



OPEN

# A pediatric emergency prediction model using natural language process in the pediatric emergency department

Arum Choi<sup>1,6</sup>, Chohee Kim<sup>2,6</sup>, Jisu Ryoo<sup>3</sup>, Jangyeong Jeon<sup>4</sup>, Sangyeon Cho<sup>4</sup>, Dongjoon Lee<sup>4</sup>, Junyeong Kim<sup>4</sup>, Changhee Lee<sup>5,7</sup>✉ & Woori Bae<sup>3,7</sup>✉

This study developed a predictive model using deep learning (DL) and natural language processing (NLP) to identify emergency cases in pediatric emergency departments. It analyzed 87,759 pediatric cases from a South Korean tertiary hospital (2012–2021) using electronic medical records. Various NLP models, including four machine learning (ML) models with Term Frequency-Inverse Document Frequency (TF-IDF) and two DL models based on the KM-BERT framework, were trained to differentiate emergency cases using clinician transcripts. Gradient Boosting, among the ML models, performed best with an AUROC of 0.715, AUPRC of 0.778, and F1-score of 0.677. DL models, especially the fine-tuned KM-BERT model, showed superior performance, achieving an AUROC of 0.839, AUPRC of 0.879, and F1-score of 0.773. Shapley-based explanations provided insights into model predictions, underlining the potential of these technologies in medical decision-making. This study demonstrates the potential of advanced DL techniques for NLP in emergency medical settings, offering a more precise and efficient approach to managing healthcare resources and improving patient outcomes.

**Keywords** Pediatric emergency department, Emergency room visits, Prediction model, Natural language process, Language model

## Emergency department overcrowding in Korea

Emergency departments (EDs) worldwide, including Korea, are facing the problem of overcrowding, resulting in the overloading of medical staff and inefficient use of ED resources<sup>1–3</sup>. Overcrowding in EDs can lead to negative outcomes such as decreased quality of care, patient safety concerns, and higher healthcare costs<sup>4</sup>. ED overcrowding is a serious problem, especially in pediatric EDs (PEDs), where timely and accurate treatment is important<sup>5</sup>. The challenge of accurately diagnosing and treating children, who are less able to communicate their symptoms than adults, is exacerbated in overcrowded and resource-constrained settings<sup>6</sup>.

## Unnecessary PED visits

PED visits among pediatric patients continue to increase, and a larger proportion of these patients have repeated visits to the PED than adults. While adults often visit the ED for chronic disease issues, children who are more susceptible to infections tend to visit the ED more frequently the younger they are<sup>7</sup>. The high rate of unnecessary, unplanned PED visits in children not only contributes to PED overcrowding but may also reflect poor quality of care<sup>8,9</sup>. Addressing these challenges is critical to improving children's health outcomes and optimizing emergency health care.

## Advantages of application of natural language process using electronic medical records

In the ED, it is crucial to accurately assess a patient's condition and determine appropriate treatment priorities. Typically, prediction models in the ED are built using patients' lab results or numerical data<sup>10,11</sup>. However, before structured patient lab results are available, clinicians describe information about a patient's condition and

<sup>1</sup>Department of Radiology, College of Medicine, The Catholic University of Korea, Seoul, Korea. <sup>2</sup>VUNO, Seoul, Korea. <sup>3</sup>Department of Emergency Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea. <sup>4</sup>Department of Artificial Intelligence, Chung-Ang University, Seoul, Korea. <sup>5</sup>Department of Artificial Intelligence, Korea University, Seoul, Korea. <sup>6</sup>Arum Choi and Chohee Kim have contributed equally to this work as first authors. <sup>7</sup>Changhee Lee and Woori Bae have contributed equally to this work as last authors. ✉email: changheele@korea.ac.kr; baewool7777@hanmail.net

symptoms in the electronic medical record (EMR) in an unstructured form in a natural language. Traditional ML methods are limited by their reliance on structured data, such as detailed (well-organized) patient descriptions and lab results. Obtaining this type of data is often a slow and resource-intensive process that demands significant clinical expertise. Applying cutting-edge natural language processing (NLP) techniques based on deep learning (DL) models to these clinician transcripts can help clinicians make real-time decisions, furnishing personalized, evidence-based recommendations that consider each patient's history and symptoms<sup>12</sup>. Using this technology, the triage process in the ED can be optimized by assessing the severity and immediacy of patient conditions, thereby improving resource allocation. This leads to shorter wait times and enhances satisfaction for patients and healthcare staff, ensuring a more effective and efficient emergency care delivery system<sup>13,14</sup>. Despite their widespread use and clear benefits, the complete utilization of DL-based NLP method in clinical settings is limited by the lack of annotated data and automated tools that are essential for effectively extracting clinical insights<sup>15</sup>.

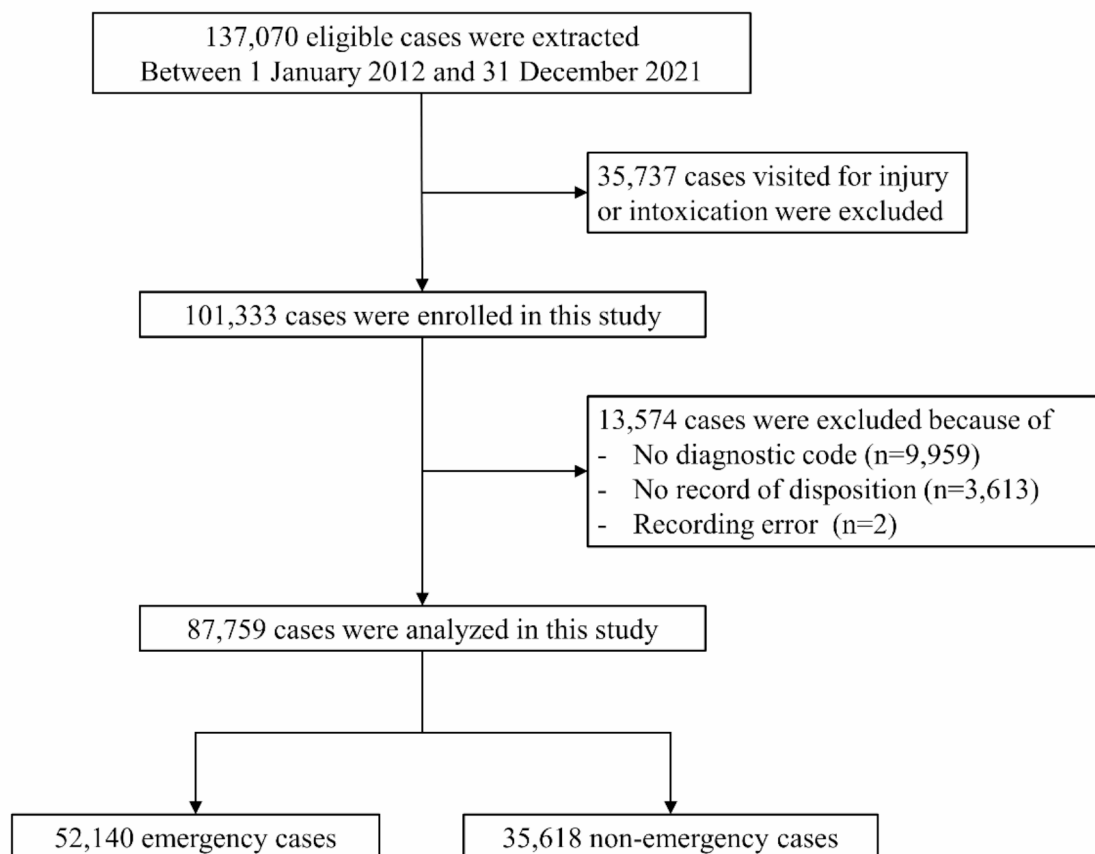
### Purpose

Overcrowding in pediatric emergency departments is often caused by unnecessary visits, leading to the inefficient use of resources and operational inefficiencies. This study aims to develop a DL-based NLP triage model for pediatric emergency patients to address this issue.

## Materials and methods

### Study population

This retrospective cohort study used EMR from the PED in a tertiary hospital in South Korea. We accessed the EMR on April 21, 2023, and information that could identify individual participants has been anonymized. The study participants were patients aged < 18 years who visited the PEDs between January 1, 2012, and December 31, 2021. The cohort was formed by excluding those who visited the PED for the treatment of injury or intoxication, missing a diagnosis code, visited the PED for purposes other than medical treatment, or had no record of disposition. The final cohort of interest included 87,759 patients, including 52,140 emergency patients (Fig. 1.).



**Fig. 1.** Flowchart of the study population.

## Emergency and non-emergency cases

Traditional triage systems such as the Canadian Triage and Acuity Scale (CTAS) or the Korean Triage and Acuity Scale (KTAS) classify the urgency of emergency department patients from level 1 (critical) to level 5 (non-urgent) based on their symptoms and vital signs. This classification determines treatment priority, with levels 1 through 3 typically considered emergency cases requiring immediate attention, while patients classified as levels 4 or 5 are considered non-emergency cases and may be treated with lower priority.

In this study, we developed a novel approach to classify emergency and non-emergency cases in the PED, addressing limitations in traditional triage systems. While conventional methods often rely on initial triage scores, such as the CTAS or KTAS, these can lead to over-triage or under-triage of patients<sup>16,17</sup>. In actual PED, many patients at triage levels 1, 2, and 3 are discharged without receiving emergency treatment, while some patients at lower triage levels still require urgent care. To address the limitations of this triage system, we defined patients who received emergency treatment as the “emergency group” and those discharged without such treatment as the “non-emergency group,” based on whether they received emergency care during their visit.

“Emergency” was defined as “a sudden, usually unexpected event that requires immediate action to minimize adverse consequences”<sup>18</sup>. Emergency cases included those who underwent blood tests, urinalysis, intravenous hydration, nebulization, immediate administration of medication in the PED, and those who were hospitalized. A detailed distribution of emergency cases according to intervention criteria is provided in Supplementary Fig. 1. “Non-emergency” refers to a clinical condition that does not require immediate medical attention, diagnosis, or treatment. Non-emergency cases included those who were discharged from the PED without immediate testing or medication or were discharged with only discharge medication.

## Preprocessing of unstructured clinical transcript

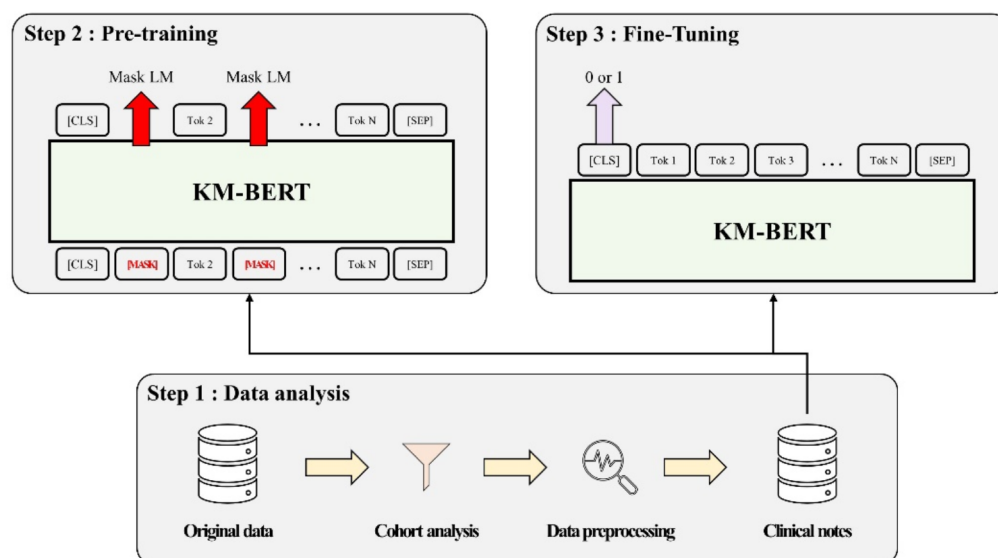
The unstructured data consists of clinicians’ written descriptions, transcribed during a patient’s visit to the ED, detailing their condition, chief complaints, and symptoms. The clinician transcripts used in our study were written without a standardized form, with medical terms presented in Korean or English and with numeric values and special characters. To enhance the effectiveness of handling clinician transcripts, we systematically separated Korean, English, numbers, and special characters before incorporating text information into the prediction models. For instance, if Korean characters are followed by English, numbers, or special characters in a clinician transcript, we insert a blank space to ensure a clear distinction among Korean, English, numbers, and special characters.

## Prediction models

Figure 2 showed a visual overview of the entire process from data extraction to model training and evaluation.

### Machine learning (ML)-based prediction model using TF-IDF

Term frequency-inverse document frequency (TF-IDF) is a traditional NLP method that serves as a statistical measure that assesses the relevance of a word to a document within a collection of documents. This is determined by multiplying two metrics: the TF score, which denotes how frequently a word appears within a document, and the IDF score, which indicates the rarity of a word across an entire set of documents. A word was considered more



**Fig. 2.** Overview of the study process from data extraction to model evaluation.

significant when assigned a higher TF-IDF value. In this study, we applied TF-IDF by treating the triage note for each patient as a document. Four ML models—one statistical method (logistic regression) and three tree-based ensemble methods (random forest, gradient boosting, and XGBoost)—were trained to predict emergency and non-emergency cases by treating the TF-IDF scores applied to the clinician transcripts as input features.

#### *Introduction to deep learning-based NLP methods*

Word2Vec is a traditional DL-based NLP method that was developed to map words into an embedded space and captures their semantic meanings by positioning words with similar meanings close to each other<sup>19</sup>. Methods based on Word2Vec have been empirically validated as useful for analyzing the relationships between medical terms and symptoms<sup>20</sup>. For example, they can help identify connections between various diagnostic terms for a specific symptom or explore associations between different treatments.

Long short-term memory (LSTM) is a type of DNN specifically designed to process sequential data. It has been widely applied to numerous NLP tasks because of its ability to extract semantically useful information from entire sentences<sup>21</sup>.

#### *Deep learning-based prediction model using KM-BERT*

We derived our DL-based prediction model using KM-BERT, a Korean language model pretrained on a collection of three types of Korean medical documents: medical textbooks, health information news, and medical research articles. In particular, we employed KM-BERT with a small vocabulary comprising 16,424 subwords. We further pretrained it on our clinician transcripts using the pretext task of masked language model (MLM). Our pretraining encourages the model to learn the semantic context by randomly masking subwords in a sentence and reconstructing the missing subwords based on the context provided by the remaining subwords. We then fine-tuned the KM-BERT model on our clinician transcripts to predict emergency and non-emergency cases, using the subwords of the clinician transcripts as input sequence. This prediction model is denoted as KM-BERT (MLM).

### **Model training and testing**

The cohort was randomly divided (64:16:20) into training, validation, and testing sets using the Python (version 3.9.13) package *scikit-learn* (version 1.2.1). The same data split was used to train and evaluate the ML-based and DL-based prediction models. For model evaluation, we used bootstrapping on the testing set with a sampling ratio of 0.75, conducting 10 random iterations to produce results with mean and standard deviations. The discriminative power of the prediction models was assessed using the following key metrics: area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and F1-score. The AUROC represents the probability that a randomly selected patient who was emergent was assigned a higher risk than a patient who was not emergent. The AUPRC is a critical metric for problems where properly classifying the positives (in our study, the emergent cases) is important. Meanwhile, the F1-score is a harmonic mean of precision (i.e., the ratio of true positives among samples with predicted positive labels) and recall (i.e., the ratio of true positives among samples with ground-truth positive labels).

To train the ML-based prediction models, we first constructed a medical dictionary to extract medical terms from the corpus of our clinician transcripts using medical dictionaries and aggregated Korean and English words with the same clinical meaning. We identified the most frequent words in the entire dataset by calculating the cumulative sum, which accounted for 85% of the overall word distribution across the clinician transcripts. This yielded a final count of 103. We then applied TF-IDF using the Python package *scikit-learn* and trained the following ML prediction models using the TF-IDF scores on the most frequent words as input features: 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 models were chosen via a grid search, that is, max depth from the candidate values {1,2,3,4,5}, the number of estimators from {100, 200, 300, 400, 500} for the tree-based ensemble models, and the regularization coefficient from {0.001, 0.01, 0.1, 1, 10, 100, 1000} for logistic regression, using the AUROC performance on the validation set.

To train the DL-based prediction model using KM-BERT, we initially pretrained KM-BERT with a small vocabulary by employing an MLM pretext task on the training set. Specifically, we processed each clinician's transcript using the subword tokenizer available in KM-BERT, utilizing the resulting subword tokens as input features. Subsequently, we added a binary classification layer consisting of one fully connected layer on top of the [CLS] token from the pretrained KM-BERT. We fine-tuned the prediction model to classify emergency and non-emergency cases. Pretraining and fine-tuning incorporated early stopping to select the model with the lowest validation loss using a batch size of 64. The learning rate was set to 5e-4 during pretraining, and for fine-tuning, it was set to 1e-5 and 1e-2 for KM-BERT and the classifier, respectively.

### **SHAP-based rationalization**

DL models have improved the performance of NLP tasks; however, their complexity makes it difficult to understand the model output. To bridge this gap, various explainability techniques have been suggested to elucidate model predictions; rationalization is a type of explainability method that provides natural language explanations<sup>22</sup>. In this work, we applied Shapley-based rationalization to provide extractive rationales, which provide words that explain why the model makes a specific prediction<sup>23</sup>. We computed the attribution of the individual words by assessing the variance in the model output when a word was included versus when it was omitted. To increase interpretability, we intentionally computed the attribution of words instead of tokens. For instance, if the model prediction is an emergent case and the word “fever” has positive attribution, this indicates that “fever” contributed to the model in making the prediction an emergent case.

Subgroup analysis

The Korean Triage and Acuity Scale (KTAS) is a Korean emergency patient classification tool that evaluates the severity and urgency of a patient's condition based on their symptoms, categorizing them into five levels to determine the priority of treatment<sup>11</sup>. The analysis targeted pediatric patients with KTAS levels from 2016 to 2021. KTAS levels 1–3 was classified as emergency, levels 4–5 were classified as non-emergency, and their performances were compared with other models.

Ethics statement

This study was approved by the Institutional Review Board (IRB) at The Catholic University of Korea (IRB approval number KC23RISI0073). The Institutional Review Board (IRB) at The Catholic University of Korea waived the requirement for informed consent. All procedures were conducted in compliance with applicable guidelines and regulations.

Results

Performance of pediatric emergency prediction models

The average length of the clinicians' transcripts was 62.5 words (SD = 36.7). We trained four topic models with TF-IDF utilizing statistical and ensemble ML models, including logistic regression, XGBoost, gradient boosting, and random forest, and two DL-based language models based on KM-BERT to predict emergent cases using clinician transcripts as input. The top-ranked words extracted by TF-IDF and their corresponding scores are presented in Appendix Table A1 and visually represented in Appendix Figure A1, providing insight into the key terms identified by our ML model. The performance results for each model are presented in Table 1. and Fig. 3. Gradient boosting shows the highest AUROC of  $0.715 \pm 0.002$ , AUPRC of  $0.778 \pm 0.001$ , and recall of  $0.626 \pm 0.003$ , while XGBoost leads to precision of  $0.741 \pm 0.001$ . Gradient boosting also achieved the highest F1-score of  $0.677 \pm 0.001$  and accuracy of  $0.649 \pm 0.001$ . Gradient boosting and XGBoost had the lowest Brier scores ( $0.209 \pm 0.000$ ), indicating the most accurate probabilistic predictions. However, the two DL-based prediction models exhibited better performance than the ML-based models with TF-IDF. Comparing the two DL-based NLP methods, the fine-tuned KM-BERT with MLM outperformed the KM-BERT for all indicators. In AUROC, it is  $0.839 \pm 0.001$ , which is higher than that of KM-BERT ( $0.788 \pm 0.002$ ), and in AUPRC, it is  $0.879 \pm 0.001$ , which is higher than that of KM-BERT ( $0.837 \pm 0.002$ ). In recall, it is  $0.724 \pm 0.002$ , slightly higher than that of KM-BERT ( $0.719 \pm 0.002$ ), while in precision, it is  $0.829 \pm 0.001$ , which is a clear difference from that of KM-BERT ( $0.775 \pm 0.002$ ). The F1-score is  $0.773 \pm 0.001$ , which is higher than that of KM-BERT ( $0.746 \pm 0.002$ ), and the accuracy is  $0.749 \pm 0.001$ , which is higher than that of KM-BERT ( $0.712 \pm 0.002$ ). KM-BERT with MLM has a Brier score of  $0.164 \pm 0.001$ , which is lower than that of KM-BERT ( $0.188 \pm 0.001$ ).

Calibration

Calibration ensures that the predicted probabilities from a prediction model align with the actual observed frequencies. In addition to comparing Brier scores, we have included calibration curves (i.e., Q-Q plots) to check the quantile relationship between the predicted probability of each prediction model and the observed event rates. If the model is perfectly calibrated, the points on the Q-Q plot will fall along a 45-degree line, indicating that the predicted probabilities accurately reflect the true event frequencies. Figure 4 confirms that each model exhibited good prediction performance. However, we observed that the topic models with TF-IDF do not fully span the range from 0 to 1, indicating that these models' calibration may be less reliable (might have a bias) when predicting probabilities close to 0 or 1.

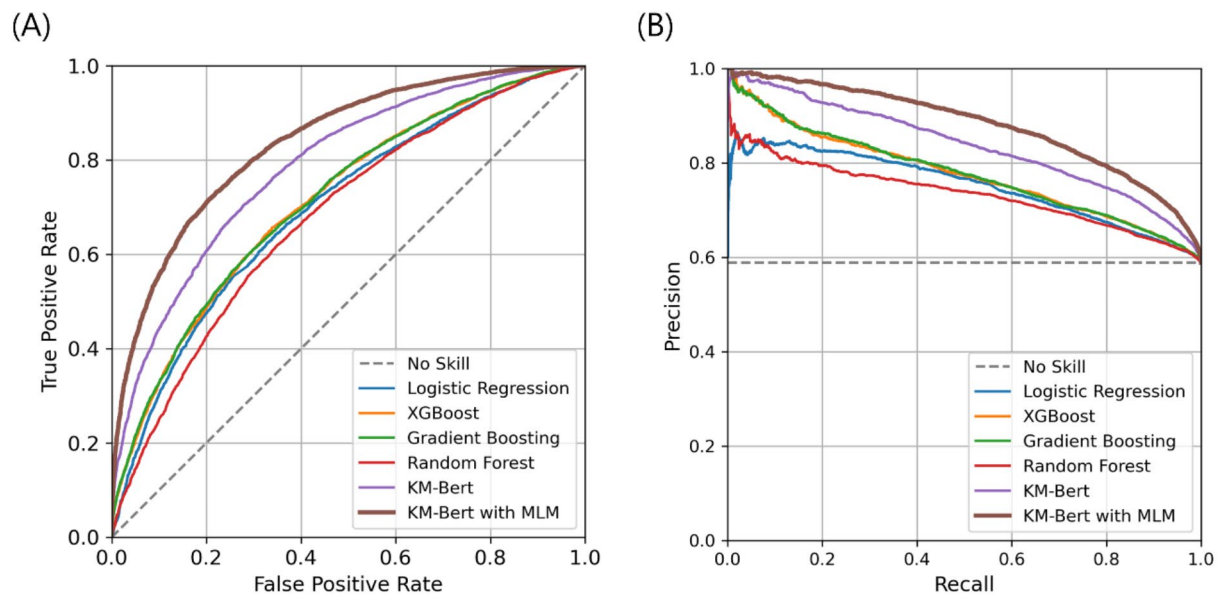
Rationalization

We analyzed the results from an ML-based prediction model using TF-IDF. The TF-IDF scores of words extracted by this model are presented in Table A1 in the appendix, while the frequency of these words is visualized in Figure A1. For rationalization, we compared passages judged by an emergency specialist and those judged by our DL-based prediction model using the same clinician transcripts. The passages in which the specialist judged

Variable	Logistic regression	XGBoost	Gradient boosting	Random forest	KM-BERT	KM-BERT with MLM*
AUROC	0.698 ± 0.002	0.714 ± 0.002	0.715 ± 0.002	0.680 ± 0.002	0.788 ± 0.002	<b>0.839 ± 0.001</b>
AUPRC	0.752 ± 0.002	0.776 ± 0.001	0.778 ± 0.001	0.735 ± 0.002	0.837 ± 0.002	<b>0.879 ± 0.001</b>
Recall	0.629 ± 0.003	0.618 ± 0.002	0.626 ± 0.003	0.625 ± 0.002	0.719 ± 0.002	<b>0.724 ± 0.002</b>
Precision	0.728 ± 0.002	0.741 ± 0.001	0.737 ± 0.001	0.716 ± 0.002	0.775 ± 0.002	<b>0.829 ± 0.001</b>
F1-score	0.675 ± 0.002	0.674 ± 0.002	0.677 ± 0.001	0.667 ± 0.001	0.746 ± 0.002	<b>0.773 ± 0.001</b>
Accuracy	0.643 ± 0.002	0.648 ± 0.002	0.649 ± 0.001	0.633 ± 0.002	0.712 ± 0.002	<b>0.749 ± 0.001</b>
Brier	0.215 ± 0.000	0.209 ± 0.000	0.209 ± 0.000	0.225 ± 0.001	0.188 ± 0.001	<b>0.164 ± 0.001</b>

**Table 1.** Performance of pediatric emergency prediction using natural language processing techniques and topic models. KM-BERT, Korean medical bidirectional encoder representations from transformers; MLM, masked language modeling; AUROC, area under the receiver operating characteristics; AUPRC, area under the precision-recall curve. Bold values indicate the best-performing model across all metrics.; \*Indicates overall best-performing model.





**Fig. 3.** Prediction Ability of ML-based and DL-based NLP Techniques for Pediatric Emergency Prediction. (A) Receiver operating characteristic curves. (B) Precision-recall curve. The corresponding values of the area under the curve for each model are shown in Table 1. KM-BERT, Korean medical bidirectional encoder representations from transformers; MLM, masked language modeling; AUROC, area under the receiver operating characteristics; AUPRC, area under the precision-recall curve.

an emergency (red in Fig. 5A) and the passages in which our prediction model judged an emergency were measured similarly (red in Fig. 5B).

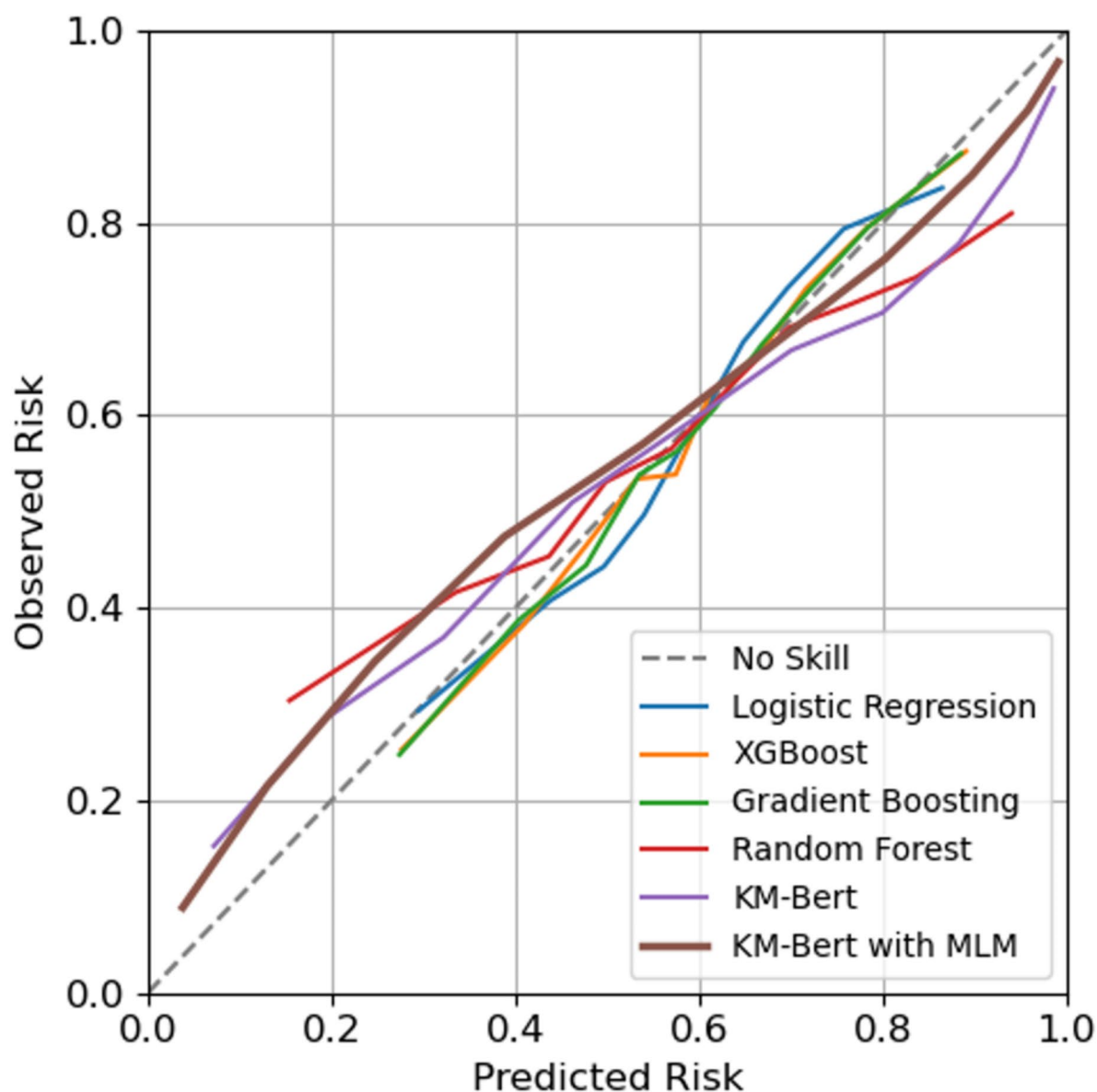
### Subgroup analysis with KTAS

Because we measured KTAS scores in patients who visited the PED in 2016, we performed a subgroup analysis using the 47,968 cases in our dataset for which KTAS scores were recorded. In this analysis, KM-BERT with MLM achieved the highest AUROC of  $0.849 \pm 0.003$ , AUPRC of  $0.896 \pm 0.003$ , recall of  $0.748 \pm 0.004$ , precision of  $0.842 \pm 0.002$ , F1-score of  $0.792 \pm 0.003$ , and accuracy of  $0.760 \pm 0.003$  and the lowest Brier score of  $0.156 \pm 0.002$ . KM-BERT also excelled, surpassing gradient boosting and KTAS with AUROC of  $0.800 \pm 0.003$ , AUPRC of  $0.861 \pm 0.003$ , recall of  $0.748 \pm 0.004$ , precision of  $0.792 \pm 0.003$ , F1-score of  $0.769 \pm 0.003$ , accuracy of  $0.724 \pm 0.003$ , and Brier score of  $0.179 \pm 0.002$ . Gradient boosting followed with AUROC of  $0.726 \pm 0.003$ , AUPRC of  $0.802 \pm 0.004$ , recall of  $0.662 \pm 0.004$ , precision of  $0.757 \pm 0.005$ , F1-score of  $0.706 \pm 0.004$ , accuracy of  $0.662 \pm 0.004$ , and Brier score of  $0.202 \pm 0.001$ , while KTAS trailed behind all models in each metric with AUROC of 0.6082, AUPRC of 0.6777, recall of 0.7324, precision of 0.6834, F1-score of 0.707, accuracy of 0.6281, and Brier of 0.2707 (Table 2; Fig. 6).

### Discussion

We assessed various models to predict emergencies in PEDs. The DL-based NLP techniques demonstrated superior reliability and prediction accuracy compared with the ML-based models using TF-IDF. In particular, KM-BERT with the MLM model exhibited exceptional performance across all metrics. Furthermore, rationalization confirmed that the emergency assessments by the experts were in close agreement with the results of the DL-based prediction model. Compared to those of the KTAS judgment scale, the DL-based prediction model's predictions were the most accurate, underscoring the high reliability and precision of the fine-tuned model in analyzing pediatric emergencies.

In the medical field, TF-IDF has been widely utilized for various purposes, such as document classification, information retrieval, and patient data analysis. The strength of this model is its ability to easily transform unstructured text data into a structured format to extract meaningful patterns and insights<sup>24,25</sup>. Additionally, ML techniques have been used to predict hospitalization during ED triage. Several studies have shown that ML models can effectively predict hospitalization by combining patient history and triage information. These models were developed using various algorithms such as logistic regression, XGBoost, and deep neural networks (DNN)<sup>26</sup>. One of studies presented a DNN-based model that predicted PED admissions and outperformed existing methods by achieving an AUC of 0.892, demonstrating the importance of text data in improving



**Fig. 4.** Calibration plots for natural language processing techniques and topic models. The observed risk compared to the predicted risk. A reference line indicates that the predicted risk and the observed emergency patient rate are exactly the same.

prediction accuracy<sup>27</sup>. However, because TF-IDF simply analyzes text based on the frequency and importance of words, it has limitations in capturing the complex context or semantic nuances of the text.

Our findings are consistent with those of other studies in the field, highlighting the critical significance of DL-based NLP methods in current medical data analysis. The performance of the fine-tuned DL models was superior to that of the ML models using TF-IDF. DL-based NLP methods have played a significant role in achieving high accuracy in predicting hospital admissions and severe diseases. These models enhance predictive performance by integrating structured and unstructured text data and effectively identifying meaningful patterns and relationships across various data types<sup>28</sup>. According to this study, cutting-edge NLP techniques with DL have been successfully used to predict ED patient dispositions based on nursing triage records. These results indicate that DL-based NLP can effectively evaluate unstructured clinical text to predict patient outcomes in the ED and that paragraph vectors can provide the most accurate predictions<sup>14</sup>.


As shown in Fig. 5, we found that what ED clinicians recorded as important in assessing the patient's emergency status was consistent with what the DL-based prediction model deemed important. The explanatory power of the DL-based prediction model can be used to identify the important factors when interviewing or examining emergency patients. In practice, inexperienced clinicians or pediatric emergency physicians may miss patients who are urgently ill despite having been interviewed in the ED, which can lead to a return visit to

(A)

[Clinician transcript]  
A **86-day-old** male with no underlying medical conditions presents to our emergency department with a nasal obstruction that has been present since the early hours of today, making it difficult to breathe. His mother said that he sneezes occasionally while his nose is stuffy, but there is no runny nose or other symptoms. He has no other symptoms of decreased urine output or digestive system symptoms. Baby has never taken antipyretic drug, visited local clinic but was referred to general hospital as he is **less than 100 days old**. **BT : 38 . 2 °C**, BW : 6 . 5 kg, 1 day fever maximum temperature : 38 . 2 °C, Last antipyretic (-), URI Sx in the family (-)

(B)

A **86-day-old** male with no underlying medical conditions presents to emergency department with **a nasal** obstruction that has been present since the early hours of today, making it difficult to breathe. His mother said that he sneezes occasionally while his nose is stuffy, but there is no runny nose or other symptoms. He has no other symptoms of decreased urine output or digestive system symptoms. Baby has never taken antipyretic drug, visited local clinic but was referred to general hospital as he is less than 100 days old. **BT : 38 . 2 °C**, BW : 6 . 5 kg, 1 day fever maximum temperature : **38 . 2 °C**, Last antipyretic (-), URI Sx in the family (-)



**Fig. 5.** Comparison of passages marked as important by clinicians and those judged important by our DL-based prediction model to assess the emergency of pediatric patients. (A) Clinician transcript with important words marked in red by the clinician; the clinician marked the important words “86 days old,” “less than 100 days old,” and “temperature of 38.2 degrees” in red letters. (B) Clinician transcript with important words colored in red and unimportant words colored in blue by the DL-based prediction model.

Variable	KTAS	Logistic regression	XGBoost	Gradient boosting	Random forest	KM-BERT	KM-BERT with MLM*
AUROC	0.610 ± 0.002	0.709 ± 0.003	0.723 ± 0.003	0.726 ± 0.003	0.691 ± 0.003	0.800 ± 0.003	<b>0.849 ± 0.003</b>
AUPRC	0.679 ± 0.003	0.780 ± 0.003	0.801 ± 0.003	0.802 ± 0.004	0.761 ± 0.004	0.861 ± 0.003	<b>0.896 ± 0.003</b>
Recall	0.733 ± 0.003	0.655 ± 0.004	0.654 ± 0.003	0.662 ± 0.004	0.650 ± 0.003	0.748 ± 0.004†	<b>0.748 ± 0.004†</b>
Precision	0.685 ± 0.003	0.754 ± 0.004	0.758 ± 0.004	0.757 ± 0.005	0.738 ± 0.004	0.792 ± 0.003	<b>0.842 ± 0.002</b>
F1-score	0.708 ± 0.002	0.701 ± 0.004	0.702 ± 0.003	0.706 ± 0.004	0.691 ± 0.003	0.769 ± 0.003	<b>0.792 ± 0.003</b>
Accuracy	0.628 ± 0.002	0.657 ± 0.003	0.660 ± 0.003	0.662 ± 0.004	0.644 ± 0.003	0.724 ± 0.003	<b>0.760 ± 0.003</b>
Brier	0.271 ± 0.001	0.209 ± 0.001	0.203 ± 0.001	0.202 ± 0.001	0.216 ± 0.001	0.179 ± 0.002	<b>0.156 ± 0.002</b>

**Table 2.** Performance of pediatric emergency prediction using natural language processing techniques and topic models compared to KTAS. KTAS, Korean Triage and Acuity Scale; KM-BERT, Korean medical bidirectional encoder representations from transformers; MLM, masked language modeling; AUROC, area under the receiver operating characteristics; AUPRC, area under the precision-recall curve. Bold values indicate the best-performing model across metrics. \*Indicates overall best-performing model; †Indicates tied best performance for Recall.

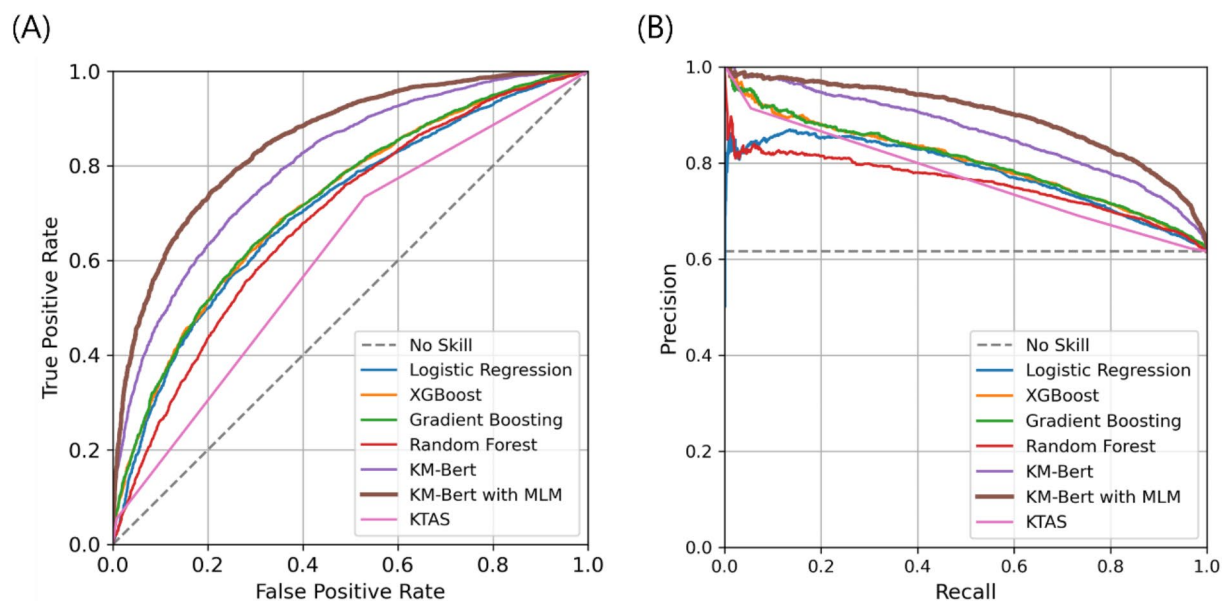
the ED. This can result in significant patient harm and inefficient use of emergency medical resources. DL-based prediction models can be used to train inexperienced clinicians and help them recognize patients in distress during patient encounters in a real-world ED setting.

Additionally, the urgency of pediatric patients presenting with PED is currently classified using various triage systems. However, as we found in a subgroup analysis in our study, traditional triage systems performed significantly worse than the DL-based prediction models at predicting emergency patients. DL-based prediction models can be used to accurately identify emergency patients so that aggressive treatment can be provided quickly. This is a desirable direction for improving clinical outcomes and ensuring patient safety.

Despite this, Word2Vec only considers neighboring words within a relatively small window, which limits its ability to capture the semantics of long sentences and the overall statistical information of the entire corpus. Although LSTMs are designed to handle long sequences, they still face challenges in learning relationships between distant words within a sentence due to the vanishing gradient problem. Furthermore, training large LSTM models is computationally infeasible because of their inherent sequential nature. These limitations make LSTMs less suitable for constructing large language models that can effectively learn meaningful semantics from extensive datasets, especially in the medical domain.

The prediction model in our study was built upon a more recent development in NLP called BERT<sup>29</sup>. BERT has gained considerable attention for overcoming the limitations of traditional DL approaches by adopting a transformer architecture as the basic building block<sup>30</sup>. In particular, BERT can provide significant performance improvements in comprehending natural languages by extracting semantically meaningful interactions in long sequences and constructing large language models with large datasets. Recently, BERT-based models have been widely utilized in the medical field<sup>31–33</sup> after pretraining on medical corpora to bridge the gap between text used in the general domain and that in the medical domain. In this study, we adopted the pretrained KM-BERT<sup>34</sup>, a BERT-based model pretrained based on the Korean medical corpus (including medical textbooks, health





**Fig. 6.** Prediction Ability of the conventional triage system (KTAS), and ML-based and DL-based NLP Techniques for Pediatric Emergency Prediction. (A) Receiver operating characteristic curves. (B) Precision-recall curve. The corresponding values of the area under the curve for each model are shown in Table 2. KTAS, Korean Triage and Acuity Scale; KM-BERT, Korean medical bidirectional encoder representations from transformers; MLM, masked language modeling; AUROC, area under the receiver operating characteristics; AUPRC, area under the precision-recall curve.

information news, and medical research articles), as the backbone of our model to handle clinical transcripts primarily written in Korean.

This study has a few limitations. First, it was conducted using a retrospective cohort of clinician transcripts from a single tertiary hospital's PED. Given a hospital's tertiary status, it is believed that there is sufficient emergency care for severely ill patients. Second, transcripts were recorded by clinicians with varying degrees of skill and experience. However, DL-based prediction models are intended to comprehend and interpret the nuances of natural language, making them capable of processing and evaluating texts with diverse styles and levels of complexity. Thirdly, the pretrained KM-BERT, which was adopted as the backbone of our model, was trained based on the Korean medical corpus. While clinical records may contain some medical terms in English, the majority of crucial information, including chief complaints and symptoms, is documented in Korean. Therefore, a pre-trained model like KM-BERT, which is capable of effectively understanding both the Korean language and Korean-based medical terminology, was most appropriate for our research objectives. Nevertheless, we also recognize the importance of medical literature and terminology written in English. Recent advancements in models trained on English corpora, such as BioMistral<sup>35</sup>, further underscore the need to incorporate multilingual approaches in future research. Additionally, a significant limitation of study is the lack of external validation data, which may result in incorporation bias in our AUROC and AUPRC metrics and lead to potential overestimation of the model's performance due to information leakage between the training and testing phases. Future studies should aim to validate these findings using external, independent datasets to confirm the generalizability and robustness of the model.

## Conclusion

Our findings highlight the increasing importance of advanced NLP techniques in the medical field, particularly during emergencies. The ability of these models to make precise and accurate forecasts can substantially benefit the management of healthcare resources, the improvement of patient care, and, ultimately, the overall efficiency of emergency medical services.

## Data availability

Deidentified data supporting the findings of this study are available from the corresponding authors upon request.

Received: 11 April 2024; Accepted: 16 January 2025

## References

- Bucci, S. et al. Emergency Department crowding and hospital bed shortage: is lean a smart answer? A systematic review. *Eur. Rev. Med. Pharmacol. Sci.* **20**, 4209–4219 (2016).
- Cha, W. C., Ahn, K. O., Shin, S. D., Park, J. H. & Cho, J. S. Emergency Department Crowding Disparity: a nationwide cross-sectional study. *J. Korean Med. Sci.* **31**, 1331–1336. <https://doi.org/10.3346/jkms.2016.31.8.1331> (2016).
- Yarmohammadian, M. H., Rezaei, F., Haghsheenas, A. & Tavakoli, N. Overcrowding in emergency departments: a review of strategies to decrease future challenges. *J. Res. Med. Sci.* **22**, 23. <https://doi.org/10.4103/1735-1995.200277> (2017).
- Improta, G. et al. A case study to investigate the impact of overcrowding indices in emergency departments. *BMC Emerg. Med.* **22**, 143. <https://doi.org/10.1186/s12873-022-00703-8> (2022).
- Timm, N. L., Ho, M. L. & Luria, J. W. Pediatric emergency department overcrowding and impact on patient flow outcomes. *Acad. Emerg. Med.* **15**, 832–837. <https://doi.org/10.1111/j.1553-2712.2008.00224.x> (2008).
- American Academy of Pediatrics Committee on Pediatric & Emergency, M. Overcrowding crisis in our nation's emergency departments: is our safety net unraveling? *Pediatrics* **114**, 878–888, (2004). <https://doi.org/10.1542/peds.2004-1287>
- Kim, B. S., Kim, J. Y., Choi, S. H. & Yoon, Y. H. Understanding the characteristics of recurrent visits to the emergency department by paediatric patients: a retrospective observational study conducted at three tertiary hospitals in Korea. *BMJ Open* **8**, e018208. <https://doi.org/10.1136/bmjopen-2017-018208> (2018).
- O'Loughlin, K. et al. Paediatric unplanned reattendance rate: A&E clinical quality indicators. *Arch. Dis. Child.* **98**, 211–213. <https://doi.org/10.1136/archdischild-2012-302836> (2013).
- Seiler, M., Furrer, P. R., Staubli, G. & Albisetti, M. Unplanned return visits to a Pediatric Emergency Department. *Pediatr. Emerg. Care* **37**, e746–e749. <https://doi.org/10.1097/PEC.0000000000001764> (2021).
- Karlafti, E. et al. Support systems of clinical decisions in the triage of the Emergency Department using Artificial Intelligence: the efficiency to support triage. *Acta Med. Lit.* **30**, 19–25. <https://doi.org/10.15388/Amed.2023.30.1.2> (2023).
- Lee, J. T., Hsieh, C. C., Lin, C. H., Lin, Y. J. & Kao, C. Y. Prediction of hospitalization using artificial intelligence for urgent patients in the emergency department. *Sci. Rep.* **11**, 19472. <https://doi.org/10.1038/s41598-021-98961-2> (2021).
- Yang, X. et al. A large language model for electronic health records. *NPJ Digit. Med.* **5**, 194. <https://doi.org/10.1038/s41746-022-00742-2> (2022).
- Mermin-Bunnell, K. et al. Use of Natural Language Processing of Patient-Initiated Electronic Health Record Messages to identify patients with COVID-19 infection. *JAMA Netw. Open* **6**, e2322299. <https://doi.org/10.1001/jamanetworkopen.2023.22299> (2023).
- Sterling, N. W., Patzer, R. E., Di, M. & Schrager, J. D. Prediction of emergency department patient disposition based on natural language processing of triage notes. *Int. J. Med. Inf.* **129**, 184–188. <https://doi.org/10.1016/j.ijmedinf.2019.06.008> (2019).
- Hossain, E. et al. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: a systematic review. *Comput. Biol. Med.* **155**, 106649. <https://doi.org/10.1016/j.compbiomed.2023.106649> (2023).
- Ukiyama, E. et al. Pediatric surgery triage: problems and improvements. *Pediatr. Int.* **54**, 501–503. <https://doi.org/10.1111/j.1442-200X.2012.03669.x> (2012).
- Lee, J. H., Jung, J. H., Noh, H. & Kim, M. J. Predictive validity of resource-adjusted Korean Triage and acuity scale in pediatric gastrointestinal tract foreign body patients. *Sci. Rep.* **14**, 19686. <https://doi.org/10.1038/s41598-024-70685-z> (2024).
- Fuchs, S. et al. Definitions and Assessment Approaches for Emergency Medical Services for children. *Pediatrics* **138** <https://doi.org/10.1542/peds.2016-1073> (2016).
- Kumari, A. & Lobiyal, D. K. Efficient estimation of Hindi WSD with distributed word representation in vector space. *J. King Saud Univ-Com.* **34**, 6092–6103. <https://doi.org/10.1016/j.jksuci.2021.03.008> (2022).
- Choi, E. et al. Multi-layer representation learning for medical concepts. *Kdd'16: Proc. 22nd Acm Sigkdd Int. Conf. Knowl. Discovery Data Min.* **1495–1504** <https://doi.org/10.1145/2939672.2939823> (2016).
- Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735> (1997).
- Gurrapu, S., Kulkarni, A., Huang, L., Lourentzou, I. & Batarseh, F. A. Rationalization for explainable NLP: a survey. *Front. Artif. Intell.* **6**, 1225093. <https://doi.org/10.3389/frai.2023.1225093> (2023).
- Lundberg, S. M. & Lee, S. I. A Unified Approach to interpreting model predictions. *Adv. Neur in* **30** (2017).
- Park, H., Lee, M. & Hwang, S. C. & Oh, S.
- Xiong, Y., Qiao, Y., Dong, S., Zhang, X. & Tan, H. 3–10 (Springer Nature Singapore).
- Hong, W. S., Haimovich, A. D. & Taylor, R. A. Predicting hospital admission at emergency department triage using machine learning. *PLoS One* **13**, e0201016. <https://doi.org/10.1371/journal.pone.0201016> (2018).
- Roquette, B. P., Nagano, H., Marujo, E. C. & Maiorano, A. C. Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. *Neural Netw.* **126**, 170–177. <https://doi.org/10.1016/j.neunet.2020.03.012> (2020).
- Stewart, J. et al. Applications of natural language processing at emergency department triage: a narrative review. *PLoS One* **18**, e0279953. <https://doi.org/10.1371/journal.pone.0279953> (2023).
- Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: pre-training of Deep Bidirectional transformers for Language understanding. *2019 Conf. North. Am. Chapter Association Comput. Linguistics: Hum. Lang. Technol. (NaacL Hlt 2019)*. **1**, 4171–4186 (2019).
- Vaswani, A. et al. Attention is all you need. *Adv. Neur in* **30** (2017).
- Huang, K., Altosaar, J., Ranganath, R. & ClinicalBERT Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342* (2019). <https://ui.adsabs.harvard.edu/abs/2019arXiv190405342H>
- Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682> (2020).
- Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **4**, 86. <https://doi.org/10.1038/s41746-021-00455-y> (2021).
- Kim, Y. et al. A pre-trained BERT for Korean medical natural language processing. *Sci. Rep.* **12**, 13847. <https://doi.org/10.1038/s41598-022-17806-8> (2022).
- Labrak, Y. et al. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373* (2024).

## Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00358602) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, RS-2021-0-01341) Artificial Intelligence Graduate School Program (Korea University, Chung-Ang University).

### Author contributions

J.K., C.L., and W.B. contributed to the study design. J.R. and W.B. contributed to acquire data. C.K., J.J., S.C., D.L., J.K., and C.L. contributed to data labeling, data curation, model development and training. A.C. and C.K. contributed to writing of this manuscript. All authors reviewed the manuscript.

### Declarations

### 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/s41598-025-87161-x>.

**Correspondence** and requests for materials should be addressed to C.L. or W.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](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 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

© The Author(s) 2025