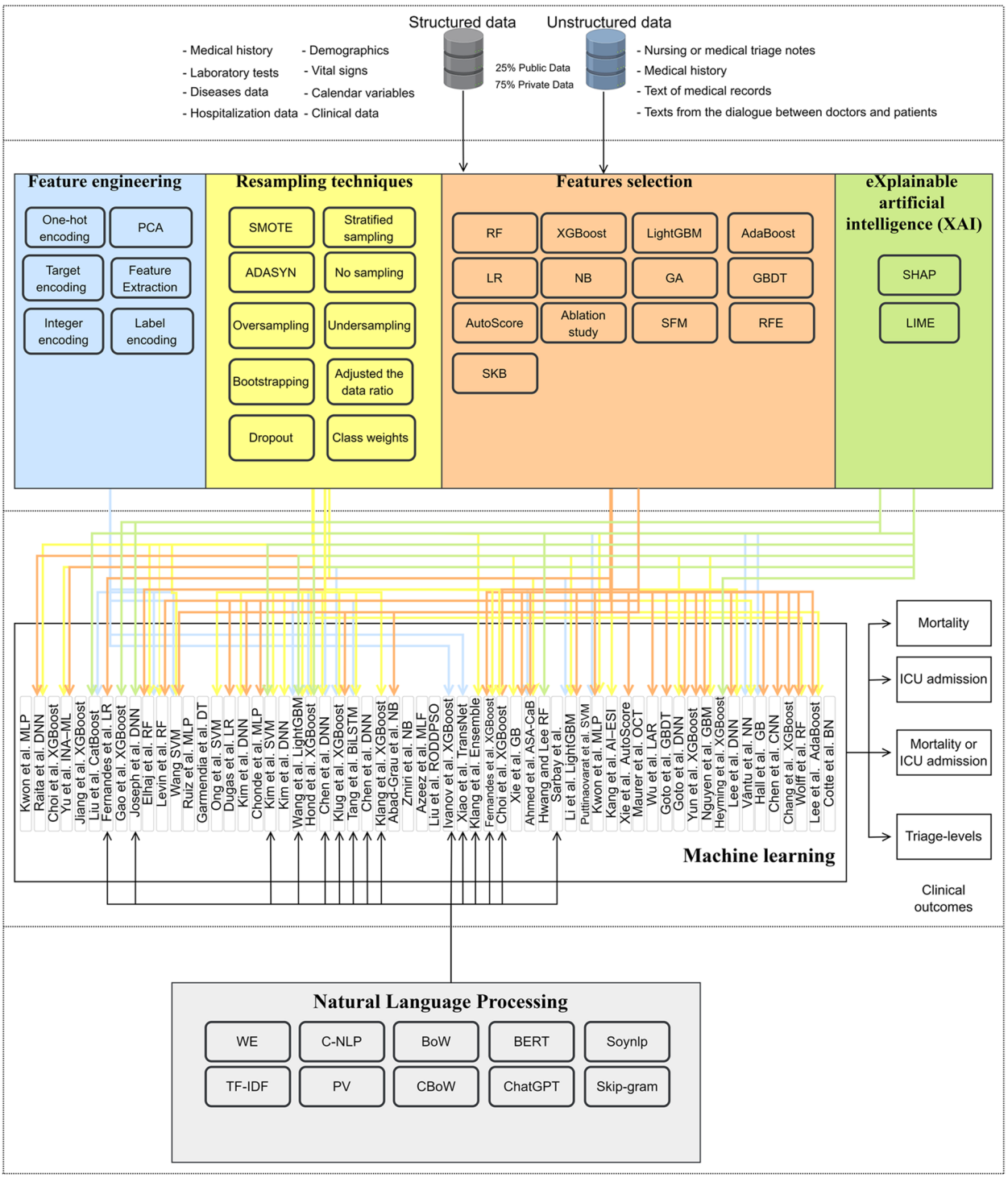

Fig. 5 Overview of the methods used in the ML workflow of the studies

as age, gender, and mode of arrival (MA)), vital signs (including systolic blood pressure (SBP), respiratory rate (RR), body temperature, oxygen saturation (SpO2), diastolic blood pressure (DBP), heart rate (HR), pulse rate, and pain scores (PS)), chief complaints (CC), and triage scores.

Notably, the predictors most frequently retained in the final models were SpO2, nursing/medical triage notes,

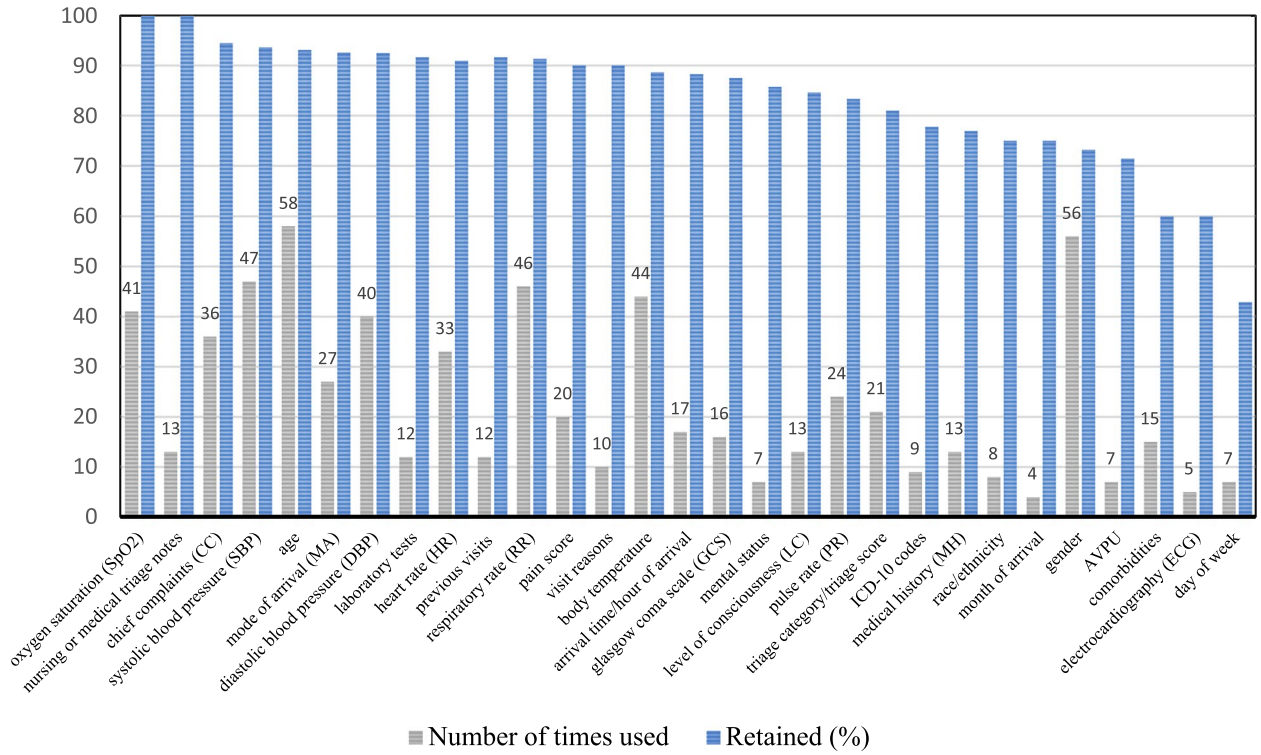

Fig. 6 Summary of the main predictor variables in modeling

CC, SBP, age, MA, DBP, laboratory test results, HR, previous visits, RR, PS, and reasons for visits. These variables are critical in ML models for patient triage because they provide a comprehensive snapshot of a patient's clinical status upon arrival. For example, SpO2 and vital signs are direct indicators of a patient's immediate health condition, helping to identify those who require urgent attention. Additionally, nursing and medical triage notes, along with chief complaints, offer valuable context and nuanced information that ML models can use to refine predictions.

Incorporating these predictors not only enhances the accuracy of triage classification but also improves predictions of critical outcomes like mortality and ICU admission. Since these variables are closely linked to key clinical indicators, their inclusion in models significantly strengthens predictive performance, ultimately supporting better decision-making in emergency care.

Evaluation metrics

The two most commonly used metrics for evaluating ML models were the ROC-AUC and sensitivity, with confidence intervals provided in Fig. 7. Only 23.4% of the studies achieved a ROC-AUC above 0.95, and 12.5% reported sensitivity higher than 0.95. Among the 11 studies (23.4%) that demonstrated the highest ROC-AUC performance, 8 were found to have a high risk of bias in their models (as detailed in Tables A3 and A4). Similarly, of the 5 studies (12.5%) reporting high sensitivity, only Li et al. [68] showed a low risk of bias (Table A3). For instance, Puttinaovarat et al. [74] exhibited a high risk of bias across all domains assessed by the PROBAST tool (Table A3).

The average ROC-AUC and sensitivity for algorithms not utilizing NLP were 0.88 and 0.80, respectively, while models incorporating NLP achieved slightly better performance with an average ROC-AUC of 0.91 and a sensitivity of 0.80. However, model performance varied significantly between studies, with ROC-AUC ranging from 0.66 to 0.99, sensitivity from 0.38 to 0.98, and accuracy from 0.56 to 0.99 (Table 5).

Notably, only 35% of the studies reported ROC-AUC values above 0.90, while the remainder ranged between 0.66 and 0.90. Among the studies with the best performance (ROC-AUC above 0.90), 76% (16 out of 21) utilized either NLP or class imbalance correction techniques. Specifically, 38% (8 out of 21) incorporated NLP, while 62% (13 out of 21) employed class balancing methods. This suggests that the use of unstructured data, such as clinical notes from triage, and the application of class balancing techniques contributed to significant improvements in the performance of ML algorithms. The most

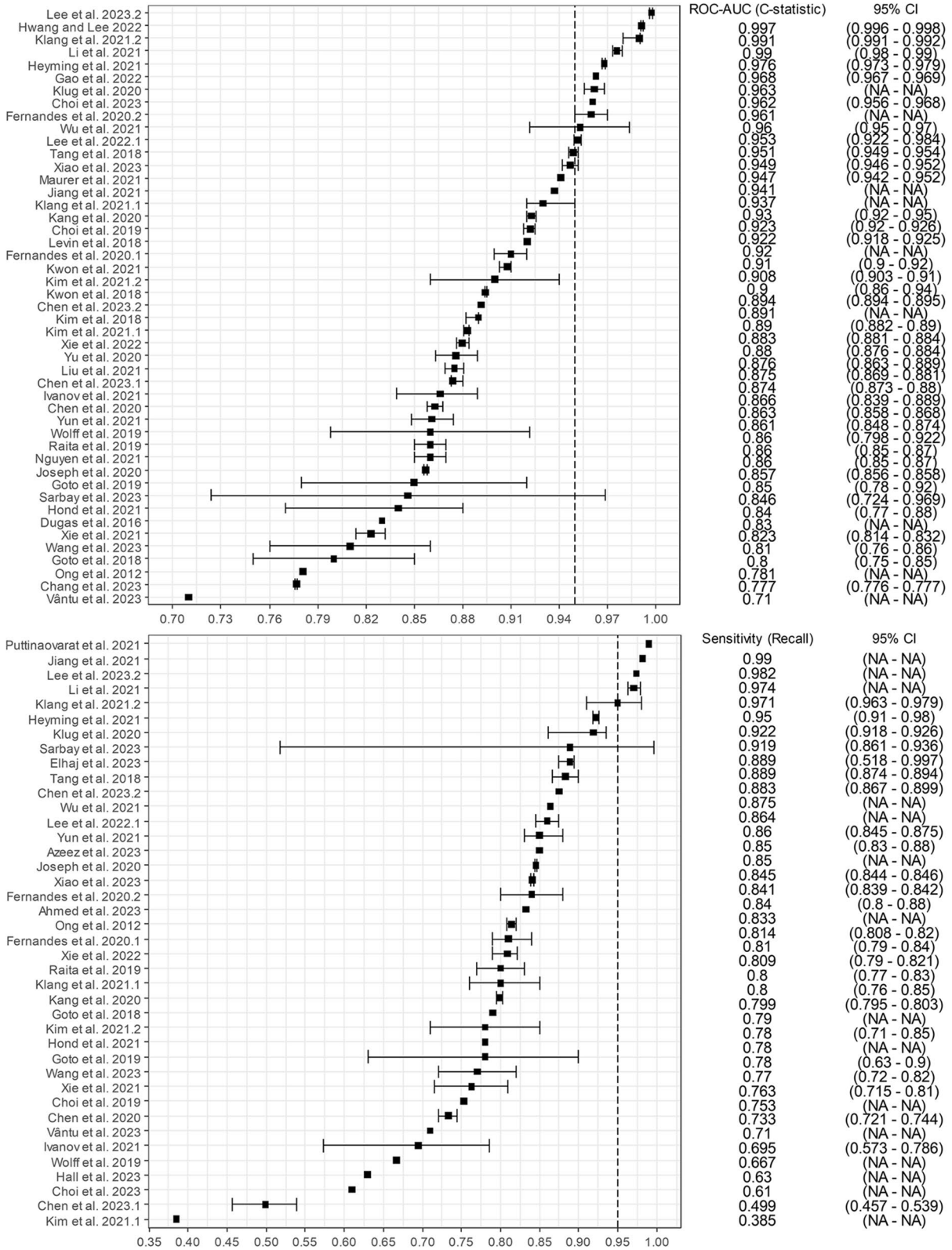

Fig. 7 Forest plot with ROC-AUC and Sensitivity (Recall) values reported in studies

commonly used performance metrics were ROC-AUC, sensitivity, precision, specificity, and accuracy. Future studies should adopt, at a minimum, these five metrics as a benchmark to ensure a more consistent, standardized, and comparable analysis of results.

Discussion

To the best of our knowledge, this is the first study to provide a comprehensive overview of the use of NLP methods to predict patient triage in EDs. Our study discusses ML and/or NLP methods for classifying patient triage in EDs, analyzing 60 studies in total. The literature consistently supports the combined use of ML and NLP methods to aid decision-making by nurses and doctors in triage, reducing wait times and length of stay in the ED, thereby improving the overall flow of emergency services.

The studies included in this review exhibit a wide diversity in methodologies, settings, and objectives related to predictive models in patient triage. Despite the differences, a common point among these studies is the pursuit of greater accuracy in predicting patient outcomes. The studies vary widely in design and scale, ranging from prospective cohorts (e.g., [9, 23, 45, 46, 76, 79]) to retrospective cohorts (e.g., [57–64]), with sample sizes spanning from 124 patients in Abad-Grau et al. [57] to over 10 million in Kwon et al. [3]. This variation in scale impacts the generalizability and robustness of findings, with larger datasets providing more reliable insights but also presenting greater complexity in data handling and analysis.

A variety of triage systems were employed across studies, including the MTS [15, 36, 46, 79], ESI [17, 59, 70, 71], and others such as the Canadian Triage and Acuity Scale [21, 39, 86] and the KTAS [6, 45, 76, 84]. These systems' application and adaptation reflect the specific needs of the healthcare settings in which they were implemented. Studies like those of Dugas et al. [24] and Goto et al. [61] used ESI to classify patients into five levels, while others focused on binary classifications for critical conditions (e.g., Ong et al. [23] and Kim et al. [25]). Studies indicate that algorithms outperform traditional triage systems (human-based), such as KTAS and ESI, in EDs for both adults and children [3, 7, 13, 17, 24, 30, 48, 64, 66, 69, 70, 73, 76, 78, 79].

Research combining ML and NLP suggests these approaches as complementary to traditional systems, due to the lack of empirical evidence from prospective studies, of which only six were identified [9, 23, 45, 46, 76, 79]. In prospective studies, the main benefits identified were: (a) superior performance compared to traditional triage systems currently used in hospitals [23, 46, 76, 79]; (b) reduced variability in risk classification when compared to assessments by nursing professionals; (c) a decrease in both under-triage [9] and over-triage, in contrast to human-based triage systems [47]; (d) the inclusion of variables in ML models that are typically not considered in traditional human-based triage systems [9]; (e) reduced workload for triage physicians in the ED [45]; (f) less time spent on triage evaluation [46]; and (g) a system to enhance clinical decision-making support [2, 46, 80, 84].

Although AI models have great potential to support triage systems, recent studies, such as Zaboli et al., [49] have shown that the current performance of models like ChatGPT is still inferior to that of nursing professionals in critical contexts. The study compared ChatGPT's performance with that of nurses in assigning severity levels in 30 clinical cases. The results showed that the nurses achieved an AUROC of 0.910 (0.757–1.000), while ChatGPT presented an AUROC of 0.669 (0.153–1.000) in predicting 72-h mortality [49], indicating the superiority of human professionals in correctly classifying the most urgent cases in the MTS system.

Of the six prospective studies, only Kim et al. [45] and Wang et al. [46] used NLP (BERT) in triage. This highlights the need for more prospective studies with ML and NLP to enhance triage and evaluate the performance of ML models in EDs. Integrated AI approaches have the potential to improve the accuracy and consistency of triage, as well as reduce human bias; however, further research is needed to validate these benefits in real-world emergency scenarios with prospective studies.

In terms of algorithms, LR is the reference model, while DNNs and decision tree-based algorithms with GB, such as XGBoost and LightGBM, demonstrate superior performance. However, most ML algorithms face limitations in explainability and are considered black boxes. Only 11% of studies used XAI to improve the interpretability of predictor variables (Table 5). Few studies have explored XAI in classification algorithms for triage, an area that requires more attention in future research. The adoption of XAI could increase confidence in the predictions of algorithms and provide insights into predictors that influence clinical outcomes, fostering better integration of AI into medical practice [52].

Feature selection is crucial for faster training, greater accuracy, and easier analysis of the modeled mechanisms [91]. In the ML workflow, it was adopted in 53% of the studies, revealing a methodological flaw in almost half of the research. The absence of this step can result in less efficient models, which are more difficult to interpret, have longer training times, and lower accuracy in predictions.

The most frequently retained predictor variables were SpO2, nursing triage notes, CC, SBP, age, MA, DBP, laboratory tests, HR, previous visits, RR, PS, and visit reasons. It is suggested that these variables be consistently used

in predicting triage-levels, mortality, and ICU admission, guiding data collection in future studies. SpO2 was the most frequently retained predictor in ML models for patient triage prediction [5, 6, 17, 25, 37, 41, 61–63]. SpO2 levels are a critical indicator of a patient's respiratory function [92]. In the ED, vital signs and SpO2 may be the most critical predictive variables for ICU admission [93]. In emergency settings, abnormal SpO2 levels can signal severe respiratory distress or hypoxemia, conditions that require immediate attention [92]. As a result, including SpO2 in ML models for triage enhances the ability to quickly identify patients with potentially life-threatening conditions, improving the accuracy of the triage process [8, 13, 75].

SBP was also a critical vital sign in most ML models for patient triage [9, 15, 69, 82, 83]. SBP serves as a key indicator of a patient's cardiovascular health, and abnormal SBP levels can signal conditions such as shock, hypertension, or hypotension, which require immediate medical intervention [94]. Age is a crucial predictor in most triage systems, as older patients often have more underlying conditions and poorer prognoses compared to younger patients [9]. Age plays a crucial role in determining a patient's risk profile, as it is closely linked to the likelihood of developing certain medical conditions and the severity of illnesses.

Triage notes were identified as key variables in all studies that utilized unstructured data [6, 10, 13, 40, 43, 67] [37, 39, 44, 46, 47, 70]. Medical and nursing notes play a crucial role as they capture detailed and contextual information about the patient's condition, which may not be fully reflected in structured data [39]. These notes often include the patient's symptoms, behavior, and clinical history, providing a more comprehensive picture for ML models to classify patient severity levels [39]. The inclusion of unstructured data, such as triage notes, significantly improved the performance of ML algorithms across all studies reviewed [6, 10, 13, 39, 70]. These notes provided critical information that complemented structured data, enhancing the model's performance. Furthermore, algorithms utilizing both the structured data and the triage notes did better than those which were trained on just the structured data or the triage notes [6, 13, 37, 41, 43, 67, 70].

Most predictive triage models present a high risk of bias, especially in model analysis, although they have a low risk of bias in predictors and participants. In contrast, models for predicting mortality and ICU admission showed lower risk of bias and better applicability. In the risk of bias assessment, ML algorithms generally have a high risk of bias in the modeling process, possibly because the PROBAST tool's risk of bias evaluation is particularly stringent, especially in the analysis of the models. Most ML models exhibited a high risk of bias due to the critical nature of the tool's assessment of predictive models. This highlights the need for future studies to minimize bias to ensure the reliability of models in clinical practice.

The complementary nature of these studies is evident in their collective contribution to understanding ED triage and predictive modeling. The initial studies used Naive Bayes [21, 23, 57, 59, 62] and LR [3, 5, 6, 21, 24, 25, 41, 61, 63] to predict triage levels, with LR serving as a reference model due to its interpretability and transparency, making it more widely used and popular. As studies evolved, models such as SVM [20, 23, 45, 62, 74] and MLP [3, 21, 22, 59, 78] began to be explored, with neural networks achieving better performance in several studies. Decision tree-based models, such as RF [6, 8, 10, 15, 17, 35, 36, 45, 67, 68, 72, 76], GB [72, 78, 86], and XGBoost [12, 15, 39, 43, 47, 70, 77, 81–85], gained prominence, with XGBoost emerging as the top-performing model in several recent studies. More advanced approaches, such as DNN, have shown superior performance in the most recent studies [5, 25, 37, 39, 45, 63, 76, 85], due to their ability to capture complex patterns in large volumes of unstructured data.

This study has several strengths: (i) It is a large-scale systematic review, comprehensively evaluating 60 articles, (ii) our search strategy did not impose date restrictions on the databases, ensuring broad coverage, (iii) we used the PROBAST tool, which assesses specific and relevant criteria for the development of predictive models, not addressed by other tools for evaluating risk of bias and applicability, and (iv) our review is the first to demonstrate how NLP methods have been applied to predict patient triage levels in EDs. In summary, the reviewed studies collectively enhance our understanding of how different triage systems, predictor variables, and ML algorithms can be effectively employed to predict patient outcomes in ED settings. Their findings highlight the importance of a tailored approach, where the choice of triage system, predictors, and algorithms should align with the specific healthcare setting's needs and the clinical objectives at hand.

Gaps and opportunities

Current research on ML algorithms in triage aims to improve classification and support healthcare professionals in prioritizing high-risk patients. Although 57 algorithms have been used, significant gaps remain. Future research directions include:

- There is little evidence from prospective studies, indicating that limited knowledge has been accumulated regarding the applicability of ML models in real-time clinical settings. Prospective validation

of ML and NLP models is necessary to evaluate real-time performance in patient triage. Incorporating chief complaints in free-text format has been shown to improve ML algorithms' performance in predicting triage levels. However, additional prospective validation is required to assess their effectiveness in supporting clinical decision-making in patient triage.

- NLP was used for preprocessing text from nursing and medical notes to utilize unstructured data as inputs for ML algorithms in patient triage prediction. This review concluded that NLP methods improved the classification capabilities of ML algorithms. Therefore, the exploration of new NLP methods, such as the Robustly Optimized BERT Pretraining Approach and ChatGPT, is crucial. The adoption of these methods in patient triage is important because they can handle complex linguistic and contextual nuances, improving accuracy in analyzing unstructured clinical data. These new approaches can enhance the early identification of disease patterns and increase efficiency in classifying and prioritizing patients in EDs.
- The adoption of feature engineering and eXplainable artificial intelligence to enhance both the performance and interpretability of ML models in predicting patient triage remains an underexplored area in the literature. While FE has been relatively overlooked in triage prediction, it holds significant potential to improve ML performance by enabling models to better capture the underlying patterns in the data [95, 96]. Recent advances in FE techniques, such as time series signatures, Fourier transformations, and entity embeddings, could be highly beneficial for generating new predictor variables that enhance the accuracy of patient triage systems. Moreover, integrating XAI into ML models is crucial, especially in patient triage, where understanding the reasoning behind predictions is essential for clinicians to trust and adopt AI-driven systems. XAI can help make the decision-making process of ML models transparent, offering interpretable insights that align with clinical expertise. This not only builds trust among healthcare professionals but also ensures that models can be effectively validated and scrutinized for biases and fairness. Incorporating XAI into patient triage systems could lead to more reliable and understandable AI solutions, ultimately improving patient outcomes while maintaining clinician confidence in the technology. Therefore, future research should prioritize exploring these underutilized areas, as they promise to enhance the practical utility and acceptance of ML models in real-world clinical settings.
- The literature on ML models in ED triage is promising in terms of enhancing the performance of traditional triage systems. However, further studies are needed to address the issue of healthcare professionals' acceptance of these technologies in integrating ML into the triage process, as well as ethical considerations. Future research should explore the long-term impact of implementing ML models in real-world emergency care settings. Identifying barriers and facilitators to acceptance will provide valuable insights for developing strategies that ensure the effective and sustainable implementation of these technologies.

Limitations

This review does not provide a meta-analysis of the evaluated studies due to the significant heterogeneity among the methods, variables, and outcomes reported in the different studies. The approaches used vary widely in terms of ML models, NLP, as well as population characteristics and predictive outcomes. This diversity makes it challenging to quantitatively combine the results into a robust statistical analysis.

Conclusion

A comprehensive systematic review of patient triage prediction using ML and/or NLP is presented. LR is the reference model, while DNN and GB-based algorithms were the best-performing models. ML algorithms showed a high risk of bias in most of the evaluated studies. Standard metrics were identified, and the most important predictors in modeling were noted. The main NLP methods used to predict patient triage, mortality, and ICU admission were summarized and discussed in terms of their results.

Our review suggests that ML models surpass traditional human-based triage systems in classifying triage levels, predicting mortality, and ICU admission. ML models can enhance triage by providing more accurate patient stratification, leading to improved outcomes in predicting mortality and ICU admission. However, adherence to PROBAST guidelines for predictive models is essential to ensure that studies present a low risk of bias.

Unstructured free-text triage notes contain rich contextual information that can capture complex patterns, such as indications of heart disease. This unstructured data can be leveraged by NLP methods to improve the accuracy of patient triage predictions. NLP methods improved the classification of algorithms by utilizing nursing notes, medical notes, and structured clinical data, compared to models that used only structured

data. FE and class balancing methods enhanced the performance of ML algorithms. However, FE and XAI were underexplored approaches in the field. Future studies should consider FE, XAI, and class imbalance correction techniques.

Abbreviations

AI	Artificial intelligence
AUC	Area under the curve
AUPRC	Area Under the Precision and Recall Curve
BiLSTM	Bidirectional Long Short-Term Memory
BoW	Bag-of-Words
BERT	Bidirectional Encoder Representations from Transformers
CBoW	Continuous Bag-of-Words
CatBoost	Categorical Boosting
CC	Chief complaints
ChatGPT	Chat Generative Pre-trained Transformer
DBP	Diastolic blood pressure
DNNs	Deep Neural Networks
ED	Emergency department
ESI	Emergency Severity Index
FE	Feature engineering
GB	Gradient Boosting
HR	Heart rate
ICU	Intensive care unit
KTAS	Korean Triage and Acuity Scale
LIME	Local Interpretable Model-agnostic Explanations
LightGBM	Light Gradient Boosting Machine
LR	Logistic regression
MA	Mode of arrival
ML	Machine learning
MLP	Multilayer Perceptron Neural Network
MTS	Manchester Triage Scale
NLP	Natural language processing
NPV	Negative predictive value
PR	Pulse rate
PRISMA	Preferred reporting items for systematic review and meta-analysis
PROBAST	Prediction model Risk of Bias Assessment Tool
PROSPERO	International Prospective Register of Systematic Reviews
PPV	Positive predictive value
PS	Pain scores
PICO-SD	Participants, Intervention, Comparison, Outcome, Study Design
RF	Random forest
RoBERTa	Robustly Optimized BERT Pretraining Approach
ROC	Receiver operating characteristic
ROC-AUC	Receiver Operating Characteristic-Area Under the Curve
RR	Respiratory rate
SBP	Systolic blood pressure
Se	Sensitivity
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic minority oversampling technique
Sp	Specificity
SpO2	Oxygen saturation
SVM	Support Vector Machine
TF-IDF	Term Frequency—Inverse Document Frequency
TTAS	Taiwan Triage and Acuity Scale
WE	Word embedding
XAI	EXplainable Artificial Intelligence
XGBoost	EXtreme gradient boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12873-024-01135-2.

Supplementary Material 1.

Acknowledgements
I would like to thank Professor Flavio S. Fogliatto for supervising the article.

Authors' contributions
BP: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Project administration and Funding acquisition. The author approved the final version of the manuscript.

Funding
This work was funded by the National Council for Scientific and Technological Development (CNPq), Brazil, through the GD doctoral scholarship [141150/2021–1].

Data availability
The datasets used and/or analyzed during this study are available in the PROSPERO repository and can be accessed online at: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=604529. Additionally, RIS metadata files from the five databases have been publicly provided at: https://new.rayyan.ai/reviews/806188/screening.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Industrial Engineering Department, Federal University of Rio Grande do Sul, Av. Osvaldo Aranha 55, Porto Alegre, RS, Brazil.

Received: 2 September 2024 Accepted: 11 November 2024
Published online: 18 November 2024

References

1. R. Sánchez-Salmerón *et al.*, "Machine learning methods applied to triage in emergency services: A systematic review," *Int. Emerg. Nurs.*, vol. 60, no. August 2021, p. 101109, Jan. 2022, https://doi.org/10.1016/j.ienj.2021.101109.
2. M. Fernandes, S. M. Vieira, F. Leite, C. Palos, S. Finkelstein, and J. M. C. Sousa, "Clinical Decision Support Systems for Triage in the Emergency Department using Intelligent Systems: a Review," Artif. Intell. Med., vol. 102, no. February 2019, p. 101762, Jan. 2020, https://doi.org/10.1016/j.artmed.2019.101762.
3. Kwon J, Lee Y, Lee Y, Lee S, Park H, Park J. Validation of deep-learning-based triage and acuity score using a large national dataset. PLoS ONE. 2018;13(10). https://doi.org/10.1371/journal.pone.0205836.
4. Zaboli A. Establishing a common ground: the future of triage systems. BMC Emerg Med. 2024;24(1):148. https://doi.org/10.1186/s12873-024-01070-2.
5. 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(1):64. https://doi.org/10.1186/s13054-019-2351-7.
6. 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(4):305. https://doi.org/10.4258/hir.2019.25.4.305.
7. Yu JY, Jeong GY, Jeong OS, Chang DK, Cha WC. Machine learning and initial nursing assessment-based triage system for emergency department. Healthc Inform Res. 2020;26(1):13. https://doi.org/10.4258/hir.2020.26.1.13.

8. H. Jiang et al., "Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease," Int. J. Med. Inform., vol. 145, no. October 2020, p. 104326, Jan. 2021, https://doi.org/10.1016/j.ijmedinf.2020.104326.
9. Liu Y, 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. Dec.2021;11(1):24044. https://doi.org/10.1038/s41598-021-03104-2.
10. Fernandes M, et al. Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing. PLoS ONE. 2020;15(3). https://doi.org/10.1371/journal.pone.0229331.
11. Hinson JS, et al. Accuracy of emergency department triage using the emergency severity index and independent predictors of under-triage and over-triage in Brazil: a retrospective cohort analysis. Int J Emerg Med. 2018;11(1):3. https://doi.org/10.1186/s12245-017-0161-8.
12. Z. Gao et al., "Developing and Validating an Emergency Triage Model Using Machine Learning Algorithms with Medical Big Data," Risk Manag. Healthc. Policy, vol. Volume 15, no. July, pp. 1545–1551, Aug. 2022, https://doi.org/10.2147/RMHP.S355176.
13. Joseph JW, et al. Deep-learning approaches to identify critically Ill patients at emergency department triage using limited information. J Am Coll Emerg Physicians Open. 2020;1(5):773–81. https://doi.org/10.1002/emp2.12218.
14. Mistry B, et al. Accuracy and reliability of emergency department triage using the emergency severity index: an international multicenter assessment. Ann Emerg Med. 2018;71(5):581-587.e3. https://doi.org/10.1016/j.annemergmed.2017.09.036.
15. H. Elhaj, N. Achour, M. H. Tania, and K. Aciksari, "A comparative study of supervised machine learning approaches to predict patient triage outcomes in hospital emergency departments," Array, vol. 17, no. October 2022, p. 100281, Mar. 2023, https://doi.org/10.1016/j.array.2023.100281.
16. A. Zaboli et al., "Assessing triage efficiency in Italy: a comparative study using simulated cases among nurses," Intern. Emerg. Med., no. 0123456789, Aug. 2024, https://doi.org/10.1007/s11739-024-03735-z.
17. Levin S, et al. 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(5):565-574.e2. https://doi.org/10.1016/j.annemergmed.2017.08.005.
18. Miles J, Turner J, Jacques R, Williams J, Mason S. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. Diagn Progn Res. 2020;4(1):16. https://doi.org/10.1186/s41512-020-00084-1.
19. Shafaf N, Malek H. Applications of machine learning approaches in emergency medicine; a review article. Arch Acad Emerg Med. 2019;7(1):34. https://doi.org/10.22037/aaem.v7i1.410.
20. Wang S. Construct an optimal triage prediction model: a case study of the emergency department of a teaching hospital in Taiwan. J Med Syst. 2013;37(9968):2–11. https://doi.org/10.1007/s10916-013-9968-x.
21. Ruiz C, Tello I, Yoo SG. Improvement of the triage process using process automatization and machine learning. Int J Appl Eng Res. 2017;12(15):4989–99.
22. Garmendia A, Rios SA, Lopez-Guede JM, Graña M. Triage prediction in pediatric patients with respiratory problems. Neurocomputing. 2019;326–327:161–7. https://doi.org/10.1016/j.neucom.2017.01.122.
23. Ong MEH, et al. 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(3):R108. https://doi.org/10.1186/cc11396.
24. Dugas AF, et al. An electronic emergency triage system to improve patient distribution by critical outcomes. J Emerg Med. Jun.2016;50(6):910–8. https://doi.org/10.1016/j.jemermed.2016.02.026.
25. Kim D, et al. A data-driven artificial intelligence model for remote triage in the prehospital environment. PLoS ONE. 2018;13(10):e0206006. https://doi.org/10.1371/journal.pone.0206006.
26. G. Feretzakis et al., "Using Machine Learning for Predicting the Hospitalization of Emergency Department Patients," in Studies in Health Technology and Informatics, 2022, pp. 405–408. https://doi.org/10.3233/SHTI220751.
27. G. Feretzakis et al., "Prediction of Hospitalization Using Machine Learning for Emergency Department Patients," in Studies in Health Technology and Informatics, 2022, pp. 145–146. https://doi.org/10.3233/SHTI220422.
28. Feretzakis G, et al. Using machine learning techniques to predict hospital admission at the emergency department. J Crit Care Med. 2022;8(2):107–16. https://doi.org/10.2478/jccm-2022-0003.
29. G. Feretzakis et al., "Predicting Hospital Admission for Emergency Department Patients: A Machine Learning Approach," in Studies in Health Technology and Informatics, 2022, pp. 297–300. https://doi.org/10.3233/SHTI210918.
30. Wu TT, Zheng RF, Lin ZZ, Gong HR, Li H. A machine learning model to predict critical care outcomes in patient with chest pain visiting the emergency department. BMC Emerg Med. 2021;21(1):112. https://doi.org/10.1186/s12873-021-00501-8.
31. Zhang L, et al. Prediction of prognosis in elderly patients with sepsis based on machine learning (random survival forest). BMC Emerg Med. 2022;22(1):26. https://doi.org/10.1186/s12873-022-00582-z.
32. Karlsson A, Stassen W, Loutfi A, Wallgren U, Larsson E, Kurland L. Predicting mortality among septic patients presenting to the emergency department–a cross sectional analysis using machine learning. BMC Emerg Med. 2021;21(1):84. https://doi.org/10.1186/s12873-021-00475-7.
33. Niemantsverdriet MSA, et al. A machine learning approach using endpoint adjudication committee labels for the identification of sepsis predictors at the emergency department. BMC Emerg Med. 2022;22(1):208. https://doi.org/10.1186/s12873-022-00764-9.
34. Deina C, Fogliatto FS, da Silveira GJC, Anzanello MJ. Decision analysis framework for predicting no-shows to appointments using machine learning algorithms. BMC Health Serv Res. 2024;24(1):37. https://doi.org/10.1186/s12913-023-10418-6.
35. Kim D, Chae J, Oh Y, Lee J, Kim IY. Automated remote decision-making algorithm as a primary triage system using machine learning techniques. Physiol Meas. 2021;42(2). https://doi.org/10.1088/1361-6579/abe524.
36. De Hond A, et al. Machine learning for developing a prediction model of hospital admission of emergency department patients: Hype or hope? Int J Med Inform. 2021;152:104496. https://doi.org/10.1016/J.IJMEDINF.2021.104496.
37. Chen C-H, Hsieh J-G, Cheng S-L, Lin Y-L, Lin P-H, Jeng J-H. Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data. Int J Med Inform. 2020;139(49). https://doi.org/10.1016/j.ijmedinf.2020.104146.
38. T.-L. Chen, J. C. Chen, W.-H. Chang, W. Tsai, M.-C. Shih, and A. Wildan Nabila, "Imbalanced prediction of emergency department admission using natural language processing and deep neural network," J. Biomed. Inform., vol. 133, no. August, p. 104171, Sep. 2022, https://doi.org/10.1016/j.jbi.2022.104171.
39. M. Chen, T. Huang, T. Chen, P. Boonyarat, and C. Chang, "Clinical narrative-aware deep neural network for emergency department critical outcome prediction," J. Biomed. Inform., vol. 138, no. January, p. 104284, 2023, https://doi.org/10.1016/j.jbi.2023.104284.
40. Klug M, et al. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score. J Gen Intern Med. 2020;35(1):220–7. https://doi.org/10.1007/s11606-019-05512-7.
41. Tang F, Xiao C, Wang F, Zhou J. Predictive modeling in urgent care: a comparative study of machine learning approaches. JAMIA Open. 2018;1(1):87–98. https://doi.org/10.1093/jamiaopen/ooy011.
42. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Informatics Assoc. 2011;18(5):544–51. https://doi.org/10.1136/amiajnl-2011-000464.
43. Klang E, et al. Predicting adult neuroscience intensive care unit admission from emergency department triage using a retrospective, tabular-free text machine learning approach. Sci Rep. 2021;11(1):1381. https://doi.org/10.1038/s41598-021-80985-3.
44. Klang E, et al. A simple free-text-like method for extracting semi-structured data from electronic health records: exemplified in prediction of in-hospital mortality. Big Data Cogn Comput. 2021;5(3):40. https://doi.org/10.3390/bdcc5030040.
45. D. Kim, J. Oh, H. Im, M. Yoon, J. Park, and J. Lee, "Automatic Classification of the Korean Triage Acuity Scale in Simulated Emergency Rooms Using Speech Recognition and Natural Language Processing: a Proof of Concept Study," J. Korean Med. Sci., vol. 36, no. 27, 2021, https://doi.org/10.3346/jkms.2021.36.e175.
46. B. Wang, W. Li, A. Bradlow, E. Bazuaye, and A. T. Y. Chan, "Improving triaging from primary care into secondary care using heterogeneous

data-driven hybrid machine learning," Decis. Support Syst., vol. 166, 2023, https://doi.org/10.1016/j.dss.2022.113899.
47. Y. Xiao, J. Zhang, C. Chi, Y. Ma, and A. Song, "Criticality and clinical department prediction of ED patients using machine learning based on heterogeneous medical data," Comput. Biol. Med., vol. 165, no. July, p. 107390, Oct. 2023, https://doi.org/10.1016/j.compbiomed.2023.107390.
48. Sarbay İ, Berikol G, Özturan İ. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): A preliminary, scenario-based cross-sectional study. Turkish J Emerg Med. 2023;23(3):156. https://doi.org/10.4103/tjem.tjem_79_23.
49. Zaboli A, Brigo F, Sibilio S, Mian M, Turcato G. Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage? Am J Emerg Med. 2024;79:44–7. https://doi.org/10.1016/j.ajem.2024.02.008.
50. F. Gao, B. Boukebous, M. Pozzar, E. Alaoui, B. Sano, and S. Bayat-Makoei, "Predictive Models for Emergency Department Triage using Machine Learning: A Systematic Review," Obstet. Gynecol. Res., vol. 05, no. 02, 2022, https://doi.org/10.26502/ogr085.
51. M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Taylor & Francis Group, 2019. Available: https://bookdown.org/max/FES/
52. A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, no. December 2019, pp. 82–115, Jun. 2020, https://doi.org/10.1016/j.inffus.2019.12.012.
53. Wolff RF, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51. https://doi.org/10.7326/M18-1376.
54. Saaiq M, Ashraf B. Modifying 'Pico' question into 'Picos' model for more robust and reproducible presentation of the methodology employed in a scientific study. World J Plast Surg. 2017;6(3):390–2.
55. M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," Int. J. Surg., vol. 88, no. March, p. 105906, Apr. 2021, https://doi.org/10.1016/j.ijsu.2021.105906.
56. Fernandez-Felix BM, López-Alcalde J, Roqué M, Muriel A, Zamora J. CHARMS and PROBAST at your fingertips: a template for data extraction and risk of bias assessment in systematic reviews of predictive models. BMC Med Res Methodol. 2023;23(1):44. https://doi.org/10.1186/s12874-023-01849-0.
57. Abad-grau M, Ierache J, Cervino C, Sebastiani P. Evolution and challenges in the design of computational systems for triage assistance. J Biomed Inform. 2008;41:432–41. https://doi.org/10.1016/j.jbi.2008.01.007.
58. Zmiri D, Shahar Y, Taieb-Maimon M. Classification of patients by severity grades during triage in the emergency department using data mining methods. J Eval Clin Pract. 2012;18(2):378–88. https://doi.org/10.1111/j.1365-2753.2010.01592.x.
59. Chonde SJ, Ashour OM, Nembhard DA, Kremer GEO. Model comparison in emergency severity index level prediction. Expert Syst Appl. 2013;40(17):6901–9. https://doi.org/10.1016/j.eswa.2013.06.026.
60. Azeez D, Ali MAM, Gan K, Saiboon I. Comparison of adaptive neuro-fuzzy inference system and artificial neutral networks model to categorize patients in the emergency department. Springerplus. 2013;2(416):1–10. https://doi.org/10.1186/2193-1801-2-416.
61. Goto T, Camargo CA, Faridi MK, Yun BJ, Hasegawa K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. Am J Emerg Med. 2018;36(9):1650–4. https://doi.org/10.1016/j.ajem.2018.06.062.
62. 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.
63. 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(1):e186937. https://doi.org/10.1001/jamanetworkopen.2018.6937.
64. Kwon J, Jeon K-H, Lee M, Kim K-H, Park J, Oh B-H. Deep learning algorithm to predict need for critical care in pediatric emergency departments. Pediatr Emerg Care. 2021;37(12):e988-94. https://doi.org/10.1097/PEC.0000000000001858.
65. Liu W, Wang Z, Liu X, Zeng N, Bell D. A novel particle swarm optimization approach for patient clustering from emergency departments. IEEE Trans Evol Comput. 2019;23(4):632–44. https://doi.org/10.1109/TEVC.2018.2878536.
66. Kang D-Y, et al. Artificial intelligence algorithm to predict the need for critical care in prehospital emergency medical services. Scand J Trauma Resusc Emerg Med. 2020;28(1):17. https://doi.org/10.1186/s13049-020-0713-4.
67. Fernandes M, et al. Risk of mortality and cardiopulmonary arrest in critical patients presenting to the emergency department using machine learning and natural language processing. PLoS ONE. 2020;15(4):e0230876. https://doi.org/10.1371/journal.pone.0230876.
68. C. Li et al., "Machine learning based early mortality prediction in the emergency department," Int. J. Med. Inform., vol. 155, no. September, p. 104570, Nov. 2021, https://doi.org/10.1016/j.ijmedinf.2021.104570.
69. Xie F, et al. Development and assessment of an interpretable machine learning triage tool for estimating mortality after emergency admissions. JAMA Netw Open. 2021;4(8):e2118467. https://doi.org/10.1001/jamanetworkopen.2021.18467.
70. Ivanov O, et al. Improving ED emergency severity index acuity assignment using machine learning and clinical natural language processing. J Emerg Nurs. 2021;47(2):265-278.e7. https://doi.org/10.1016/j.jen.2020.11.001.
71. Heyming TW, Knudsen-Robbins C, Feaster W, Ehwerhemuepha L. Criticality index conducted in pediatric emergency department triage. Am J Emerg Med. 2021;48:209–17. https://doi.org/10.1016/j.ajem.2021.05.004.
72. Nguyen M, et al. Developing machine learning models to personalize care levels among emergency room patients for hospital admission. J Am Med Informatics Assoc. 2021;28(11):2423–32. https://doi.org/10.1093/jamia/ocab118.
73. Maurer LR, et al. Trauma outcome predictor: an artificial intelligence interactive smartphone tool to predict outcomes in trauma patients. J Trauma Acute Care Surg. 2021;91(1):93–9. https://doi.org/10.1097/TA.0000000000003158.
74. Puttinaovarat S, Pruitikanee S, Kongcharoen J, Horkaew P. Machine learning based emergency patient classification system. Int J Online Biomed Eng. 2021;17(05):133. https://doi.org/10.3991/ijoe.v17i05.22341.
75. Yun H, Choi J, Park JH. Prediction of critical care outcome for adult patients presenting to emergency department using initial triage information: an XGBoost algorithm analysis. JMIR Med Inform. 2021;9(9):e30770. https://doi.org/10.2196/30770.
76. Hwang S, Lee B. Machine learning-based prediction of critical illness in children visiting the emergency department. PLoS ONE. 2022;17(2):e0264184. https://doi.org/10.1371/journal.pone.0264184.
77. Lee S, Kang WS, Seo S, Kim DW, Ko H. Model for predicting in-hospital mortality of physical trauma patients using artificial intelligence techniques: nationwide population-based study in Korea. J Med Internet Res. 2022;24(12):1–18. https://doi.org/10.2196/43757.
78. Xie F, et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. Sci Data. 2022;9(1):658. https://doi.org/10.1038/s41597-022-01782-9.
79. Cotte F, et al. Safety of triage self-assessment using a symptom assessment app for walk-in patients in the emergency care setting: observational prospective cross-sectional study. JMIR mHealth uHealth. 2022;10(3):e32340. https://doi.org/10.2196/32340.
80. Vântu A, Vasilescu A, Băicoianu A. Medical emergency department triage data processing using a machine-learning solution. Heliyon. 2023;9(8):e18402. https://doi.org/10.1016/j.heliyon.2023.e18402.
81. A. Ahmed, M. Al-Maamari, M. Firouz, and D. Delen, "An Adaptive Simulated Annealing-Based Machine Learning Approach for Developing an E-Triage Tool for Hospital Emergency Operations," Inf. Syst. Front., no. 0123456789, Sep. 2023, https://doi.org/10.1007/s10796-023-10431-4.
82. Lee S, et al. An artificial intelligence model for predicting trauma mortality among emergency department patients in South Korea: retrospective cohort study. J Med Internet Res. 2023;25:e49283. https://doi.org/10.2196/49283.
83. Choi A, et al. Development of a machine learning-based clinical decision support system to predict clinical deterioration in patients visiting the emergency department. Sci Rep. 2023;13(1):8561. https://doi.org/10.1038/s41598-023-35617-3.
84. Chang H, et al. Clinical support system for triage based on federated learning for the Korea triage and acuity scale. Heliyon. 2023;9(8):e19210. https://doi.org/10.1016/j.heliyon.2023.e19210.
85. Chen Y-HJ, et al. An AI-enabled dynamic risk stratification for emergency department patients with ECG and CXR integration. J Med Syst. 2023;47(1):81. https://doi.org/10.1007/s10916-023-01980-x.

86. Hall JN, Galaev R, Gavrilov M, Mondoux S. Development of a machine learning-based acuity score prediction model for virtual care settings. BMC Med Inform Decis Mak. 2023;23(1):200. https://doi.org/10.1186/s12911-023-02307-z.
87. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, vol. 47, no. 4. in Wiley Series in Probability and Statistics, vol. 47. Wiley, 2013. https://doi.org/10.1002/9781118548387.
88. Moons KGM, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. 2019;170(1):W1. https://doi.org/10.7326/M18-1377.
89. Biswas SS. Role of Chat GPT in Public health. Ann Biomed Eng. 2023;51(5):868–9. https://doi.org/10.1007/s10439-023-03172-7.
90. 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.
91. J. Li et al., "Feature Selection: A Data Perspective," ACM Comput. Surv., vol. 50, no. 6, Jan. 2016, https://doi.org/10.1145/3136625.
92. A. Madevska Bogdanova, B. Koteska, T. Vićentić, S. D. Ilić, M. Tomić, and M. Spasenović, "Blood Oxygen Saturation Estimation with Laser-Induced Graphene Respiration Sensor," J. Sensors, vol. 2024, pp. 1–10, Jan. 2024, https://doi.org/10.1155/2024/4696031.
93. Chen Y, et al. Machine learning model identification and prediction of patients' need for icu admission: a systematic review. Am J Emerg Med. 2023;73:166–70. https://doi.org/10.1016/j.ajem.2023.08.043.
94. Razo C, et al. Effects of elevated systolic blood pressure on ischemic heart disease: a burden of proof study. Nat Med. 2022;28(10):2056–65. https://doi.org/10.1038/s41591-022-01974-1.
95. M. Kuhn and K. Johnson, "3.4 Resampling," in Feature Engineering and Selection: A Practical Approach for Predictive Models, Taylor & Francis Group, 2019. Available: https://bookdown.org/max/FES/resampling.html#rolling-origin-forecasting
96. T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. vanden Broucke, "Special issue on feature engineering editorial," Mach. Learn., no. 0123456789, Aug. 2021, https://doi.org/10.1007/s10994-021-06042-2.

Publisher's Note