scientific reports

Check for updates

OPEN

Artificial intelligence for severity triage based on conversations in an emergency department in Korea

Jae Won Seo[1,5], Sung-Joon Park[2,5], Young Jae Kim[3], Jung-Youn Kim[2], Kwang Gi Kim[1,4]✉ & Young-Hoon Yoon[2]✉

In the fast-paced emergency departments, where crises unfold unpredictably, the systematic prioritization of critical patients based on a severity classification is vital for swift and effective treatment. This study aimed to enhance the quality of emergency services by automatically categorizing the severity levels of incoming patients using AI-powered natural language processing (NLP) algorithms to analyze conversations between medical staff and patients. The dataset comprised 1,028 transcripts of bedside conversations within emergency rooms. To verify the robustness of the models, we performed tenfold cross-validation. Based on the area under the receiver operating characteristic curve (AUROC) values, the support vector machine achieved the best performance among the term frequency-inverse document frequency-based conventional machine learning models with an AUROC of 0.764 (95% CI 0.019). Among the neural network models, multilayer perceptron performed with an AUROC of 0.759 (± 0.024). This research explored methods for automatically classifying patient severity using real-world conversations, including those with nonsensical and confused content. To achieve this, artificial intelligence algorithms that consider the frequency and order of words used in the conversation were employed alongside neural network models. Our findings have the potential to significantly contribute to alleviating overcrowding in emergency departments of hospitals, with future extensions involving highly efficient large language models. The results suggest that a fluid and immediate response to urgent situations, a reduction in patient waiting time, and effectively addressing the special circumstances of the emergency room environment can be achieved using this approach.

Keywords Emergency room, Triage, Classification, Natural language processing, Artificial intelligence

Emergency departments face a diverse influx of patients, ranging from minor cases to life threatening emergencies, which require prompt and comprehensive assessments by medical professionals. Despite the escalating demand for emergency medical services in Korea, the supply of emergency medical professionals has not kept up with the increasing demand[1]. This trend has led to overcrowding within emergency departments, disrupting the healthcare system, prolonging patient waiting times, and compromising the quality of emergency care. Overcrowding poses a grave concern in emergency medicine, resulting in delays in the treatment for severely ill patients, with potentially fatal consequences. Beyond overcrowding, traditional triage systems frequently encounter challenges related to triage errors, including over-triage which is assign higher-than-necessary urgency, and under-triage failing to recognize truly urgent cases. These errors can lead to resource misallocation, compromised patient safety, and further strain on already limited emergency medical resources[2–7]. Thus, ensuring the efficient allocation of limited medical resources to address the needs of a large patient volume requires the swift and accurate identification of patient severity levels. To overcome these obstacles, there are systematic triage systems that reflect the characteristics of each country and region[8–11]. In Korea, the Korean Triage and Acuity Scale (KTAS) was developed by the Ministry of Health and Welfare in 2012 based on the Canadian Triage Acuity Scale, and has been implemented since 2016[12].

[1]Department of Health Sciences and Technology, GAIHST, Gachon University, Incheon 21999, Korea. [2]Department of Emergency Medicine, Korea University College of Medicine, Seoul 02841, Korea. [3]Department of Gachon Biomedical & Convergence Institute, Gachon University Gil Medical Center, Incheon 21565, Korea. [4]Department of Biomedical Engineering, College of IT Convergence, Gachon University, Seongnam-Si 13120, Korea. [5]Jae Won Seo and Sung-Joon Park have contributed equally to this work and share the first authorship. ✉email: kimkg@gachon.ac.kr; yyh71346@naver.com

The rapid advancement of artificial intelligence (AI) has consistently demonstrated impressive performance, particularly in natural language processing (NLP) research involving textual and time series data[13–15]. AI models have demonstrated remarkable efficacy, particularly in the medical field, where numerous studies have utilized emergency department data to predict patient prognosis and classify severity based on patient information, including vital signs and self-reported pain levels[16–18]. However, previous studies utilizing the KTAS classification have primarily relied on simulated data rather than real-time conversation data. The pioneering work by Choi et al. utilized NLP to predict KTAS levels based on triage notes recorded by nursing professionals, demonstrating the potential of machine learning approaches for severity classification in Korean emergency departments[19]. Chang et al. further advanced this field by developing a clinical support system for KTAS based on federated learning[20]. Additionally, a recent systematic review by Porto highlighted significant opportunities for further research in applying machine learning and NLP to emergency department triage, underscoring the relevance of our approach[21]. Moreover, one study achieved a notable AUROC of 0.90 by classifying severity based on voice data from medical staff-patient conversations. However, their approach relied on simulated, rather than actual, interactions[22,23]. To the best our knowledge, no study has utilized real bedside conversations collected in the emergency department of hospitals in Korea for patient severity triage.

In this study, we automatically classified the severity of patients using only the content of multilateral conversations conducted at the bedside. To this end, we used AI-based NLP algorithms, both traditional machine learning algorithms and neural network-based algorithms, to analyze the effect of the nature of the conversations. Our objective was to investigate the effectiveness of NLP AI algorithms on anomalous real clinical data, rather than on simulated data that NLP AI algorithms are typically trained on. While comparing the predictive performance of models using structured data such as vital signs versus unstructured conversation data would provide additional insights into the relative value of different data types for triage prediction, this initial study focuses specifically on establishing the feasibility of using actual bedside conversations for severity classification. Such comparative analysis represents a valuable direction for future research that could further optimize triage decision support systems.

Materials and methods

Materials

This prospective observational study was conducted at three regional EDs of Korea University Hospital from June 2022 to December 2022. Korea University Anam Hospital and Guro Hospital are regional EDs in Seoul and Korea University Ansan Hospital is a regional ED in Ansan, Gyeonggi-do, a metropolitan area. The annual number of patients visiting the emergency department in all three hospitals was approximately 150,000. In this study, voice recordings were acquired from the initial stage of the study patients visiting the emergency department until the patients were discharged from the EDs. These data were then re-transcribed by a trained recorder, and based on these transcripts, the medical staff participating in the study checked the transcripts for abnormalities. These transcripts were also re-labeled as pre-interview stage (so-called "triage" in the medical field), initial consultation, medication and examination, explanation, and discharge to generate data. The analyzed data comprised 1,048 clinician-patient and companion conversations.

The severity classification, performed in the triage during the first visit to the emergency department, is crucial for determining the need for treatment and the formulation of a treatment plan. We specifically focused on conversations that clinicians identified as those that occurred during the triage. In Korea, it is legally mandated to establish and operate triage stations to ensure that patients undergo triage before entering the ED. In most hospitals, triage is performed by nurses, and in the three hospitals included in this study, nurses also carried out the triage process. The KTAS is classified from 1 to 5 depending on the severity of the patient with KTAS 1 indicating urgent life-threatening situations and KTAS 5 indicating minimal severity. During the triage process, informed consent was obtained from patients, and voice recordings of conversations between patients and medical staff were collected using a recording device. Since patients classified as KTAS 1–2 often required immediate medical intervention, obtaining consent was challenging, making voice data collection difficult. Consequently, KTAS 1–2 patients were excluded from the analysis. The severity of the data used was based on the KTAS, which utilizes data corresponding to stages 3, 4, and 5. The KTAS scores were continuously reevaluated by medical staff and updated according to changes in patient conditions during their stay in the emergency department. In this study, only KTAS scores evaluated in the triage were considered. We performed a binary classification considering KTAS stage 3 as severe and KTAS stages 4 and 5 as mild, leading to significant findings. The characteristics of our datasets, a result of our thorough analysis, are presented in Table 1.

Study design

In this study, AI algorithms were categorized into two broad categories. The first category, "conventional machine learning," included algorithms that require manual processes, such as feature selection, and make decisions based on predefined functions derived from these features. These algorithms are primarily used for structured data processing and typically involve manual steps, such as feature engineering, which include tasks such as data preprocessing, feature extraction, and feature selection. They use specific functions derived from the selected features to make decisions and continue to be widely used in many studies as they typically require less time for training than deep learning models and perform uniformly well on smaller datasets[24–27]. In this study, we aimed to classify patient severity through conversations by applying support vector machine (SVM), logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB), which are among the most commonly used classifiers in existing machine learning algorithms. The second category, "deep learning," was based on artificial neural networks. These models can effectively process large amounts of data by automatically generating features and making decisions through deep networks. Due to this advantage, they are adept at analyzing complex and lengthy data, showing higher performance than traditional machine learning, and have been widely used in NLP

Total N distinct conversations		Severe		Mild
		KTAS 3	KTAS 4	KTAS 5
		753	221	74
Data set	Train	602	177	59
	Validation	75	22	8
	Test	76	22	7
Age	20–29	76	41	19
	30–39	85	34	8
	40–49	87	30	11
	50–59	132	38	10
	60–69	181	43	13
	>70	192	35	13
Gender	Female	355	99	36
	Male	398	122	38
Average N words per conversation (±SD)		923.54 (±549.47)	805.95 (±491.69)	602.81 (±399.19)

Table 1. Description of emergency conversations datasets. SD, standard deviation.

tasks[28–33]. In this study, deep learning models, such as multilayer perceptron (MLP), bidirectional long short-term memory (BiLSTM), and convolutional neural network (CNN), were used to evaluate their effectiveness using conversational data of varying lengths containing transcripts of multi-party conversations between patients, clinicians, and companions. Our selection of machine learning and deep learning models was guided by both theoretical considerations and empirical evidence from the literature. The traditional machine learning algorithms (SVM, LR, RF, XGB) were chosen based on their established performance in text classification tasks and their ability to handle high-dimensional, sparse feature spaces typical of NLP applications. As highlighted in a recent systematic review by Porto[34], XGBoost and deep learning approaches have demonstrated superior performance for patient triage prediction in emergency departments. The neural network models (MLP, BiLSTM, CNN) were selected for their proven effectiveness in capturing sequential dependencies and contextual information in text data, which is particularly valuable when analyzing the complex linguistic patterns in clinical conversations. A flow chart of this study is provided in Fig. 1. The code available at https://github.com/Jaewon-Seo97/er_conversations_ktas_v1.git.

Conventional machine learning models

Conventional machine learning models (e.g. SVM, LR, RF, and XGB) typically utilize feature extraction methods from the input raw data. In this study, we employed the Term Frequency–Inverse Document Frequency (TF–IDF) vectorization technique, which quantifies the importance of words in a document relative to a corpus by weighting terms based on their frequency in an individual document offset by their frequency across the entire dataset. This approach helps highlight diagnostically significant terms while down-weighting common words that carry less clinical relevance. A critical method in NLP, TF–IDF is a numerical measure that reflects the importance of each word within a given document, relative to a collection of documents. A practical application of TF–IDF involves assessing the importance of a term in a document by considering its frequency in that document and its rarity across the corpus. The technique was chosen for this study based on the assumption that patient severity would lead to specific patterns and effects in the words used during the conversations, including those related to pain, symptoms, and questions. Transcripts of multi-party conversations and the frequency of the words used were utilized to vectorize each word. The Scikit-learn library was used to calculate the TF–IDF values, which follow slightly modified formulas for Term frequency (TF) and Inverse document frequency (IDF), as detailed below[35].

TF is a practical measure of the frequency of a word within a conversation, which is calculated by dividing the number of occurrences of the word by the total number of words in the conversation. This is a useful tool for identifying frequently used words in a conversation. For the i-th word in the j-th conversation, let n_{ij} be the number of occurrences and $\sum_k n_{kj}$ be the total number of words in the conversation, $\mathrm{TF_{ij}}$ is represented using Eq. (1):

$$\mathrm{TF_{ij}} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

IDF assesses how uncommon a particular term is in the entire corpus. It is calculated by taking the logarithm of the ratio of the number of conversations containing that term to the total number of conversations. This allows us to weigh terms down to a standard across all conversations, regardless of severity, where C represents the total number of conversations, and *Ni* denotes the number of documents containing the i-th word:

$$\mathrm{IDF_i} = \log\left(\frac{1+D}{1+N_i}\right) + 1 \tag{2}$$

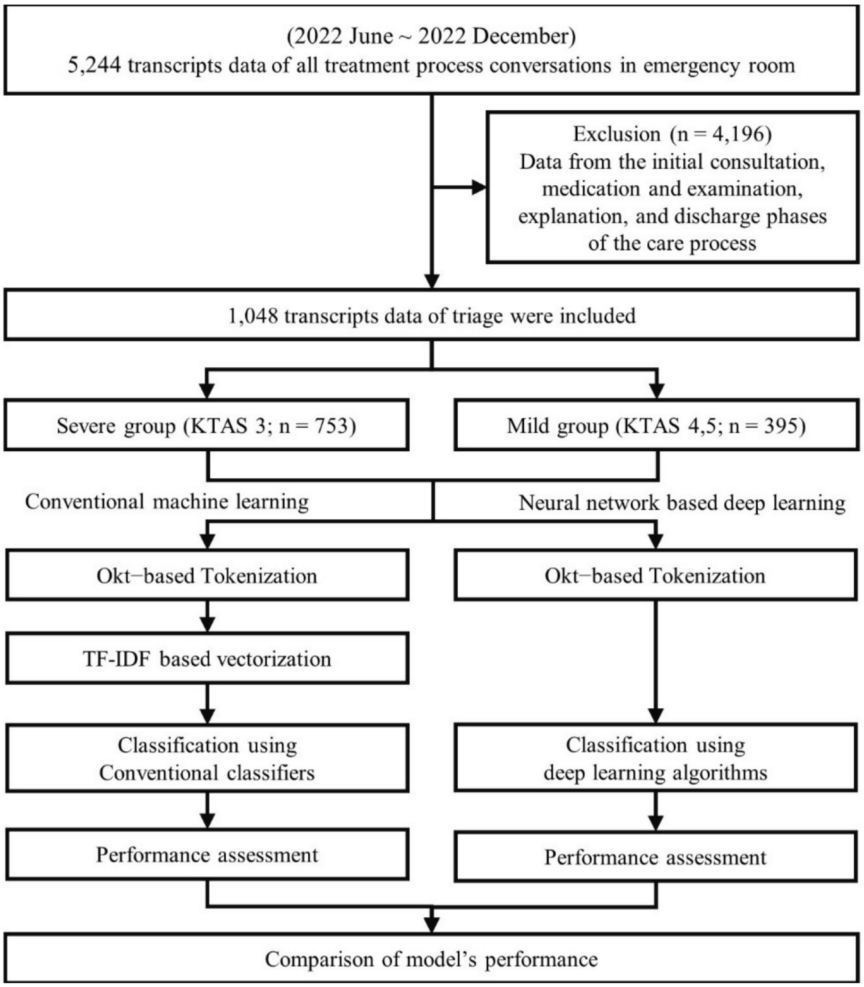

Fig. 1. The study process.

The TF–IDF score for a term in a conversation is derived by multiplying the TF and IDF values of that term:

$$\mathrm{TF-IDF_{ij}} = \mathrm{TF_{ij}} \times \mathrm{IDF_I} \tag{3}$$

As a preprocessing step, we performed morphological tokenization using the Open Korea Text (Okt) morphological analyzer of KoNLPy, a Python open-source library. Subsequently, the extracted features were used to learn each classification model by applying four machine learning classifiers. For each of the machine learning classifiers, hyper-parameters were optimized using a grid search approach. The tuned hyper-parameters for each algorithm are provided in Supplemental Table 1.

Neural network based deep learning models

In recent years, the performance of deep learning algorithms based on artificial neural networks has improved exponentially. In particular, deep learning models have proven to have valid applications in NLP by outperforming conventional machine learning on unstructured data. In this study, we applied MLP, BiLSTM, and CNN models based on artificial neural networks to extract and learn features suitable for patient severity classification by considering contextual content and sequence in long conversations with multiple speakers. The neural network models are trained using tensorflow framework, and.

MLP is the basic form of an artificial neural network and consists of an input layer, one or more hidden layers, and an output layer. Because MLPs use nonlinear activation functions, they can effectively learn nonlinear relationships between input features. We believe this would be advantageous for capturing patterns in conversations and modeling complex interactions, which are important for severity classification. Text-based data typically has higher-dimensional features compared to structured data, and MLPs can effectively handle these higher-dimensional features. These models are able to learn higher-level abstract representations of text in hidden layers beyond simple TF–IDF vectors in the input layer.

BiLSTM is an advanced type of recurrent neural network (RNN) designed to capture dependencies in sequential data by processing input sequences in both forward and backward directions. It consists of two LSTM networks: one that processes sequences from beginning to end (forward LSTM) and one that processes from

end to beginning (backward LSTM)[36]. This bidirectional processing allows each word's preceding and following context to be considered. Since the data utilized in this study includes patient and companion responses to clinicians' questions in a multi-party conversation or clinicians' judgments based on patient and companion's symptom descriptions, each utterance highly depends on the context of the previous or subsequent conversation. Therefore, we used the BiLSTM model because utilizing this bidirectional contextual information allows for more accurate severity classification.

CNN is a class of deep neural networks known primarily for image processing. However, these models are also very effective for specific natural language processing tasks. CNN excels at detecting localized patterns within lengthy data. Using filters to extract features from short-term particles in conversations, they can effectively learn which words or phrases are essential in determining severity. Moreover, through convolutional operations, they can recognize specific patterns regardless of where they are in a single conversation. The data for this study is from a real-world emergency room conversation, and the critical information distinguishing severity can occur anywhere in the conversation. Given these characteristics, the CNN structure has the advantage of being able to detect significant patterns regardless of where the word is located, which was the main reason for utilizing this model in the present study.

Results

To train the AI models, we separated the data into three sets (train, validation, and test) in an 8:1:1 ratio. The test set, carefully separated from the training data, was not used to train the models, ensuring the validity of our analysis. To further confirm the robustness of our models, we performed a tenfold cross-validation. The test set was not used for training, and we utilized 105 data entries (76 for KTAS 3, 22 for KTAS 4, and 7 for KTAS 5) to ensure that each class was equally represented. The We calculated the AUROC, recall, accuracy, precision, and F1-score of the conventional machine learning-based models (e.g. SVM, LR, RF, and XGB) and deep learning based neural network models (e.g. MLP, BiLSTM, and CNN)[37]. Table 2 shows the confusion matrix-based performance values obtained to evaluate and compare the models' average performance from tenfold cross-validation. Each result of the tenfold model performance is shown in Table S2 (Supplementary 2).

The SVM (0.764; 95% CI 0.019) and LR (0.763; 95% CI 0.016) based on conventional machine learning achieved the highest AUROC values, indicating that these two models were effective in classification compared to other models for the data used in this experiment. Among the deep learning-based neural networks, MLP (0.759; 95% CI 0.023) achieved the highest AUROC, while RF (0.718; 95% CI 0.024), XGB (0.711; 95% CI 0.022), and CNN (0.735; 95% CI 0.022) had relatively low AUROC values. Figure 2 shows a box plot comparing the performance of each model based on a tenfold cross-validation.

Discussion

The performance evaluation of machine learning models for emergency department triage reveals important insights into the effectiveness of different algorithmic approaches for patient severity classification. This comprehensive analysis examines traditional machine learning techniques against neural network architectures while addressing the unique challenges of processing real-world clinical conversations.

Model performance evaluation

The performance evaluation of the models used in this study showed that among the existing machine learning models using TF-IDF-based vectorization based on AUROC values, SVM and LR achieved the highest AUROC values (0.764 [95% CI 0.019] and 0.763 [95% CI 0.016], respectively). Because the data used in this study is

Model		AUROC (±95% CI)	*Recall (±95% CI)	Accuracy (±95% CI)	Precision (±95% CI)	F1-score (±95% CI)
Machine learning	SVM	0.764	0.916	0.761	0.787	0.654
		(0.746–0.783)	(0.891–0.941)	(0.745–0.778)	(0.779–0.795)	(0.632–0.676)
	LR	0.763	0.988	0.750	0.746	0.544
		(0.747–0.779)	(0.979–0.996)	(0.741–0.759)	(0.740–0.752)	(0.521–0.568)
	RF	0.718	0.964	0.751	0.757	0.582
		(0.694–0.742)	(0.953–0.975)	(0.741–0.762)	(0.748–0.766)	(0.553–0.611)
	XGB	0.711	0.903	0.745	0.778	0.631
		(0.689–0.734)	(0.879–0.927)	(0.736–0.754)	(0.772–0.785)	(0.617–0.645)
Neural network	MLP	0.759	0.809	0.740	0.826	0.682
		(0.736–0.783)	(0.779–0.838)	(0.721–0.759)	(0.814–0.837)	(0.662–0.702)
	BiLSTM	0.741	0.846	0.746	0.812	0.670
		(0.707–0.775)	(0.793–0.898)	(0.718–0.774)	(0.793–0.831)	(0.638–0.702)
	CNN	0.735	0.787	0.723	0.821	0.667
		(0.713–0.757)	(0.751–0.823)	(0.702–0.744)	(0.809–0.834)	(0.646–0.687)

Table 2. Average performance results from tenfold cross-validation according to the models. SVM, support vector machine; LR, logistic regression; RF, random forest; XGB, extreme gradient boosting; MLP, multi-layer perceptron; BiLSTM, bidirectional long short-term memory, convolutional neural network; CNN.

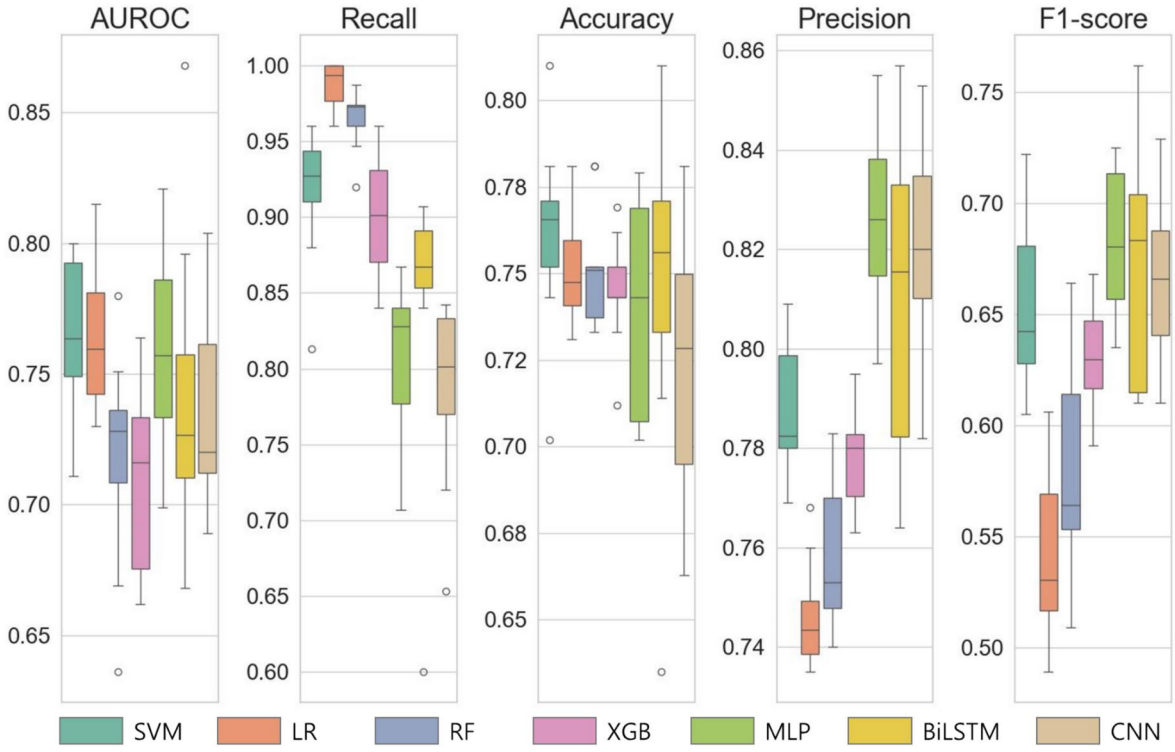

Fig. 2. Evaluation results from each model with 95% confidence intervals.

highly imbalanced, the precision and F1-score should also be considered when evaluating the performance of the two classes. However, the LR model exhibited the lowest precision and F1-score performance, indicating that the LR model's structure, which specializes in linear separation, was inefficient due to the complexity of the real-world clinical-based data used in the experiment. Furthermore, the neural network-based models (MLP, BiLSTM, and CNN), which are more effective for non-linear and complex data compared to traditional machine learning models, demonstrated relatively consistent overall performances due to the nature of the data. In particular, the models that performed above 0.80 for both recall and precision were MLP (recall 0. 809 [95% CI 0.030], precision 0.826 [95% CI 0.011]) and BiLSTM (recall 0.846 [95% CI 0.053], precision 0.812 [95% CI 0.019]), both of which are deep learning-based neural network models that are effective for complex and lengthy data and predicted relatively evenly across all classes.

The AUROC values achieved by our models (ranging from 0.711 to 0.763) reflect the inherent challenges of analyzing unstructured, real-world clinical conversations compared to more structured healthcare data. Several factors contribute to these performance metrics: First, emergency conversations contain significant noise, including interruptions, emotional responses, and non-clinical content that can obscure relevant clinical information. Second, the linguistic variability across different physicians, patients, and companions introduces heterogeneity that challenges standardized analysis. Third, unlike simulated conversations or structured clinical notes used in previous studies, our dataset captures the authentic complexity and messiness of real emergency interactions, including confused responses from distressed patients and conversational detours. Finally, our relatively modest sample size of 1,048 conversations limits the model's opportunity to learn the full spectrum of linguistic patterns associated with different severity levels.

Comparison with related studies

Triage is considered a pivotal way to prevent overcrowding in emergency departments, and some AI-based studies for automatic severity classification have been conducted worldwide. The application of machine learning in emergency departments extends beyond severity classification to encompass various aspects of emergency care using structured data. Recent studies have demonstrated promising results in predicting patient arrivals, optimizing resource allocation, and improving triage accuracy across different healthcare systems. Chang et al. developed a clinical support system for triage based on federated learning specifically for the KTAS, demonstrating how collaborative AI approaches can enhance triage while maintaining patient privacy[20]. Similarly, Choi et al.'s pioneering work with the KTAS system established foundational approaches for machine learning-based severity prediction using structured clinical data[19]. Other researchers have explored integrated approaches combining multiple data streams to enhance predictive performance in emergency settings[34]. Some Korean studies were conducted to classify severity using only conversations between patients and medical staff. However, these differed from our study in their purpose and data used. Cho et al.[38] showed similarities in their utilization of conversation data collected from actual clinical sites. However, they extracted STT and patient information based on Korean speech data to create an EHR for KTAS classification. In contrast,

our study classified severity on the basis of conversational texts from patients, companions, and clinicians to create a system that enables instant classification using only conversational content. Lee et al.[23] and Kim et al.[22] achieved a higher performance (AUROC: 0.89 vs. 0.90) by utilizing AI algorithms to analyze patient information based on the conversations in Korean data. However, a critical limitation of their studies was that the data comprised recorded clinician-patient conversations in a simulated setting, representing a potentially significant divergence from data collected in actual emergency departments. In contrast, the data used in the present study represents real clinical conversations, which contains many unpredictable variables, such as interruptions in the flow of conversation, irrelevant answers to medical staff questions, and varying length distributions for a single conversation. These diverse factors can significantly impact the predictions.

Challenges of korean language processing

Korean is an agglutinative language, one of the most morphologically rich and typologically diverse languages. NLP using Korean is more challenging due to the presence of adverbs, inconsistent word spacing, and various expressions of predicates that have the same meaning [39]. Due to these difficulties, this study has limitations. For example, this study did not include a detailed classification of KTAS scores 4 and 5, and our models need to be more robust to be utilized in real emergency settings. Although we primarily aimed to accurately triage mild cases and prevent overcrowding in emergency departments, our models show relatively low performance in the F1-score, which measures accuracy for each class. This is due to the imbalance of severity classes in our collected data, which reflects the real-world situation, and is a limitation of our study. We selected the Open Korean Text (OKT) analyzer, an open-source tool that efficiently tokenizes and tags parts of speech optimized for Korean, including compound word analysis and conjugation processing essential for medical terminology. While OKT accommodates common speech patterns and clinical terminology used in emergency settings, regional linguistic differences exist throughout Korea, and regions with unique dialects may require additional fine-tuning for optimal performance.

Future research directions

Future research should prioritize external validation to establish the generalizability of our conversation-based severity classification approach. While our current study demonstrates promising results within the three Korea University hospital systems, validation across diverse healthcare settings remains essential yet challenging. Collecting conversations in clinical environments faces substantial barriers, including privacy regulations, technical difficulties in recording clear audio in noisy emergency departments, and resource-intensive transcription processes.

A limitation of our current approach is the lack of explainability analysis. Implementing Explainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations) values would provide valuable insights into which conversation elements most strongly influence severity predictions. Such analysis would enhance clinical interpretability and reveal diagnostic linguistic patterns specific to different severity levels. Future iterations of this research will incorporate these explainability approaches to better understand the decision-making process of our models and identify the most clinically relevant conversational features. We also plan to explore Large Language Models (LLMs), transformer-based deep learning architectures trained on vast text corpora that can understand and process natural language with remarkable capabilities. LLM algorithms based on Korean medical data could be processed to handle homophones from different patients and incorporate long contextual data to significantly improve model performance. Additionally, we aim to expand our research to include multimodal approaches that utilize clinical information such as vital signs to improve prediction performance.

The implementation of AI-based triage systems in clinical practice raises important ethical considerations that must be carefully addressed. Primary concerns include maintaining patient privacy during conversation recording and analysis, ensuring that algorithmic decisions don't exacerbate existing healthcare disparities, and defining appropriate human oversight of AI recommendations. Our work represents an important advancement in applying NLP to authentic clinical scenarios, establishing a foundation for future refinements that could incorporate multimodal data to enhance predictive accuracy in emergency triage.

Conclusions

In this study, we used an AI algorithm to classify the severity of patients based on real multilateral dialogues between clinicians, patients, and companions collected within emergency department of hospitals in Korea. We applied conventional machine learning (e.g. SVM, LR, RF, and XGB) using the TF-IDF technique, which assigns importance to each word based on the frequency of occurrence of the word in the conversation. Furthemore, deep learning-based models (e.g. MLP, BiLSTM, and CNN), which effectively extract long contextual information, were also applied, and the results were analyzed through performance evaluation of the models. The performance evaluation results showed that the TF-IDF-based SVM model achieved the highest performance; however, it was slightly lower than the results reported in previous studies on severity classification based on conversations within emergency department of Korean hospitals.

Notably, this study classified patient severity based on in situ data collected from actual conversations in emergency departments. Unlike previous studies that primarily relied on simulated conversations or structured clinical data, our approach leverages the authentic, often messy, complexities of real-world clinical interactions. By presenting a novel data set for NLP analysis, the results presented in this study provide valuable insights that could help facilitate the effective triaging of patients under time-sensitive conditions in the emergency department of hospitals in the future.

Data availability

The data that support the findings of this study are available from The Open AI Dataset Project (AI-Hub, S. Korea). Restrictions apply to the availability of these data, which are accessible through AI-Hub (https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71,433). Access to the data requires registration on the platform and compliance with the data request procedures, conditions, and methods specified by AI-Hub. Data are available from the corresponding author upon reasonable request and with permission from AI-Hub.

Code availability

The code available at https://github.com/Jaewon-Seo97/er_conversations_ktas_v1.git. In our experiments, we used Python 3.8, and the following open-source libraries: tensorflow = 2.10.0, joblib = 1.2.0, JPype1 = 1.4.1, konlpy = 0.6.0, h5py = 3.11.0, xgboost = 1.7.2, pandas = 1.5.2, numpy = 1.24.1, scikit-learn = 1.3.1.

Received: 31 December 2024; Accepted: 23 April 2025
Published online: 15 May 2025

References

1. Choi, S. K. The view of emergency medicine physician over the Korean emergency medical system; problems and improvements. *Public Health Aff.* **3**, 177–183 (2019).
2. Busti, C., Marchetti, R. & Monti, M. Overcrowding in emergency departments: Strategies and solutions for an effective reorganization. *Ital. J. Med.* https://doi.org/10.4081/itjm.2024.1714 (2024).
3. Foglia, E. et al. Overcrowding and boarding time: Emergency department performance and impacting factors. *Value Health* **26**, S442–S442 (2023).
4. McCarthy, M. L. Overcrowding in emergency departments and adverse outcomes. *BMJ* **342**, d2830. https://doi.org/10.1136/bmj.d2830 (2011).
5. Pearce, S., Marr, E., Shannon, T., Marchand, T. & Lang, E. Overcrowding in emergency departments: An overview of reviews describing global solutions and their outcomes. *Intern. Emerg. Med.* **19**, 483–491. https://doi.org/10.1007/s11739-023-03477-4 (2024).
6. Loftus, T. J. et al. Overtriage, undertriage, and value of care after major surgery: An automated, explainable deep learning-enabled classification system. *J. Am. Coll. Surg.* **236**, 279–291. https://doi.org/10.1097/XCS.0000000000000471 (2023).
7. Huabbangyang, T. et al. Associated factors of under and over-triage based on the emergency severity index; a retrospective cross-sectional study. *Arch. Acad. Emerg. Med.* **11**, e57. https://doi.org/10.22037/aaem.v11i1.2076 (2023).
8. Gratton, R. J. et al. Acuity assessment in obstetrical triage. *J. Obstet. Gynaecol. Can.* **38**, 125–133. https://doi.org/10.1016/j.jogc.2015.12.010 (2016).
9. Zachariasse, J. M. et al. Validity of the manchester triage system in emergency care: A prospective observational study. *PLoS ONE* **12**, e0170811. https://doi.org/10.1371/journal.pone.0170811 (2017).
10. Zachariasse, J. M. et al. Performance of triage systems in emergency care: A systematic review and meta-analysis. *BMJ Open* **9**, e026471. https://doi.org/10.1136/bmjopen-2018-026471 (2019).
11. Huibers, L., Giesen, P., Wensing, M. & Grol, R. Out-of-hours care in western countries: Assessment of different organizational models. *BMC Health Serv Res* **9**, 105. https://doi.org/10.1186/1472-6963-9-105 (2009).
12. Park, J. & Lim, T. Korean triage and acuity scale (KTAS). *J. Korean Soc. Emerg. Med.* **28**, 547–551 (2017).
13. Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **82**, 3713–3744. https://doi.org/10.1007/s11042-022-13428-4 (2023).
14. Raparthi, M. et al. Advancements in natural language processing-a comprehensive review of AI techniques. *J. Bioinf. Artif. Intell.* **1**, 1–10 (2021).
15. Sawicki, J., Ganzha, M. & Paprzycki, M. The state of the art of natural language processing—a systematic automated review of NLP literature using NLP techniques. *Data Intell.* **5**, 707–749 (2023).
16. Le Glaz, A. et al. Machine learning and natural language processing in mental health: Systematic review. *J. Med. Internet. Res.* **23**, e15708. https://doi.org/10.2196/15708 (2021).
17. Sheikhalishahi, S. et al. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Med. Inform* **7**, e12239. https://doi.org/10.2196/12239 (2019).
18. Mueller, B., Kinoshita, T., Peebles, A., Graber, M. A. & Lee, S. Artificial intelligence and machine learning in emergency medicine: A narrative review. *Acute Med. Surg.* https://doi.org/10.1002/ams2.740 (2022).
19. Choi, S. W., Ko, T., Hong, K. J. & Kim, K. H. Machine learning-based prediction of Korean triage and acuity scale level in emergency department patients. *Healthc. Inform. Res.* **25**, 305–312. https://doi.org/10.4258/hir.2019.25.4.305 (2019).
20. Chang, H. et al. Clinical support system for triage based on federated learning for the Korea triage and acuity scale. *Heliyon* **9**, e19210. https://doi.org/10.1016/j.heliyon.2023.e19210 (2023).
21. Porto, B. M. Improving triage performance in emergency departments using machine learning and natural language processing: A systematic review. *BMC Emerg. Med.* **24**, 219. https://doi.org/10.1186/s12873-024-01135-2 (2024).
22. Kim, D. et al. Automatic classification of the Korean triage acuity scale in simulated emergency rooms using speech recognition and natural language processing: A proof of concept study. *J. Korean Med. Sci.* https://doi.org/10.3346/jkms.2021.36.e175 (2021).
23. Lee, S. et al. Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage. *Am. J. Emerg. Med.* **77**, 29–38. https://doi.org/10.1016/j.ajem.2023.11.063 (2024).
24. HaCohen-Kerner, Y., Miller, D. & Yigal, Y. The influence of preprocessing on text classification using a bag-of-words representation. *PLoS ONE* https://doi.org/10.1371/journal.pone.0232525 (2020).
25. Chen, T. Q. & Guestrin, C. XGBoost: A scalable tree boosting system. *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016). https://doi.org/10.1145/2939672.2939785
26. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
27. Joachims, T. *Machine Learning: ECML-98* (Springer, 1998).
28. Adhikari, A., Ram, A., Tang, R. & Lin, J. Rethinking Complex Neural Network Architectures for Document Classification. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Naacl Hlt 2019), Vol. 1*, 4046–4051 (2019).
29. Athanasios, T., Spyros, B., Panagiota, T. & Athanasios, V. An exploration on text classification using machine learning techniques. *25th Pan-Hellenic Conference on Informatics with International Participation (Pci2021)*, 247–249 (2021). https://doi.org/10.1145/3503823.3503869
30. Duque, A. B., Santos, L. L. J., Macêdo, D. & Zanchettin, C. Squeezed Very Deep Convolutional Neural Networks for Text Classification. *Artificial Neural Networks and Machine Learning - Icann 2019: Theoretical Neural Computation, Pt I* **11727**, 193–207 (2019). https://doi.org/10.1007/978-3-030-30487-4_16

31. Malhotra, R. & Singh, P. Recent advances in deep learning models: A systematic literature review. *Multimed. Tools Appl.* **82**, 44977–45060. https://doi.org/10.1007/s11042-023-15295-z (2023).
32. Ramlakhan, S. et al. Understanding and interpreting artificial intelligence, machine learning and deep learning in Emergency Medicine. *Emerg. Med. J.* **39**, 380–385. https://doi.org/10.1136/emermed-2021-212068 (2022).
33. Wang, J., Wang, Z. Y., Zhang, D. W. & Yan, J. Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2915–2921 (2017).
34. Porto, B. M. & Fogliatto, F. S. Enhanced forecasting of emergency department patient arrivals using feature engineering approach and machine learning. *BMC Med. Inform. Decis. Mak.* **24**, 377. https://doi.org/10.1186/s12911-024-02788-6 (2024).
35. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
36. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**, 2673–2681. https://doi.org/10.1109/78.650093 (1997).
37. Powers, D. M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint* (2020). arXiv:2010.16061
38. Cho, A. et al. Effect of applying a real-time medical record input assistance system with voice artificial intelligence on triage task performance in the emergency department: Prospective interventional study. *JMIR Med. Inform.* https://doi.org/10.2196/39892 (2022).
39. Shin, D., Kam, H. J., Jeon, M. S. & Kim, H. Automatic classification of thyroid findings using static and contextualized ensemble natural language processing systems: Development study. *JMIR Medi. Inform.* (2021). https://doi.org/10.2196/30223

Acknowledgements

This research was supported by a grant of Korea University (Grant no. K2509731). The funding source had no role in the design of this study; data collection, analysis, and interpretation; and decision to publish or preparation of the manuscript. Moreover, this work was supported by the GRRC program of Gyeonggi province [GRRC-Gachon2023(B01), Development of AI-based medical imaging technology], and by the Gachon Program (GCU-202307640001).

Author contributions

JWS: Writing–original draft preparation, conceptualization, methodology, visualization, formal analysis, investigation, software, data curation. SJP: Writing–original draft preparation, conceptualization, formal analysis, investigation, data curation. YJK: Conceptualization, investigation, validation, data curation, writing–review & editing. JYK: Conceptualization, validation, data curation, investigation. KGK: Conceptualization, validation, resources, funding acquisition, supervision, writing–review & editing. YHY: Conceptualization, validation, resources, supervision, writing–review & editing. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study was reviewed and approved by the Institutional Review Board of Korea University Hospital (IRB no: Guro Hospital, 2022GR0156; Ansan Hospital, 2022AS0132; Anam Hospital, 2022AN0288). All experimental protocols were reviewed and approved by the Korea University Hospital IRB, and the study was performed in accordance with the relevant guidelines and regulations, including the Declaration of Helsinki.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-99874-0.

Correspondence and requests for materials should be addressed to K.G.K. or Y.-H.Y.