BMC Emergency Medicine

RESEARCH

Open Access

Using machine learning and natural language processing in triage for prediction of clinical disposition in the emergency department

Yu-Hsin Chang[1], Ying-Chen Lin[2], Fen-Wei Huang[1], Dar-Min Chen[3], Yu-Ting Chung[3], Wei-Kung Chen[1*] and Charles C.N. Wang[4*]

Abstract

Background Accurate triage is required for efficient allocation of resources and to decrease patients' length of stay. Triage decisions are often subjective and vary by provider, leading to patients being over-triaged or under-triaged. This study developed machine learning models that incorporated natural language processing (NLP) to predict patient disposition. The models were assessed by comparing their performance with the judgements of emergency physicians (EPs).

Method This retrospective study obtained data from patients visiting EDs between January 2018 and December 2019. Internal validation data came from China Medical University Hospital (CMUH), while external validation data were obtained from Asia University Hospital (AUH). Nontrauma patients aged ≥ 20 years were included. The models were trained using structured data and unstructured data (free-text notes) processed by NLP. The primary outcome was death in the ED or admission to the intensive care unit, and the secondary outcome was either admission to a general ward or transferal to another hospital. Six machine learning models (CatBoost, Light Gradient Boosting Machine, Logistic Regression, Random Forest, Extremely Randomized Trees, and Gradient Boosting) and one Logistic Regression derived from triage level were developed and evaluated using EPs' predictions as reference.

Result A total of 17,2101 and 41,883 patients were enrolled from CMUH and AUH, respectively. EPs achieved F1 core of 0.361 and 0.498 for the primary and secondary outcomes, respectively. All machine learning models achieved higher F1 scores compared to EPs and Logistic Regression derived from triage level. Random Forest was selected for further evaluation and fine-tuning, because of its robust calibration and predictive performance. In internal validation, it achieved Brier scores of 0.072 and 0.089 for the primary and secondary outcomes, respectively, and 0.076 and 0.095 in external validation. Further analysis revealed that incorporating unstructured data significantly enhanced the model's performance. Threshold adjustments were applied to improve clinical applicability, aiming to balance the trade-off between sensitivity and positive predictive value.

*Correspondence:
Wei-Kung Chen
ercwk@mail.cmuh.org.tw
Charles C.N. Wang
cnwang@asia.edu.tw
Full list of author information is available at the end of the article

Conclusion This study developed and validated machine learning models that integrate structured and unstructured triage data to predict patient dispositions, distinguishing between general ward and critical conditions like ICU admissions and ED deaths. Integrating both structured and unstructured data significantly improved model performance.

Keywords Natural Language Processing, Triage, Emergency Department, Disposition Prediction, Machine Learning

Introduction

The demand for medical resources in emergency departments (EDs) is increasing rapidly in Taiwan. In 2000, approximately 3.3 million individuals visited EDs, and this number had increased to more than 4.3 million in 2019. This increase can be attributed to the expanding elderly population, the implementation of National Health Insurance, and improvements in medical resource accessibility. Efficiently managing undifferentiated patients is critical to improving efficiency and decreasing length of stay in EDs, thereby maximizing the use of limited resources [1, 2].

Triage plays a crucial role as the first step in prioritizing patients visiting Eds, improving patient safety and alleviating overcrowding in the ED [3, 4]. In 2010, Taiwan adopted the five-level Taiwan Triage and Acuity Scale (TTAS), which is a modified version of the Canadian Triage and Acuity Scale [5]. Scores on the TTAS are significantly correlated with hospitalization and medical resource consumption rates [6]. However, subjective clinical judgment based on nurse's experience and external environment during triage contributes to high interrater variation among triage nurses, resulting in some patients being over-triaged or under-triaged [7–9]. Under-triage, or the inability to accurately prioritize patients with severe, urgent conditions, can lead to delays in time-sensitive interventions, increasing the risk of clinical deterioration, morbidity, and mortality [10–12]. Conversely, over-triage results in unnecessary allocation of emergency resources, potentially delaying care for more critically ill patients [13]. Additionally, the largest patient group is triage level 3, but this group shows wide variability, from mild to severe cases, complicating the sorting process.

Artificial intelligence (AI) has been used to predict the prognosis of emergency patients, assisting in clinical decision-making processes across various scenarios [14, 15]. Machine learning models for ED triage have been developed to predict clinical outcomes [16–20]. Machine learning models for ED triage have been trained using the emergency severity index, the Korean Triage and Acuity System, and the Canadian Triage and Acuity Scale. Machine learning models for ED triage primarily focus on structured data, such as age, vital signs, comorbidity, and sex. Free-text notes recorded by experienced nurses during triage provide concise yet crucial information about patients. Several studies have applied natural language processing (NLP) techniques to analyze free-text notes made by nurses. Machine learning models that incorporate both structured and unstructured data have been demonstrated to have the highest prediction performance [21–24]. However, most of the AI models utilizing structured and unstructured information of triage in the literature focus on predicting a single outcome, such as either ICU admission or hospital admission.

In the present study, we aimed to address this gap by developing a model capable of predicting patient disposition to either ward or critical outcome (either ICU admission or death in ED) by using both triage information from the TTAS and bilingual triage notes in a combination of Chinese and English. By leveraging a broader range of patient information, this dual outcome prediction provides a more comprehensive triage decision-making tool, allowing for nuanced patient allocation that meets various levels of care demands within the hospital setting. To evaluate the performance and ensure the generalizability of the model, we employed both internal and external validation methods. We also requested EPs to make disposition decisions based solely on triage data, without access to additional information such as blood tests, imaging, or detailed medical histories—similar to how the AI models operated. This allowed us to directly compare the predictive abilities of the models with those of EPs, using the EPs' judgments as a benchmark to explore the differences in their predictive performance. We anticipate that machine learning models trained with a combination of structured and unstructured data can accurately leverage both types of information, including NLP-processed unstructured data, to predict dispositions.

Method

Study design and participants

This study was retrospective. Data were obtained from two hospitals in Taichung, Taiwan, between January 2018 and December 2019. The data used for establishing the model and performing internal validation were obtained from China Medical University Hospital (CMUH), which is a 1,700-bed, tertiary teaching hospital. Approximately 160,000 patients visit the ED of CMUH annually. The data employed for external validation were obtained

from Asia University Hospital (AUH), which is a regional, acute-care hospital with 482 beds. Approximately 36,000 patients visit the ED of AUH annually. China Medical University Hospital is a high-level emergency hospital equipped with 24/7 consultation services from various specialties, including surgery, internal medicine, orthopedics, neurosurgery, neurology, obstetrics and gynecology, anesthesiology, and pediatrics [25, 26]. It can provide round-the-clock care for acute stroke patients, acute coronary syndrome patients, major trauma cases, high-risk pregnancies, and neonatal care. On the other hand, Asia University Hospital is a mid-level emergency hospital with high-level capabilities in handling acute coronary syndrome [25]. It offers 24/7 consultation services with specialists in internal medicine, surgery, and orthopedics and is also capable of providing 24-hour care for acute coronary syndrome patients [26].

The TTAS triage system consists of 2 main categories: traumatic and nontraumatic. The nontraumatic category is further divided into 13 subcategories that cover 125 chief complaints (pulmonary, cardiovascular, digestive, neurological, musculoskeletal, genitourinary, ophthalmologic, dermatologic, obstetric, psychiatric, substance misuse, general, and ear, nose, and throat–related). The computerized TTAS system determines the appropriate triage level for a patient by considering several factors. These include (a) whether the case involves trauma; (b) the chief complaint reported by the patient; (c) first-order modifiers such as information about mechanism of injury, pain severity, and vital signs, including degree of respiratory distress, systemic blood pressure, diastolic blood pressure, heart rate, consciousness level, and body temperature. If these variables do not provide enough information to determine the triage level accurately, the system uses second-order modifiers.

Patients without trauma aged 20 years or older and presenting at the ED between January 2018 and December 2019 were enrolled. Patients were excluded if they (1) were discharged against the medical advice of the EP, (2) escaped (left after evaluation by doctor but before disposition), (3) left without being seen, (4) cancelled their registration, (5) had an ambiguous disposition, such as being documented as "other disposition", "Receiving treatment at ED", or having an unrecorded disposition or (6) had an out-of-hospital cardiac arrest.

We specifically developed electronic questionnaires for this study and conducted a random selection of 625 patients from the CMUH cohort's testing dataset. As in Supplementary file 1, each of the 25 questionnaires was designed to include triage data for 25 patients. It is important to note that the information provided in these questionnaires remain consistent with the information used for machine learning, encompassing both structured and unstructured data (Supplementary File 1), with the exception of patient sex, which has been removed to mitigate concerns regarding the potential disclosure of personal information. These tailored questionnaires were then privately sent to individual EPs, and each EP was required to complete them independently, basing their predictions solely on the information provided within the questionnaire. Among the 34 attending EPs who were qualified as emergency specialists by the Taiwan Society of Emergency Medicine at CMUH, two EPs directly involved in the study were excluded from participating. From the remaining pool, 28 EPs were randomly selected to answer the questionnaires.

Data collection and processing

Both structured and unstructured information recorded by triage nurses were stored in electronic medical records (EMRs). Structured information on age, sex, body mass index, vital signs, consciousness level, use and type of indwelling tube (e.g., central venous catheter, endotracheal tube, tracheostomy tube, arterial catheter, nasopharyngeal tube, Foley catheter, and drainage tube), whether the patient was transferred and from which facility, mode of arrival, request for an ED bed, comorbidities, pregnancy status, frequency of ED visits (> 2 times a week or > 3 times a month), 72-hour unscheduled returns.

Unstructured data included clinical notes of chief complaint, and the triage dependence. Clinical notes for chief complaints were written in short sentences or words in both Chinese and English. While measuring vital signs during triage, nurse gathers information from the patient or, if needed, from accompanying family members or friends. Examples of clinical notes include statements like "abdominal pain and diarrhea started a few days ago," "redness and pus in both hips," and "generalized body pain, facial droop since a few days ago, and lower limb weakness after getting up at around 6 AM". Based on the gathered information, the nurse selects the appropriate category from a computerized triage classification system to determine the patient's final triage level. Triage dependence involves triage nurses quickly generating specific descriptive phrases by selecting options from a computerized list, covering the patient's system classification, main symptoms, and key findings, including specific vital signs or pain scores. For example, "patient belongs to the nervous system category, presenting with dizziness/vertigo, positional, without other neurological symptoms," or "patient belongs to the respiratory system category, presenting with shortness of breath, mild respiratory distress (SpO_2: 92–94%)." This process directly correlates with the determination of the triage level.

Final dispositions (e.g., admitted to intensive care unit (ICU) or ward, discharged, discharged against medical advice, expired, or escaped) were also recorded in EMRs. We handled categorical variables by converting them using one-hot encoding. This approach ensures that the categorical data are represented in a binary format, suitable for input into the machine learning models.

Natural language processing

NLP is increasingly being used in the health care sector. In NLP, sophisticated algorithms and machine learning techniques are used to search, analyze, and interpret massive volumes of patient data and to extract valuable insights and meaningful concepts from clinical notes that were previously considered lost due to the textual nature of the data [27].

We processed the unstructured data using a series of NLP techniques. First, we performed data cleansing, which involved removing irrelevant information, standardizing formats, and handling missing or noisy data. This step included removing stopwords, punctuation, and irrelevant characters, as well as performing tokenization and lemmatization to standardize the text. Chinese word segmentation was performed in Jieba in accurate mode [28]. To improve the accuracy of the segmentation, we incorporated a specialized medical dictionary containing relevant medical terms (e.g., disease names, symptoms, and treatments) into Jieba, ensuring the accurate segmentation of medical terminology. The key variables in this stage were the words, terms, and segmented phrases, which represented important clinical terminology. which represented important clinical terms. The detailed list of each Chinese term and phrase extracted from unstructured data, along with their English translations, is provided in Supplementary Table 1. These variables were then encoded using one-hot encoding, where each word in the vocabulary was represented by a binary vector with a '1' indicating the presence of that word in the text and a '0' indicating its absence. This transformed the textual data into numerical vectors suitable for model input.

Additionally, we integrated structured data (e.g., patient demographics, lab results) with the encoded text features. These structured variables were preprocessed as follows: numerical variables (e.g., age, lab test results) were normalized, while categorical variables (e.g., gender, diagnosis category) were encoded using one-hot encoding. We concatenated the encoded text features and structured data into a single feature vector, which was then used as input to our machine learning model.

Training pipeline

The model training process is illustrated in Fig. 1. To conduct internal validation, we partitioned the CMUH

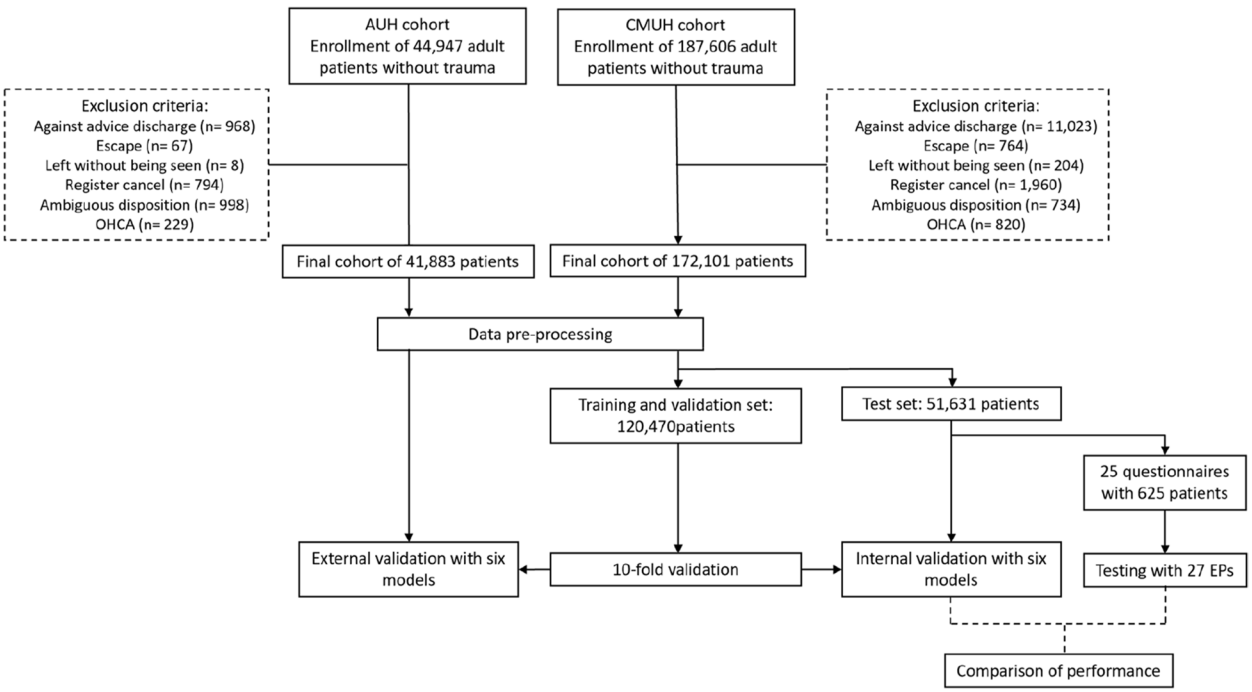

Fig. 1 Flow chart of patient enrollment and model establishment at AUH and CMUH. AUH, Asian University Hospital; CMUH, China Medical University Hospital; EP emergency physician; OHCA out-of-hospital cardiac arrest

cohort, allocating 70% for training and 30% for testing purposes. Within the training data set, we employed 10-fold cross validation to evaluate the model's performance. For missing or outlier values, we used MICE (Multiple Imputation by Chained Equations) for imputation in both the test set and external validation cohort. This method iteratively imputes missing values by creating multiple predictions based on other variables in the dataset. Each missing value is estimated by drawing from a distribution that considers the relationships between variables, improving robustness and minimizing potential bias introduced by simpler imputation methods [29]. Outlier values were defined as follows: systemic blood pressure > 300 mmHg or < 30 mmHg, diastolic blood pressure > 300 mmHg, systemic blood pressure lower than diastolic blood pressure, heart rate > 300 beats/min or < 20 beats/min, respiratory rate > 60 breaths/min, body temperature > 45°C or < 30°C, and body mass index > 150 kg/m² or < 5 kg/m². Regarding to class imbalance, we did not apply resampling methods. Instead, we adjusted sample weights according to each model to achieve comparable recall and specificity across the models for comparison purposes. The testing data set served to internally validate the previously trained models, with the true outcomes blinded during this evaluation phase to prevent any potential bias. AUH cohort was used for external validation to evaluate the model's generalizability.

We constructed multiclass classifiers (3 classes) using the following 6 commonly employed machine learning models: Categorical Boosting (CatBoost), Light Gradient Boosting Machine (LGB), Logistic Regression (LR), Random Forest, Extremely Randomized Trees, and Gradient Boosting (GB) [30–35]. These models were chosen because they represent a diverse range of machine learning techniques, including decision trees (Random Forest, Extremely Randomized Trees), boosting methods (CatBoost, LGB, Gradient Boosting), and linear models (LR), allowing us to compare the performance of different approaches on the same dataset to identify the best-performing model. These models are open source and are widely used in machine learning models for predicting medical issues [36–41]. Besides, a logistic regression (LR) model was trained using only the TTAS level (LR-TTAS model) and was compared with other models and EPs. The model outputs are transformed into predicted classes using the Argmax function, which selects the class with the highest predicted probability as the final prediction. The model with the best overall performance in both discrimination and calibration was selected and then used for fine tuning and assessing the influence of various types of input data (structured and unstructured) on performance.

We employed a wrapper method for feature selection to identify relevant features, reducing dimensionality and enhancing model efficiency by eliminating redundant data, minimizing overfitting, and decreasing computational costs. This was followed by random search for hyperparameter fine-tuning, utilizing 5-fold cross-validation. Random search offers advantages over grid search by exploring a broader range of hyperparameter combinations in less time. Random search allows for a more efficient and often more effective sampling of the hyperparameter space, increasing the likelihood of finding near-optimal configurations without exhaustive testing [42].

Analysis

The performance of each model was assessed by calculating the area under the receiver operating characteristic curve (AUROC) using a one-versus-rest approach. Additional metrics included F-1 score, accuracy, sensitivity, specificity, and positive and negative predictive values. To further ensure the objectivity of the evaluation, the analysis was conducted using automated processes to minimize subjective influence. Additionally, we incorporated the DeLong test to evaluate the significance of performance differences between the models [43, 44]. We have also included the calibration plots and Brier scores to evaluate the performance. Calibration plots were employed to assess how well the predicted probabilities align with the actual outcomes, providing a visual representation of the model's calibration. The Brier score was used as a quantitative measure to evaluate the accuracy of the probabilistic predictions, with lower scores indicating better overall model performance, incorporating both discrimination and calibration [45].

Based on the sample size calculation, selecting 64 patients would achieve the required statistical power (Power = 0.80), significance level (alpha = 0.05), and a medium effect size (Effect size = 0.5) for a population of 172,101. However, we chose 625 patients and distributed their data across 25 questionnaires. This approach ensures a representative patient population in the study questionnaires while also reducing the burden on respondents, thereby improving the quality and completeness of the questionnaire responses. Regarding the performance of EPs, since physicians were instructed to predict categories (discharge, admission to ward, or admission to ICU) rather than probabilities or risks, AUROC is not suitable for evaluation. Therefore, other metrics mentioned above will be used for assessment.

In addition to using importance rankings, we applied SHAP (SHapley Additive exPlanations) values to

enhance the interpretability of our model [46]. SHAP values clarify each feature's contribution to predictions, with the summary plot illustrating how feature variations influence model outcomes. This approach bridges medicine and data science by providing the explanation behind the model's outputs, allowing us to verify that the model's operation aligns with clinical practice and knowledge [47].

Outcomes

The primary outcome was critical disposition, which was death in ED (either in-hospital cardiac arrest in the ED with/without return of spontaneous circulation or critical discharge) or direct ICU admission. The secondary outcome was general ward admission or transferal to another hospital.

Result

A total of 187,606 nontrauma patients aged 20 years or older visited the ED of the CMUH during the study period. After exclusions, the final cohort consisted of 172,101 patients (Fig. 1). The patients were randomly divided into 2 groups: the training group containing 120,470 patients, and the testing group comprising 51,631 patients. Next, 25 questionnaires

Table 1 Demographic characteristics of patients who visited the ED of CMUH ($N = 172101$)

Variables		Variables	
Age, mean ± SD	52.32 (20.13)	Comorbidity, No. (%)	
Sex-female, No. (%)	91961 (54.11%)	Diabetes mellitus	28693 (16.67)
Body mass index, mean ± SD	24.43 (35.37)	Hypertension	46255 (26.88)
Vital signs, mean ± SD		Congestive heart failure	1851 (1.08)
Respiratory rate (per min)	20.29 (2.12)	Ischemic heart disease	5115 (2.97)
SBP (mmHg)	135.21 (25.77)	End-stage renal disease	5572 (3.24)
DBP (mmHg)	84.59 (17.01)	Liver cirrhosis	2834 (1.65)
Heart rate (bpm),	91.86 (21.05)	COPD	1793 (1.04)
Shock index	0.71 (0.35)	Malignancy	18361 (10.67)
Body temperature (C)	36.96 (2.17)	Pregnancy, No. (%)	2554 (1.48)
Consciousness, No. (%)		Intensive ED visits, No. (%)	
Alert	7173 (4.17)	Over twice in a week	12986 (7.55)
Non-alert	164928 (95.83)	Over 3 times in a month	7742 (4.50)
Indwelling tube, No. (%)		72-hour ED return, No. (%)	3181 (1.87)
CVC	52 (0.03)	24-hour ED return, No. (%)	996 (0.58)
Endo	721 (0.42)	System of complaints, No. (%)	
Tracheostomy	1133 (0.67)	Gastrointestinal-related	48565 (28.22)
A-line	13 (0.01)	Neurological-related	29124 (16.92)
Drainage tube	232 (0.14)	General and others	25337 (14.72)
Nasogastric tube	3441 (2.02)	Cardiovascular-related	20488 (11.90)
FOLEY catheter	2392 (1.41)	Respiratory-related	14318 (8.32)
Transferred, No. (%)	18616 (10.82)	Urological-related	9875 (5.74)
Usage of ambulance, No. (%)	21427 (12.45)	Dermatologic-related	9081 (5.28)
Triage, No. (%)		Musculoskeletal-related	5735 (3.33)
1	8061 (4.68)	Ear, Nose, and Throat	4936 (2.87)
2	35050 (20.37)	Other	4643 (2.70)
3	115202 (66.94)	Disposition, No. (%)	
4	12925 (7.51)	Primary outcome[a]	6427 (3.73)
5	863 (0.50)	Secondary admission[b]	39796 (23.12)
Request for an ED Bed, No (%)	30325 (17.62)	Discharge	125878 (73.14)
Fever, No. (%)	32210 (18.72)		

CMUH China Medical University Hospital, *COPD* Chronic obstructive pulmonary disease, *CVC* Central venous catheter, *DBP* Diastolic blood pressure, *ED* Emergency department, *SBP* Systolic blood pressure, *SD* Standard deviation

[a] Admission to intensive care unit or in-hospital cardiac arrest in emergency department

[b] Admission to general ward or transfer to other hospital

were sent to 28 EPs, and 27 complete questionnaires were returned. The median post-residency clinical experience of the EPs was 8 years (interquartile range: 5.25–10.75 years). The patients' demographic characteristics are listed in Table 1. The primary outcome was observed in 6,427 patients (3.73%), and the secondary outcome was observed in 39,796 patients (23.1%). Table 2 shows that the proportion of patients with primary or secondary outcomes was inversely related to triage levels, with the highest proportions of both outcomes observed at TTAS level 1. Conversely, discharge rates rose as TTAS levels increased. Notably, the proportion of deaths in the ED was remarkably low, remaining below 1% across all TTAS levels except for level 1, where it reached 2.91%.

A total of 44,947 nontrauma patients were identified from AUH, and after exclusions, we enrolled 41,883 patients with a mean age of 53.1 (Fig. 1). In this cohort, the proportion of patients at triage levels 1 and 2 was lower than at CMUH, while the proportion of less urgent patients (those at levels 4 and 5) was higher (Supplementary Table 2). Additionally, the proportions of patients with endotracheal tubes and tracheostomies were lower, and no patients were admitted with arterial lines or central venous catheters. The rate of primary outcome was also lower at AUH. The gastrointestinal, neurologic, general, cardiovascular, and respiratory-related systems were the most common systems of complaint at both CMUH and ANH. As shown in Supplementary Table 3, the missing and outlier values in both CMUH and AUH datasets are primarily concentrated in the vital signs categories, with rates not exceeding 0.5%. The only exception is the BMI in the AUH dataset, which has a high missing rate of 33.9%.

Model performance

In internal validation, the performance of each machine learning model is summarized in Table 3. For the primary outcome, although the boosting models demonstrated significantly better AUROC (Supplementary Table 4) and F1 scores for predicting the primary outcome compared to other models, Fig. 2A indicates that these three models (LGBM, CatBoost and GB) exhibit a notable overestimation of risk in their calibration. In contrast, while Random Forest did not achieve the highest AUROC, it yielded the highest F1 score of 0.500 and the lowest Brier score of 0.072. Besides, most models achieved higher F1 scores compared to the F1 score of 0.361 for EPs. Given the lower prevalence of the primary outcome, most models demonstrated higher specificity and negative predictive value. For the secondary outcome, compared to EPs and LR-TTAS, all models improved performance with Random Forest standing out with the highest AUC if 0.847 and the lowest Brier score of 0.089. LR-TTAS showed the poorest performance with a F1 score of 0 in the primary outcome and 0.171 in the secondary outcome. As shown in Supplementary Table 4, the DeLong test indicated that the AUROC of Random Forest was significantly higher than that of the other models. As seen in Fig. 2B, LR exhibits the largest deviation from perfect calibration, consistently overestimating risk across the probability range.

The results of external validation are shown in Table 4, and overall performance, such as sensitivity and specificity, was lower than in internal validation. For the primary outcome, although AUROC of Random Forest was not outstanding and was lower than that of the boosting models and even logistic regression (as shown in Supplementary Table 4), it achieved the highest F1 score of 0.420, along with robust specificity and NPV. Notably,

Table 2 Distribution of disposition by TTAS level in the CMUH dataset (Total patients= 172101)

	TTAS				
	Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Discharge	1028 (12.8)	17736 (50.6)	93926 (81.5)	12337 (95.5)	851 (98.6)
Primary outcome	2888 (35.8)	2664 (7.6)	854 (0.7)	19 (0.2)	2 (0.2)
Admission to ICU	2710 (33.6)	2618 (7.5)	836 (0.7)	19 (0.2)	2 (0.2)
Death in ED[a]	235 (2.91)	88 (0.25)	31 (0.03)	0 (0.0)	0 (0.0)
Secondary outcome	4145 (51.4)	14650 (41.8)	20422 (17.7)	569 (4.4)	10 (1.2)
Admission to ward	4060 (50.4)	14467 (41.3)	20062 (17.4)	554 (4.3)	10 (1.2)
Transfer to other hospital	85 (1.0)	183 (0.5)	360 (0.3)	15 (0.1)	0 (0.0)
Total	8061	35,050	115,202	12,925	863

CMUH China Medical University Hospital, *ED* Emergency department, *ICU* Intensive care unit, *TTAS* Taiwan Triage and Acuity Scale

[a] Death in ED includes patients who died despite resuscitation efforts in the ED or those were not resuscitated because of having a signed Do Not Resuscitate order

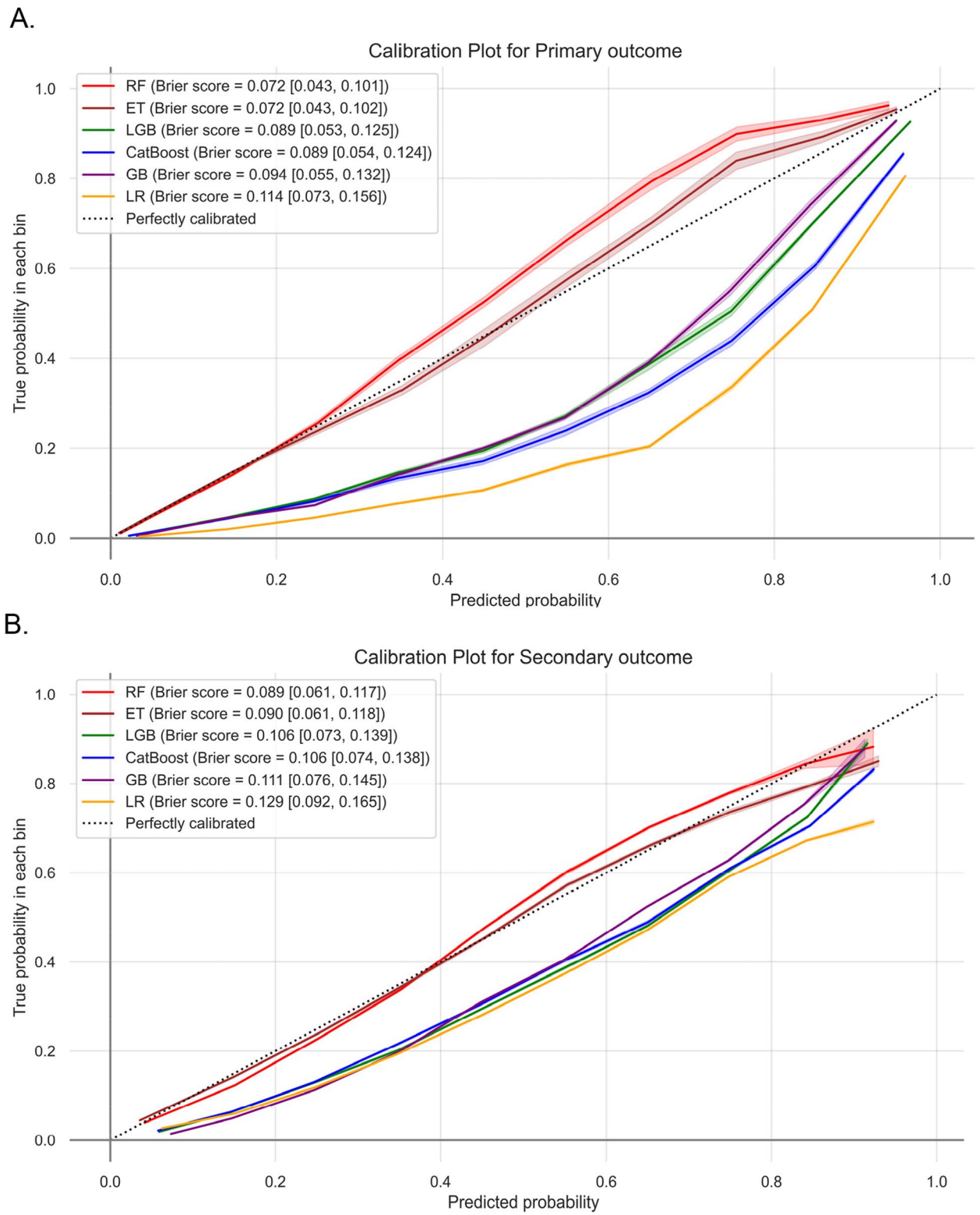

Fig. 2 This figure presents calibration plots for multiple predictive models assessing their performance in predicting (**A**) primary and (**B**) secondary outcomes in the CMUH test set. Each model's Brier score, indicating the accuracy of probabilistic predictions, is displayed along with 95% CI. The shaded areas represent the 95% CI for each model's calibration curve. CatBoost, Categorical Boosting; CI, confidence intervals; CMUH, China Medical University Hospital; ET, Extremely Randomized Trees; GB, Gradient Boosting; LGB, Light Gradient Boost Machine; LR, Logistic Regression; RF, Random Forest; LR-TTAS, Logistic Regression-Taiwan Triage Acuity scale

Table 3 Prediction ability of seven machine learning models and reference for internal validation at CMUH

	AUROC	*p*-value*	BS	F1 score	Accuracy	Sensitivity	PPV	Specificity	NPV
Primary outcome									
LGB	0.937	-	0.089	0.494	0.954	0.589	0.425	0.968	0.983
CatBoost	0.935	0.06	0.089	0.479	0.950	0.602	0.398	0.964	0.984
GB	0.932	0.06	0.094	0.479	0.951	0.584	0.406	0.966	0.983
LR	0.928	< 0.05	0.114	0.415	0.924	0.703	0.294	0.933	0.987
RF	0.926	0.50	0.072	0.500	0.955	0.591	0.433	0.969	0.983
ET	0.919	< 0.05	0.072	0.484	0.953	0.575	0.417	0.968	0.983
LR-TTAS	0.858	< 0.05	0.095	0.000	0.962	0.000	0.000	1.000	0.962
EPs	N/A	N/A	N/A	0.361	0.942	0.550	0.268	0.954	0.986
Secondary outcome									
RF	0.847	-	0.089	0.580	0.782	0.655	0.519	0.820	0.889
ET	0.843	< 0.05	0.089	0.575	0.782	0.643	0.521	0.824	0.886
LGB	0.841	0.06	0.106	0.600	0.786	0.701	0.525	0.811	0.901
CatBoost	0.839	< 0.05	0.106	0.600	0.791	0.682	0.535	0.824	0.897
GB	0.829	< 0.05	0.111	0.586	0.782	0.675	0.518	0.814	0.894
LR	0.800	< 0.05	0.129	0.553	0.758	0.654	0.480	0.789	0.885
LR-TTAS	0.666	< 0.05	0.117	0.171	0.771	0.103	0.499	0.969	0.784
EPs	N/A	N/A	N/A	0.498	0.674	0.626	0.413	0.691	0.842

AUROC Area under receiver-operator curve, *BS* Brier score, *CMUH* China Medical University Hospital, *EPs* Emergency physicians, *ET* Extremely Randomized Trees, *GB* Gradient Boosting, *LGB* Light Gradient Boosting Machine, *LR* Logistic Regression, *LR-TTAS* Logistic Regression-Taiwan Triage Acuity scale, *NPV* Negative predictive value, *PPV* Positive predictive value, *RF* Random Forest

*The *p*-values in the table were derived using the DeLong test, which involved pairwise comparisons of each model's AUROC with that of the model ranked directly above it in performance

it also recorded the lowest Brier score among all models. Figure 3A shows that in the AUH primary outcome calibration, Random Forest and Extremely Randomized Trees have the closest calibration to the ideal line. In contrast, other models, such as Logistic Regression, tend to shift more toward the overestimation quadrant, partly reflected by their higher Brier scores. In the secondary outcome, LGB and CatBoost models reached the highest AUROC and F1 score; however, Fig. 3B displayed a more pronounced overestimation risk of these boosting models, leading to less favorable Brier scores than Random Forest and Extremely Randomized Tree. Compared to the other models, the LR-TTAS model also had the lowest AUROC and F1 score for both primary and secondary outcome.

After using a comprehensive evaluation that included Brier Score, AUROC, and F1 score, we selected Random Forest as the model for further analysis. Although it does not have the highest AUROC, Random Forest demonstrated strong F1 scores in both internal and external validation, and its calibration—except for the ET model—outperformed the other models. Therefore, considering both reliable calibration and discrimination, Random Forest achieved the best overall performance.

We further analyzed how different feature types impact the performance of the Random Forest model. As shown in Fig. 4, combining the unstructured and structured set improves model performance for both primary and secondary outcomes. Table 5 provides detailed results, showing that using all features significantly increases the AUROC, as confirmed by the DeLong test. For example, the AUROC for the secondary outcome increased from 0.761 to 0.847, and the F1 score improved from 0.478 to 0.580. While the improvements in NPV and specificity for predicting the primary outcome were relatively modest, other metrics such as F1 score, sensitivity, and PPV showed substantial enhancement.

Through feature selection, we eliminated 64 features, including those processed through NLP, resulting in a final set of 103 features. Additionally, we fine-tuned the hyperparameters using random search, with the results presented in Supplementary Table 5. The optimal values for hyperparameters of Random Forest model included a maximum depth of 19, a minimum sample leaf of 2, a minimum sample split of 7, and 178 estimators. In Table 6, we adjusted thresholds across three categories to simulate practical clinical applications and evaluate predictive outcomes. The optimal threshold was identified based on Youden's J Statistic, aiming to maximizes the difference between the True Positive Rate (sensitivity) and the False Positive Rate (1-specificity). However, a lower threshold for primary outcomes led to a high rate

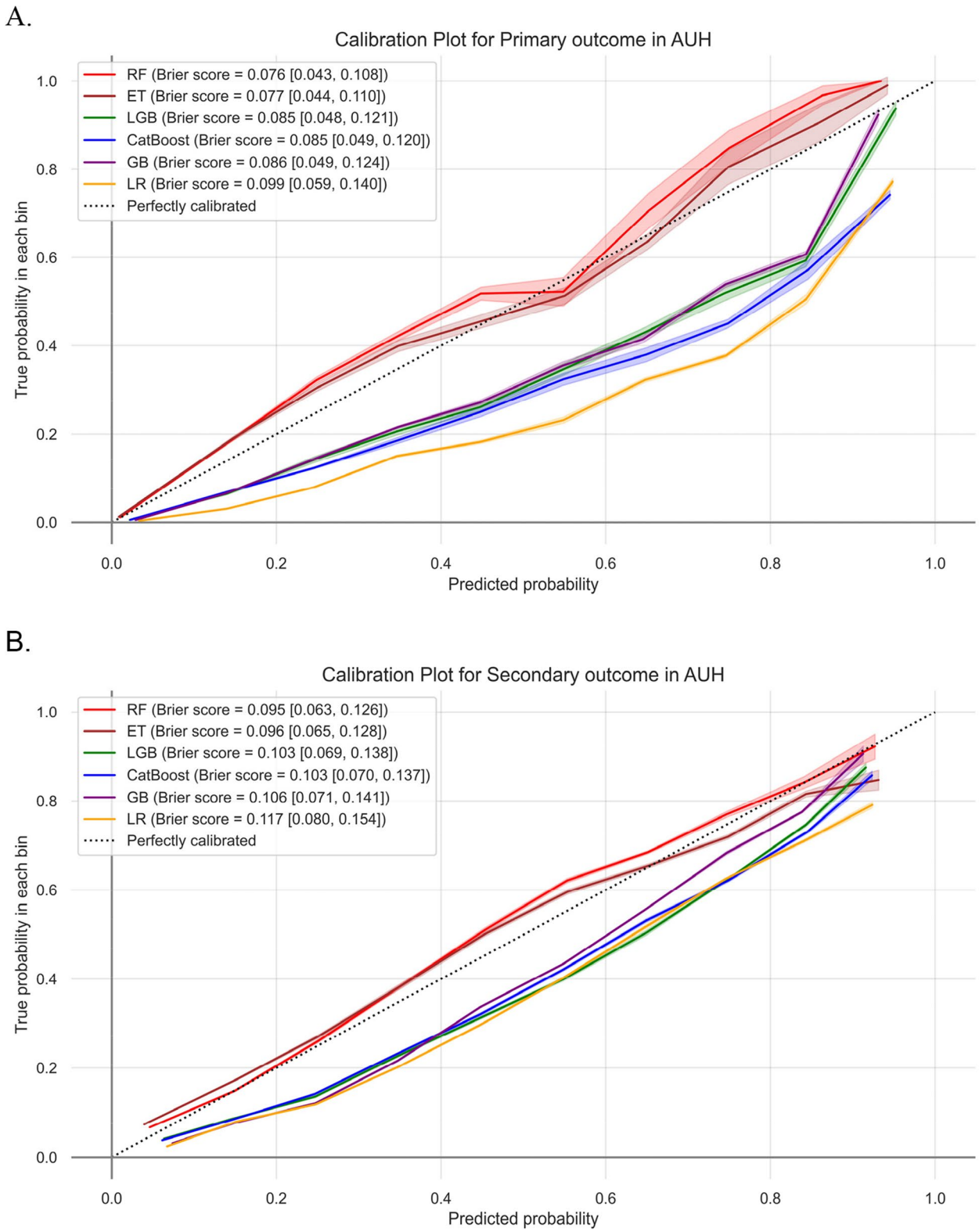

Fig. 3 This figure presents calibration plots for multiple predictive models assessing their performance in predicting (**A**) primary and (**B**) secondary outcomes in the AUH cohort. Each model's Brier score, indicating the accuracy of probabilistic predictions, is displayed along with 95% CI. The shaded areas represent the 95% CI for each model's calibration curve. AUH, An-Nan Hospital; CatBoost, Categorical Boosting; CI, confidence intervals; ET, Extremely Randomized Trees; GB, Gradient Boosting; LGB, Light Gradient Boost Machine; LR, Logistic Regression; RF, Random Forest; LR-TTAS, Logistic Regression-Taiwan Triage Acuity scale

Table 4 Prediction ability of seven machine learning models and reference for external validation at AUH

	AUROC	*p*-value*	BS	F1 score	Accuracy	Sensitivity	PPV	Specificity	NPV
Primary outcome									
GB	0.933	-	0.086	0.409	0.962	0.429	0.391	0.979	0.982
LGB	0.932	0.45	0.085	0.409	0.962	0.423	0.396	0.979	0.982
CatBoost	0.929	0.08	0.085	0.410	0.961	0.442	0.382	0.977	0.982
LR	0.928	0.57	0.099	0.393	0.949	0.534	0.311	0.962	0.985
RF	0.916	< 0.05	0.076	0.420	0.962	0.449	0.395	0.978	0.982
ET	0.905	< 0.05	0.077	0.406	0.961	0.432	0.384	0.978	0.982
LR-TTAS	0.896	< 0.05	0.090	0.000	0.969	0.000	0.000	1.000	0.969
Secondary outcome									
LGB	0.815	-	0.103	0.573	0.791	0.618	0.535	0.842	0.882
CatBoost	0.815	0.40	0.103	0.573	0.798	0.599	0.550	0.856	0.879
RF	0.807	< 0.05	0.095	0.552	0.790	0.572	0.534	0.854	0.872
GB	0.805	0.33	0.106	0.560	0.792	0.582	0.539	0.854	0.874
ET	0.802	0.18	0.096	0.541	0.789	0.550	0.533	0.859	0.867
LR	0.792	< 0.05	0.117	0.544	0.771	0.601	0.496	0.821	0.875
LR-TTAS	0.676	< 0.05	0.113	0.136	0.770	0.080	0.972	0.456	0.782

AUH Aisa University Hospital, *AUROC* Area under receiver-operator curve, *BS* Brier score, *ET* Extremely Randomized Trees, *GB* Gradient Boosting, *LGB* Light Gradient Boosting Machine, *LR* Logistic Regression, *LR-TTAS* Logistic Regression-Taiwan Triage Acuity scale, *NPV* Negative predictive value, *PPV* Positive predictive value, *RF* Random Forest

*The p-values in the table were derived using the DeLong test, which involved pairwise comparisons of each model's AUROC with that of the model ranked directly above it in performance

of false predictions. For instance, out of 250 patients, while 8 patients were correctly identified as having a primary outcome, 35 patients were incorrectly classified as such. While raising the threshold for primary outcomes may reduce screening rates, it significantly decreases the likelihood of false alarms. For example, when the threshold is adjusted to 0.20, only 5 patients are correctly diagnosed with a primary outcome, but the number of false positives decreases to just 6. For discharge predictions, the overall accuracy was high, with error rates ranging around 14–25 cases. For secondary outcomes, predictions varied depending on the threshold combinations.

Feature importance

Figure 5 shows the top predictors for both outcomes were largely similar, including transferred, age, shock index, and BMI. To further explore our model's interpretability, we used SHAP values. As illustrated in Supplementary Fig. 1A, lower values in "Triage Level" and "Glasgow Coma Scale" increase the model's predicted likelihood of the primary outcome. Additionally, patients who arrive by ambulance, are transferred, or require a bed in the ED are also associated with a higher predicted likelihood of the primary outcome. For the secondary outcome, similar features were influential, with additional factors such as older age, presence of fever, elevated body temperature, increased heart rate, and higher shock index positively correlating with the model's prediction of the primary outcome (Supplementary Fig. 1B).

Discussion

We compared the predictive performance of machine learning models with assessments made by EPs. The machine learning models accurately predicted both the primary, and the secondary outcome with high AUROC and F1 scores, demonstrating their potential to assist triage nurses in determining patient priorities by predicting final outcomes upon leaving the ED. The results also demonstrate that integrating structured and unstructured data can significantly enhance the predictive performance of the models. This approach reflects real-world clinical practice and could optimize triage decisions in emergency settings.

In clinical practice, EPs make disposition decisions based on a range of factors that can vary by country and the level of the hospital. For example, patients with the same triage level may receive different dispositions from EPs at CMUH compared to those at AUH, reflecting localized decision-making processes. This variability, along with differences in admission criteria between tertiary hospitals like CMUH and regional hospitals, likely contributed to the decline in model performance observed during external validation, especially regarding secondary outcomes. To mitigate overfitting in our retrospective study, we employed several strategies.

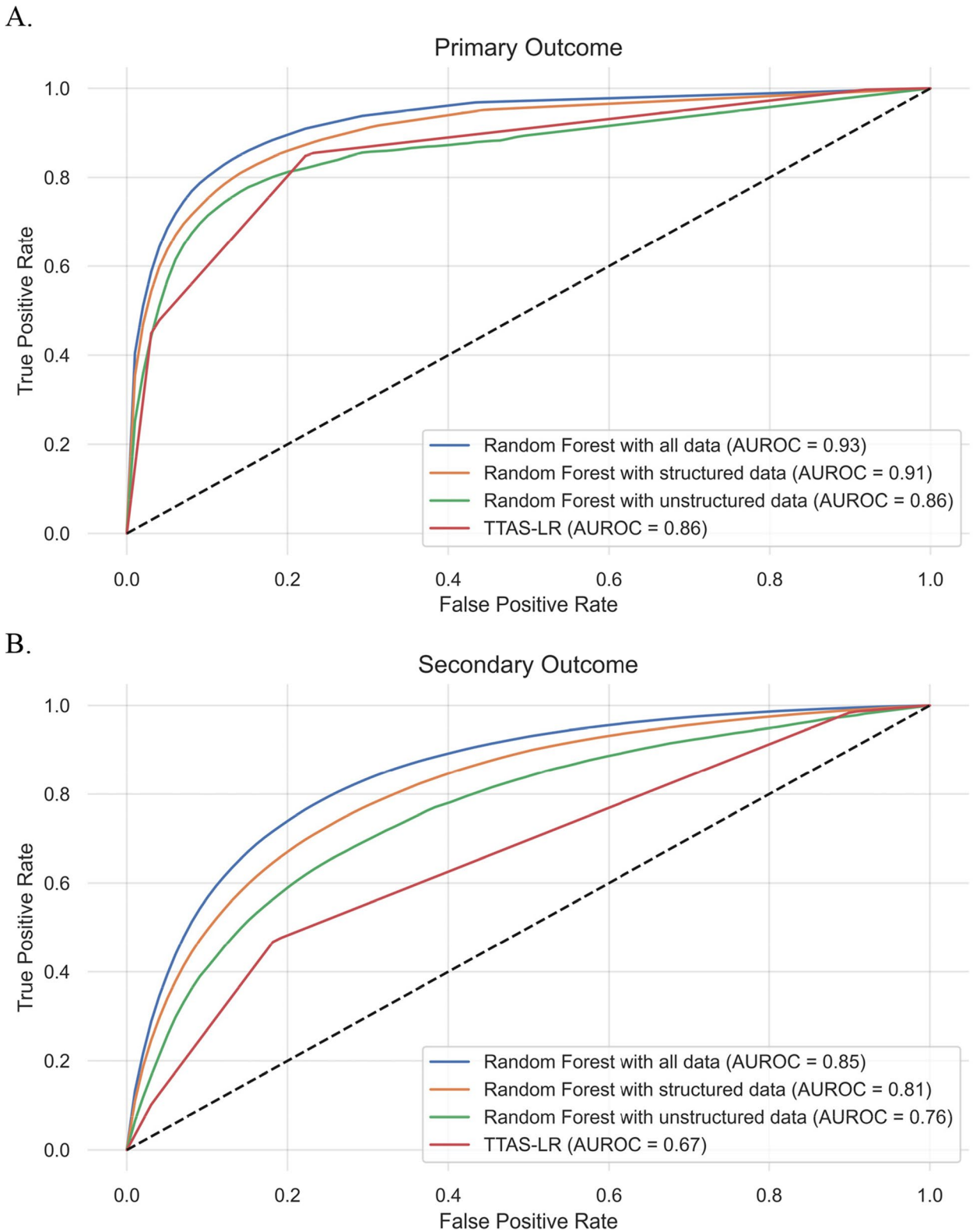

Fig. 4 Receiver operating characteristic curves of physician assessments and LR-TTAS and Random Forest models with different training input, including both structured and unstructured data and only one of them. **A** Prediction of primary outcome. **B** Prediction of secondary outcome. AUROC, area under receiver operating characteristic curve; LR, Logistic Regression; TTAS, Taiwan Triage and Acuity Scale

Table 5 Performance comparison of Random Forest using different feature sets in the CMUH dataset

	AUC	*p*-value*	F1 score	Accuracy	Sensitivity	PPV	Specificity	NPV
Primary outcome								
All features	0.926	-	0.500	0.955	0.591	0.433	0.969	0.983
Structured set	0.906	< 0.05	0.468	0.951	0.561	0.402	0.967	0.982
Unstructured set	0.861	< 0.05	0.406	0.941	0.526	0.331	0.958	0.981
LR-TTAS	0.858	< 0.05	0.000	0.962	0.000	0.000	1.000	0.962
Secondary outcome								
All features	0.847	-	0.580	0.782	0.655	0.519	0.820	0.889
Structured set	0.812	< 0.05	0.534	0.756	0.611	0.475	0.799	0.874
Unstructured set	0.761	< 0.05	0.478	0.745	0.509	0.450	0.815	0.848
LR-TTAS	0.666	< 0.05	0.171	0.771	0.103	0.499	0.969	0.784

AUROC Area under receiver-operator curve, *CMUH* China Medical University Hospital, *LR-TTAS* Logistic Regression-Taiwan Triage Acuity scale, *NPV* Negative predictive value, *PPV* Positive predictive value

*The *p*-values in the table were derived using the DeLong test, which involved pairwise comparisons of each model's AUROC with that of the model ranked directly above it in performance

Table 6 Threshold-specific predictions for discharge, primary, and secondary outcomes per 250 emergency department visits in the CMUH test cohort

Threshold			Prediction per 250 ED visits								
Discharge	Primary	Secondary	Discharge			Primary			Secondary		
			True	False	F1	True	False	F1	True	False	F1
0.74[a]	0.04[a]	0.22[a]	145	14	0.849	8	35	0.309	23	25	0.427
0.72	0.06	0.22	149	15	0.857	8	25	0.364	26	27	0.475
0.72	0.08	0.20	149	15	0.857	7	19	0.406	30	30	0.508
0.72	0.10	0.18	149	15	0.857	7	15	0.439	32	32	0.528
0.66	0.12	0.22	157	19	0.875	7	13	0.461	30	24	0.537
0.62	0.14	0.24	162	22	0.883	6	11	0.478	29	20	0.540
0.58	0.14	0.28	166	25	0.889	6	11	0.479	26	16	0.522
0.66	0.16	0.18	158	19	0.875	6	9	0.492	32	26	0.561
0.62	0.18	0.20	162	22	0.883	6	8	0.497	31	21	0.561
0.60	0.20	0.20	165	24	0.887	5	6	0.502	30	20	0.564

CMUH China Medical University Hospital, *ED* Emergency department

[a] The optimal thresholds were determined based on Youden's J Statistic

First, we used 10-fold cross-validation to prevent the model from being too tailored to any single subset of data, improving generalization. Second, we applied feature selection to identify and retain only the most relevant features, reducing dimensionality and helping the model focus on key predictive variables. This not only improves model efficiency but also reduces the risk of overfitting by eliminating noise and irrelevant data. Last, we adjusted sample weights in the models to handle class imbalance, ensuring that the model did not overfit to the majority class at the expense of the minority class.

The F-1 score for EP assessments were 0.361, and 0.498 for the primary outcome, and secondary outcome, respectively. We intentionally did not include the same patient across different questionnaires to compare consistency among physicians, recognizing the inherent variability in clinical judgment, which aligns with real-world clinical practice. This approach also helped prevent discussions among EPs that could potentially lower the reliability and validity of their responses. By standardizing the format of all data presented, we ensured that the information provided to the physicians was consistent and complete, thereby minimizing external influences on

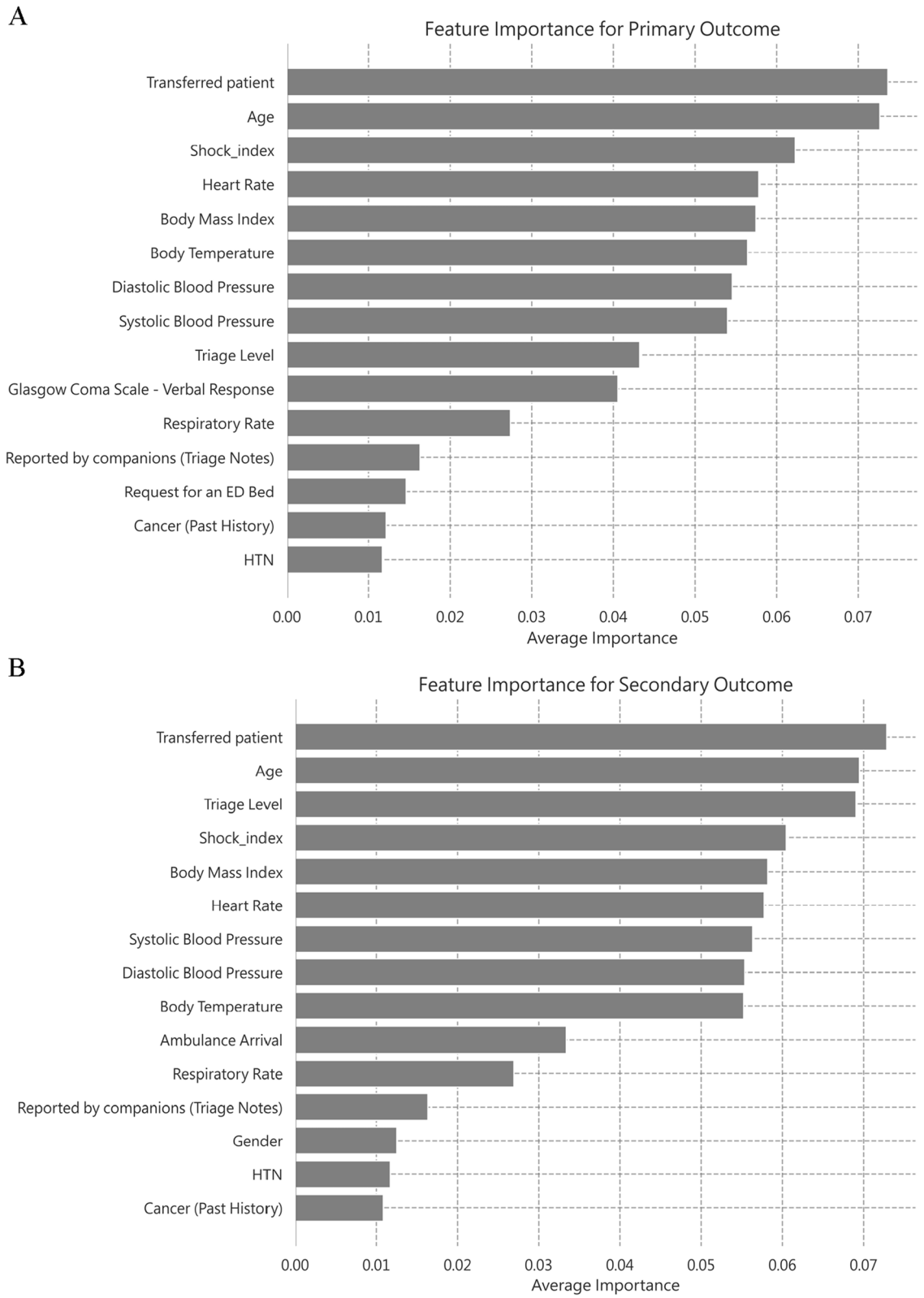

Fig. 5 Fifteen most influential features for Random Forest model prediction of (**A**) primary outcome and (**B**) secondary outcome. ED, emergency department; HTN, hypertension

their decision-making. Zlata et al. proposed that to predict admission (without further differentiating between ICU or ward) or discharge status, EPs use triage information (vital signs, nursing notes, and EMRs) without physically examining patients. The AUROC obtained was 0.703 (95% CI: 0647 − 0.760) with a sensitivity of 51.8% and a specificity of 88.9% [48]. Of note, it is important to recognize that clinicians in practice make decisions based on a comprehensive evaluation of the patient's condition, not solely on triage information. For machine learning models, rare or unique cases pose a challenge due to their infrequent occurrence, making accurate predictions difficult. In contrast, clinicians excel in these situations as they can leverage their extensive clinical experience and expertise, integrating multiple aspects of a patient's presentation—such as medical history, symptoms, and test results—to make comprehensive judgments. This allows them to apply past experiences flexibly and use intuition and clinical reasoning to make more accurate decisions, even in uncommon scenarios, an ability that current machine learning models struggle to fully replicate.

Our findings indicate that the LR-TTAS model was unable to accurately distinguish between patients requiring general ward admission and those requiring ICU admission. Consequently, the predictive performance for the primary outcome was relatively poor in both the internal and external validation. This may have been due to an overlap in the distribution of TTAS levels between the 2 outcome groups. Therefore, relying solely on TTAS level as a training feature led to inaccurate classification.

While simpler models, such as LR, offer higher interpretability, they often come at the cost of predictive accuracy, especially in high-dimensional data settings like ours. In contrast, complex models, such as Random Forest, excel in handling high-dimensional data and capturing complex feature interactions, providing superior accuracy. Given our goal to provide precise and timely predictions in critical environments such as the ED, we prioritized models that offer higher performance. To balance the trade-off between accuracy and interpretability, we applied model-agnostic interpretability tools, including feature importance ranking and SHAP values. These tools help identify key features influencing predictions and enhance interpretability by illustrating each feature's contribution to individual predictions, thereby improving transparency and trust in our model's output. The SHAP summary plot further confirmed that the model's decision-making aligns well with clinical understanding, reinforcing confidence in its predictions.

In the Random Forest for predicting primary and secondary outcomes, referral from other healthcare institutions is a key predictive feature, as CMUH, a medical center, frequently receives severe or rare cases requiring advanced care and hospitalization. Age is widely recognized as a risk factor for higher morbidity and mortality, aligning with its predictive value for ICU admission and some studies also found age was a relative important factors when predicting either ward or ICU admission [16, 49–51]. Other features of vital signs, such as heart rate, blood pressure, body temperature and shock index, are key components of various warning systems, including NEWS, MEWS, and REMS [52–54].

When using the model, it is essential to evaluate the impact of misclassifying patients with mild conditions as severe and vice versa, as each type of misclassification could differently affect the overall ED system and patient safety. For example, as shown in Table 6, using the optimal thresholds may result in an excess of falsely classified critical patients, which could excessively consume resources in the resuscitation area and delay treatment for truly urgent cases. After increasing the threshold for primary outcomes, although this may result in missing one to three primary outcome patients per 250 cases, it can significantly reduce false alarms and help prevent the critical area from becoming overwhelmed. Any severe cases not flagged by the model can still be identified through triage nurses or further clinical assessments and examinations. At the same time, lowering the threshold for discharge predictions allows for more efficient identification of mild cases. However, excessively decreasing this threshold could lead to a substantial influx of patients requiring intensive care within the mild area, potentially necessitating additional support from other areas in the ED, which may also cause operational challenges. Therefore, while the model can be trained to identify an optimal threshold using various techniques, in practice, thresholds should be adjusted to align with the specific clinical needs and resource availability of each setting.

Comparison to other studies

Several studies have attempted to predict admission outcomes in EDs by using NLP [23, 24, 55, 56]; however, none of the studies distinguished between admission to the ICU and to a ward. Resource consumption differs between patients admitted to a ward and those admitted to the ICU. Therefore, rather than solely judging admission or not, it is also critical to identify those severely ill patients with demand for intensive care. Raita and Kwon et al. found that machine learning models outperformed traditional triage systems in predicting various outcomes, including discharge and admission to an ICU or ward [16]. Our models offer the potential for optimization by integrating NLP, thus enabling better differentiation of critically ill patients. Structured data, such as vital signs and chronic disease information, provides quantitative

and standardized details, while unstructured data, like chief complaints, offers qualitative insights (e.g., descriptions of pain as dull, sharp, or tearing). Combining both types allows for a more comprehensive understanding of the patient's condition. Unstructured data can provide context that might be missing from structured data; for instance, a triage note mentioning chest pain with cold sweats can add valuable context to elevated heart rate and blood pressure readings, leading to more accurate predictions. Of note, there are also some limitations of current NLP techniques, particularly in processing medical text. Medical jargon, abbreviations, and domain-specific terminology can present challenges for standard NLP models, often leading to misinterpretations or incomplete data extraction. Additionally, the presence of mixed-language notes, as often occurs in multilingual healthcare settings, further complicates the text processing. To mitigate these issues, we applied domain-specific tokenizers and customized dictionaries to better capture medical terminology.

In recent years, large language models (LLMs) have rapidly evolved and been applied to the field of triage. Generally, performance varies significantly across different models or versions [57, 58]. Limited research exists on using LLMs to further predict patient outcomes. Arian et al. evaluated 30 patients using ChatGPT-3.5 to determine triage levels and assess its outcome prediction capabilities [59]. For hospital admission, time-dependent conditions, and critical situations, the AUROC ranged from 0.78 to 0.81, while performance in predicting 72-hour mortality was the lowest (AUROC of 0.669). They converted all triage data for each patient into a single vignette for the LLM, allowing the model to make a final judgment. In contrast, we process structured data and clinical notes separately, and some differences exist in the data used. This approach may contribute to the variations in performance outcomes. How to integrate LLMs into our pipeline to optimize model performance is a crucial topic in the future.

Our research has several limitations. First, several potentially confounding factors—such as medication use, socioeconomic status, smoking status, betel nut chewing status, and alcohol consumption—were not considered. These data are usually not obtained during triage due to limited time and privacy concerns. Second, regarding missing data, BMI was the only feature with more than 20% missing values. We addressed this by using imputation methods to fill in the missing values, aiming to minimize any impact of incomplete data on model performance. Third, we encountered data imbalance in both internal and external validation datasets. Although resampling techniques were not applied, we addressed the imbalance by adjusting sample weights. Given our large dataset, this approach preserved the original data distribution, reduced computational complexity, and minimized overfitting risks, thereby enhancing model performance, and maintaining representativeness across different data classes. However, rare outcomes, such as in-hospital mortality, represented a small subset of the already limited primary outcomes. Their low frequency in the dataset may have hindered the model's ability to effectively learn patterns associated with these cases. Fourth, the outcomes used in this study may not fully capture all dimensions of triage system effectiveness. For example, our study focused on immediate clinical outcomes such as hospitalization and ICU admission, but did not account for other critical measures like patient prognosis (e.g., survival rates, length of hospital stay), ED waiting times, or resource utilization efficiency. Last, we included only two neighboring hospitals, which may limit the generalizability of our models. Including a wider range of hospitals from different regions and varying levels of care would be essential for a more comprehensive evaluation of the models' generalizability and performance across diverse settings.

Future works

We will further expand the study to include more hospitals at different levels. By incorporating data from a broader range of hospitals, we can avoid training on a single data source, thereby enhancing the model's generalizability. This would enable more precise and reliable integration into diverse clinical environments.

Additionally, to assess the model's impact on actual clinical outcomes, such as patient length of stay and overall mortality, we suggest that real-world implementation into the existing triage system would be necessary. This integration could reduce judgment errors and improve efficiency, enabling more detailed and precise triage. However, future challenges include cross-departmental collaboration to integrate the system with existing hospital workflows, ensuring seamless data transfer for timely and accurate real-time decision support. Additionally, variability in decision-making criteria across different hospitals may reduce the system's effectiveness, as it might not align with each institution's unique guidelines. Lastly, gaining the trust of clinical staff is also essential, as healthcare professionals may be hesitant to rely on AI-based predictions that differ from their clinical judgment.

Conclusion

This study developed and validated machine learning models, using both internal and external data, integrating structured and unstructured triage information to predict patient dispositions, specifically distinguishing

between general ward admissions and critical conditions, including ICU admissions and ED deaths. The integration of both structured and unstructured data significantly enhanced model performance, aligning with the complexity of real-world clinical decision-making.

Overall, AI-assisted triage systems show great potential for improving efficiency, patient safety, and resource allocation in emergency departments, with the possibility for broader application across diverse healthcare settings.

Abbreviations

ED	Emergency Department
TTAS	Taiwan Triage and Acuity Scale
AI	Artificial Intelligence
NLP	Natural Language Processing
EP	Emergency Physician
CMUH	China Medical University Hospital
AUH	Asia University Hospital
ICU	Intensive Care Unit
EMR	Electronic Medical Record
CatBoost	Categorical Boosting
LR	Logistic Regression
GB	Gradient Boosting
AUROC	Area Under the Receiver Operating Characteristic Curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12873-024-01152-1.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Acknowledgements
Not applicable.

Authors' contributions
CYH, CWK, and WCCN designed the study, oversaw the analysis, and were involved in drafting and revising the manuscript. CYH secured institutional review board approval. CWK was responsible for data collection at CMUH, while CYT and CDM gathered data from Asia University Hospital. LYC and CYH was in charge of developing the machine learning algorithms and data analysis. HFW and CYH carried out the data-preprocessing. CYH assumes overall responsibility for the integrity of the paper. All authors have read and given their approval for the final version of the manuscript.

Funding
This work was supported by Asia University (grant number: ASIA-112-CMUH-07).

Data availability
The data collected and analyzed in the course of this study are not publicly accessible because of confidentiality obligations stipulated by Institutional Review Board (IRB) guidelines. Nevertheless, access to these datasets can be arranged upon justified request. For further information, please reach out to Dr. Yu-Hsin Chang.

Declarations

Ethics approval and consent to participate
The Institutional Review Board of China Medical University approved this study and waived the need for individual informed consent (CMUH109-REC1-196). This waiver was granted in accordance with 45 CFR 46.116(f), as the research involved no more than minimal risk to participants, and the waiver did not adversely affect their rights and welfare.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Emergency Medicine, China Medical University Hospital, No. 2, Yude Rd., North Dist, Taichung City 40447, Taiwan. [2]Institute of Information Science and Engineering, National Yang Ming Chiao Tung University, No. 1001, Daxue Rd. East Dist, Hsinchu City 300093, Taiwan. [3]Department of Emergency Medicine, Asia University Hospital, No. 222, Fuxin Rd., Wufeng Dist, Taichung City 413505, Taiwan. [4]Department of Bioinformatics and Medical Engineering, Asia University, No. 500, Liufeng Rd., Wufeng Dist, Taichung City 413305, Taiwan.

Received: 27 May 2024 Accepted: 4 December 2024
Published online: 18 December 2024

References

1. Oredsson S, et al. A systematic review of triage-related interventions to improve patient flow in emergency departments. Scand J Trauma Resusc Emerg Med. 2011;19:43.
2. Yarmohammadian MH, et al. Overcrowding in emergency departments: a review of strategies to decrease future challenges. J Res Med Sci. 2017;22:23.
3. Arya R, et al. Decreasing length of stay in the emergency department with a split emergency severity index 3 patient flow model. Acad Emerg Med. 2013;20(11):1171–9.
4. Saghafian S, et al. Complexity-augmented triage: a Tool for improving patient safety and operational efficiency. Manuf Service Oper Manage. 2014;16(3):329–45.
5. Chiu HY, et al. Current trends in emergency triage in Taiwan the five-level triage system. Hu Li Za Zhi. 2008;55:87–9.
6. Ng CJ, et al. Validation of the Taiwan triage and acuity scale: a new computerised five-level triage system. Emerg Med J. 2011;28(12):1026–31.
7. 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):565–74. e2.
8. McLeod SL, et al. Interrater reliability, accuracy, and triage time pre- and post-implementation of a real-time electronic triage decision-support Tool. Ann Emerg Med. 2020;75(4):524–31.
9. Chang W, et al. Using the five-level Taiwan triage and acuity scale computerized system: factors in decision making by emergency department triage nurses. Clin Nurs Res. 2017;26(5):651–66.
10. Seiger N, et al. Undertriage in the Manchester triage system: an assessment of severity and options for improvement. Arch Dis Child. 2011;96(7):653–7.
11. Hitchcock M, et al. Triage: an investigation of the process and potential vulnerabilities. J Adv Nurs. 2014;70(7):1532–41.
12. Haas B, et al. Survival of the fittest: the hidden cost of undertriage of major trauma. J Am Coll Surg. 2010;211(6):804–11.
13. 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.
14. Hu CA, et al. Using a machine learning approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan. BMJ Open. 2020;10(2):e033898.
15. Rahmatinejad Z, et al. A comparative study of explainable ensemble learning and logistic regression for predicting in-hospital mortality in the emergency department. Sci Rep. 2024;14(1):3406.
16. Raita Y, et al. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care. 2019;23(1):64.

17. Mowbray F, et al. Predicting hospital admission for older emergency department patients: insights from machine learning. Int J Med Inf. 2020;140:104163.
18. Rendell K, et al. The Sydney Triage to Admission Risk Tool (START2) using machine learning techniques to support disposition decision-making. Emerg Med Australas. 2019;31(3):429–35.
19. Kwon JM, et al. Validation of deep-learning-based triage and acuity score using a large national dataset. PLoS ONE. 2018;13(10):e0205836.
20. Yu JY, et al. Machine learning and initial nursing assessment-based triage system for emergency department. Healthc Inf Res. 2020;26(1):13–9.
21. Choi SW, et al. Machine learning-based prediction of Korean triage and Acuity scale level in emergency department patients. Healthc Inf Res. 2019;25(4):305–12.
22. 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):e0229331.
23. Handly N, et al. Evaluation of a hospital admission prediction model adding coded chief complaint data using neural network methodology. Eur J Emerg Med. 2015;22(2):87–91.
24. Arnaud E, et al. Deep Learning to Predict Hospitalization at Triage: Integration of Structured Data and Unstructured Text. IEEE International Conference on Big Data (Big Data). 2020. pp. 4836–41.
25. Emergency Capability Levels of Hospitals in Taiwan. 2024 June 4th, 2024. https://www.google.com/url?client=internal-element-cse&cx=012254495936870409035:lzvyrg0mtim&q=https://www.mohw.gov.tw/dl-87431-90c637cf-e9b3-4a38-94e5-5610e0db641e.html&sa=U&ved=2ahUKEwidwpGYqp2IAxXwGlkFHaqYLG0QFnoECAQQAQ&usg=AOvVaw2K2u_lXGzUuljcEgpEOoHq. Cited 2024 August 30th.
26. Taiwan Hospital Emergency Medical Capability Rating Scale. 2012. Cited 2024 August 30th. https://www.mohw.gov.tw/dl-6282-a951816e-6ae8-42be-8f29-0ca99bd3b2d7.html.
27. Hossain E, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. Comput Biol Med. 2023;155:106649.
28. fxsjy. *jieba*. Available from: https://github.com/fxsjy/jieba. Cited 2017 Sep. 14.
29. Azur MJ, et al. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40–9.
30. Prokhorenkova L et al. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst, 2018;31:6639–49.
31. Ke G et al. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst, 2017;30:3149 - 3157.
32. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression. Volume 398, 3rd ed. Hoboken, New Jersey: Wiley; 2013.
33. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
34. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3–42.
35. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
36. Zhao QY, et al. Development and validation of a machine-learning model for prediction of Extubation failure in Intensive Care Units. Front Med (Lausanne). 2021;8:676343.
37. Chen T, et al. Prediction of Extubation Failure for Intensive Care Unit patients using light gradient boosting machine. IEEE Access. 2019;7:150960–8.
38. Taylor RA, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach. Acad Emerg Med. 2016;23(3):269–78.
39. 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.
40. Chen W, et al. The effects of emergency department crowding on triage and hospital admission decisions. Am J Emerg Med. 2020;38(4):774–9.
41. 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;35:220–7.
42. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13(2):281–305.
43. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
44. Sun X, Xu W. Fast implementation of DeLong's Algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett. 2014;21(11):1389–93.
45. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3.
46. Lundberg SM, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67.
47. Moore A, Bell M. XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction: a UK Biobank cohort study. Clin Med Insights Cardiol. 2022;16:11795468221133612.
48. Vlodaver ZK, et al. Emergency medicine physicians' ability to predict hospital admission at the time of triage. Am J Emerg Med. 2019;37(3):478–81.
49. Park J, et al. A nationwide analysis of intensive care unit admissions, 2009–2014–The Korean ICU National Data (KIND) study. J Crit Care. 2018;44:24–30.
50. Fenn A, et al. Development and validation of machine learning models to predict admission from emergency department to inpatient and intensive care units. Ann Emerg Med. 2021;78(2):290–302.
51. Wilhelms SB, Wilhelms DB. Emergency department admissions to the intensive care unit–a national retrospective study. BMC Emerg Med. 2021;21:1–9.
52. Burch V, Tarr G, Morroni C. Modified early warning score predicts the need for hospital admission and inhospital mortality. Emerg Med J. 2008;25(10):674–8.
53. McGinley A, Pearse RM. A national early warning score for acutely ill patients. British Medical Journal Publishing Group; 2012.
54. Olsson T, Terént A, Lind L. Rapid Emergency Medicine Score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients. J Intern Med. 2004;255(5):579–87.
55. Tahayori B, Chini-Foroush N, Akhlaghi H. Advanced natural language processing technique to predict patient disposition based on emergency triage notes. Emerg Med Australas. 2020;33(3):480–4.
56. Zhang X, et al. Prediction of Emergency Department Hospital Admission based on Natural Language Processing and neural networks. Methods Inf Med. 2017;56(5):377–89.
57. Masanneck L, et al. Triage performance across large Language models, ChatGPT, and untrained doctors in Emergency Medicine: comparative study. J Med Internet Res. 2024;26:e53297.
58. Colakca C, et al. Emergency department triaging using ChatGPT based on emergency severity index principles: a cross-sectional study. Sci Rep. 2024;14(1):22106.
59. Zaboli A, et al. Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage? Am J Emerg Med. 2024;79:44–7.

Publisher's Note