communications medicine

A Nature Portfolio journal

Article

https://doi.org/10.1038/s43856-025-01052-w

# A foundational triage system for improving accuracy in moderate acuity level emergency classifications

Check for updates

Tuo Liu[1,2,9], Yang Gu[3,4,9], Hongyi Chen[1,2,9], Yan Zhang[3,4,9], Leqi Zheng[1,2], Xuanqi Huang[1,2], Yanjun Xu[3,4], Cai Wen[3,4], Mansheng Chen[1,2], Jiaqi Lin[5], Dongguo Huang[3,4], Feixia Chen[3,4], Yulan Zhong[3,4], Hui Chen[3,4], Yanfeng Guo[6], Mei Lu[6], Guangwei Zhang[1,2], Hao Wu[3,4], Changdong Wang[1,2] ✉, Xiaotu Xi[7,8] ✉, Li Li[3,4] ✉ & Tao Yu[3,4] ✉

## Abstract

**Background** Triage is an essential part of Emergency Medicine, which may be assisted by AI models due to limited availability of medical staff. However, AI models for aiding triage have difficulty in identifying levels that are difficult or ambiguous for human clinicians to distinguish. This study aims to develop a more reliable triage model that improves the accuracy of classification, especially for cases with moderate acuity.

**Methods** We developed a new triage model called KUTS, a foundational classification model for emergency triage, which leverages a knowledge prompt-tuning encoder and an uncertainty-based classifier. KUTS takes tabular data and chief complaints as input, and then assign different acuity levels to patients based on their condition acuity. We trained and tested the model on multiple real-world emergency department datasets.

**Results** Here we show that on the most difficult level for human to distinguish (moderate acuity level), our KUTS substantially outperforms the previous shallow single-modal methods, deep single-modal methods and deep multi-modal methods on AUC score, by an average of 0.19, 0.35, and 0.13 (from 0.76, 0.60, 0.82 to 0.95), respectively. Besides, on all the triage levels, our KUTS also outperforms the previous shallow single-modal methods, deep single-modal methods and deep multi-modal methods on AUC score, by an average of 0.14, 0.20 and 0.06 (from 0.82, 0.76, 0.90 to 0.96), respectively.

**Conclusions** KUTS provides a foundational framework and paradigm for the study of emergency triage, and facilitates the development of more efficient triage systems.

## Plain language summary

Emergency departments are often overcrowded, and deciding the urgency of patients' needs —called triage—is a key part of managing limited resources. However, some cases are hard to judge, even for experienced medical staff. This study developed a new computer model to help hospitals better sort patients based on how serious their condition is. The model uses both patient information and the reason they came to the hospital to make its decision. It was tested on real hospital data and was especially good at identifying patients whose conditions were moderate—neither clearly serious nor clearly mild—where mistakes often happen. This model could help hospitals make faster and more accurate decisions, leading to better care for patients and improved use of medical resources in busy emergency rooms.

Over the past decade, the number of patients visiting emergency departments has grown worldwide, and emergency departments (EDs) have faced numerous challenges. In the past five years, there has been a consistent increase in ED attendances by non-urgent patients, which has exacerbated the overload in these departments and hindered the timely allocation of valuable medical resources to those with urgent needs[1]. Although some studies have explored the idea of placing primary care professionals in hospital EDs to manage patients with non-urgent health issues, the effectiveness of such interventions in alleviating the burden on EDs remains unproven[2]. As a result, triage plays a critical role in ED processes, as it prioritizes patients based on the severity of their conditions, ensuring timely and appropriate care.

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. [2]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China. [3]Department of Emergency Medicine, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China. [4]Institute of Cardiopulmonary Cerebral Resuscitation, Sun Yat-Sen University, Guangzhou, China. [5]School of Mathematics (Zhuhai), Sun Yat-sen University, Guangzhou, China. [6]Department of Emergency Medicine, Guangdong Provincal Hospital of Chinese Medicine, Guangzhou, China. [7]The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangdong Provincial Hospital of Chinese Medicine, Guangzhou, China. [8]Guangdong Provincial Key Laboratory of Research on Emergency in TCM, Guangzhou, China. [9]These authors contributed equally: Tuo Liu, Yang Gu, Hongyi Chen, Yan Zhang. ✉e-mail: wangchd3@mail.sysu.edu.cn; xxttm@gzucm.edu.cn; lil3@mail.sysu.edu.cn; yut@mail.sysu.edu.cn

There are several challenges in the current triage system. Typically, experienced and knowledgeable nurses play a key role in quickly collecting patient complaints and vital signs, enabling the identification of life-threatening conditions within limited time frames. However, given the constraints of healthcare human resources, particularly in regions with limited medical facilities, experienced nurses are not always available[3,4]. Additionally, inadequate training for triage nurses can lead to a decrease in both the efficiency and accuracy of the process[5–7]. Even nurses with years of experience may also suffer from triage fatigue[8]. Given these challenges, we believe that developing a robust triage tool to support nurses in making triage decisions is essential, which can reduce the amount of time triage takes and the overall time between presentation and evaluation.

Various traditional triage methods, such as the Australasian Triage Scale (ATS)[9], the Canadian Triage and Acuity Scale (CTAS)[10], the Emergency Severity Index (ESI)[11], the Manchester Triage System (MTS)[12], and the Chinese Emergency Triage Expert Consensus (CETEC)[13], have been used worldwide in EDs[14] to help triage economically and straightforwardly. Despite their recognized importance, traditional triage methods are not without limitations. ESI triage, for example, is estimated to be misdiagnosed in 34% of attendances[15]. Studies have also shown that in the absence of daily triage audits by physicians, the error rate in nurse triage can reach as high as 23.3%[16]. Furthermore, many critically ill patients may not exhibit obvious symptoms upon arrival at the emergency department, and their conditions may rapidly deteriorate during hospitalization[17–20]. The above circumstances pose challenges to the efficiency and accuracy of triage. Therefore, there is a compelling need to alleviate the burden on healthcare professionals and optimize care delivery in emergency settings.

There is also another challenge in the modern triage process, that is the mistriage of the moderate acuity patients. Sax et al. found that among the 1,713,260 ED encounters that mistriage occurred, patients with ESI[11] level 3 had the highest proportion of undertriage[15]. We observed a similar phenomenon when analyzing the emergency triage process based on the CETEC triage scale[13], which is currently the most widely used triage system in Chinese hospitals. Our analysis of the SYSMH-S dataset, which includes clinical data from patients attending the emergency department of South Campus of Sun Yat-sen Memorial Hospital, revealed that moderate acuity (level 3) patients were most frequently misclassified during triage. Further statistical results are presented in Results. Mistriage of moderate acuity level patients may potentially result in grave consequences. Mistriage higher than needed acuity level results in a waste of medical resources, whereas mistriage lower than needed acuity level can delay treatment and pose life-threatening risks. Therefore, the accurate distinction of moderate acuity level emergency cases is crucial. However, currently human nurses do not have an effective way to solve this problem.

In recent years, many researchers have studied the effectiveness of shallow machine learning methods in the emergency triage[21], including Decision Tree[22], Support Vector Machine[23], Naive Bayes[24], Random Forest[25] and XGBoost[26]. These methods accept tabular data of patients, that is, an array of continuous numbers or categorical numbers, and then get the predicted triage results from the machine learning models, which is proved to be more accurate than the ESI[27]. More and more deep learning methods also have been proposed to assist emergency triage recently, including multilayer perceptron (MLP)[21], TabTransformer[28], FT-Transformer[29], BERT[30], HAIM[31] and IRENE[32]. These deep learning methods use either a single modality or multiple modalities as input to predict the corresponding acuity level. The above previous AI methods for aiding triage, although some performed well on average across all levels, did not excel in accurately identifying patients of moderate acuity level. Therefore, they also failed to solve the problem that human clinicians find challenging.

In this paper, we present KUTS, a knowledge prompt-tuning uncertainty-inspired triage system based on pretrained language model. To the best of our knowledge, KUTS is the first triage approach that uses a knowledge prompt-tuning pretrained language model to cope with multimodal triage data, and is the first triage approach that uses an uncertainty-based classifier to deal with samples with high uncertainty. Thanks to these two components, KUTS can effectively distinguish Level 3 patients, solving the problem that human clinicians find really challenging and laborious. KUTS is primarily composed of the knowledge-based encoder and the uncertainty-based classifier. The encoder module in our system comprises two key stages: prompt construction and feature generation. In the prompt construction stage, we develop a prompt for the pretrained language model (PLM) by combining patient information and relevant expert knowledge. This prompt is divided into two parts: the patient information prompt (PIP) and the knowledge prompt (KP). Initially, we create PIP templates to convert tabular patient data and chief complaints into a standardized textual format, facilitating understanding by the PLM. Subsequently, we implement a conditional judgment mechanism to incorporate patient-specific KPs based on their medical conditions. Concatenating PIP and KP forms the final prompt, which is then fed into the PLM to extract patient features. In the uncertainty-based classifier module, we employ an uncertainty-based classifier to derive the triage decision along with an associated uncertainty score, which obtains the evidence feature $E$ of patients and then calculates the belief masses and corresponding uncertainty score using Dirichlet distribution. Finally, we get the predicted triage result from the classifier, along with the uncertainty score of the patient. We validated the effectiveness of our model on four real-world datasets. The experimental results demonstrate that KUTS not only performs excellently across all levels but also substantially surpasses the previous methods in the moderate acuity levels that are challenging for human clinicians to distinguish, effectively solving this challenging problem.

## Methods

### Data introduction

Our data consists of tabular data and chief complaints. All triage data was collected as patients arrive in the emergency department. The tabular data consists of demographics (age and gender), vital signs, and method of admission. Vital signs consist of seven parameters, namely body temperature, heart rate, respiration, systolic pressure, diastolic pressure, SPO2 and consciousness. These measurements are taken and recorded by emergency department nurses during the triage process when a patient presents in the emergency department, all of which are mandatory according to the triage system. Method of admission refers to the way patients come to the emergency department of a hospital, which is also mandatory according to the triage system. As for the chief complaints, they are also mandatory according to the triage system, and consist of a series of symptoms chosen by nurses from a predefined presenting symptoms list. We set the maximum length of the chief complaint to 80. If a patient's chief complaint had more than 80 words, we only took the first 80 words; Otherwise, zero padding was used to satisfy the length requirement. Some patient records had missing vital signs or chief complaints, and these samples were transferred to the missing sample dataset. When conducting overall performance comparison, we only use complete samples as input, while incomplete samples are considered only in the study on handling missing data.

### Study design and setting

**Study design**. This study is a retrospective cohort study aimed at developing and evaluating an emergency department (ED) triage model, with a focus on improving the classification accuracy for moderate acuity patients (CETEC Levels 3). The study uses historical clinical data collected from multiple EDs to develop and assess the performance of the model. It is worth mentioning that despite the use of expert knowledge, our study is still a retrospective analysis, based on historical patient records and established clinical guidelines. The expert knowledge incorporated into our model is derived from the Chinese Emergency Triage Expert Consensus (CETEC)[13], as shown in Fig. 1a, serving as a standardized reference rather than real-time expert input. No prospective data collection or expert intervention occurred during the study.

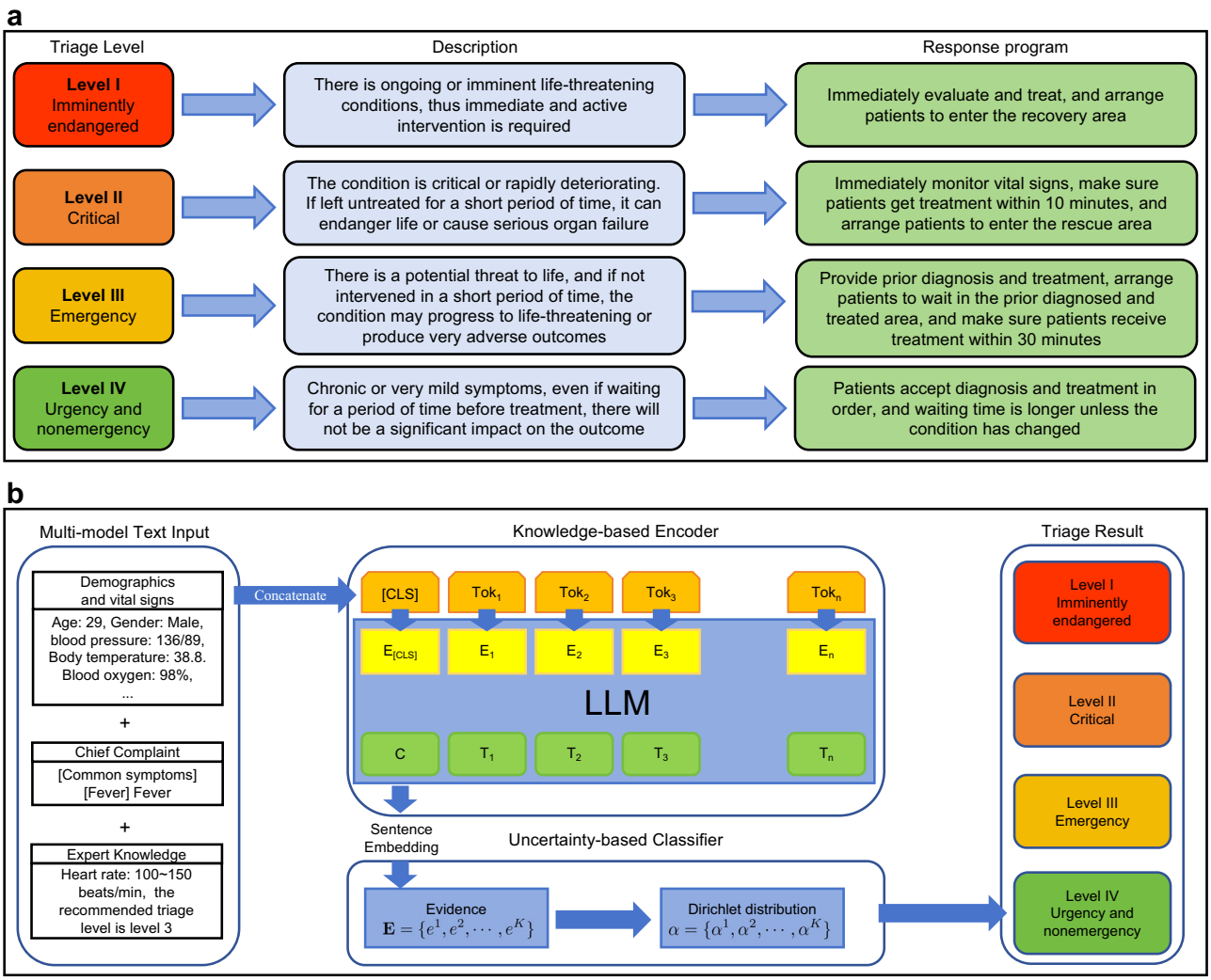


**Fig. 1 | 4-level triage standard of CETEC and the overview of KUTS. a** CETEC, A 4-level triage scale, developed by the National Health Commission of China, was used to classify patients into Levels 1–4, namely "Imminently endangered", "Critical", "Emergency", and "Urgency and nonemergency", respectively, from the acutest to the mildest. **b** The framework of Knowledge-based Uncertainty-inspired Triage System (KUTS), which transforms patient's information and external expert knowledge into prompts for a pretrained language model (PLM), finetunes the PLM to obtain patient features, and then uses an uncertainty-based classifier to obtain the triage level. It mainly consists of two modules: Knowledge-based encoder module and uncertainty-based classifier module.

The gold standard used for training and testing KUTS is the final triage result provided by doctors. The initial triage classification, assigned by nurses, was later revised by physicians based on their diagnosis of the patient's condition, providing a more authoritative final triage level. We use this gold standard as labels to train our model. During the testing phase, we compare the model's predictions with the gold standard to calculate various evaluation metrics.

**Setting**. The study was conducted across three emergency departments (EDs) of three major medical centers in Guangzhou, China, as well as one open-access database. Specifically, it included 98,719 patients from Sun Yat-sen Memorial Hospital, South Campus (SYSMH-S), collected between January 1, 2022, and August 1, 2023; 10,252 patients from Sun Yat-sen Memorial Hospital, North Campus (SYSMH-N), collected between January 1, 2022, and July 31, 2023; and 4783 patients from Guangdong Traditional Chinese Medicine Hospital (GTCMH), collected between November 1, 2023, and November 15, 2023. In addition, the study utilized the MIMIC-IV-ED database, which includes data from 369,934 patients who visited the emergency department of Beth Israel Deaconess Medical Center between 2008 and 2019. These datasets allow for a comprehensive evaluation of the model's performance across different patient populations and triage systems.

## Participants

Inclusion criteria. The study included both adult and pediatric patients who presented to the emergency department (ED). To be eligible, patients were required to have complete data necessary for triage classification, including demographics, vital signs, method of admission, and chief complaint.

Exclusion criteria. Patients with incomplete data–such as missing vital signs or chief complaints–were excluded from the performance analysis. However, these patients were still included in the analysis of missing data.

## Data collection and processing

Data sources. Clinical data was collected from Sun Yat-sen Memorial Hospital (both campuses), Guangdong Traditional Chinese Medicine Hospital, and the MIMIC-IV-ED database.

Data collected. The collected data consisted of two main components. First, the tabular data included demographics such as age and gender, as well as vital signs including temperature, heart rate, respiration rate, systolic and diastolic blood pressure, SPO2, level of consciousness, and the method of admission. Second, chief complaints were recorded as text data, with each

entry limited to a maximum length of 80 words. Complaints exceeding this length were truncated, while shorter ones were zero-padded.

Data preprocessing. Incomplete records (missing vital signs or chief complaints) were excluded from the performance comparisons but were analyzed separately in the missing data section. A specific analysis was conducted on handling missing data.

### Model development and evaluation

Model training. We present KUTS, a knowledge prompt-tuning uncertainty-inspired triage system based on a pretrained language model. The model leverages knowledge prompt-tuning and uncertainty-inspired techniques to improve triage classification accuracy, especially for moderate acuity patients (Specifically, CETEC Level 3).

Performance evaluation. Performance was evaluated using metrics such as accuracy, precision, recall, F1 score, and AUC, with particular focus on the model's ability to classify moderate acuity patients. We compare the model's predictions with the gold standard to calculate these various evaluation metrics.

Comparison with existing systems. KUTS was compared with existing machine learning-based systems, specifically for its ability to accurately classify moderate acuity patients.

**Ethical approval**. Our study has been approved by the Medical Ethics Committee of Sun Yat-sen Memorial Hospital, Sun Yat-sen University. The ethics approval number is SYSKY-2023-375-02. Ethics approval waived the requirement for consent based on the study proposal and nature of use of the patient data.

### Baseline models

We included five baseline models for performance comparisons, including shallow machine learning approaches without the use of deep neural networks utilizing only tabular data (denoted as Shallow single-modal methods), deep learning methods utilizing only tabular data or only chief complaint (denoted as Deep single-modal methods), and multi-modal deep learning methods utilizing both tabular data and chief complaint (denoted as Deep multi-modal methods). Implementation details are discussed below.

**Shallow single-modal methods**. There are a number of existing shallow machine learning methods using tabular data during the triage. The baseline models employed in this study encompass several shallow machine learning techniques, namely Decision Tree[22], Support Vector Machine (SVM)[23], Naive Bayes[24], Random Forest[25], XGBoost[26]. Decision Tree operates as a simple baseline model that partitions the feature space based on the most discriminative features. SVM is a classification algorithm designed to ascertain the decision boundary that maximizes the margin between distinct classes. This characteristic renders SVM particularly robust in classification tasks, especially within high-dimensional spaces. The margin, in this context, pertains to the distance between the decision hyperplane and the closest data points belonging to each class. The Naive Bayes algorithm, based on Bayesian inference, assumes feature independence to compute posterior probabilities of class labels given observed features. Random Forest operates as an ensemble learning approach that constructs multiple decision trees. Each tree undergoes training using a bootstrap sample of the training data and a randomized subset of metrics. These trees are constructed simultaneously, and predictions are obtained by aggregating the collective votes of each individual tree. Another ensemble learning method, XGBoost, represents an advanced implementation of gradient boosted decision trees. It sequentially constructs decision trees while optimizing a differentiable loss function through gradient descent. This sequential nature of XGBoost enables fine-tuning and gradual enhancement of the model's predictive capabilities. In practical implementation, structured continuous data and categorical data are concatenated into a feature vector, which serves as the input for the aforementioned algorithms.

**Deep single-modal methods**. In our triage task, we have developed metric-only deep learning architectures based on multilayer perceptron (MLP)[33], TabTransformer[28] and FT-Transformer (Feature Tokenizer + Transformer)[29]. MLP, a classical and widely-used model, learns parametric embeddings of data. Initially, the structured categorical metrics data undergo embedding before entering into MLP. Following this, structured continuous metrics data, combined with the transformed categorical data are fed into the MLP. In this architecture, we employ an MLP comprising 3 hidden layers, with ReLU activation functions between each pair of adjacent layers to introduce nonlinearity.

TabTransformer and FT-Transformer, on the other hand, represent innovative approaches in deep learning for the tabular domain leveraging Transformer architecture[34]. TabTransformer integrates column embedding for categorical metrics followed by a sequence of $N$ transformer modules. The output of the final layer Transformer, combined with normalized continuous metrics, undergo further processing by multi-layer perceptron to generate predictions. FT-Transformer adopts similar architecture with TabTransformer but distinguishes itself by integrating feature tokenizers for both categorical and continuous metrics. The feature tokenizer module converts each input metric to embeddings, which are then stacked with a [CLS] (classification token)[30] into the input matrix for transformer modules. Subsequently, the final representation of the CLS is then passed through a fully connected layer to obtain the prediction results.

We also included a text-only pretrained Bidirectional Encoder Representations from Transformer (BERT)[30]. Each chief complaint is separated into sentences. The input embeddings are obtained by the sum of the corresponding token embeddings, segmentation embeddings and position embeddings. CLS is positioned at the beginning of token embeddings while SEP (special token) is utilized for sentence separation. BERT learns contextual embeddings through an encoder-only multi-layer bidirectional Transformer.

**Deep multi-modal methods**. We compare KUTS with two established deep multi-modal methods for classification: HAIM[31] and IRENE[32]. These two frameworks are designed to handle diverse patient data modalities, including images, structured tabular data, and unstructured natural language. However, in our triage task, we solely utilize the tabular and text components of these frameworks. Both HAIM and IRENE follow a similar methodology, wherein they encode tabular data and tokenize text data using feature extractors (linear projection or MLP) and pretrained models such as BERT[30], respectively. Nevertheless, HAIM merges embeddings from different data modalities by flattening, normalizing, and concatenating them into a one-dimensional embedding, whereas IRENE obtains a fused embedding by sequentially inputting the two embeddings into bidirectional multi-modal attention blocks and self-attention blocks. In terms of prediction, HAIM employs XGBoost to classify the fused embedding, while IRENE utilizes an MLP in the classification head of the output from the last self-attention block.

### KUTS

**Overview of KUTS**. We propose Knowledge-based Uncertainty-inspired Triage System (KUTS), which transforms patient's information and external expert knowledge into prompts for a pretrained language model (PLM), finetunes the PLM to obtain patient features, and then uses an uncertainty-based classifier to obtain the triage level. It mainly consists of two modules: Knowledge-based encoder module and uncertainty-based classifier module. The overall framework is shown in Fig. 1b.

The knowledge-based encoder module consists of the prompt construction stage and the feature generation stage. For the prompt construction stage, we first generate the prompt for PLM, which consists of two parts: The patient information prompt (PIP) and the knowledge prompt (KP), as shown in Fig. 2a. Firstly, we construct a set of PIP templates to transform

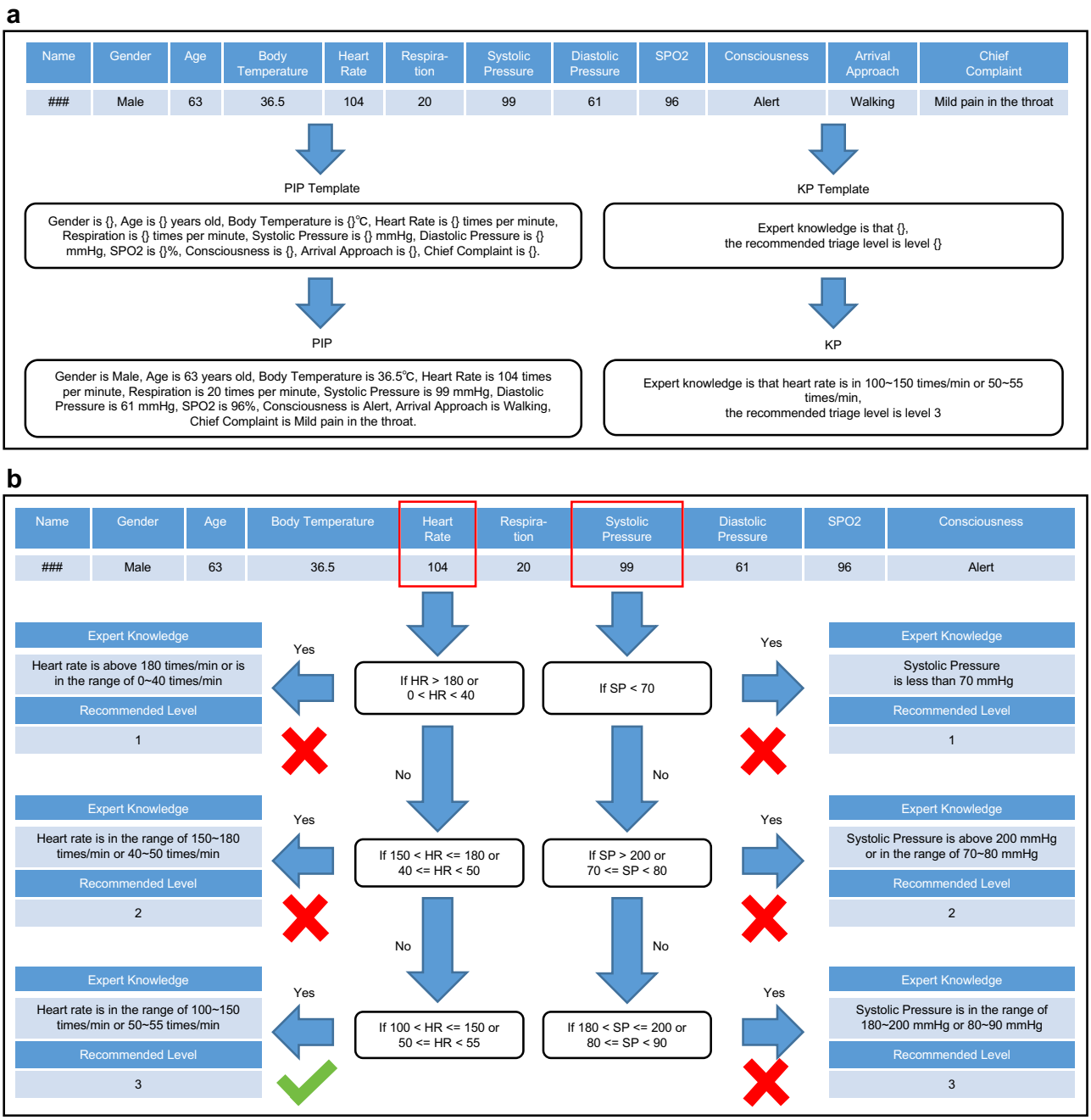


**Fig. 2 | Examples of the generation of PIP, KP, expert knowledge and recommended triage level. a** An example of the generation of PIP and KP. We construct a collection of patient information templates and expert knowledge templates, which transform the patient information and external expert knowledge into textual prompts. **b** An example of the generation of the expert knowledge and recommended triage level. On SYSMH-S, SYSMH-N and GTCMH, we use Chinese Emergency Triage Expert Consensus[13] as the source of our expert knowledge and recommended level, while on MIMIC-IV-ED, we use Emergency Severity Index (ESI)[11] as the source of our expert knowledge and recommended level. The figure only presents the conditional judgment of two vital signs as example, and our implementation adopts judgment of more signs.

patient's tabular data and chief complaints into a unified textual prompt. By doing this we can transform patient's information into a form that can be understood by a PLM. Then, we introduce a conditional judgment design that incorporates specific KP for each patient based on their specific conditions. We concatenate PIP and KP as the final prompt, and then input it to the PLM to obtain the patient's feature. For the uncertainty-based classifier module, we use an uncertainty-based classifier to generate the final triage result with an uncertainty score, which leads to a more reliable decision making without losing any accuracy. We will provide more details on the method in the following sections.

**Pretrained language model in KUTS**. The KUTS model is built on the T5-base (Text-to-Text Transfer Transformer) model, which is a pre-trained language model developed by Google Research[35]. T5 is designed to handle a wide range of natural language processing tasks by treating all tasks as a text-to-text problem, where both the input and the output are formatted as text sequences. This flexible approach allows the T5 model to be fine-tuned for various downstream tasks.

In the case of KUTS, we leverage the T5-base model as the backbone for triage level prediction. T5-base was pretrained on a large corpus of diverse text data, providing a robust understanding of language patterns and

structures. To tailor the model for emergency department triage tasks, we perform **supervised fine-tuning (SFT)** on labeled datasets. By leveraging the pretrained capabilities of T5-base and fine-tuning it on specialized datasets, KUTS benefits from both the general language understanding of T5 and the expert knowledge embedded in emergency triage data.

**Knowledge-based encoder**. Prompt learning has achieved great success in NLP, and many researchers have applied it to the field of medicine[36,37]. In this section, we construct a collection of patient information templates and expert knowledge templates, which transform the patient information and external expert knowledge into textual prompts. This not only eliminates semantic differences between them but also utilizes generic knowledge in pretrained language models.

Patient information prompts. On our triage datasets, patient information consists of two parts: tabular data including patient's age, gender and vital signs, and textual data, which is patient's chief complaint. In order to turn these information into a textual prompt for PLM, we construct a collection of PIP templates, which helps the model learn various aspects of the patients. Specifically, PIP can transform patient's tabular data and chief complaints into a unified textual prompt in a certain format, as shown in Fig. 2a. For a patient $p$, we can fill his/her information into the corresponding fields in the template to obtain: *Gender is {}, Age is {} years old, Body Temperature is {}*°C, *Heart Rate is {} times per minute, Respiration is {} times per minute, Systolic Pressure is {} mmHg, Diastolic Pressure is {} mmHg, SPO2 is {}%, Consciousness is {}, Arrival Approach is {}, Chief Complaint is {}*. By filling the specific content into {} of the above template, we have obtained a comprehensive description of this patient.

With constructing the patient's description using PIP, we can leverage PLM to extract rich semantics from patient's information, which will help capture patient's main feature. PIP can transform the tabular data into textual data, which also increases data utilization efficiency and accelerates model convergence.

Knowledge prompts. Using expert knowledge as side information in triage systems can substantially improve their performance. However, due to the vast amount of expert knowledge, it is impossible to include all expert knowledge in every patient's prompt. Therefore, we propose a conditional judgment design to decide whether a specific knowledge text should be incorporated into the patient's KP.

An example of the generation of KP is shown in Fig. 2b. Specifically, for an expert knowledge text: *Heart rate is in the range of 100–150 times/min or 50–55 times/min, the recommended triage level is Level 3*, we will check if the patient's heart rate is within this range (100–150 times/min or 50–55 times/min). If it is in this range, then we add this expert knowledge text to KP. For another knowledge text: *Systolic blood pressure is in the range of 180–200 mmHg or 80–90 mmHg, the recommended triage level is 3*, similarly, we will check if the patient's systolic blood pressure is within this range (180–200 mmHg or 80–90 mmHg). If it is in then we add this text to KP. If the patient meets the requirements of multiple expert knowledge texts, the recommended triage level is the acutest level corresponding to each of these expert knowledge texts.

By implementing the conditional judgment design, we can decide whether a specific expert knowledge text should be added into the patient's KP, which will reduce model burden while improve model accuracy.

Fused prompts. After obtaining PIP and KP, we concatenate them in a certain format as the input of PLM. By inputting the PIP and KP into PLM, we can integrate the encoding task into a full language environment and use the powerful ability of PLM to extract patient's essential features.

### Uncertainty-based classifier

Framework of the uncertainty-based classifier. Uncertainty-inspired methods have been applied to some medical problems and achieved success[38–40]. The uncertainty-based classification methods introduce the concept of uncertainty score, which enhances the reliability of classification outcomes and meanwhile boosts model accuracy. Leveraging patient's feature from knowledge-based encoder, the final decision and corresponding overall uncertainty score were obtained by the uncertainty-based classifier, which was mainly composed of three steps. Notably, this was a $K$-level triage classification.

Firstly, we obtain the evidence feature $E$ of patients by applying Softplus activation function to ensure the feature values are larger than 0:

$$E = \text{Softplus}(F), \tag{1}$$

where $F$ was the patient's feature we obtained from the knowledge-based encoder. Secondly, we parameterize $E$ to Dirichlet distribution as follows:

$$\boldsymbol{\alpha} = E + 1, \text{ i.e., } \alpha_k = e_k + 1, \tag{2}$$

where $\alpha_k$ and $e_k$ are the $k$-th category Dirichlet distribution parameters and evidence, respectively. Thirdly, we calculate the belief masses and corresponding uncertainty score as:

$$b_k = \frac{e_k}{S} = \frac{\alpha_k - 1}{S}, u = \frac{K}{S}, \tag{3}$$

where $S = \sum_{k=1}^{K}(e_k + 1) = \sum_{k=1}^{K}\alpha_k$ is the Dirichlet intensities. The relationship described in Equation 3 indicates that the probability allocated to category $k$ is directly proportional to the observed evidence for category $k$. Conversely, a decrease in the total evidence obtained corresponds to an increase in overall uncertainty.

**Definition of Dirichlet distribution**. The Dirichlet distribution is a continuous multivariate probability distribution that is often used to model distributions over probability vectors or random variables that represent proportions. It is parameterized by a vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$, where $K$ is the number of categories. Each $\alpha_k$ represents a concentration parameter associated with category $k$.

The probability density function (PDF) of the Dirichlet distribution for a $K$-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_K)$ is given by:

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_i\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}x_i^{\alpha_k - 1}, \tag{4}$$

where $\Gamma(\cdot)$ is the gamma function.

The parameters $\boldsymbol{\alpha}$ control the shape of the distribution. When all $\alpha_i$ are equal, the Dirichlet distribution is symmetric, and when some $\alpha_i$ are larger than others, the distribution becomes skewed towards the corresponding dimensions.

Loss function. The design of our loss function is inspired by the approach of focal loss[41] and uncertainty-based loss[42], which can be presented as:

$$L_{KUTS} = \alpha(1 - \exp(-L_{TUN}))^{\gamma}L_{TUN}, \tag{5}$$

where $\alpha$ is a balance factor, which is adjustable according to the frequency of categories, and $\gamma$ is a regulatory factor, usually ranging from [0, 5]. Considering the severe category imbalance on the triage datasets, we choose focal loss as the backbone of our loss function, which can effectively help the model handling class imbalance and focus on hard examples, so that the model's performance will be improved while its sensitivity to hyperparameters will be reduced. $L_{TUN}$ is the temperature uncertainty-based loss. Following[42], it can be presented as:

$$L_{TUN} = L_{UN} + L_{TCE}, \tag{6}$$

where $L_{UN}$ is the uncertainty-based loss, and $L_{TCE}$ is the temperature cross entropy loss. $L_{UN}$ is formed by weighted addition of $L_{UN\text{-}CE}$ and $L_{KL}$, which can be presented as:

$$L_{UN} = L_{UN-\mathrm{CE}} + \lambda * L_{KL}, \tag{7}$$

where $L_{UN\text{-}CE}$ is used to ensure that the correct predictions of each sample produce more evidence than other levels, and $L_{KL}$ is used to ensure that incorrect predictions will produce less evidence. $\lambda$ is a gradually increasing balance factor to prevent the model from paying too much attention to $KL$ divergence in the early stages of training, which may lead to a lack of good exploration of the parameter space and result in a flat and uniform distribution of network output. $L_{UN\text{-}CE}$ can be presented as:

$$\begin{aligned} L_{UN-\mathrm{CE}} &= \int \left[ \sum_{k=1}^{K} -y_k \log(p_k) \right] \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} dp_k \\ &= \sum_{k=1}^{K} y_k (\psi(S_k) - \psi(\alpha_k)), \end{aligned} \tag{8}$$

where $\psi()$ is the digamma function, and $B()$ is the multinomial beta function for the concentration parameter $\alpha$. $L_{KL}$ can be presented as:

$$L_{KL} = \log\left( \frac{\Gamma\left(\sum_{k=1}^{K} \hat{\alpha}_k\right)}{\Gamma(K) \prod_{k=1}^{K} \Gamma\left(\sum_{k=1}^{K} \hat{\alpha}_k\right)} \right) + \sum_{k=1}^{K} (\hat{\alpha}_k - 1) \left[ \psi(\hat{\alpha}_k) - \psi \sum_{k=1}^{K} \hat{\alpha}_k \right], \tag{9}$$

where $\Gamma()$ is a gamma function, and $\hat{\alpha} = y + (1 - y) \odot \alpha$ is an adjustment parameter for the Dirichlet distribution, which can avoid punishing the evidence of the groundtruth class to 0.

$L_{TCE}$ is the temperature cross entropy loss, which is introduced in order to ensure the confidence of parameterized features during the training process. It can be presented as:

$$L_{TCE} = -\sum_{k=1}^{K} y_k \log\left( \frac{b_k}{\tau} \right), \tag{10}$$

where $b_k$ is the confidence mass of level $k$, and $\tau$ is the temperature coefficient used to adjust the confidence value distribution, which is initialized to 0.01 and gradually increases to 1 to prevent a lower confidence level for the belief mass distribution in the initial stage of training. Although uncertainty-based loss $L_{UN}$ can guide the optimization of models based on Dirichlet concentration parameterized feature distributions, the Dirichlet concentration will change the original feature distribution, which may lead to a decrease in the classifier's confidence in parameterized features. Thus we use $L_{TCE}$ to directly guide model optimization based on parameterized features, which can effectively avoid the drawback of $L_{UN}$.

**Training and classification.** We employ the T5 model architecture[35], which is an encoder-decoder-based pretrained language model using mask prediction as the pretraining task. We mainly use T5-encoder as our pretrained encoder, get the patient's embedding from the fused prompt, and then classify it with the uncertainty-based classifier.

### Implementation details

During the training phase, we utilized AdamW[43] as the default optimizer, as we empirically observed that it yielded superior performance on both baseline models and KUTS. On SYSMH-S, we initially set the learning rate to $5 \times 10^{-5}$ and employed cosine weight decay for weight regularization. The model underwent 20 epochs of training, with the initial learning rate reduced to $5 \times 10^{-6}$ towards the end of training. The batch size was set to 256. On SYSMH-N, we initially set the learning rate to $1 \times 10^{-6}$ and employed cosine weight decay for weight regularization. The model underwent 100 epochs of training, with the initial learning rate reduced to $1 \times 10^{-7}$ towards the end of training. The batch size was set to 64. On GTCMH, we initially set the learning rate to $5 \times 10^{-6}$ and employed cosine weight decay for weight regularization. The model underwent 100 epochs of training, with the initial learning rate reduced to $5 \times 10^{-7}$ towards the end of training. The batch size was set to 32 for SYSMH-S. On MIMIC-IV-ED, we initially set the learning rate to $5 \times 10^{-5}$ and employed cosine weight decay for weight regularization. The model underwent 20 epochs of training, with the initial learning rate reduced to $5 \times 10^{-6}$ towards the end of training. The batch size was set to 256.

Throughout training, we assessed model performance on the validation set and computed the validation AUC score after each epoch. The model checkpoint with the highest validation AUC score was saved and subsequently tested on the testing set. Our model implementation was carried out using PyTorch[44].

In the process of training the model, we did not fix the random seeds and trained 10 times under the same hyperparameter settings, while recording the AUC score of the model on the test set for each training. It is worth noting that in order to test the performance of the model more fairly, the training set, validation set, and testing set for each training session were also randomly divided. Subsequently, we used these 10 AUC scores to calculate the average as the final performance of the model, and calculated its 95% confidence interval.

To demonstrate the statistical significance of our experimental results, we first repeated the experiments for KUTS and the best performing baseline (that is, IRENE) 10 times with different random seeds. Then, we used independent two-sample t-test (two-sided) to compare the mean performance of IRENE and the best baseline results, and calculated P values.

### Statistics and reproducibility

The data processing and statistical analyses were conducted using the Python3 programming languages. Our experimental code has been made public, and the results can be reproduced by running the code using the parameters provided in the paper. Please refer to the Section "Code availability" for details.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Results

### Dataset characteristics for multi-modal triage

We conducted a study across three emergency departments (EDs) of two major medical centers in Guangzhou, China, and one open-access database. Clinical data, collected by well-trained triage nurses at Sun Yat-sen Memorial Hospital and Guangdong Traditional Chinese Medicine Hospital, included vital signs and chief complaints. CETEC, A 4-level triage scale, developed by the National Health Commission of China, was used to classify patients into Levels 1–4, as shown in Fig. 1a. Besides, it is worth noting that the clinical data collected by well-trained triage nurses at Sun Yat-sen Memorial Hospital and Guangdong Traditional Chinese Medicine Hospital were ultimately labeled based on the final classification assigned by the physician after reviewing the patient. The basic characteristics of the datasets are summarized in Table 1.

The first dataset, SYSMH-S, was from the South campus ED of Sun Yat-sen Memorial Hospital, covering 98,719 patients between 1 January 2022 and 1 August 2023. This site serves a balanced population and has a high number of patients at all triage levels, providing a robust dataset for model development. The second dataset, SYSMH-N, was from the North campus ED of Sun Yat-sen Memorial Hospital, with 10,252 patients enrolled between 1 January 2022 and 31 July 2023. This area has an aging population, and 15.88% of the population is over 65 years old. The number of patients at Levels 3 and 4 is similar, and the study will explore improving the model's ability to differentiate between these levels. The third dataset, GTCMH, was from ED of Guangdong Traditional Chinese Medicine

**Table 1 | Basic characteristics of the four datasets. Data are n (%), or median (IQR)**

| | SYSMH-S | SYSMH-N | GTCMH | MIMIC-IV-ED |
|---|---|---|---|---|
| Total number of participants | 98,719 | 10,252 | 4783 | 369,934 |
| Gender | | | | |
| 7D2Male | 49,468 (50.11%) | 4165 (40.63%) | 1869 (39.08%) | 167,519 (45.28%) |
| 7D2Female | 49,251 (49.89%) | 6087 (59.37%) | 2914 (60.92%) | 202,415 (54.72%) |
| Age | 39 (29,55) | 46 (31,65) | 51 (33,67) | 50 (32,65) |
| Time of triage/presentation | | | | |
| 7D200:00-05:59 | 8927 (9.04%) | 878 (8.56%) | NA | NA |
| 7D206:00-11:59 | 32,192 (32.61%) | 3079 (30.03%) | NA | NA |
| 7D212:00-17:59 | 30,996 (31.40%) | 3034 (29.59%) | NA | NA |
| 7D218:00-23:59 | 26,604 (26.95%) | 3261 (31.81%) | NA | NA |
| States of consciousness | | | | |
| 7D2Alertness | 98,178 (99.46%) | 10,224 (99.74%) | 4761 (99.54%) | NA |
| 7D2Somnolence | 123 (0.13%) | 5 (0.05%) | 6 (0.13%) | NA |
| 7D2Slight coma | 68 (0.07%) | 3 (0.03%) | 7 (0.15%) | NA |
| 7D2Deep coma | 347 (0.35%) | 19 (0.19%) | 9 (0.19%) | NA |
| Methods of admission | | | | |
| 7D2Walk-in | 88,666 (89.82%) | 8309 (87.06%) | 4328 (90.49%) | 233,683 (63.17%) |
| 7D2Ambulance | 2412 (2.44%) | 109 (1.14%) | 101 (2.11%) | 124,752 (33.72%) |
| 7D2Wheelchair | 3681 (3.73%) | 476 (4.99%) | 203 (4.24%) | 0 (0.00%) |
| 7D2Assisted entry | 2303 (2.33%) | 483 (5.06%) | 40 (0.84%) | 0 (0.00%) |
| 7D2Transfer from another hospital | 1653 (1.67%) | 167 (1.75%) | 111 (2.32%) | 0 (0.00%) |
| 7D2Helicopter | 0 (0.00%) | 0 (0.00%) | 0 (0.00%) | 62 (0.02%) |
| 7D2Other | 4 (0.00%) | 0 (0.00%) | 0 (0.00%) | 1112 (0.30%) |
| 7D2Unknown | 0 (0.00%) | 0 (0.00%) | 0 (0.00%) | 10,325 (2.79%) |
| Systolic blood pressure, mmHg | 124 (112,138) | 126 (112,143) | 119 (107,137) | 133 (120,148) |
| Diastolic blood pressure, mmHg | 77 (69,85) | 72 (64,81) | 74 (68,80) | 77 (68,87) |
| Temperature, °C | 36.3 (36.2,36.6) | 36.4 (36.3,36.8) | 36.5 (36.3,37.3) | 36.7 (36.4,37.0) |
| SpO2, % | 98 (97,99) | 98 (97,99) | 98 (97,99) | 99 (98,100) |
| Respiration rate, bpm | 18 (16,18) | 20 (18,20) | 20 (18,20) | 18 (16,18) |
| Heart rate, bpm | 86 (78,98) | 88 (78,99) | 88 (77,100) | 84 (72,96) |
| Acuity level | | | | |
| 7D21 | 482 (0.49%) | 29 (0.28%) | 25 (0.52%) | 11,399 (3.08%) |
| 7D22 | 3245 (3.29%) | 321 (3.13%) | 139 (2.91%) | 121,685 (32.89%) |
| 7D23 | 12,075 (12.23%) | 4842 (47.23%) | 1034 (21.62%) | 208,774 (56.44%) |
| 7D24 | 82,917 (83.99%) | 5060 (49.36%) | 3585 (74.95%) | 27,059 (7.32%) |
| 7D25 | NA | NA | NA | 1017 (0.28%) |

The acuity level of SYSMH-S, SYSMH-H and GTCMH is based on National Health Commission of the People's Republic of China. The acuity level of MIMIC-IV-ED is based upon acuity utilizing the Emergency Severity Index (ESI)[11] 5-level triage system. NA=not applicable.

Hospital, enrolling 4783 patients between 1 November 2023 and 15 November 2023. The patient numbers are smaller, with a pyramidal distribution across triage levels. The fourth dataset was the MIMIC-IV-ED, an open-access database containing admissions of 369,934 patients from the Beth Israel Deaconess Medical Center ED between 2008 and 2019[45,46]. This database uses a 5-level Emergency Severity Index (ESI) triage system[11].

Pediatric patients were also included due to the importance of pediatric triage. For incomplete data (missing vital signs or chief complaints), samples were excluded from performance comparisons but analyzed in the missing data section.

### Analysis of mistriage in moderate acuity patients

In the SYSMH-S dataset, which includes clinical data from 98,719 emergency department visits, we identified 4290 instances of mistriage. Among these cases, the actual level 3 patients, based on the final classification assigned by the physician after reviewing the patient, constituted the largest proportion constituted the largest proportion, accounting for 2586 cases (60.3%). Specifically, these patients, who should have been assigned to level 3, were incorrectly triaged to level 1 (1.0%), level 2 (10.6%), and level 4 (88.4%), respectively. These distributions are illustrated in Fig. 3a, b. These findings emphasize the need for improved strategies to accurately triage moderate acuity patients, as mistriage to higher acuity levels leads to a waste of medical resources, while mistriage to lower acuity levels can delay critical treatment and pose serious risks to patient safety.

### Overall performance comparison

On all the four hospital datasets, our KUTS achieved an average AUC (Area Under the Curve) score of 0.960 (95% CI: 0.956, 0.965) on SYSMH-S, 0.961 (95% CI: 0.916, 0.980) on SYSMH-N, 0.950 (95% CI: 0.893, 0.978) on GTCMH, and 0.885 (95% CI: 0.879, 0.888) on MIMIC-IV-ED, respectively,

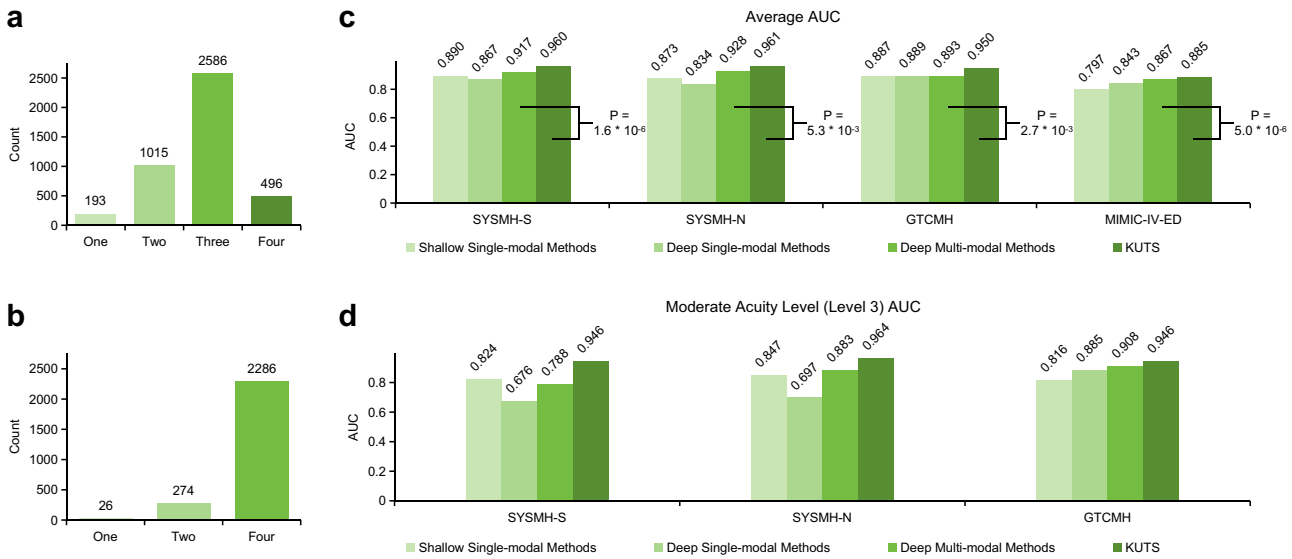


**Fig. 3 | The analysis of mistriage in moderate acuity patients and the performance comparison between KUTS and baseline models. a** The distribution of mistriaged patients across different CETEC acuity levels on the SYSMH-S dataset. **b** The distribution of actual CETEC level 3 patients misassigned to different CETEC acuity levels on the SYSMH-S dataset. **c** Comparison of the experimental results from the shallow single-modal methods, deep single-modal methods, deep multi-modal methods and KUTS on the four triage datasets. **d** Comparison of the experimental results of Level 3 from the shallow single-modal methods, deep single-modal methods, deep multi-modal methods and KUTS on the three 4-triage-level datasets.

**Table 2 | Average AUC scores of KUTS and baselines on the SYSMH-S, SYSMH-N, GTCMH and MIMIC-IV-ED datasets**

| Datasets | | SYSMH-S | SYSMH-N | GTCMH | MIMIC-IV-ED |
|---|---|---|---|---|---|
| Shallow single-model methods | Decision tree | 0.659 [0.649, 0.667] | 0.709 [0.665, 0.759] | 0.680 [0.644, 0.720] | 0.561 [0.559, 0.564] |
| | Support vector machine | 0.836 [0.831, 0.843] | 0.797 [0.750, 0.835] | 0.838 [0.816, 0.858] | 0.797 [0.750, 0.835] |
| | Naive bayes | 0.852 [0.845, 0.857] | 0.829 [0.780, 0.845] | 0.856 [0.831, 0.893] | 0.684 [0.676, 0.696] |
| | Random forest | 0.872 [0.865, 0.878] | 0.855 [0.812, 0.901] | 0.877 [0.856, 0.901] | 0.706 [0.701, 0.712] |
| | XGBoost | 0.890 [0.885, 0.894] | 0.873 [0.827, 0.907] | 0.887 [0.850, 0.907] | 0.756 [0.750, 0.760] |
| Deep single-model methods | FT-transformer | 0.699 [0.607, 0.758] | 0.659 [0.620, 0.703] | 0.695 [0.616, 0.780] | 0.627 [0.609, 0.640] |
| | Tab transformer | 0.728 [0.690, 0.754] | 0.692 [0.639, 0.727] | 0.751 [0.683, 0.803] | 0.722 [0.701, 0.735] |
| | MLP | 0.785 [0.748, 0.827] | 0.731 [0.664, 0.806] | 0.716 [0.497, 0.846] | 0.678 [0.544, 0.727] |
| | BERT-single | 0.867 [0.846, 0.882] | 0.834 [0.796, 0.854] | 0.889 [0.861, 0.917] | 0.843 [0.839, 0.846] |
| Deep multi-model methods | HAIM | 0.903 [0.888, 0.917] | 0.891 [0.860, 0.911] | 0.864 [0.710, 0.949] | 0.866 [0.857, 0.878] |
| | IRENE | 0.917 [0.895, 0.929] | 0.928 [0.887, 0.946] | 0.893 [0.850, 0.927] | 0.867 [0.859, 0.871] |
| Our model | KUTS | 0.958 [0.954, 0.961] | 0.961 [0.941, 0.981] | 0.950 [0.893, 0.978] | 0.885 [0.879, 0.888] |
| Improve-ment | Average | 18.58% | 21.57% | 18.05% | 22.19% |
| | Minimum | 4.69% | 3.56% | 6.38% | 2.08% |

The baseline models include the shallow single-modal methods (including decision tree, support vector machine, naive Bayes, random forest and XGBoost), the deep single-modal methods (including FT-Transformer, TabTransformer, MLP and BERT) and the multi-modal methods (including HAIM and IRENE). The average improvement (The second to last row) represents the average improvement of KUTS compared with the baselines on this evaluation metric, while the minimum improvement (last row) represents the improvement of KUTS compared with the best performing baseline on this evaluation metric. The evaluation metric in this table is AUC, with 95% confidence intervals in brackets.

as shown in Fig. 3c and Table 2, substantially outperforming the baselines. On the other 6 evaluation metrics (sensitivity, specificity, PPV, NPV, F1 score, F2 score), our KUTS also performed much better than baseline models (Supplementary Tables 5–8). Next, we will conduct a specific analysis of the performance of KUTS and the baseline models on each dataset.

On the SYSMH-S dataset, KUTS achieved AUC scores of 0.991, 0.955, 0.946, and 0.948 on four levels respectively (Supplementary Table 1). Other metrics (sensitivity, specificity, PPV, NPV, F1, F2) were mostly superior to baseline models, with specificity and NPV being exceptions (Supplementary Table 5). As for the SYSMH-N dataset, KUTS achieved AUC scores of 0.950, 0.948, 0.964 and 0.983 on four levels respectively (Supplementary Table 2). KUTS also outperformed baselines in sensitivity, specificity, PPV, NPV, F1, and F2, with F1 and F2 being 13.50% and 14.30% higher than the best baseline model (Supplementary Table 6). When it comes to the GTCMH dataset, KUTS scored AUC values of 0.948, 0.943, 0.946 and 0.963 on four levels respectively (Supplementary Table 3). KUTS outperformed baselines in most metrics, except NPV, where it was slightly lower than HAIM (Supplementary Table 7).

To further evaluate the generality of KUTS, we also conducted experiments on a ESI-guided[11] dataset namely MIMIC-IV-ED. Different from the three Chinese hospital triage datasets, MIMIC-IV-ED has 5 triage levels. On MIMIC-IV-ED, our KUTS achieved an average AUC score of 0.885 (95% CI: 0.879, 0.888), outperforming all the baseline models as well (Supplementary Table 4). Specifically, it achieved an AUC score of 0.928,

**Table 3 | Level 3 AUC Scores of KUTS and baselines on the SYSMH-S, SYSMH-N, GTCMH and MIMIC-IV-ED datasets**

| Datasets | | SYSMH-S | SYSMH-N | GTCMH | MIMIC-IV-ED |
|---|---|---|---|---|---|
| Shallow single- model methods | Decision tree | 0.624 [0.614, 0.635] | 0.699 [0.686, 0.711] | 0.658 [0.643, 0.679] | 0.543 [0.541, 0.545] |
| | Support vector machine | 0.754 [0.750, 0.759] | 0.750 [0.738, 0.762] | 0.731 [0.705, 0.775] | 0.750 [0.738, 0.762] |
| | Naive bayes | 0.766 [0.760, 0.776] | 0.754 [0.738, 0.766] | 0.768 [0.716, 0.790] | 0.523 [0.500, 0.555] |
| | Random forest | 0.808 [0.801, 0.817] | 0.835 [0.827, 0.844] | 0.795 [0.767, 0.815] | 0.641 [0.638, 0.645] |
| | XGBoost | 0.824 [0.817, 0.832] | 0.847 [0.838, 0.857] | 0.816 [0.794, 0.831] | 0.663 [0.660, 0.666] |
| Deep single- model methods | FT- transformer | 0.437 [0.335, 0.618] | 0.474 [0.381, 0.574] | 0.656 [0.623, 0.681] | 0.622 [0.611, 0.626] |
| | Tab transformer | 0.439 [0.347, 0.556] | 0.487 [0.434, 0.530] | 0.655 [0.598, 0.693] | 0.651 [0.648, 0.655] |
| | MLP | 0.556 [0.465, 0.669] | 0.556 [0.470, 0.665] | 0.736 [0.690, 0.759] | 0.630 [0.521, 0.658] |
| | BERT-single | 0.676 [0.607, 0.739] | 0.697 [0.612, 0.789] | 0.885 [0.860, 0.909] | 0.752 [0.735, 0.759] |
| Deep multi- model methods | HAIM | 0.734 [0.685, 0.785] | 0.789 [0.719, 0.829] | 0.908 [0.879, 0.936] | 0.785 [0.781, 0.789] |
| | IRENE | 0.788 [0.699, 0.829] | 0.883 [0.816, 0.914] | 0.813 [0.735, 0.876] | 0.789 [0.740, 0.803] |
| Our Model | KUTS | 0.941 [0.939, 0.945] | 0.978 [0.960, 0.986] | 0.946 [0.889, 0.970] | 0.805 [0.800, 0.808] |
| Improve-ment | Average | 47.58% | 42.54% | 25.08% | 22.63% |
| | Minimum | 14.81% | 9.17% | 4.19% | 2.03% |

The baseline models include the shallow single-modal methods (including decision tree, support vector machine, naive Bayes, random forest and XGBoost), the deep single-modal methods (including FT-Transformer, TabTransformer, MLP and BERT) and the multi-modal methods (including HAIM and IRENE). The average improvement (The second to last row) represents the average improvement of KUTS compared with the baselines on this evaluation metric, while the minimum improvement (last row) represents the improvement of KUTS compared with the best performing baseline on this evaluation metric. The evaluation metric in this table is AUC, with 95% confidence intervals in brackets.

0.849, 0.805, 0.921, and 0.920 on five levels, respectively. The sensitivity, specificity, PPV, NPV, F1 score and F2 score of the KUTS model (Supplementary Table 8) are 0.717, 0.743, 0.709, 0.790, 0.709 and 0.715, respectively, mostly outperforming the baseline models.

### Validation of KUTS ability to accurately predict moderate acuity level

To validate whether KUTS improves prediction accuracy for the moderate acuity level compared to the baseline, we summarized and analyzed the performance of KUTS and the baseline models in predicting whether patients belong to Level 3 across four datasets. As shown in Fig. 3d and Table 3, KUTS achieved AUC scores of 0.941, 0.978, 0.946, and 0.805 on the SYSMH-S, SYSMH-M, GTCMH, and MIMIC-IV-ED datasets, respectively, substantially outperforming all baselines. On these four datasets, KUTS achieved average improvements of 47.58, 42.54, 25.08, and 22.63% over the baseline models. Even compared to the best-performing baseline, KUTS achieved improvements of 14.81, 9.17, 4.19, and 2.03%, respectively. These experimental results strongly demonstrate that KUTS effectively addresses the previous methods' shortcomings in diagnosing the moderate acuity level.

### Validation of KUTS ability to handle missing data

To demonstrate the excellent ability of our model to handle missing data, we conducted the following studies on training with missing data. Since missing data study requires a larger number of missing samples than complete samples, we conducted experiments on the SYSMH-N dataset, which contained 9544 complete samples and 51,330 incomplete samples. A complete sample refers to a sample with meaningful values in all of the inputs including gender, age, body temperature, heart rate, respiration, systolic pressure, diastolic pressure, SPO2, consciousness, arrival approach and chief complaint, while an incomplete sample refers to a sample with missing values in at least one of the above entries.

We added different numbers of incomplete samples to the complete dataset and accordingly conducted multiple experiments. Under different ratios of missing data to complete data, we tested the performance of KUTS and IRENE (best performing baseline model) on SYSMH-N. The results are shown in Fig. 4a and Supplementary Table 9-10. Compared with datasets composed of complete samples, the performance of the model does degrade on datasets containing missing data. This is very reasonable because missing data contains less information and poorer data quality, which may affect the training effectiveness of the model and thus reduce its performance. In the case of ratio of 1:1, we can observe that compared with the performance decrease of IRENE (7.33%, from 0.928 to 0.860), the performance of KUTS has decreased less (4.58%, from 0.961 to 0.917). Considering performance on missing data, KUTS achieved an average AUC (Area Under the Curve) score of 0.917 (95% CI: 0.901, 0.929), outperforming IRENE by 6.63% (0.860 (95% CI: 0.835, 0.885)). On other ratios (Supplementary Table 9) and with other evaluation metrics (Supplementary Table 10), KUTS also substantially outperformed IRENE, which showed that compared with the baseline models, KUTS could better handle missing data and achieve a better performance. This is of great significance because in real-life emergency triage scenarios, it is often impossible to collect complete information of a patient, in which case KUTS can play a better role than the previous methods.

### KUTS's generalizability and robustness to domain shift

To verify the generalization performance of our KUTS model, we conducted a domain shift study as follows: Training the model on a dataset from one hospital, testing it on datasets from other hospitals, and observing the performance of the model. Here, due to substantial differences between the datasets of Chinese hospitals and MIMIC-IV-ED (with different categories and triage criteria), we used the datasets of SYSMH-S, SYSMH-N, and GTCMH from three Chinese hospitals in the domain shift experiments. Specifically, SYSMH-S served as the primary source of training data due to its substantial dataset size and comprehensive patient records, providing a rich foundation for model training. Conversely, SYSMH-N and GTCMH were utilized solely for testing purposes, as they are smaller in scale, offering valuable insights into the model's generalizability across diverse clinical settings.

The experimental results are shown in Table 4 and Supplementary Table 11. The model was trained exclusively on the SYSMH-S dataset, which was split into training and validation sets (8:2 ratio). After training, we selected the model with the best performance on the validation set and evaluated it on the SYSMH-N and GTCMH test sets. The model trained on SYSMH-S achieved an average AUC score of 0.950 on SYSMH-N, which decreased by 1.14% (from 0.961 to 0.950) compared with the performance of the model trained and tested both on SYSMH-N. In comparison, the best baseline (IRENE) achieved 0.910 AUC on SYSMH-N, decreasing by 1.94%. This suggests KUTS outperforms the baseline in domain shift experiments. Additional evaluation metrics are available in the Supplementary Table 11.

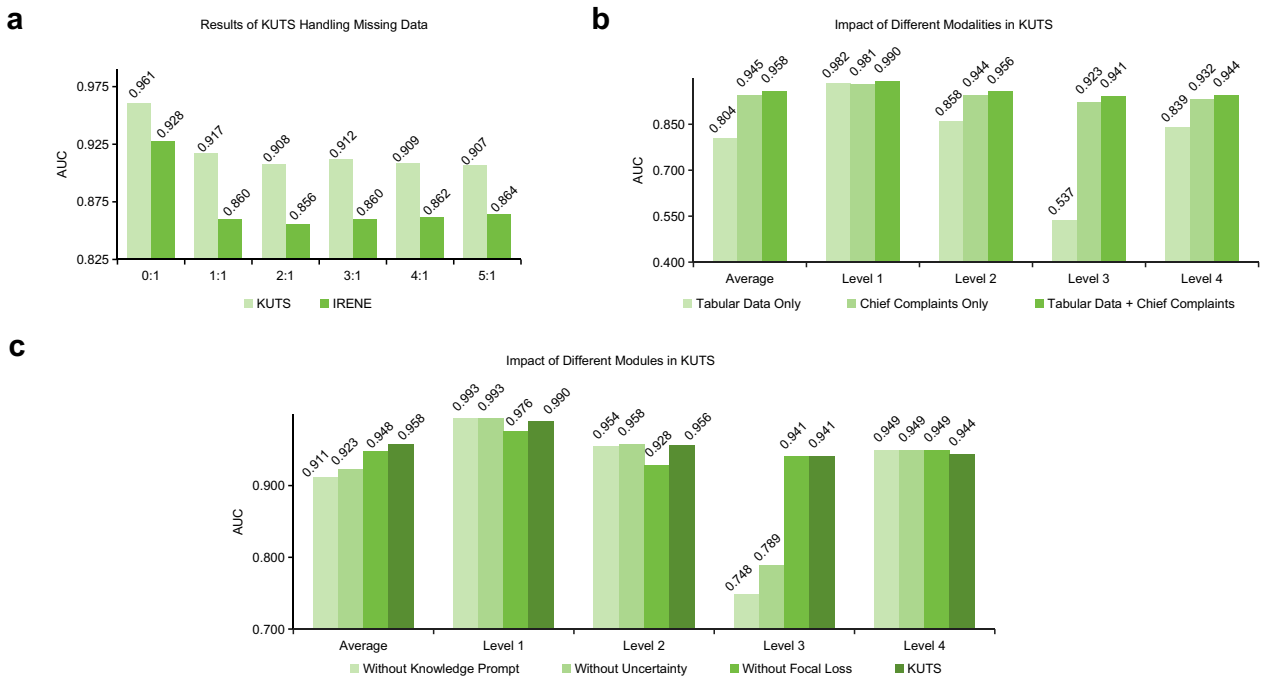


**Fig. 4 | The results of handling missing data by KUTS and IRENE, and the results of ablation experiments on SYSMH-S. a** The results of handling missing data by KUTS and IRENE. "Ratio" represents the ratio of missing samples to complete samples. The evaluation metric is AUC score. **b** Impact of different modalities. Blue bars stand for the AUC Scores of KUTS when using only tabular data as input. Yellow bars stand for the AUC Scores of KUTS when using only chief complaints as input. Green bars stand for the AUC Scores of KUTS when using both tabular data and chief complaints as input. **c** Impact of different modules. Blue bars stand for the AUC Scores of KUTS when excluding knowledge prompt from model structure. Yellow bars stand for the AUC Scores of KUTS when excluding uncertainty classifier from model structure. Purple bars stand for the AUC Scores of KUTS when excluding focal loss function from model structure. Green bars stand for the AUC Scores of KUTS.

**Table 4 | AUC scores of domain shift experiments on SYSMH-N and GTCMH. IRENE is the best-performing baseline**

| SYSMH-N | | Average | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|---|---|
| KUTS | Same | 0.961 | 0.946 | 0.936 | 0.978 | 0.985 |
| | Different | 0.950 | 0.967 | 0.965 | 0.951 | 0.916 |
| IRENE | Same | 0.928 | 0.963 | 0.912 | 0.883 | 0.955 |
| | Different | 0.910 | 0.946 | 0.910 | 0.848 | 0.935 |
| GTCMH | | Average | Level 1 | Level 2 | Level 3 | Level 4 |
| KUTS | Same | 0.950 | 0.948 | 0.943 | 0.946 | 0.963 |
| | Different | 0.914 | 0.980 | 0.944 | 0.856 | 0.877 |
| IRENE | Same | 0.893 | 0.913 | 0.910 | 0.813 | 0.934 |
| | Different | 0.772 | 0.944 | 0.900 | 0.375 | 0.871 |

**Same** means training and testing both on SYSMH-N/GTCMH, and **Different** means training on SYSMH-S and testing on SYSMH-N/GTCMH.

Considering that SYSMH-S and SYSMH-N come from different wards of the same hospital and may have similar data characteristics, we further tested the model trained on SYSMH-S on the GTCMH dataset from Guangdong Traditional Chinese Medicine Hospital, which is another hospital completely independent of Sun Yat-sen Memorial Hospital. The model achieved an average AUC score of 0.914, which decreased by 3.79% (from 0.950 to 0.914) compared with the performance of the model trained and tested both on GTCMH. This decrease is expected due to the domain shift between hospitals. IRENE, however, achieved only 0.772 AUC on GTCMH, getting a 13.55% decrease. This demonstrates that KUTS performs substantially better than baselines in a true domain shift scenario. Other evaluation metrics also support this conclusion, as shown in the Supplementary Table 11.

The success of the domain shift experiments has confirmed a promising prospect: Training our model with extensive emergency triage data from well-equipped hospitals in major cities and then deploying it in smaller hospitals with fewer medical resources. This approach enables the transfer of medical capabilities from large to small hospitals. As a result, patients can receive timely and accurate triage services in smaller hospitals, reducing the risk of unnecessary or delayed medical treatments.

### Impact of different modalities and modules in KUTS

To fully explore the effects of various modalities and modules, we conducted comprehensive ablation experiments and present their results. This section mainly analyzes the ablation experiments on SYSMH-S, with the AUC score presented in Fig. 4b, c and Supplementary Table 12, and other evaluation metrics presented in Supplementary Table 13. The results of ablation experiments on other datasets are shown in Supplementary Table 14-19. Firstly, we investigated the impact of different modalities, including tabular data (row 1) and chief complaint (row 2). We observed that excluding tabular data from the input would result in the 1.56% performance loss (from 0.960 to 0.945) of the model (row 1), which proved that the tabular data actually boosted the triage performance of our model. Apart from the tabular data, we also studied the impact of chief complaint (row 2). We observed that including chief complaint brought a 19.40% performance gain (from 0.804 to 0.960), which means that the chief complaint also plays a very important role among the input modalities. We also observed that the improvement of model performance by chief complaints was much greater than that of tabular data, which also indicated that the role of chief complaints was greater in the two input modalities. This is also confirmed in the subsequent attention visualization case study. In addition, another consequence of excluding chief complaint from the input was that the AUC score of the third level dropped substantially (from 0.946 to 0.537). This also indicated that the chief complaint plays an important role in the model's

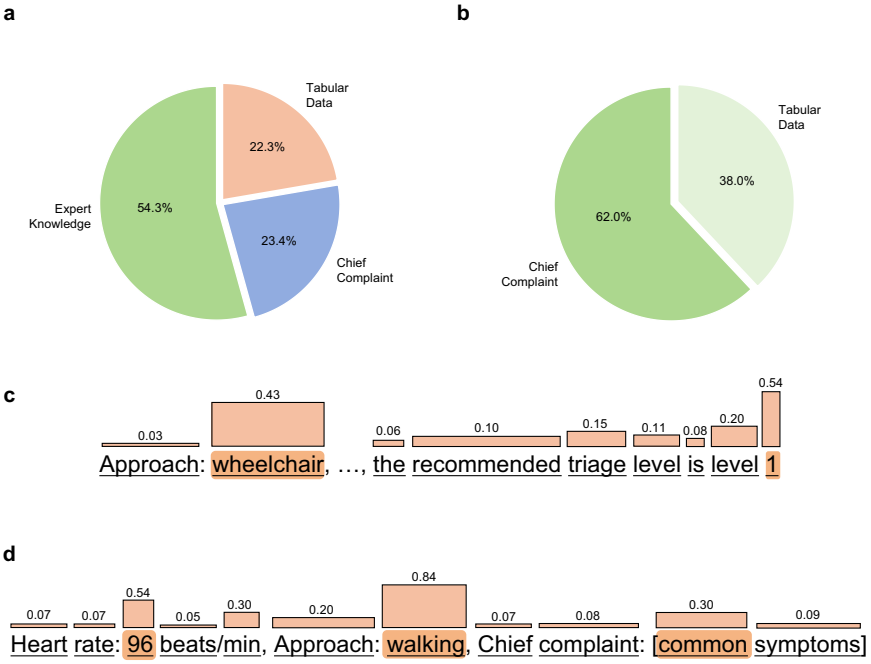


**Fig. 5 | Attention analysis of two different triage cases. a**, **b** Attention allocated to different types of data from two triage cases, including tabular data, chief compliant and expert knowledge. Here the tabular data consists of demographic data and vital signs. **c**, **d** Attention allocated to each word of partial input texts, with some of important words highlighted. **c** is the case guided by expert knowledge. And **d** is the case without knowledge guidance.

judgment of Level 3, which was in line with reality, since Level 3 was a mild emergency level, and sometimes there was not much difference from Level 4 patients in terms of vital signs alone. Therefore, it is necessary to leverage chief complaint to better differentiate between mild Level 3 and 4 patients.

Then, we investigated the impact of different modules, including knowledge prompt (row 3), uncertainty classifier (row 4) and focal loss (row 5). We saw that adding expert knowledge as a prompt was indeed more effective than directly feeding patient information text to a pretrained language model, since the knowledge-prompt brought a 5.38% performance improvement (from 0.911 to 0.960). The attention visualization results can also prove this observation. This improvement brought by the knowledge-prompt verified the advantage of the knowledge-based prompt-tuning technology of the pretrained language model, which leveraged external knowledge bases and provided a constraint to the pretrained language model, increasing its accuracy and reliability. Furthermore, we investigated the impact of the uncertainty-based classifier[42]. We observed that the uncertainty-based classifier remarkably improved the triage performance because removing uncertainly loss function reduced model performance by 3.85% (from 0.960 to 0.923). It is reasonable because the uncertainty-based model dynamically integrates different modalities at an evidence level instead of typically integrating them into a unified comprehensive representation, which increased the model's understanding of different modalities and improved its performance. We can observe that the uncertainty-based classifier brought a 19.90% performance improvement (from 0.789 to 0.946) on Level 3, while in the mean time brought a performance reduction of 0.20% and 0.31% on Level 1 and Level 2, respectively. This is because the uncertainty-based classifier can make model focus more on samples with high uncertainty, which are often difficult to classify. Therefore, the model performed much better on these samples, leading to an improvement in accuracy. However, for samples that are easy to classify, due to the low uncertainty of their classification results, the model may not pay special attention, so the accuracy on these categories may decrease slightly. Lastly, we investigated the impact of Focal Loss[41]. We observed that Focal Loss brought about the 1.27% performance improvement (from 0.948 to 0.960) to our model, which proved the advantage of this loss function. It allowed the model to focus more on categories that are difficult to distinguish, thus effectively alleviating the problem of class imbalance.

### Attention visualization results

Figure 5 presents attention visualization results for two triage cases: One guided by expert knowledge and the other without such guidance. In Fig. 5a, b, attention weights are depicted for each category of data employed by KUTS to inform decisions. In the first case in Fig. 5a, expert knowledge substantially influences the triage process, constituting 54.3% of the total weight, with chief complaint ranking as the second most influential data at 23.4%. This outcome is consistent with the findings depicted in Fig. 5b, where chief complaint carries 62.0% of the total weight, exceeding tabular data. Figure 5c, d illustrate partial input texts with attention weights assigned to each word, emphasizing several most salient words in each case. Figure 5c highlights "1", referenced in the expert knowledge as "..., the recommended triage level is Level 1" along with "wheelchair", observed in the patient's approach of arrival at the hospital. The heightened attention to "wheelchair" is reasonable, given its association with paralysis or other medical conditions. Combing these insights and expert guidance, our model's prediction aligns with the recommended triage level, and they are both consistent with true label. Conversely, in Fig. 5d, representing the case lacking expert guidance, KUTS assigns considerable weight to the fields of approach, heart rate, and chief complaint, all of which registering as slight or common symptoms. In both cases, our model catches the salient features most related to triage outcomes.

## Discussion

KUTS is a foundational classification model for emergency triage that demonstrates strong adaptability to different populations, triage standards, and application scenarios. Our experiments evaluated KUTS on four datasets from distinct hospitals and regions—SYSMH-S, SYSMH-N, GTCMH, and MIMIC-IV-ED—each reflecting varied demographic characteristics. KUTS consistently achieved superior results compared with baseline methods, indicating robust generalization to diverse patient populations. Furthermore, the model performed well across datasets using different triage standards: CETEC, which divides patients into four levels based solely on clinical condition, and ESI, a five-level system that also considers anticipated medical resource use. Despite the conceptual differences between CETEC and ESI, KUTS significantly outperformed existing approaches on both, confirming its flexibility in adapting to distinct triage guidelines.

A major advantage of KUTS lies in its reliance on basic and easily collectible input data, such as vital signs and chief complaints. In contrast, some recent approaches, such as the one proposed by Williams et al.[47], depend on detailed patient histories, which require clinician-patient interaction to obtain. This dependency not only increases operational complexity but may also fail in pre-hospital or emergency scenarios where patients are unresponsive. KUTS avoids this issue by only requiring readily available inputs. Given the increasing accessibility of devices capable of capturing vital signs, KUTS offers a low-cost and practical solution for a wide range of emergency scenarios, including resource-limited settings.

Our experimental results also show that KUTS outperforms previous AI triage methods across all datasets, with an average improvement of 5%. The model achieved an AUC increase of 4.69% on SYSMH-S, 3.56% on SYSMH-N, 6.38% on GTCMH, and 2.08% on MIMIC-IV-ED, as shown in Table 2. These gains can be attributed to two key design features: the use of a knowledge prompt-tuning pretrained language model and an uncertainty-based classifier. The pretrained model captures rich semantic information from large medical corpora, while the knowledge prompts inject expert-derived guidance during fine-tuning. In contrast, many earlier models are trained from scratch without external knowledge, limiting their ability to generalize beyond the training data. The uncertainty-based classifier further enhances performance by quantifying prediction confidence, enabling better classification of ambiguous or borderline cases—something traditional cross-entropy-based classifiers cannot provide.

One of the most compelling benefits of KUTS is its improved accuracy in classifying Level 3 patients, a group historically difficult for both human clinicians and previous AI models to categorize. Level 1 (most severe), Level 2 (relatively severe), and Level 4 (least severe) cases are generally easier to distinguish, but Level 3 patients, who may present with intermediate or ambiguous symptoms, are prone to misclassification. Clinicians often err on the side of caution, assigning higher severity levels to such patients, which can lead to resource waste. KUTS addresses this challenge effectively, achieving notable AUC improvements for Level 3 classification: 14.81% on SYSMH-S, 9.17% on SYSMH-N, and 4.19% on GTCMH. This performance is largely due to KUTS's ability to focus on uncertain cases through its uncertainty-based design, while also drawing on expert-informed semantic understanding from the knowledge-based encoder.

KUTS also demonstrates significant advantages in handling missing data, a common challenge in emergency settings. Traditional machine learning models typically require complete feature vectors and often rely on imputation methods to fill in missing values—approaches that are both time-consuming and prone to introducing errors. In contrast, KUTS leverages the flexibility of text-based inputs and the contextual reasoning capabilities of pretrained language models. When certain attributes are missing, they can simply be omitted from the input text without degrading model performance. This feature is particularly valuable in real-world emergency environments, especially during pre-hospital care or mass casualty incidents, where full patient information may not be immediately available. By eliminating the need for imputation and reducing the preprocessing burden, KUTS streamlines the triage process and enhances both efficiency and accuracy.

The strong generalization ability of KUTS is further validated through domain shift experiments, where the model consistently outperformed baselines when tested across different institutions. This is primarily attributed to its architecture, which combines pretrained language models with knowledge prompt-tuning. The pretrained models benefit from large-scale medical corpora, capturing deep contextual understanding, while expert knowledge prompts help align the model with the specific requirements of triage tasks. This dual mechanism enables KUTS to adapt well to unseen data distributions, a crucial advantage for deployment in diverse healthcare environments. The success of the domain shift experiments highlights a promising future application: training KUTS on comprehensive data from well-resourced urban hospitals and deploying it in under-resourced rural hospitals. Such an approach could help democratize access to high-quality triage, improve healthcare equity, and enhance outcomes by enabling early and accurate prioritization of care.

While KUTS shows considerable promise, it is important to acknowledge its limitations. The model has been validated on multiple datasets with different triage scales and settings, yet its performance in other real-world emergency contexts remains to be further examined. Variations in hospital infrastructure, medical protocols, and available resources could impact its applicability. Additionally, although we evaluated KUTS on CETEC and ESI, other triage systems such as the Manchester Triage System (MTS), Australasian Triage Scale (ATS), and Canadian Triage and Acuity Scale (CTAS) warrant exploration in future work. Despite its strong average performance, no model—including KUTS—can guarantee perfect predictions. Emergency triage decisions are inherently complex and patient conditions highly variable. Thus, we recommend that KUTS be used as a decision-support tool, augmenting but not replacing clinical expertise.

Looking ahead, KUTS may also be adapted for pre-hospital emergency care scenarios, such as disaster response or ambulance triage. Its lightweight input requirements and robustness to missing data make it a suitable candidate for deployment in challenging environments. Furthermore, future research could explore the integration of multimodal data, such as images or video, to enhance diagnostic accuracy. For example, photographs of injuries or facial expression analysis might provide additional insights to inform triage decisions. These avenues represent exciting opportunities to further expand the impact and effectiveness of KUTS in real-world emergency medicine.

## Data availability

Restrictions apply to the availability of SYSMH-S, SYSMH-N and GTCMH, which were used with permission of the participants for the current study. De-identified data may be available for research purposes from the corresponding authors on reasonable request. MIMIC-IV-ED[45] is a publicly available dataset that can be accessed on https://physionet.org/content/mimic-iv-ed/2.2/, and its Excel version can be found in Supplementary Data 1. All source data underlying the graphs and charts presented in the main figures and tables source data for the figures can be found in Supplementary Data 2. Each sheet in the file corresponds to a specific figure or table in the main text, with sheet names matching the figure or table labels (e.g., Fig. 1, Table 1).

## Code availability

The custom code is available at https://github.com/TobyLieu/KUTS. To be more comprehensive, the corresponding system based on KUTS is constructed, which can be found and tested in https://yx-nscc-sub.cn/triage/. The username and the password are guest and kuts1234, respectively. The link on the website provides a demonstration video that can guide users on how to use this system.

Received: 18 June 2024; Accepted: 18 July 2025;
Published online: 31 July 2025

## References

1. Zaboli, A. et al. Emergency departments in contemporary healthcare: are they still for emergencies? an analysis of over 1 million attendances. *Healthcare* 12, 2426 (2024).
2. Gonçalves-Bradley, D. et al. Primary care professionals providing non-urgent care in hospital emergency departments. *Cochrane Database Syst. Rev.* 11, CD002097 (2018).
3. Ullman, A. J. & Davidson, P. M. Patient safety: the value of the nurse. *Lancet* **397**, 1861–1863 (2021).
4. Kuehn, B. M. WHO: strengthen nurse workforce. *JAMA* **323**, 1886–1886 (2020).
5. Oh, W.-O. & Jung, M.-J. Triage-clinical reasoning on emergency nursing competency: a multiple linear mediation effect. *BMC Nurs.* **23**, 274 (2024).

6. Hu, F. et al. The impact of simulation-based triage education on nursing students' self-reported clinical reasoning ability: a quasi-experimental study. *Nurse Educ. Pract.* **50**, 102949 (2021).
7. Zaboli, A. et al. Enhancing triage accuracy: the influence of nursing education on risk prediction. *Int. Emerg. Nurs.* **75**, 101486 (2024).
8. Reay, G., Smith-MacDonald, L., Then, K. L., Hall, M. & Rankin, J. A. Triage emergency nurse decision-making: incidental findings from a focus group study. *Int. Emerg. Nurs.* **48**, 100791 (2020).
9. Australasian College for Emergency Medicine. Guidelines on the implementation of the Australasian triage scale in emergency departments, V6 G24 (2023).
10. Warren, D. W. et al. Revisions to the Canadian triage and acuity scale paediatric guidelines (PaedCTAS). *Can. J. Emerg. Med.* **10**, 224–232 (2008).
11. Wuerz, R. C., Milne, L. W., Eitel, D. R., Travers, D. & Gilboy, N. Reliability and validity of a new five-level triage instrument. *Acad. Emerg. Med.* **7**, 236–242 (2000).
12. Mackway-Jones, K., Marsden, J. & Windle, J.*Emergency triage: Manchester triage group* (Wiley, 2013).
13. Donglei, S., Xiaoying, L. & Ying, Z. et al. Emergency triage expert consensus [in Chinese]. *Chin. J. Emerg. Med.* **27**, 599–604 (2018).
14. Zachariasse, J. M. et al. Performance of triage systems in emergency care: a systematic review and meta-analysis. *BMJ Open* **9**, e026471 (2019).
15. Sax, D. R. et al. Evaluation of the emergency severity index in US emergency departments for the rate of mistriage. *JAMA Netw. Open* **6**, e233404–e233404 (2023).
16. Zaboli, A. et al. Daily triage audit can improve nurses' triage stratification: a pre–post study. *J. Adv. Nurs.* **79**, 605–615 (2023).
17. Brouns, S. H. et al. Performance of the Manchester triage system in older emergency department patients: a retrospective cohort study. *BMC Emerg. Med.* **19**, 1–11 (2019).
18. Mistry, B. et al. Accuracy and reliability of emergency department triage using the emergency severity index: An international multicenter assessment. *Ann. Emerg. Med.* **71**, 581–587 (2018).
19. Blomaard, L. C. et al. Geriatric screening, triage urgency, and 30-day mortality in older emergency department patients. *J. Am. Geriatr. Soc.* **68**, 1755–1762 (2020).
20. Christ, M., Grossmann, F., Winter, D., Bingisser, R. & Platz, E. Modern triage in the emergency department. *Dtsch. Ärzteblatt Int.* **107**, 892 (2010).
21. Raita, Y. et al. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit. Care* **23**, 1–13 (2019).
22. Song, Y.-Y. & Ying, L. Decision tree methods: applications for classification and prediction. *Shanghai Arch. psychiatry* **27**, 130 (2015).
23. Hearst, M., Dumais, S., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Appl.* **13**, 18–28 (1998).
24. Rish, I. et al. An empirical study of the naive Bayes classifier. In *IJCAI Proc. Workshop on Empirical Methods in Artificial Intelligence*, vol. 3, 41–46 (Citeseer, 2001).
25. Biau, G. & Scornet, E. A random forest guided tour. *Test* **25**, 197–227 (2016).
26. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794 (SIGKDD, 2016).
27. 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.* **71**, 565–574 (2018).
28. Huang, X., Khetan, A., Cvitkovic, M. & Karnin, Z. S. TabTransformer: tabular data modeling using contextual embeddings. *CoRR* **abs/2012.06678** (2020).
29. Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting deep learning models for tabular data. In *Proc. Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, 18932–18943 (NIPS, 2021).
30. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* **abs/1810.04805** (2018).
31. Soenksen, L. R. et al. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digit. Med.* **5**, 149 (2022).
32. Zhou, H.-Y. et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng.* **7**, 743–755 (2023).
33. Taud, H. & Mas, J.-F. Multilayer perceptron (MLP). *Geomatic Approaches for Modeling Land Change Scenarios* 451–455 (MLP, 2018).
34. Vaswani, A. et al. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008 (NIPS, 2017).
35. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020).
36. Savage, T., Nayak, A., Gallo, R., Rangan, E. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digit. Med.* **7**, 20 (2024).
37. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* 1–9 (2024).
38. Ren, K. et al. Uncertainty-informed mutual learning for joint medical image classification and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 35–45 (Springer, 2023).
39. Linmans, J., Elfwing, S., van der Laak, J. & Litjens, G. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. *Med. Image Anal.* **83**, 102655 (2023).
40. Chua, M. et al. Tackling prediction uncertainty in machine learning for healthcare. *Nat. Biomed. Eng.* **7**, 711–718 (2023).
41. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proc. IEEE International Conference on Computer Vision*, 2980–2988 (ICCV, 2017).
42. Wang, M. et al. Uncertainty-inspired open set learning for retinal anomaly identification. *Nat. Commun.* **14**, 6757 (2023).
43. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *Proc. 7th International Conference on Learning Representations* (ICLR, 2019).
44. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, 8024–8035 (NIPS, 2019).
45. Johnson, A. et al. MIMIC-IV. *PhysioNet* 49–55 (2020).
46. Johnson, A. E. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).
47. Williams, C. Y. et al. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Netw. Open* **7**, e248895–e248895 (2024).


## Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant No.2024YFA1011900, by Open Project Program of Guangxi Key Laboratory of Digital Infrastructure under Grant No.GXDIOP2024011, by the Guangzhou Science and Technology Plan (No.SL2023A03J01292), by the Guangxi Key R&D Program (No.2024AB08144), and by Guangdong Provincial Key Laboratory of Research on Emergency in TCM (No.2023B1212060062).

## Author contributions

T.L.: study conceptualization and design, construction of model, technical implementation, data analysis, statistical analysis, manuscript drafting; Y.G.: study conceptualization and design, data preparation, resources;

H.C.: study conceptualization and design, construction of model, data analysis; Y.Z.: study conceptualization and design, data preparation, statistical analysis; L.Z. and X.H.: data analysis, statistical analysis; Y.X. and C.W.: data preparation; M.C. and J.L.: statistical analysis; D.H., F.C., Y.Z., H.C., Y.G., and ML.: resources; G.Z.: data analysis; H.W.: resources; C.W.: study conceptualization and design, construction of model, manuscript drafting; X.X., L.L.: study conceptualization and design, data preparation, resources; T.Y.: study conceptualization and design, data preparation, resources, manuscript drafting.

## Competing interests
The authors declare no competing interests.

## Additional information
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s43856-025-01052-w.

**Correspondence** and requests for materials should be addressed to Changdong Wang, Xiaotu Xi, Li Li or Tao Yu.

**Peer review information** *Communications Medicine* thanks Thomas Beaney and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025