BMC Emergency Medicine

RESEARCH

Open Access

# Deep learning-based natural language processing for critical care identification in pediatric emergency department

Jiyoung Agatha Kim[1†], Sangyeon Cho[2†], Minyoung Hwang[3], Dongjoon Lee[4], Jangyeong Jeon[2], Rhoan Lee[5], Changhee Lee[3], Junyeong Kim[2*†] and Woori Bae[1*†]

## Abstract

**Background** The initial assessment of pediatric emergency patients is challenging due to diverse clinical presentations and the limited applicability of existing mortality scoring systems that require detailed laboratory results often unavailable at first encounter. Deep learning (DL)–based Natural Language Processing (NLP) offers the ability to extract complex patterns within textual data attained at the first encounter with the physician and can enhance the accuracy and timely prediction of critically ill patients in the Pediatric emergency department (PED). This study identified whether DL–based NLP techniques, which integrate structured data (vital signs) with unstructured data (clinical text written by physicians), can accurately identify patients requiring critical care in PEDs.

**Methods** Four machine learning (ML) models using term frequency-inverse document frequency-logistic regression, extreme gradient boosting, gradient boosting, and random forest, and two DL–based NLP models based on the Korean Medical Bidirectional Encoder Representations from Transformers (KM-BERT) framework were tested on 87,748 children, including 599 (0.7%) critical cases and 14,095 (16.1%) admission cases in a tertiary hospital, from January 1, 2012, to December 31, 2021.

**Results** KM-BERT, using a Masked Language Model (MLM), outperformed all ML-based models in predicting critically ill patients, achieving the highest Area Under the Receiver Operating Curve (AUROC) of 0.941 ± 0.015 and an accuracy of 0.952 ± 0.002. It also demonstrated superior performance in predicting hospitalized patients, with an AUROC of 0.879 ± 0.004 and an accuracy of 0.798 ± 0.002.

**Conclusion** These findings suggest that DL-based NLP approaches, particularly KM-BERT with MLM, may improve the early identification of critically ill and hospitalized pediatric patients in the PED compared with conventional ML-based

†Jiyoung Agatha Kim and Sangyeon Cho contributed equally to this work.

†Junyeong Kim and Woori Bae contributed equally to this work.

*Correspondence:
Junyeong Kim
junyeongkim@cau.ac.kr
Woori Bae
baewool7777@hanmail.net

Full list of author information is available at the end of the article

© The Author(s) 2026. **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/.

models. Further prospective, multi-center studies are required to validate these findings and establish their suitability for routine clinical integration.

**Keywords** Pediatric emergency department, Critical care, Pediatrics, Natural language processing, Deep learning, Artificial intelligence

## Introduction

Pediatric emergency department (PED) visits in Korea have tripled over the past decade, reaching approximately 1,350,000 visits in 2019, with a 0.6% incidence in critically ill patients [1]. Critically ill patients benefit from early detection and intervention [2]. Accordingly, there is growing interest in improving the quality of care for critically ill patients in the PED. Therefore, the timely identification of critically ill patients and the application of appropriate interventions are pivotal roles of PEDs.

The initial assessment of patients in the PEDs is an especially sensitive process owing to the heterogeneity of presentation, with a diverse range of clinical conditions, particularly for those requiring intensive care [3]. Previously reported scoring models, such as the Pediatric Risk of Mortality (PRISM) III/IV and Pediatric Index of Mortality (PIM)-3, have been developed to predict outcomes and mortality in the pediatric settings [4]. However, these scoring systems are insufficient in the PED setting because of their reliance on precise vital signs and blood test results, which are often unavailable during the initial patient encounter [5, 6]. Thus, decisions regarding interventions are primarily dependent on the physicians' clinical judgments at this stage of care [7].

Machine learning (ML) has been introduced to assist in accurately predicting diseases in the ED for achieving prompt clinical intervention [8]. In previous studies, ML models were used to assist in the stratification of patients for clinical prioritization based on their acuity level at the triage stage [9]. However, the predictive performance of this study was limited because of its dependence on basic structured information from patients for model development.

Unstructured clinical texts written by physicians on the initial patient encounter offer a more detailed and nuanced assessment of the patient compared to structured data. Incorporating information acquired from unstructured clinical text can enhance the accuracy and timeliness of prediction models, particularly in the pediatric population, where diverse disease entities, overlapping symptoms, and varied clinical presentations present significant diagnostic challenges. Deep Learning (DL)–based Natural Language Processing (NLP) enables the extraction of complex patterns and contexts from textual data. Therefore, this study aimed to develop a prediction model to identify patients who required critical care using a DL–based NLP applied to clinical texts written by physicians on the initial patient encounter. Furthermore, we sought to quantify the relative importance of the predictor variables in determining critical cases and hospitalizations.

## Methods

### Study design and setting

This retrospective study used electronic medical record data obtained from a PED at a tertiary hospital in South Korea. The study included patients aged < 18 years who visited the PED between January 1, 2012, and December 31, 2021. The electronic medical record was assessed on April 21, 2023, and the clinical text written by physicians on the initial patient encounter were de-identified and made available for research. The exclusion criteria were patients who visited because of intoxication or injury, patients with missing diagnosis codes, patients with no information about disposition, or patients with errors in their records. Information obtained at the initial stage of the patients' clinical history taking was analyzed, including patient age, sex, means of transportation, initial vital signs (temperature, pulse rate, and respiratory rate), disposition, and unstructured clinical texts written by physicians.

### Outcomes

The primary outcome of this study was the prediction of patients who required critical care. Patients requiring critical care were defined as (1) those who were admitted or transferred to the intensive care unit (ICU) and (2) children who died in the PED. The secondary outcome was all PED patients who were hospitalized or were transferred to another hospital for admission, including those classified under the primary outcome.

### Data preprocessing

To account for age-related variations in vital signs, the pulse and respiratory rates were first standardized using age-specific Z-scores, whereas temperature was standardized across all age groups. The clinical text written by physicians includes a description of the patient's condition, primary complaints, history, and symptoms at the time of presentation to the PED. In this study, the clinical texts extracted were written with significant linguistic heterogeneity, characterized by medical terminologies presented in either English or Korean, as well as the use of abbreviations, symbols, and acronyms. To optimize the reliability of the subsequent prediction models, a systematic text preprocessing approach was implemented

that involved rigorously segregating linguistic components. Specifically, we parsed Korean and English lexical elements, numerical components, and special characters before integrating them into prediction models. Figure 1 presents a visual overview of the prediction model architecture and data processes.

### Machine learning–based prediction model

Two distinct prediction tasks were defined: identifying patients requiring critical care (primary outcome) and predicting hospital admissions (secondary outcome). For each task, four ML models, a statistical method (Logistic Regression) [10] and three tree–based ensembles (Random Forest, Extreme Gradient Boosting (XGBoost) and Gradient Boosting) [11–13], were trained using only structured data (patient age, sex, mode of arrival, vital signs, and disposition) to predict primary and secondary outcomes, respectively. To leverage the information contained within clinical texts written by physicians, the texts were incorporated using the classic NLP method known as Term Frequency-Inverse Document Frequency (TF-IDF) [14], which quantifies the importance of words within a document relative to a larger corpus. By calculating the TF-IDF scores for words within each patient's clinical texts, a numerical representation of the clinical narrative was created, highlighting terms with high relevance. These TF-IDF scores were augmented as input features to train the four ML–based prediction models. This process enabled us to evaluate the prediction efficiency of ML–based models the primary and secondary outcomes based on both the information extracted from the structured data and the physicians' textual assessments.

### Deep learning-based prediction model

For predicting the primary and secondary outcomes, a DL–based model was conducted using KM-BERT, a domain specific Korean language model originally pretrained on a broad Korean medical corpus encompassing medical textbooks, health related news articles, and scientific research papers [15]. We used KM-BERT, equipped with a 16,424 subword vocabulary, and performed additional pretraining on clinical text written by

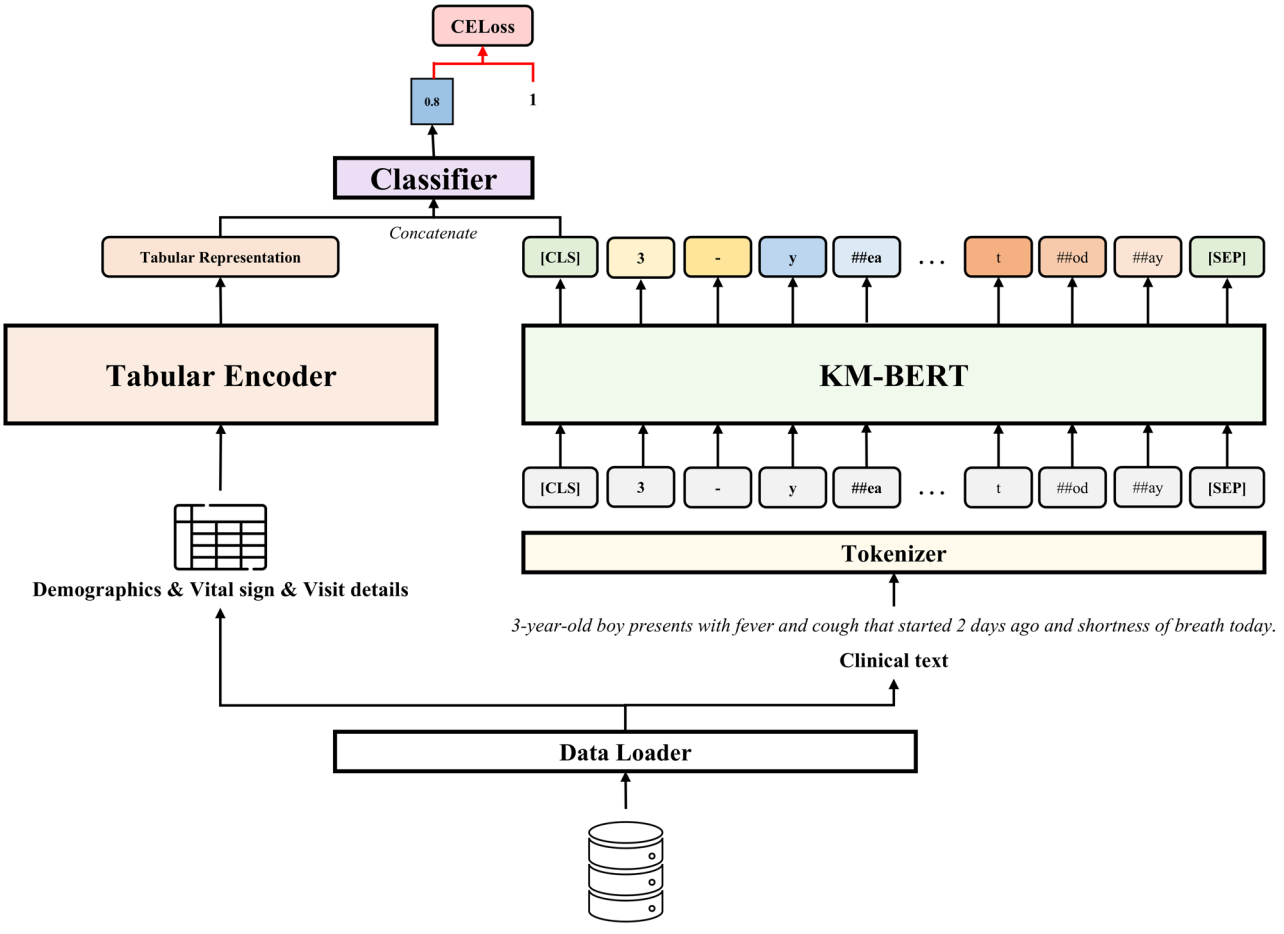



**Fig. 1** Overview of prediction model architecture and data processes. To optimize the reliability of the subsequent prediction models, a systematic text preprocessing approach was implemented that involved rigorously segregating linguistic components. We parsed Korean and English lexical elements, numerical components, and special characters before integrating them into prediction models

physicians using the masked language modeling (MLM) task [16]. In this phase, the selected subwords in each sentence were randomly masked, and the model learned to recover them by leveraging the contextual information from the surrounding subwords. This approach enables the model to gain a nuanced understanding of the semantic contexts specific to clinical texts written by physicians. Next, we fine-tuned the pretrained KM-BERT model with the clinical texts to predict primary and secondary outcomes using the subwords of the given clinical texts as input features. The prediction model resulting from this process is referred to as KM-BERT (MLM).

To effectively integrate both structured and unstructured data in DL–based critical illness prediction models, we designed a novel architecture by introducing a multilayer perceptron as a Tabular Encoder and combined it with the pretrained KM-BERT. This approach allowed us to leverage the strengths of each model type: the Tabular Encoder processed structured information such as patient demographics and vital signs, generating meaningful latent representations, whereas KM-BERT extracted rich contextual embeddings from unstructured clinical texts. By concatenating the output embeddings from both encoders, we create a unified feature representation that captures the complementary information present in both data modalities. This combined representation was then used as an input to a prediction layer, enabling the model to identify patients requiring critical care and predict hospital admissions for primary and secondary outcomes, respectively.

#### *Model training and testing*

Using scikit-learn (v1.2.1) in Python 3.9.13, we randomly allocated the cohort into training, validation, and test sets in a 64:16:20 ratio. To maintain the validity of our empirical comparisons, we used the same data partition for each machine learning and deep learning model. A bootstrapping technique was used for the test set to evaluate the model. A 0.75 sampling ratio was applied, and 10 random iterations were performed to obtain reliable statistical results. These results presented both the mean and standard deviation, thus providing a comprehensive characterization of the model's performance. Model discrimination was evaluated using seven principal metrics: area under the receiver operating characteristic curve (AUROC) [17], area under the precision recall curve (AUPRC), accuracy, recall, precision, specificity, and F1-score [18]. The predictive power was evaluated using the Brier score [19].

ML–based models were trained by identifying the most salient linguistic features, which were determined through the cumulative distribution of words in clinical texts written by physicians. Our approach yielded the top 103 words, which spanned 85% of the total word distribution. TF-IDF was then used to select high frequency words using the Python package *scikit-learn* and subsequently trained four ML–based prediction models on top of the TF-IDF scores: Logistic Regression, Random Forest, and Gradient Boosting using the Python package *scikit-learn* and XGBoost using the Python package *xgboost* (version 1.7.6). The hyperparameters of the ML–based models were optimized using a grid search. For tree–based ensemble models, the maximum depth values tested were 1, 2, 3, 4, and 5, whereas the number of estimators ranged between 100, 200, 300, 400, and 500. For logistic regression, we investigated the regularization coefficient of 100 values evenly spaced on a logarithmic scale between $10^{-4}$ and $10^{4}$. The optimization process was based on AUROC performance, which was evaluated using the validation set. For the DL–based model, we used KM-BERT and conducted pretraining using an MLM task applied to our training set, focusing on limited vocabulary. Preprocessing involved tokenizing each clinical text using KM-BERT's subword tokenizer, with the resulting subword tokens serving as input features. The model architecture was augmented with a binary classification layer, specifically a single fully connected layer, using the CLS token representation from the pretrained KM-BERT as the input. To mitigate overfitting and ensure optimal model performance, we selected the model with the lowest validation loss using a batch size of 64 and incorporating early stopping in both the pretraining and fine-tuning phases. The learning rate was set to 5e-4 for pretraining and 1e-5 and 1e-2 for KM-BERT and the classifier, respectively, for fine-tuning.

#### *Evaluating word contribution to model predictions*

We developed a sliding window leave one out method to quantify the predictive impact of individual words. The method slides a fixed size window across the text word by word. To focus the analysis on semantic content, regular expressions were used to exclude noninformative tokens such as punctuation. At each step, words within the window were temporarily masked to measure the resulting shift in the model's prediction score. Each word level importance score, a measure of its contextual influence, is the average of the scores from all windows in which it appeared. This score represents the word's contribution to increasing the predicted risk; each point on the plot corresponds to the score for a word in a single sample, where higher values signify a greater contribution to elevating risk in that context. To prevent infrequent words from dominating the rankings, we constrained the analysis in two ways. First, it was restricted to the meaningful word set used to train our ML models, which was initially identified through TF-IDF. Second, only the words that appeared at least 10 times in the entire dataset were included in the final analysis and visualization.

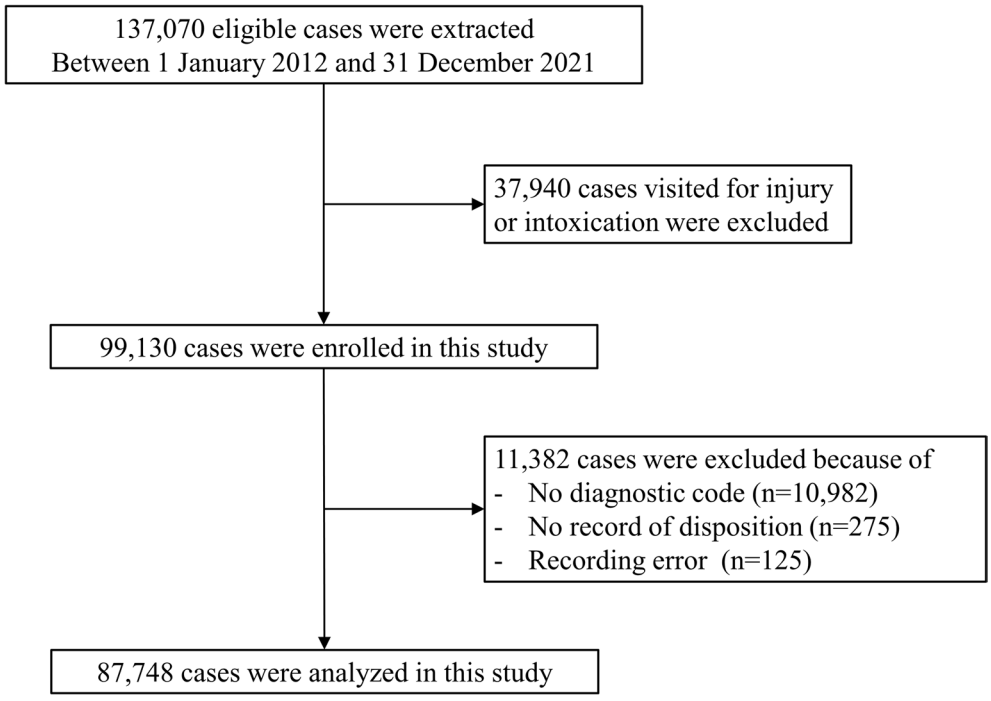



**Fig. 2** Flowchart of the study population

**Table 1** Structured data and outcomes in 87,748 pediatric patients visiting to the PED

| Variables/category | | Value |
|---|---|---|
| Age, median (IQR), month | | 47.0 (9.0–85.0) |
| Female | | 39,504 (45.0) |
| Vital signs, median (IQR) | Temperature, °C | 37.3 (35.6–39.0) |
| | Pulse rate, beats/min | 110.0 (86.0-134.0) |
| | Respiratory rate, breaths/min | 22.0 (18.0–26.0) |
| Mode of arrival | Self-referred | 76,779 (87.5) |
| | Referred from clinic | 7546 (8.6) |
| | Outpatient department | 3423 (3.9) |
| Means of transportation | Walk-in | 83,791 (95.5) |
| | 119 Ambulance | 6535 (7.4) |
| | Hospital Ambulance | 3423 (3.9) |
| | Air Ambulance | 20 (0.2) |
| | Public services | 7 (0.01) |
| | Others | 8 (0.01) |
| Outcomes | Critical care | 599 (0.7) |
| | Hospitalization | 14,095 (16.1) |

All values are frequencies (%) except where otherwise indicated. PED, Pediatric emergency department

## Results

The database included 137,070 cases from January 1, 2012, to December 31, 2021. After excluding 37,940 patients who visited for injury or intoxication, 10,982 patients without a diagnostic code, 275 patients with no record of disposition, and 125 patients with recording errors, 87,748 patients were enrolled in this study (Fig. 2).

Table 1 presents the characteristics of the PED patients. The median (interquartile range) age was 47.0 (9.0–85.0) months, and 39,504/87,748 (45.0%) were female. There were 599/87,748 (0.7%) patients requiring critical care, and 14,095/87,748 (16.1%) were hospitalized.

Table 2 presents predictive performance for identifying critically ill patients using both structured and unstructured data. Among the topic models incorporating structured and unstructured data with Term Frequency-Inverse Document Frequency (TF-IDF), logistic regression showed the highest area under the receiver operating characteristic curve (AUROC) of 0.904 ± 0.016, while Extreme Gradient Boosting (XGBoost) showed the

**Table 2** Comparing predictive performance for critically ill patient using structured and unstructured data

| Variable | Structured data | | Structured and unstructured data | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Logistic regression | XGBoost | Logistic regression | XGBoost | Gradient boosting | Random forest | KM-BERT | KM-BERT with MLM |
| AUROC | 0.851 ± 0.015 | 0.871 ± 0.013 | 0.904 ± 0.016 | 0.898 ± 0.012 | 0.888 ± 0.010 | 0.854 ± 0.024 | 0.917 ± 0.012 | 0.941 ± 0.015 |
| AUPRC | 0.112 ± 0.024 | 0.117 ± 0.029 | 0.116 ± 0.026 | 0.151 ± 0.034 | 0.094 ± 0.019 | 0.115 ± 0.031 | 0.162 ± 0.030 | 0.218 ± 0.038 |
| Recall | 0.705 ± 0.037 | 0.709 ± 0.029 | 0.712 ± 0.041 | 0.697 ± 0.027 | 0.710 ± 0.043 | 0.716 ± 0.040 | 0.709 ± 0.042 | 0.716 ± 0.063 |
| Precision | 0.031 ± 0.004 | 0.036 ± 0.005 | 0.054 ± 0.007 | 0.060 ± 0.007 | 0.046 ± 0.005 | 0.029 ± 0.004 | 0.069 ± 0.008 | 0.097 ± 0.015 |
| Specificity | 0.849 ± 0.003 | 0.873 ± 0.003 | 0.916 ± 0.003 | 0.926 ± 0.002 | 0.899 ± 0.002 | 0.838 ± 0.004 | 0.935 ± 0.002 | 0.955 ± 0.002 |
| F1-score | 0.059 ± 0.007 | 0.069 ± 0.009 | 0.101 ± 0.012 | 0.110 ± 0.011 | 0.086 ± 0.009 | 0.056 ± 0.008 | 0.125 ± 0.013 | 0.171 ± 0.024 |
| Accuracy | 0.848 ± 0.003 | 0.871 ± 0.003 | 0.914 ± 0.003 | 0.924 ± 0.002 | 0.898 ± 0.002 | 0.837 ± 0.004 | 0.949 ± 0.002 | 0.952 ± 0.002 |
| Brier | 0.006 ± 0.001 | 0.007 ± 0.001 | 0.006 ± 0.001 | 0.006 ± 0.001 | 0.008 ± 0.001 | 0.006 ± 0.001 | 0.008 ± 0.001 | 0.006 ± 0.001 |

XGBoost: extreme gradient boosting; KM-BERT: Korean Medical Bidirectional Encoder Representations from Transformers; KM-BERT with MLM: Korean Medical Bidirectional Encoder Representations from Transformers with Masked Language Model; AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision recall curve

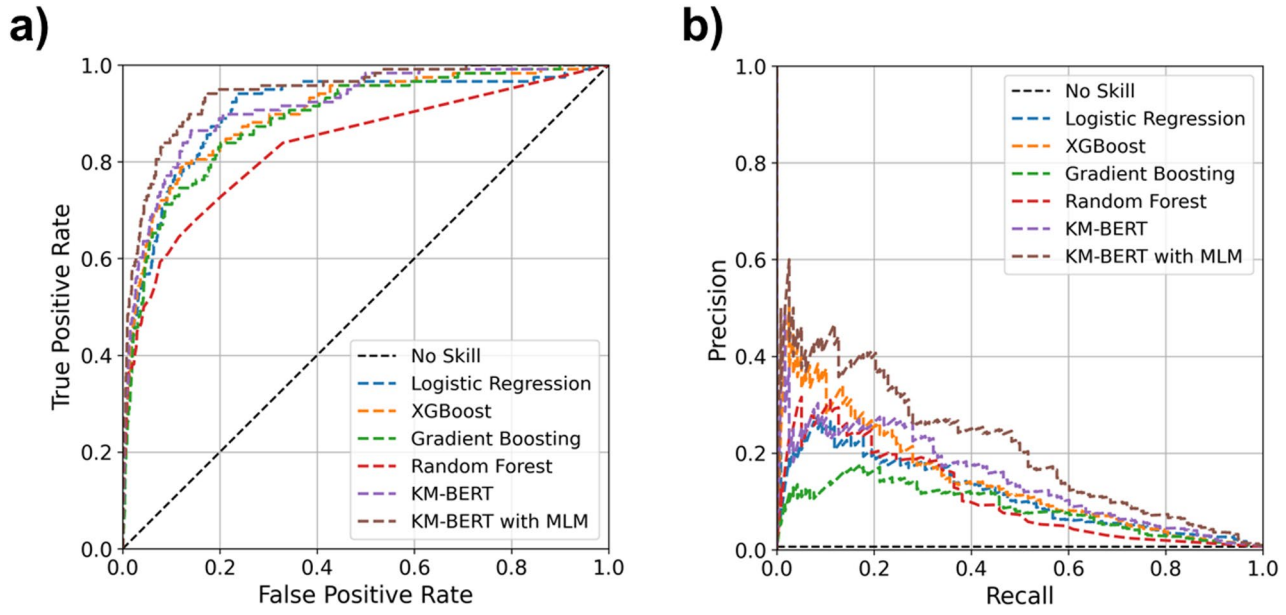



**Fig. 3** Comparison of predictive power of ML–based and DL–based models for predicting critically Ill patients. **a**) Receiver operating characteristic curves. **b**) Precision recall curve. The area under the curve values associated with each model are summarized in Table 2. KM-BERT: Korean Medical Bidirectional Encoder Representations from Transformers; KM-BERT with MLM: Korean Medical Bidirectional Encoder Representations from Transformers with Masked Language Model; XGBoost: extreme gradient boosting

highest area under the precision recall curve (AUPRC) of 0.151 ± 0.034. XGBoost also tops in precision of 0.060 ± 0.007, specificity of 0.926 ± 0.002, F-1 score of 0.110 ± 0.011, and accuracy of 0.924 ± 0.002. The Korean Medical Bidirectional Encoder Representations from Transformers (KM-BERT) with Masked Language Model (MLM) showed superior performance compared to the other models. It revealed an AUROC of 0.941 ± 0.015 and AUPRC of 0.218 ± 0.038. The recall was 0.716 ± 0.063, while the precision was 0.097 ± 0.015. Specificity was 0.955 ± 0.002, and F1-score was 0.171 ± 0.024. The AUROC and AUPRC curves are shown in Fig. 3.

Table 3 presents the predictive performance of the structured and unstructured data for hospitalized patients. Among the ML–based models incorporating structured and unstructured data with TF-IDF, Gradient boosting outperformed the other models, with an AUROC of 0.802 ± 0.004. Random forest had the highest AUPRC of 0.535 ± 0.011. XGboost had the highest recall (0.805 ± 0.007). Gradient boosting had the highest precision (0.286 ± 0.003), specificity (0.615 ± 0.004), F1-score (0.421 ± 0.004), and accuracy 0.645 ± 0.004. Random forest had the lowest Brier score (0.104 ± 0.001). The KM-BERT model, fine-tuned with the MLM, outperformed all the other models. KM-BERT with MLM showed an AUROC of 0.879 ± 0.004 and AUPRC of 0.659 ± 0.013. KM-BERT with MLM showed a precision of 0.431 ± 0.005, specificity of 0.797 ± 0.003, an F1-score of 0.560 ± 0.005, and accuracy of 0.798 ± 0.002. KM-BERT was superior in recall (0.808 ± 0.009 vs. 0.801 ± 0.009).

**Table 3** Comparing predictive performance for hospitalized patient using structured and unstructured data

| Variable | Structured data | | | | Structured and unstructured data | | | |
|---|---|---|---|---|---|---|---|---|
| | Logistic regression | XGBoost | Logistic regression | XGBoost | Gradient boosting | Random forest | KM-BERT | KM-BERT with MLM |
| AUROC | 0.730 ± 0.004 | 0.764 ± 0.005 | 0.767 ± 0.005 | 0.801 ± 0.005 | 0.802 ± 0.004 | 0.793 ± 0.005 | 0.842 ± 0.006 | 0.879 ± 0.004 |
| AUPRC | 0.441 ± 0.009 | 0.494 ± 0.009 | 0.475 ± 0.010 | 0.534 ± 0.010 | 0.532 ± 0.009 | 0.535 ± 0.011 | 0.591 ± 0.014 | 0.659 ± 0.013 |
| Recall | 0.804 ± 0.007 | 0.809 ± 0.011 | 0.799 ± 0.009 | 0.805 ± 0.007 | 0.802 ± 0.007 | 0.804 ± 0.008 | 0.808 ± 0.009 | 0.801 ± 0.009 |
| Precision | 0.225 ± 0.003 | 0.241 ± 0.004 | 0.253 ± 0.003 | 0.284 ± 0.004 | 0.286 ± 0.003 | 0.272 ± 0.004 | 0.351 ± 0.004 | 0.431 ± 0.005 |
| Specificity | 0.468 ± 0.004 | 0.511 ± 0.006 | 0.547 ± 0.003 | 0.611 ± 0.004 | 0.615 ± 0.004 | 0.587 ± 0.003 | 0.713 ± 0.003 | 0.797 ± 0.003 |
| F1-score | 0.352 ± 0.004 | 0.371 ± 0.005 | 0.384 ± 0.004 | 0.420 ± 0.004 | 0.421 ± 0.004 | 0.406 ± 0.005 | 0.489 ± 0.005 | 0.560 ± 0.005 |
| Accuracy | 0.522 ± 0.004 | 0.559 ± 0.005 | 0.587 ± 0.003 | 0.642 ± 0.004 | 0.645 ± 0.004 | 0.622 ± 0.003 | 0.728 ± 0.003 | 0.798 ± 0.002 |
| Brier | 0.114 ± 0.001 | 0.108 ± 0.001 | 0.111 ± 0.002 | 0.104 ± 0.002 | 0.104 ± 0.002 | 0.104 ± 0.001 | 0.098 ± 0.002 | 0.088 ± 0.002 |

XGBoost: extreme gradient boosting; KM-BERT: Korean Medical Bidirectional Encoder Representations from Transformers; KM-BERT with MLM: Korean Medical Bidirectional Encoder Representations from Transformers with Masked Language Model; AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision recall curve

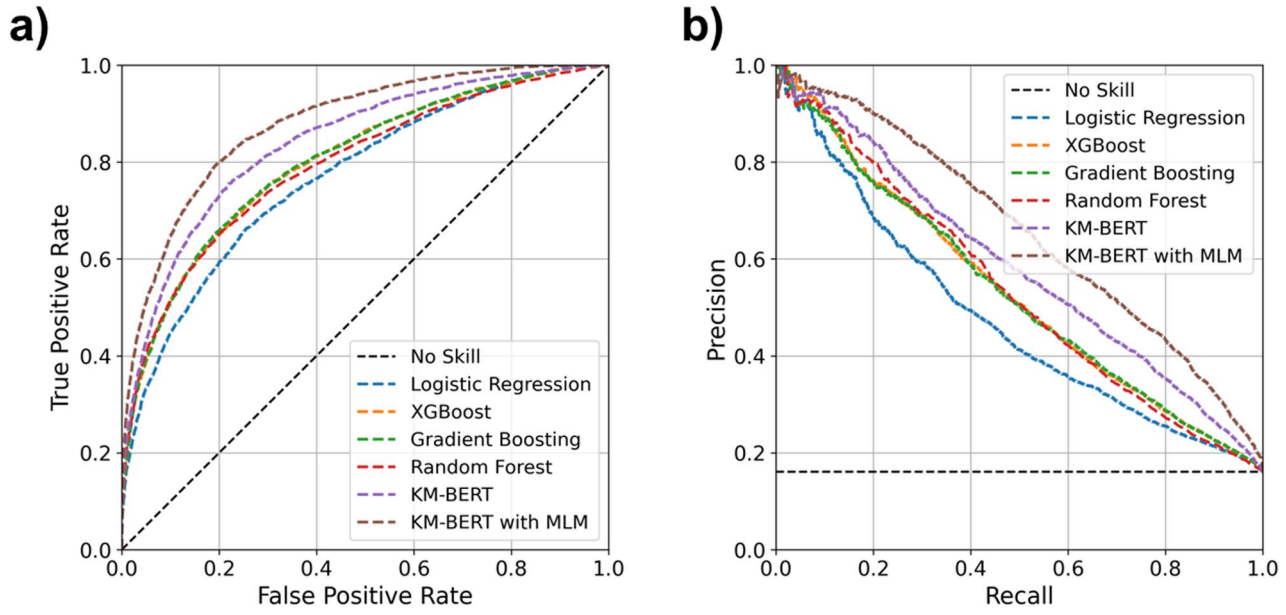



**Fig. 4** Comparison of predictive power of ML–based and DL–based models for predicting hospitalized patients. **a**) Receiver operating characteristic curves. **b**) Precision recall curve. KM-BERT fine-tuned with MLM outperformed all other models across most metrics, demonstrating superior discrimination and calibration compared with ML-based models using structured and unstructured data. KM-BERT: Korean Medical Bidirectional Encoder Representations from Transformers; KM-BERT with MLM: Korean Medical Bidirectional Encoder Representations from Transformers with Masked Language Model; XGBoost: extreme gradient boosting

The Brier score was lower in the KM-BERT with MLM group (0.088 ± 0.002 vs. 0.098 ± 0.002). The AUROC and AUPRC curves are shown in Fig. 4.

The distribution of the prediction probabilities of all models for predicting critically ill and hospitalized patients is shown in Fig. 5. For critically ill patients, KM-BERT with MLM had the highest median (interquartile range) predictive probability of 0.112 (0.018–0.323), followed by KM-BERT with 0.049 (0.013–0.353) and random forest with 0.032 (0.006–0.148). A similar trend was observed for hospitalized patients, with KM-BERT with MLM having the highest median probability of 0.358 (0.129–0.758), followed by KM-BERT at 0.279 (0.085–0.669) and Random Forest at 0.292 (0.134–0.524). The detailed values in Fig. 5 are presented in Supplementary Tables 1 and 2.

Figure 6 illustrates the beeswarm plot of the feature importance based on the window leave one out scores across the total population. Features such as *swelling*, *itching*, *allergy*, and *stiffness* were among the most influential words that contributed to driving model predictions.

## Discussion

### Principal findings

This study evaluated the ability of ML– and DL–based NLP models to predict two clinical outcomes of the PED using clinical texts: critically ill patients and

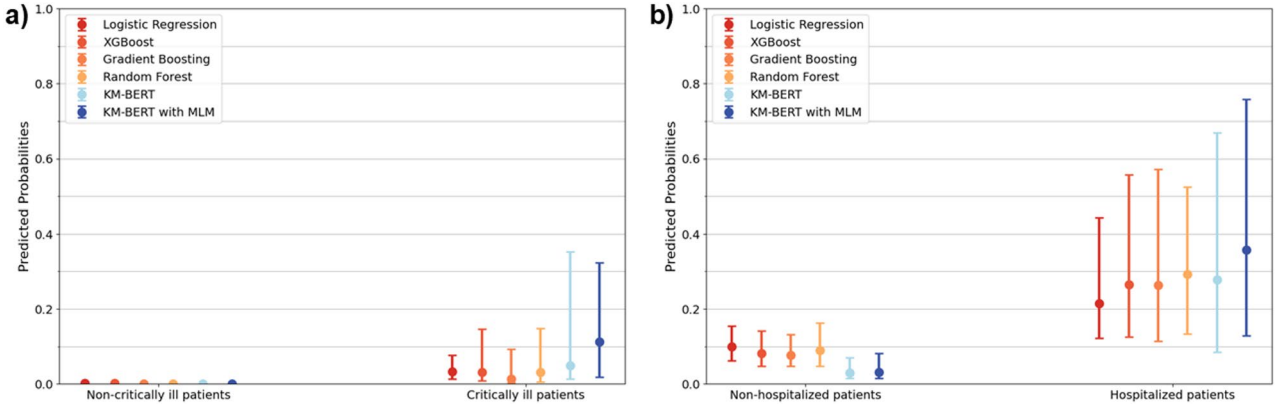



**Fig. 5** Predicted probabilities of ML–based and DL–based models. **a**) Critically ill patients. **b**) Hospitalized patients. KM-BERT with MLM demonstrated the highest median predictive probability, followed by KM-BERT and random forest for both critically ill and hospitalized patients. KM-BERT: Korean Medical Bidirectional Encoder Representations from Transformers; KM-BERT with MLM: Korean Medical Bidirectional Encoder Representations from Transformers with Masked Language Model; XGBoost: extreme gradient boosting

hospitalizations. KM-BERT with the MLM model outperformed the ML–based NLP models for all metrics in predicting critically ill patients. Our findings suggest that DL-based NLP may be a feasible approach for the early identification of critically ill patients in emergency settings.

### Comparison with previous work

Prompt recognition and intervention are essential to improving survival outcomes in critically ill pediatric patients. Early detection and prompt intervention can reverse acute organ dysfunction, reduce the likelihood of intensive care unit (ICU) admission, shorten ICU and hospital stays, and improve both short- and long-term survival [20–24]. Current intensive care scoring models such as PRISM III/IV, PIM-3, and Pediatric Logistic Organ Dysfunction (PELOD)-2 rely primarily on structured data, including vital signs and laboratory results. Another limitation is the use of test results that are difficult to obtain for children in PED settings. Consequently, their feasibility in real time pediatric emergency settings remains limited, particularly when rapid and comprehensive risk assessment is required.

Prior research has explored topic modeling approaches to identify critically ill pediatric patients. One prospective study applied ML techniques and achieved an AUROC of 0.85 for critically ill patients and 0.80 for hospitalized patients [9]. However, this study did not incorporate unstructured clinical data, which may have limited the predictive accuracy. Another study implemented a DL–based prediction algorithm on PED patients, obtaining an AUROC of 0.908 for critically ill patients and 0.782 for hospitalized patients [25]. This study incorporated topic modeling but excluded unstructured data.

Other studies have suggested that NLP-integrated models can meaningfully support triage and early decision-making in emergency settings. For example, Brann et al. reported that a model using free-text triage notes and clinical data available at triage predicted sepsis with an AUROC of 0.94 and concluded that such an approach may facilitate timely and reliable alerting for intervention [26]. In another retrospective study, Chang et al. showed that machine-learning models integrating structured and unstructured triage data outperformed emergency physicians and a triage-level reference model for disposition prediction, while Fernandes et al. demonstrated that adding chief-complaint text and clinical variables improved prediction of ICU admission compared with triage priority alone [27, 28]. Another DL–based NLP study has primarily focused on automated documentation, medical coding, and clinical information extraction in the emergency department [29]. These findings suggest that NLP-based models may serve as feasible decision-support tools for improving workflow, risk recognition, and resource allocation at the initial encounter.

Despite the increasing use of ML and DL in health care, NLP applications in PED triage remain poorly explored. By leveraging state of the art NLP models trained on unstructured clinical text, our study extends the previous work by demonstrating that DL methods can provide meaningful clinical insights that contribute to patient risk stratification in the PEDs.

### Strengths and limitations of the study

This study improved the identification of critically ill and hospitalized patients compared with previous topic modeling approaches using DL–based NLP to analyze unstructured clinical texts. In particular, KM-BERT with the MLM model outperformed all other models in terms of AUROC, AUPRC, precision, specificity, and accuracy. There are several potential explanations for the incremental improvements in the prediction performance

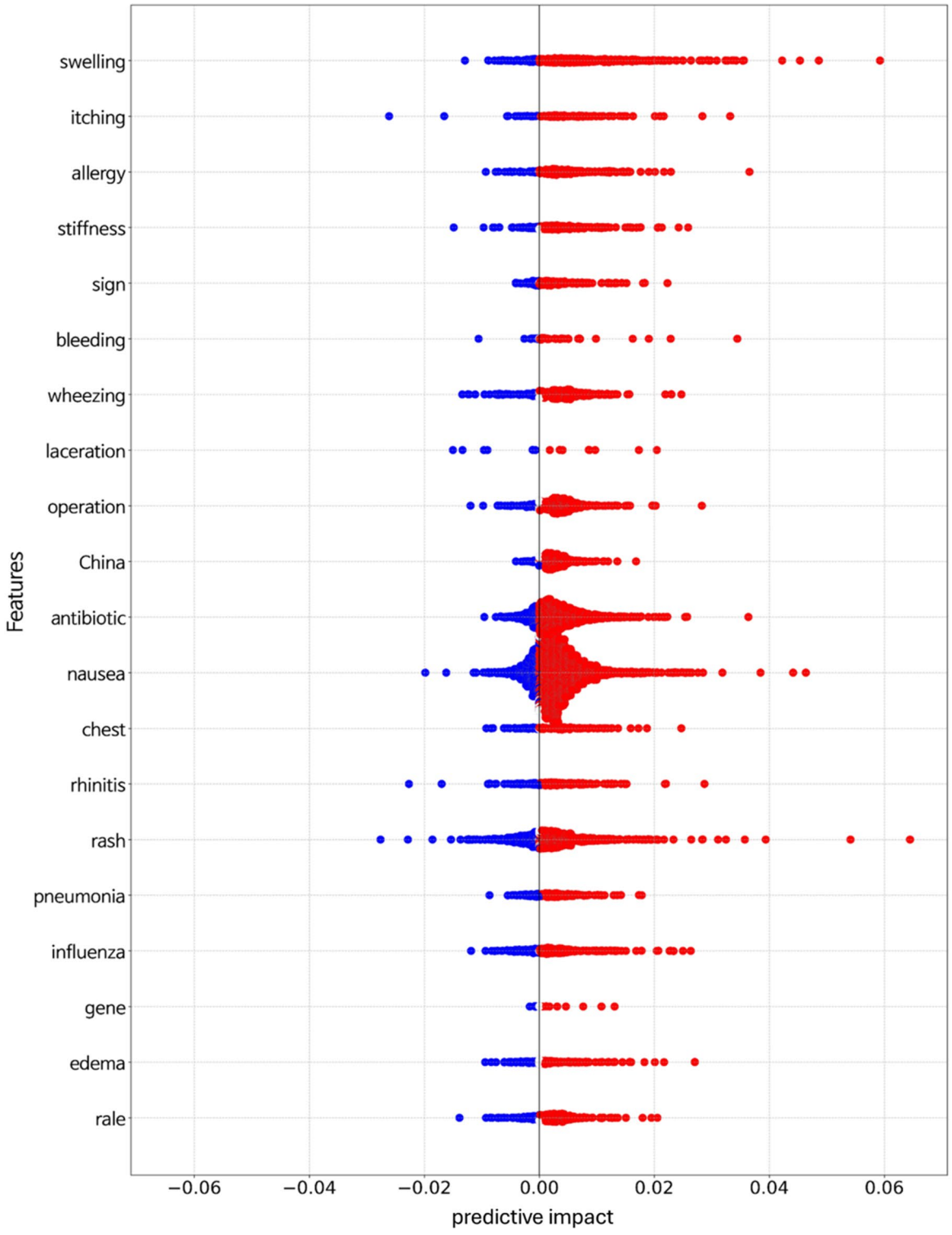



**Fig. 6** Top twenty predictors with the highest impact in predicting critically ill patients in sliding window leave one out analysis

observed with our DL–based NLP approaches. DL models can capture high order nonlinear interactions within unstructured clinical texts that traditional modeling approaches cannot process effectively. DL–based NLP offers scalability and adaptability within the broader framework of health information technology. The integration of advanced NLP techniques enables a more refined interpretation of unstructured text, thereby

further enhancing real-time risk stratification and facilitating rapid decision making. These findings, along with the recent advancements in NLP applications for clinical use, provide promising evidence that artificial intelligence assisted clinical assessments can optimize the early identification of critically ill pediatric patients and improve emergency care outcomes.

Figure 5 shows the predicted probability distributions generated by various ML– and DL–based NLP techniques for identifying critically ill and hospitalized patients. The difference in the median predicted probabilities between the two groups provided insights into the discriminative power of the models. Notably, the DL–based models, particularly KM-BERT with MLM, consistently yielded the highest probability values, indicating their superior predictive performance for identifying critically ill patients. The inclusion of rich contextual information from physician notes and other textual data likely contributes to the superior performance of this prediction model. Compared with conventional scoring models and recent structured data–based ML models, our approach captures nuanced clinical information more accurately, enabling real-time risk assessment at the earliest stage of patient care. This highlights the potential of DL–based NLP to enhance emergency decision-making and patient outcomes. Furthermore, by identifying the key features linked to early intervention in critical care, health care professionals can allocate resources and enhance training for health care students, thereby ensuring the effective monitoring of critically ill patients (Fig. 5). This study highlights the potential application of DL–based NLP methods to streamline the ED process by predicting critically ill patients at the physician-patient contact point. This is particularly noteworthy, given that existing conventional scoring systems cannot be immediately integrated into the clinician assessment stage. By leveraging DL–based NLP to interpret unstructured clinical text, our approach can provide an objective and scalable tool that enhances diagnostic precision and supports prompt data-driven decision-making in pediatric emergency care. Sharing the results obtained from DL–based NLP can enhance communication between medical staff members. Furthermore, streamlined supportive management can be achieved through real-time, continuous reevaluation of every new piece of information obtained during the patient's stay. Future studies should investigate the potential implementation of NLPs within the ED workflow and comprehend the effectiveness of DL–based NLP methods in predicting critically ill patients before implementation in clinical practice.

Although KM-BERT showed the best performance among the evaluated models in our study, whether this level of performance is sufficient for independent clinical use remains uncertain. This was a single center retrospective study that focused on critically ill patients defined only by clinical outcomes. Due to the retrospective nature of our study, there is a high risk of selection bias [30].

Additionally, current guidance for clinical prediction models emphasizes that model quality should be judged not only by discrimination, but also by calibration, clinical usefulness, and generalizability in the intended setting; importantly, internal validation alone is not sufficient to establish safe clinical applicability [31, 32]. Within this framework, our findings should be interpreted as encouraging rather than definitive. Our findings support the potential role of KM-BERT as a clinical decision-support system, rather than as a replacement for clinician judgment, and further prospective validation is needed before independent clinical use can be recommended.

A further limitation of this study is the substantial class imbalance. This has been well recognized in clinical prediction research, as rare clinical outcomes are inherently affected by outcome class imbalance [33, 34]. This issue is particularly relevant in emergency department settings such as the present study, where critically ill patients comprised only 0.7% of the cohort. Several prior emergency department prediction studies have explicitly addressed class imbalance using methods such as random oversampling, random under sampling, and other imbalance-aware modelling strategies such as synthetic minority over-sampling techniques [34–36]. Nevertheless, these techniques may themselves introduce additional bias into NLP-based prediction models and therefore require careful application. Recent methodological studies have shown that random oversampling, random under sampling, and related class-imbalance correction methods can distort the effective training distribution, lead to overestimation of risk, and degrade calibration even when some classification metrics appear to improve [37–39]. Accordingly, in the final methodology of the present study, we intentionally did not apply oversampling methods or class reweighting. This decision was based on our view that a risk prediction model intended for real-world administrative use should provide well-calibrated probabilities, as reflected by the Brier score. Techniques that artificially alter the training distribution or penalize the majority class may inherently distort the predicted probabilities, rendering the predicted risk scores less reflective of actual real-world prevalence. Until more widely accepted clinical guidance becomes available on how class imbalance should be addressed in such settings, the evaluation and interpretation of these models should remain cautious rather than definitive.

Another important limitation is that the model was based exclusively on clinical notes documented at the initial physician encounter, which may not adequately

capture the dynamic evolution of patient status during the PED stay. This concern is supported by recent work showing that real-time multimodal prediction models in the emergency department achieved improved predictive performance when continuously updated clinical data were incorporated at shorter intervals [40]. Even so, while improved accuracy with temporally updated data may be expected, this approach differs from the purpose of the present study, which was to evaluate whether clinically meaningful predictions could be generated rapidly using information available at the time of the initial assessment.

In addition, inter-physician variability in the phrasing, structure, and completeness of clinical notes may have introduced heterogeneity affecting NLP model performance. There is existing evidence that suggests that documentation differences are influenced more by clinical context and specialty-specific language patterns than by physician-level variation alone [41]. While this may indicate that inter-physician variability is less concerning within the same medical specialty, these factors should still be carefully considered in future studies.

Taken together, these considerations suggest that future model development should incorporate longitudinal, real-time clinical information, including serial vital signs, subsequent physician documentation, and early laboratory results, to improve robustness, temporal responsiveness, and clinical applicability. As future research accumulates, we believe that the paradigm of patient classification will change as NLPs become more integrated into clinical practice.

## Conclusions

This study suggests that DL-based NLP approaches may improve the early identification of critically ill and hospitalized pediatric patients. These findings support the potential value of advanced predictive approaches to aid early triage in the pediatric emergency department and improve clinical decision-making and resource allocation. Nevertheless, the results should be considered preliminary rather than definitive, as they do not show that the KM-BERT with MLM model is ready for safe or immediate implementation in real-world emergency department practice. Further prospective, multi-center studies are needed to validate its performance across diverse settings and confirm its reliability, safety, and practical utility before routine clinical integration.

### Abbreviations

| | |
|---|---|
| AUROC | Area under the receiver operating characteristic curve |
| AUPRC | Area under the precision recall curve |
| DL | Deep learning |
| ICU | Intensive care unit |
| KM-BERT | Korean Medical Bidirectional Encoder Representations from Transformers |
| KM-BERT with MLM | Korean Medical Bidirectional Encoder Representations from Transformers with Masked Language Model |
| NLP | Natural Language Processing |
| PIM-3 | Pediatric Index of Mortality-3 |
| PELOD-2 | Pediatric Logistic Organ Dysfunction-2 |
| PRISM III/IV | Pediatric Risk of Mortality III/IV |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| XGBoost | Extreme gradient boosting |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12873-026-01560-5.

Supplementary Material 1

### Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2024-00358602); and by the Ministry of Science and ICT (MSIT), Korea, under the National Program for Excellence in SW, supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP) in 2026 (2025-0-00032). The authors wish to acknowledge the financial support of the Catholic Medical Center Research Foundation in the program year of 2025.

### Author contributions

JAK, SC, JK and WB participated in conceptualizing the study, oversaw the analysis, and were involved in drafting and revising the manuscript. CL, MH, DJ and JJ collected data and developed the machine learning algorithms and data analysis. SC, MH and RL carried out the data-preprocessing. JK and WB 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 research was supported by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP (Institute of Information Communications Technology Planning Evaluation) in 2026 (2025-0-00032).

### Data availability

Data are available from the corresponding author on reasonable request.

## Declarations

### Ethical approval

This study was approved by the Institutional Review Board of the Catholic University of Korea (IRB approval number: KC23RISI0073) in accordance with the Declaration of Helsinki and waived the need for individual informed consent.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

### Author details

[1]Department of Emergency Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea
[2]Department of Artificial Intelligence, Chung-Ang University, Seoul, Korea
[3]Department of Artificial Intelligence, Korea University, Seoul, Korea
[4]VUNO, Seoul, Korea
[5]LG CNS, Seoul, Korea

Received: 21 November 2025 / Accepted: 24 March 2026
Published online: 16 June 2026

**References**

1. Noh H. A nationwide study on emergency department utilization among pediatric patients in Korea: a special focus on pediatric emergency medical centers. Soonchunhyang Med Sci. 2024:13–22.
2. Gunnerson KJ, Bassin BS, Havey RA, Haas NL, Sozener CB, Medlin RP Jr., Gegenheimer-Holmes JA, Laurinec SL, Boyd C, Cranford JA, et al. Association of an Emergency Department-Based Intensive Care Unit With Survival and Inpatient Intensive Care Unit Admissions. JAMA Netw Open. 2019;2(7):e197584.
3. Bradman K, Borland M, Pascoe E. Predicting patient disposition in a paediatric emergency department. J Paediatr Child Health. 2014;50(10):E39–44.
4. Ramazani J, Hosseini M. Prediction of ICU mortality in critically ill children: Comparison of SOFA, GCS, and FOUR score. Med Klin Intensivmed Notfmed. 2019;114(8):717–23.
5. Pines JM, Garson C, Baxt WG, Rhodes KV, Shofer FS, Hollander JE. ED crowding is associated with variable perceptions of care compromise. Acad Emerg Med. 2007;14(12):1176–1181.
6. Pines JM, Hilton JA, Weber EJ, Alkemade AJ, Al Shabanah H, Anderson PD, Bernhard M, Bertini A, Gries A, Ferrandiz S, et al. International perspectives on emergency department crowding. Acad Emerg Med. 2011;18(12):1358–70.
7. Dunser MW, Noitz M, Tschoellitsch T, Bruckner M, Brunner M, Eichler B, Erblich R, Kalb S, Knoll M, Szasz J, et al. Emergency critical care: closing the gap between onset of critical illness and intensive care unit admission. Wien Klin Wochenschr. 2024;136(23–24):651–61.
8. Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. N Engl J Med. 2023;388(13):1201–8.
9. Goto T, Camargo CA Jr, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning–based prediction of clinical outcomes for children during emergency department triage. JAMA Netw open. 2019;2(1):e186937.
10. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression. Wiley; 2013.
11. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
12. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001:1189–232.
13. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining: 2016; 2016:785–794.
14. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag. 1988;24(5):513–23.
15. Kim Y, Kim JH, Lee JM, Jang MJ, Yum YJ, Kim S, Shin U, Kim YM, Joo HJ, Song S. A pre-trained BERT for Korean medical natural language processing. Sci Rep. 2022;12(1):13847.
16. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019:4171–4186.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
18. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:201016061. 2020.
19. Glenn WB. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3.
20. Giacomini M, Iapichino G, Cigada M, Minuto A, Facchini R, Noto A, Assi E. Short-term noninvasive pressure support ventilation prevents ICU admittance in patients with acute cardiogenic pulmonary edema. Chest. 2003;123(6):2057–61.
21. Haas NL, Medlin RP Jr., Cranford JA, Boyd C, Havey RA, Losman ED, Rice MD, Bassin BS. An emergency department-based intensive care unit is associated with decreased hospital length of stay for upper gastrointestinal bleeding. Am J Emerg Med. 2021;50:173–7.
22. Du J, Gunnerson KJ, Bassin BS, Meldrum C, Hyzy RC. Effect of an emergency department intensive care unit on medical intensive unit admissions and care: A retrospective cohort study. Am J Emerg Med. 2021;46:27–33.
23. Mitarai T, Gordon AJ, Nudelman MJR, Urdaneta AE, Nesbitt JL, Niknam K, Graber-Naidich A, Wilson JG, Kohn MA. Association of an Emergency Critical Care Program With Survival and Early Downgrade Among Critically Ill Medical Patients in the Emergency Department. Crit Care Med. 2023;51(6):731–41.
24. Schober A, Sterz F, Herkner H, Locker GJ, Heinz G, Fuhrmann V, Sitzwohl C, Weiser C, Wallmuller C, Stratil P, et al. Post-resuscitation care at the emergency department with critical care facilities–a length-of-stay analysis. Resuscitation. 2011;82(7):853–8.
25. Kwon JM, Jeon KH, Lee M, Kim KH, Park J, Oh BH. Deep Learning Algorithm to Predict Need for Critical Care in Pediatric Emergency Departments. Pediatr Emerg Care. 2021;37(12):e988–94.
26. Brann F, Sterling NW, Frisch SO, Schrager JD. Sepsis Prediction at Emergency Department Triage Using Natural Language Processing: Retrospective Cohort Study. JMIR AI. 2024;3(1):e49784.
27. Chang Y-H, Lin Y-C, Huang F-W, Chen D-M, Chung Y-T, Chen W-K, Wang CC. Using machine learning and natural language processing in triage for prediction of clinical disposition in the emergency department. BMC Emerg Med. 2024;24(1):237.
28. Fernandes M, Mendes R, Vieira SM, Leite F, Palos C, Johnson A, Finkelstein S, Horng S, Celi LA. 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.
29. Maleki Varnosfaderani S, Forouzanfar M. The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century. Bioeng (Basel). 2024;11(4):337.
30. Stewart J, Lu J, Goudie A, Arendts G, Meka SA, Freeman S, Walker K, Sprivulis P, Sanfilippo F, Bennamoun M. Applications of natural language processing at emergency department triage: A narrative review. PLoS ONE. 2023;18(12):e0279953.
31. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, Harrell FE Jr., Martin GP, Moons KGM, van Smeden M, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:e074819.
32. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, Bonnett LJ, Collins GS. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820.
33. Salmi M, Atif D, Oliva D, Abraham A, Ventura S. Handling imbalanced medical datasets: review of a decade of research. Artif Intell Rev. 2024;57(10):273.
34. Look CS, Teixayavong S, Djarv T, Ho AF, Tan KB, Ong ME. Improved interpretable machine learning emergency department triage tool addressing class imbalance. Digit Health. 2024;10:20552076241240910.
35. Chen TL, Chen JC, Chang WH, Tsai W, Shih MC, Wildan Nabila A. Imbalanced prediction of emergency department admission using natural language processing and deep neural network. J Biomed Inf. 2022;133:104171.
36. Porto BM. Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review. BMC Emerg Med. 2024;24(1):219.
37. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inf Assoc. 2022;29(9):1525–34.
38. Piccininni M, Wechsung M, Van Calster B, Rohmann JL, Konigorski S, van Smeden M. Understanding random resampling techniques for class imbalance correction and their consequences on calibration and discrimination of clinical risk prediction models. J Biomed Inf. 2024;155:104666.
39. Yang C, Fridgeirsson EA, Kors JA, Reps JM, Rijnbeek PR. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. J Big Data. 2024;11(1):7.
40. Choi A, Lee K, Hyun H, Kim KJ, Ahn B, Lee KH, Hahn S, Choi SY, Kim JH. A novel deep learning algorithm for real-time prediction of clinical deterioration in the emergency department for a multimodal clinical decision support system. Sci Rep. 2024;14(1):30116.
41. Miller K, Moon S, Fu S, Liu H. Contextual variation of clinical notes induced by EHR migration. In: AMIA Annual Symposium Proceedings. 2024:1155.

## Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.