J Korean Med Sci. 2021 Jul 12;36(27):e175
https://doi.org/10.3346/jkms.2021.36.e175
eISSN 1598-6357·pISSN 1011-8934

JKMS

Original Article
Medical Informatics

Automatic Classification of the Korean Triage Acuity Scale in Simulated Emergency Rooms Using Speech Recognition and Natural Language Processing: a Proof of Concept Study

Dongkyun Kim ,[1*] Jaehoon Oh ,[2*] Heeju Im ,[3] Myeongseong Yoon ,[2] Jiwoo Park ,[2] and Joohyun Lee [1]

[1]Department of Electrical and Electronic Engineering, Hanyang University, Ansan, Korea
[2]Department of Emergency Medicine, College of Medicine, Hanyang University, Seoul, Korea
[3]Department of Artificial Intelligence, Hanyang University, Seoul, Korea

Received: Jan 21, 2021
Accepted: Jun 7, 2021

Address for Correspondence:
Joohyun Lee, PhD
Department of Electrical and Electronic Engineering, Hanyang University, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan 15588, Republic of Korea.
E-mail: joohyunlee@hanyang.ac.kr

*Dongkyun Kim and Jaehoon Oh are co-first authors and contributed equally to this work.

ORCID iDs
Dongkyun Kim
https://orcid.org/0000-0002-6355-0101
Jaehoon Oh
https://orcid.org/0000-0001-8055-1467
Heeju Im
https://orcid.org/0000-0002-0839-3895
Myeongseong Yoon
https://orcid.org/0000-0003-1019-3347
Jiwoo Park
https://orcid.org/0000-0002-4857-851X
Joohyun Lee
https://orcid.org/0000-0002-7698-1568

ABSTRACT

Background: Rapid triage reduces the patients' stay time at an emergency department (ED). The Korean Triage Acuity Scale (KTAS) is mandatorily applied at EDs in South Korea. For rapid triage, we studied machine learning-based triage systems composed of a speech recognition model and natural language processing-based classification.
Methods: We simulated 762 triage cases that consisted of 18 classes with six types of the main symptom (chest pain, dyspnea, fever, stroke, abdominal pain, and headache) and three levels of KTAS. In addition, we recorded conversations between emergency patients and clinicians during the simulation. We used speech recognition models to transcribe the conversation. Bidirectional Encoder Representation from Transformers (BERT), support vector machine (SVM), random forest (RF), and k-nearest neighbors (KNN) were used for KTAS and symptom classification. Additionally, we evaluated the Shapley Additive exPlanations (SHAP) values of features to interpret the classifiers.
Results: The character error rate of the speech recognition model was reduced to 25.21% through transfer learning. With auto-transcribed scripts, support vector machine (area under the receiver operating characteristic curve [AUROC], 0.86; 95% confidence interval [CI], 0.81–0.9), KNN (AUROC, 0.89; 95% CI, 0.85–0.93), RF (AUROC, 0.86; 95% CI, 0.82–0.9) and BERT (AUROC, 0.82; 95% CI, 0.75–0.87) achieved excellent classification performance. Based on SHAP, we found "*stress*", "*pain score point*", "*fever*", "*breath*", "*head*" and "*chest*" were the important vocabularies for determining KTAS and symptoms.
Conclusion: We demonstrated the potential of an automatic KTAS classification system using speech recognition models, machine learning and BERT-based classifiers.

Keywords: Triage; Classification; Machine Learning; Natural Language Processing; Deep Learning

Funding
This study was supported by the National Research Foundation of Korea (2019R1F1A1063502). This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2020-0-01343, Artificial Intelligence Convergence Research Center [Hanyang University ERICA]). This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) and funded by the Korean government (MSIT) (grant No. 2019M3E5D1A01069363).

Disclosure
The authors have no potential conflicts of interest to disclose.

Author Contributions
Conceptualization: Oh JH, Lee JH. Data curation: Kim DK, Oh JH, Yoon M, Park JW. Formal analysis: Kim DK, Im HJ, Oh JH, Lee JH. Funding acquisition: Oh JH, Lee JH. Investigation: Kim DK, Im HJ, Yoon M. Methodology: Kim DK, Im HJ, Lee JH. Software: Kim DK, Im HJ. Validation: Kim DK, Oh JH, Yoon M, Lee JH. Visualization: Kim DK. Writing - original draft: Kim DK, Oh JH, Lee JH. Writing - review & editing: Oh JH, Lee JH.

INTRODUCTION

The number of patients visiting emergency departments (ED) reached 10 million per year in South Korea as of 2019.[1] If the ED is overcrowded, many critical patients cannot be examined and treated by the best and quickest method in the right place and with sufficient medical resources.[2,3] To solve this problem, the Korean Triage and Acuity Scale (KTAS) has been mandatorily applied at the ED in South Korea since 2016. The KTAS, which is based on the Canadian Triage and Acuity Scale (CTAS), consists of the following processes: 1) Check the '*critical first look*' as soon as a patient arrives at the ED, 2) Screen for infectious disease, 3) Conduct triage assessment, 4) Select and document the presenting complaint of the patient, 5) Consider modifiers (e.g., consciousness, vital sign, pain, and other histories), and 6) Assign the triage level. In the KTAS decision process, there are 17 major classes, 167 main symptoms, and if we also consider modifiers, there are approximately 2,700 cases for the 5-level assignment decision tree. The goal is to triage a patient within 10 to 15 minutes of arrival.[4,5] The 5-level triage tool, KTAS, is the most considerable predictor affecting the disposition of ED patients. The KTAS is associated with the average length of stay and mortality rate.[5,6] The KTAS is a reliable triage tool, and KTAS scores determined by emergency nurses and experts are consistent in most cases.[7,8] Also, the KTAS is preferred over triaging patients by Emergency Severity Index.[9]

This study was conducted to determine whether artificial intelligence can accurately classify KTAS levels and symptoms by extracting multivariate information from conversations between clinicians and emergency patients. Artificial intelligence has already been widely used and has shown remarkable performance in diverse areas such as medical imaging and diagnosis. The collaborative process of clinical science and data science is essential for accurate medical decision-making.[10] Recent studies applied machine learning (ML) to predict the severity and hospitalization of patients at ED based on electronic health record (EHR). The machine learning approaches achieved high performance in predicting hospitalization for adult, children, and pediatric asthma patients.[11-13] Although multivariate logistic regression has been predominant in the field of medical statistics, in predicting the disposition of emergency patients, non-linear ML models, such as tree-based ensemble models, and deep neural networks, performed better than the logistic regression model.[14] In addition, if natural language data is analyzed along with EHR, machine learning could more accurately predict the emergency patients' severity.[15,16] Although this machine learning approach showed high classification performance, it was challenging to interpret the decision of ML classifiers as in other areas. Several studies measured the feature importance for interpreting ML models. The mean impurity difference (MID) of complete RF nodes could be used to identify important features.[17,18] However, MID overestimates high cardinality features and cannot expand model interpretation to the test dataset. Other studies interpreted the model with the post-hoc analysis, permutation feature importance (PFI).[19-21] PFI is a model-agnostic methodology that can be used for interpretation regardless of the model type, and it allows model interpretation on both training data and test data. However, PFI always yields different results due to the random shuffling of the data. Recently, Shapley Additive exPlanations (SHAP) has been used to overcome this weakness and quantify the contribution of each feature in ML classification.[22,23] SHAP is an advanced algorithm that mathematically demonstrates the consistency and fairness of feature significance.

This preliminary study aimed to develop an automatic triage system using speech recognition and natural language processing (NLP) for the Korean language to save clinicians' precious time at ED. This study did not use EHR data as inputs of models, and the KTAS and symptoms were classified by analyzing bilateral conversations between clinicians and patients. This study used speech recognition models to transcribe the conversation and developed ML and deep learning-based classifiers that predict the KTAS and symptoms. This study also identified important words for the classification by measuring the game-theoretic feature importance, SHAP.

METHODS

Study design

We conveniently selected 762 retrospective cases among conscious patients who visited the ED of Hanyang University Hospital between September and December 2019. We simulated the triage situation of each case based on the clinical records and recorded dialog during simulation. Each dialog represents a case between an emergency patient and a medical clinician. These data include necessary information for KTAS classification, such as signs of infectious diseases, symptoms, vital signs, pain scores, and other histories. This study's primary outcomes were based on three KTAS levels (2, 3, and 4) and six main symptoms (chest pain, dyspnea, fever, stroke, abdominal pain, and headache) that human clinicians have classified correctly. The six main symptoms accounted for 54.5% of internal medicine patients in the emergency medical center.[24] For the speech recognition and classification task, we randomly split the overall dataset into a training dataset (80%) and a test dataset (20%). **Fig. 1** shows our experimental system consists of four steps: 1) Voice data collection, 2) Automatic speech recognition, 3) Deep learning and ML-based classification, and 4) Analysis of word importance.

Collection of voice data and human-transcribed scripts

We recruited twenty volunteers consisting of four emergency medical technicians, each having a certification of a KTAS classifier, and sixteen students from a university located in the capital area of South Korea and a university hospital in Seoul. The volunteers were older than 18 years of age and in good health status. Volunteers were excluded if they have a vocal cord or pulmonary/heart disease. The volunteers performed the given roles as if they were in a real case. Two commercial recorders (VTR6600; Philips, Amsterdam, Netherlands) were used to record the simulated dialogues, and each device was placed in front of the medical clinician or patient. After the voice data were collected, we recruited four undergraduate students from the Division of Electrical Engineering at Hanyang University, and they manually transcribed the simulated dialogues to generate human-transcribed scripts.

Automatic transcription: speech-to-text process

To automatically transcribe the dialogs, we trained speech recognition models (IBM-Custom-Speech and Microsoft-Custom-Speech) through a transfer learning algorithm. We converted a test set of dialogs into text documents and calculated the character error rate (CER), a general metric used to evaluate a Korean speech recognition model. The performance of a speech recognition model is better as the CER value decreases. The CER is a function defined as follows:

$$CER = \frac{S + D + I}{N} \quad (1)$$

N is the total number of words in the ground-truth script, S, D, and I are substituted, deleted, and inserted characters in a speech recognition result to obtain the ground-truth script.

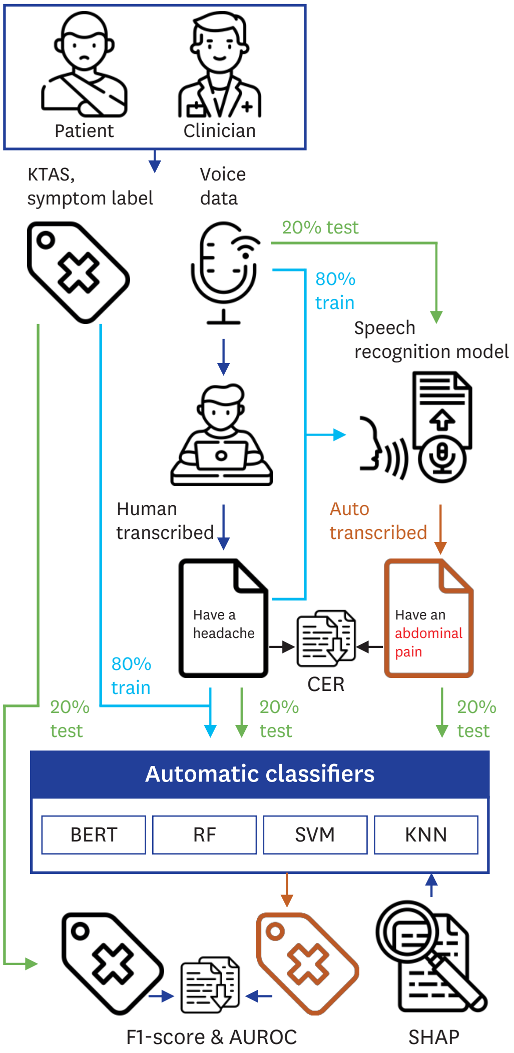

Fig. 1. Overall flow of the classification system in this study.
KTAS = Korean Triage and Acuity Scale, CER = character error rate, BERT = Bidirectional Encoder Representations from Transformers, SVM = support vector machine, KNN = k-nearest neighbors, RF = random forest, AUROC = area under the receiver operating characteristic curve, SHAP = Shapley Additive exPlanations.

Then, we statistically compared the error of speech recognition models using the McNemar's test.[25] We calculated significant probabilities with a 2-tailed test. *P* values < 0.05 were considered statistically significant.

Training of the Bidirectional Encoder Representations from Transformers (BERT) model

The BERT is a deep learning-based NLP technique developed by Google, which showed superior performance in many NLP tasks.[26] BERT consists of layers called bidirectional transformers that enable a classifier to grasp the context of the input sentences.[26] As the original BERT model mainly focuses on English documents, we used the Korean Language Model (KorBERT) provided by ETRI (No. 2013-2-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services), which pre-trained a large number of Korean texts. Based on the KorBERT transformer layer, we developed a deep learning classifier called BERT-KTAS that simultaneously predicts KTAS and symptoms using text. We tokenized the whole transcripts as morphemes and used them as inputs of BERT-KTAS. BERT-KTAS refines the input tokens with its built-in embedding layer and then extracts information from inputs using transformer layers, finally performs classification by fully connected layers. We placed two fully connected layers behind the KorBERT transformers, one for KTAS classification and another for symptom classification. Each fully connected layer receives the encoded information from transformer layers and performs multi-class classification (three categories for KTAS levels and six categories for symptoms). In the training process, we tuned BERT-KTAS's hyperparameters (e.g., learning rate schedule, batch size, and train epochs) to maximize the area under the receiver operating characteristic curve (AUROC) using *Grid Search* and *Warmup Learning Rate Control*.[27] Further details of the structure and training processes of BERT-KTAS are provided in **Supplementary Method 1**.

Training of machine learning models

To compare with the BERT-KTAS, we also developed ML classifiers. ML is a high-level concept that includes deep learning, but in this study, ML refers to traditional ML models: support vector machine (SVM), k-nearest neighbors (KNN), and random forest (RF), which have been widely used in classification tasks. To enable these ML models to classify natural language, the pre-processing such as *Feature selection* and *Word feature weighting* had been preceded. First, we selected outcome-relevant words from the training dataset by maximal χ^2 statistics.[28] In our study, the domain of outcome was the set of all KTAS and symptom labels, $\mathcal{L}$, and the domain of inputs was the set of words in transcripts, $\mathcal{W}$. χ^2 quantifies the difference of a word $W \in \mathcal{W}$ distribution for each label $L \in \mathcal{L}$. We chose the first 100 words (noun, verb, adjective, and adverb) in the order of the magnitude of the χ^2 statistic.

$$\chi^2(W,L) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (2)$$

N is the number of all documents in the training dataset, A is the number of documents that include word W and belong to label L, B is the number of documents that include word W and do not belong to label L, C is the number of documents that do not include word W and belong to label L, and D is the number of documents that do not include word W and do not belong to label L.

Next, we converted the selected words to numeric vectors using Okapi BM25. Okapi BM25 showed better performance than other frequency-based word weighting techniques in the document clustering and categorization experiment.[29,30] In the BM25 equation, TF(q_i) represents the frequency of the word (q_i) in the text file (d_j) and IDF(q_i) calculates the weight of the word (q_i)'s sparsity in the entire text.

$$\mathrm{IDF}(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad (3)$$

$$\mathrm{TF}(qi) = \frac{3 \times f(q_i, d_j)}{1 + f(q_i, t_j) + \frac{|d_j|}{avgdl}} \quad (4)$$

$$\mathrm{BM25}(d^j, q_i) = IDF(q_i) \times TF(q_i) \quad (5)$$

N is the number of documents in the dataset, q_i is the i-th word in the set of selected words by χ^2 statistics, $n(q_i)$ is the number of documents containing q_i, $f(q_{ij}, d)$ is the q_i's term frequency in document d_j, $|d_j|$ is the length of document d_j in words, and *avgdl* is the average document length in the dataset. We fitted the χ^2 statistics and BM25 on the training dataset and transformed the test dataset using fitted expression. As a result, each document was converted to a 100-dimensional numerical vector and SVM, KNN, and RF classified these vectors. Unlike the BERT model, since ML cannot perform multi-output classification, KTAS levels and symptoms cannot be predicted simultaneously with one model. Thus, models for KTAS and symptoms were separately developed and evaluated. Also, since SVM is only capable of binary classification, we developed three SVMs for KTAS levels and six SVMs for symptoms. To derive predictions from SVMs, we selected the class with the highest prediction confidence among SVMs. We tuned the hyperparameters of each classifier to maximize the AUROC using 5-fold *cross-validation* and *grid search*. Further details of ML training process are explained in **Supplementary Method 2**.

Statistical analysis

We trained the classification models on the human-transcribed data and evaluated the models on the human-transcribed data and auto-transcribed data. We calculated the macro average of Recall (True positive/[True positive + False positive]), Precision (True Positive/[True positive + False negative]), F1-score (2 * Recall * Precision/[Recall + Precision]) and the AUROC to quantify the classification performance.[31] Macro averaging is a method of averaging the evaluation metric results over an entire class. The 95% confidence interval (CI) was calculated using bootstrapping.[32] To statistically compare the receiver operating characteristic (ROC) curves between BERT and other machine learning models, we conducted the DeLong's test.[33] We calculated significant probabilities with a 2-tailed test. P values < 0.05 were considered statistically significant.

The models we used, such as SVM, RF, and BERT-KTAS, perform nonlinear operations, and this property complicates the models, making them difficult to be interpreted. Thus, we interpreted our classifiers using a game theory-based analysis method called SHAP, which is mathematically based on Shapely value, which is very consistent in the result and fair to identify feature importance. The Shapely is a function defined as follows:

$$\text{Shapely value}(i) = \sum_{S \in N\{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!}[v(S \cup \{i\}) - v(S)] \quad (6)$$

N is the set of features (words), n is the total number of features, S is a set (coalition) of features, and the function $v(S)$ is the classifier's output when the coalition includes the features in S. The Shapley value is a game-theoretic metric that calculates the contribution of each participant for cooperation outcomes.[34] SHAP interprets a machine learning model by approximating the shapely value of input variables.[35] We approximated the Shapely value using coefficients of a linear model for ML and the permutationally distorted input word to approximate the Shapely value for BERT-KTAS. We conducted this study using NumPy,[36] Scikit-learn,[37] PyTorch in Python language.[38]

Ethics statement

This study was approved by the Institutional Review Board at Hanyang University Hospital (HYUH 2020-02-008-003). We designed a prospective and preliminary simulation study with scenarios developed from retrospective medical chart reviews. The study was carried out at the simulation room of Hanyang University Hospital (Seoul, Republic of Korea) in April 2020. The simulation participants were well informed of this study before the experiment and provided written consent.

RESULTS

The baseline characteristics of participants who provided these cases are shown in **Table 1**. Of the 762 patients, KTAS level 2, level 3, and level 4 were 205, 353, and 204, respectively. There was no case of stroke in level 4 of KTAS.

Speech recognition CER

The CERs of speech recognition models are shown in **Table 2**. IBM-Custom-Speech and Microsoft-Custom-Speech provide the transfer learning function, enabling us to further train the model based on the human-transcribed scripts. Through transfer learning, IBM, and Microsoft's CER were 25.21% and 30.8%, respectively. The results of McNemar's test show that the error rates between the two speech recognition models are significantly different ($P < 0.05$).

Classification over human-transcribed scripts

This result showed the classifiers' performance on human-transcribed test datasets. This experiment was the same as the experiment with an ideal speech recognition model with zero CER. **Table 3** shows the result of classification on the human-transcribed documents.

Table 1. The demographic characteristics of study population

Variables	KTAS			
	Level 2 (n = 205)	Level 3 (n = 353)	Level 4 (n = 204)	Total (n = 762)
Age, yr	56.32 ± 22.8	53.69 ± 20.53	33.92 ± 16.57	49.13 ± 22.17
Women	100 (48.78)	185 (52.41)	130 (63.73)	415 (54.46)
Symptom				
Chest pain	55 (26.83)	49 (13.88)	8 (3.92)	112 (14.70)
Dyspnea	35 (17.07)	60 (17.00)	27 (13.24)	122 (16.01)
Fever	14 (6.83)	68 (19.26)	31 (15.20)	113 (14.83)
Stroke	20 (9.76)	64 (18.13)	0 (0.00)	84 (11.02)
Abdominal pain	38 (18.54)	54 (15.30)	68 (33.33)	160 (21.00)
Headache	43 (20.98)	58 (16.43)	70 (34.31)	171 (22.44)

Values are presented as mean ± standard deviation or number (%).
KTAS = Korean triage and acuity system.

Table 2. The comparison of speech recognition models

Model	IBM-Custom-Speech	Microsoft-Custom-Speech	*P* value[a]
CER (%)	25.21%	30.80%	< 0.001

CER = character error rate.
[a] *P* value < 0.05 were considered statistically significant and derived from the McNemar's test.

Table 3. Model performance with human-transcribed test dataset[a]

Model	Recall	Precision	F1-score	AUROC	*P* value[b]
KTAS classifier					
BERT-KTAS	0.72 (0.65–0.80)	0.73 (0.66–0.80)	0.73 (0.66–0.80)	0.85 (0.79–0.90)	
SVM	0.78 (0.71–0.85)	0.78 (0.70–0.85)	0.78 (0.70–0.85)	0.90 (0.86–0.94)	0.022
K-NN	0.74 (0.67–0.81)	0.76 (0.69–0.83)	0.75 (0.68–0.82)	0.90 (0.85–0.94)	0.014
RF	0.74 (0.67–0.81)	0.75 (0.68–0.82)	0.75 (0.68–0.81)	0.90 (0.85–0.93)	0.016
Symptom classifier					
BERT-KTAS	0.95 (0.91–0.98)	0.93 (0.89–0.97)	0.94 (0.90–0.98)	0.99 (0.97–1.00)	
SVM	0.93 (0.88–0.97)	0.93 (0.87–0.97)	0.93 (0.88–0.97)	1.00 (0.99–1.00)	0.162
K-NN	0.94 (0.90–0.98)	0.94 (0.89–0.98)	0.94 (0.90–0.98)	0.99 (0.98–1.00)	0.148
RF	0.94 (0.89–0.98)	0.95 (0.90–0.98)	0.94 (0.90–0.98)	1.00 (1.00–1.00)	0.110

BERT = Bidirectional Encoder Representations from Transformers, KTAS = Korean Triage and Acuity System, SVM = support vector machine, K-NN = k-nearest neighbors, RF = random forest, AUROC = area under the receiver operating characteristic curve.
[a] All performances are shown as mean (95% confidence interval); [b] *P* value < 0.05 were considered statistically significant and derived from the Delong's test.

Table 4. Model performance with IBM's auto-transcribed test dataset[a]

Model	Recall	Precision	F1-score	AUROC	*P* value[b]
KTAS classifier					
BERT-KTAS	0.62 (0.54–0.70)	0.68 (0.59–0.77)	0.65 (0.57–0.73)	0.82 (0.75–0.87)	
SVM	0.73 (0.66–0.79)	0.73 (0.66–0.80)	0.73 (0.66–0.79)	0.86 (0.81–0.90)	< 0.001
K-NN	0.73 (0.65–0.80)	0.72 (0.64–0.79)	0.72 (0.65–0.79)	0.89 (0.85–0.93)	< 0.001
RF	0.63 (0.55–0.71)	0.67 (0.58–0.75)	0.65 (0.57–0.73)	0.86 (0.82–0.9)	< 0.001
Symptom classifier					
BERT-KTAS	0.87 (0.82–0.92)	0.88 (0.83–0.92)	0.87 (0.82–0.92)	0.98 (0.96–0.99)	
SVM	0.90 (0.83–0.94)	0.90 (0.84–0.95)	0.90 (0.83–0.94)	0.99 (0.98–1.0)	< 0.001
K-NN	0.85 (0.78–0.91)	0.85 (0.79–0.91)	0.85 (0.78–0.91)	0.99 (0.98–1.0)	< 0.001
RF	0.86 (0.84–0.95)	0.87 (0.80–0.94)	0.87 (0.81–0.93)	0.99 (0.99–1.0)	< 0.001

BERT = Bidirectional Encoder Representations from Transformers, KTAS = Korean Triage and Acuity System, SVM = support vector machine, K-NN = k-nearest neighbors, RF = random forest, AUROC = area under the receiver operating characteristic curve.
[a]All performances are shown as mean (95% confidence interval); [b]*P* value < 0.05 were considered statistically significant and derived from the Delong's test.

For the KTAS level, the SVM (AUROC, 0.9; 95% CI, 0.86–0.94), KNN (AUROC, 0.9; 95% CI, 0.85–0.94) and RF (AUROC, 0.9; 95% CI, 0.86–0.94) achieved higher performance than BERT-KTAS (AUROC, 0.85; 95% CI, 0.79–0.9) and *P* values of difference between the BERT-KTAS and the other models' AUROC (SVM, KNN, and RF) were 0.022, 0.014, and 0.016, respectively. However, for the symptom class, all performances were over 0.99 in AUROC and there is no statistically significant difference between them (all $P > 0.05$).

Classification over auto-transcribed scripts from speech recognition models

We also conducted the same experiment using the auto-transcribed scripts from speech recognition models. In this experiment, we could see the effect of CER of speech recognition models on the classification tasks. The performance of the overall classifiers decreased as CER increased. **Table 4** shows the result of classification on the transcription of the IBM speech recognition model. In KTAS classification, the KNN achieved the highest AUROC (0.89; 95% CI, 0.85–0.93). In symptom classification, performance of all classifiers was similar, with AUROC greater than 0.98, but SVM was the highest in F1-score (AUROC, 0.9; 95% CI, 0.83–0.94). With the IBM speech recognition model, all *P* values of difference between BERT-KTAS and other models' AUROC were less than 0.05 in both KTAS and symptom classification. The list of classification performance with Microsoft speech recognition models is provided in **Supplementary Table 1**.

SHAP importance of models

Table 5 shows the top 5 medicine or symptom related nouns, verbs, adjectives, adverbs, and numbers for each model. The output of SHAP comes out as Korean morpheme stems, but we expressed it in words with appropriate endings. These are roughly the most important words of each classification model in each classification task, in descending order. BERT-KTAS mainly considered "*hurt like it bursts*," "*onset*," "*hands are numb*," "*symptom*," and "*time*" as important features for KTAS classification and considered "*cold*," "*hurt like it breaks*," "*stomach*," "*respiration*," and "*very*" for symptoms classification. Although BERT can capture the context, the vocabulary mainly used is less related to the actual KTAS and Symptom. The list of important words for ML was different from BERT-KTAS's result. When ML performed KTAS classification, "*stress*," "*chest*," "*pain score point*," and "*allergy*" were important. In the symptom classification, words directly related to symptom classes such as "*fever*," "*breath*," "*head*," "*chest*," and "*stomach*" were important for ML.

Table 5. Ranking of important words of classification models

Model	BERT-KTAS	SVM	KNN	RF
KTAS	터질 것 같다 (hurt like it bursts)	스트레스 (stress)	스트레스 (stress)	앓다 (suffer)
	발생 (onset)	가슴 (chest)	가슴 (chest)	2 (pain score 2 point)
	손이 저리다 (hands are numb)	어깨 (shoulder)	어깨 (shoulder)	내원 (visit to hospital)
	증상 (symptom)	목 (neck)	어지럼증 (dizziness)	4 (pain score 4 point)
	시간 (time)	4 (pain score 4 point)	토하다 (vomit)	알레르기 (allergy)
Symptom	춥다 (cold)	열이 나다 (have a fever)	열이 나다 (have a fever)	머리 (head)
	깨질 것 같다 (hurt like it breaks)	숨 (breath)	가슴 (chest)	배 (stomach)
	배 (stomach)	머리 (head)	숨이차다 (breathless)	통증 (pain)
	호흡 (respiration)	가슴 (chest)	숨 (breath)	가슴 (chest)
	너무 (very)	가래 (phlegm)	열 (fever)	기침 (cough)

KTAS = Korean Triage and Acuity System, SVM = support vector machine, KNN = k-nearest neighbors, RF = random forest, BERT = Bidirectional Encoder Representations from Transformers, SHAP = Shapley Additive exPlanations.
[a]The words are in descending order of the mean of the absolute values of the SHAP; [b]Each word is represented in Korean, and the translated vocabulary is represented in parentheses.

DISCUSSION

This preliminary study focused on the automatic KTAS classification using speech recognition, NLP, and artificial intelligence. Recent studies of the ML-based triage system showed high performance with text and numerical data in EHR.[15,16] Compared to these studies, we used voice data to triage emergency patients. We collected voice data from triage simulation and evaluated transfer-learned speech recognition models. According to a study, the error rate of the commercial Korean speech recognition models ranges from 16.29% to 61.43%.[39] We obtained relatively low CER using transfer-learning. We developed a deep learning-based BERT-KTAS to classify the KTAS and main symptoms. This model automatically extracts useful information for the KTAS and symptom classification from input texts. Over auto-transcribed documents, BERT-KTAS achieved the AUROC 0.82 for KTAS classification. For the comparison with BERT-KTAS, we also developed ML-based classifiers with maximal χ^2 statistics and BM25. With these processes, useful information for classification could be extracted before ML training. All the ML-based classifiers achieved AUROC greater than 0.86 with the auto-transcribed documents. In general, the performance with AUROC greater than 0.8 is evaluated as excellent discrimination.[40]

Deep learning is a neural network that includes multiple hidden layers and has the capacity to learn a complex pattern through the hierarchical nonlinear operation of sequential layers. However, the high complexity of the model tends to make itself overfit the training data, resulting in poor generalization ability in the test data.[41] We improved the generalization performance of BERT-KTAS by applying the network dropout technique and a pre-trained model.[42] Nevertheless, the prediction performances of BERT-KTAS were slightly lower than that of ML models on the auto transcribed data. The BERT pre-trained corpus contained 4.7 billion morphemes extracted from newspaper articles and encyclopedias. However, the corpus analyzed in the dialogue files of this study is from medical conversations between the patients and the clinicians, which is a specific and different domain from general articles and encyclopedias, and the vocabulary dictionary of BERT focused on general vocabularies rather than medical vocabularies. Therefore, the advantages of the pre-trained model were not fully exerted in KTAS prediction. On the other hand, ML models generally work better than deep learning on small datasets, with the help of curated features.[43] Besides, the proposed ML models established word selection criteria directly from the prepared dataset through pre-processing. Therefore, the ML models were able to achieve better classification performance than BERT-KTAS in our study with relatively small datasets. Although BERT-KTAS achieved slightly lower AUROC than ML-based classifiers, BERT-KTAS does not require pre-processing necessary for ML-based classifiers because BERT has an embedding process built into the model. We expect that the generalization performance of BERT-KTAS will surpass ML in future large-scale problems with more classes and datasets.

The purpose of SHAP was to measure the contribution of word to KTAS or symptom classification by calculating the effect of each word on the model's prediction probability. The mathematical characteristics of SHAP (local accuracy, missingness, and consistency) support that this is a fair way to interpret the classification models using feature importance.[35] According to the SHAP results, for the symptom classification, the vocabulary meaning the patient's symptom, such as "*fever*," "*breath*," "*head*," "*chest*," and "*stomach*," were the important vocabulary. These words are directly related to the primary symptom outcomes of our study. However, the important vocabularies for the KTAS classification were different from the factors considered in actual human-aided KTAS. As mentioned above, since dictionary and pre-trained BERT were not targeted for medical domains, many medical vocabularies were split into sub-word tokens and could not be appropriately used for KTAS classification.[44]

There are several limitations to our study. First, although we have improved the performance of the speech recognition models through transfer learning, the CER of actual patients' voices could be higher. Therefore, future studies should collect more voice data and reduce the CER of speech recognition models through transfer learning. Second, our data contained only six main symptoms categories and three KTAS levels, so it is still insufficient to introduce this system to actual emergency departments. As our small-scale study was limited to simulated data as the first study, in the future, we will study our classification process to a large-scale dataset with actual patients and diverse KTAS levels and main symptoms to show the practicality of our classification process. We will also collect side information of patients (e.g., postures or moods) and analyze how the accuracies of classifiers can be increased by using these data as additional input features. Lastly, many factors considered in human-aided KTAS, such as patient first looks, respiration status, hemodynamic status, neurological status, were not identified as important words for classification in the SHAP results. To further improve the classification performance and model interpretation, studies on pre-processing technology to extract the actual KTAS related factors is needed.

In conclusion, this preliminary study showed the potential of developing an automatic classification system that directly classifies the KTAS level and symptoms from the conversation between patients and clinicians. This concept study will be helpful to reduce clinicians' workload at the ED. In a future study, we will first extract the factors related to KTAS from dialogues by developing a deep learning algorithm and then conduct classification based on these extracted factors.

SUPPLEMENTARY MATERIALS

Supplementary Method 1

Training process of Bidirectional Encoder Representations from Transformers (BERT)-Korean Triage and Acuity Scale (KTAS)

Click here to view

Supplementary Method 2

Training process of benchmark machine learning (ML)

Click here to view

Supplementary Table 1

Model performance with Microsoft's auto-transcribed test dataset

Click here to view

REFERENCES

1. National Emergency Medical Center (Korea). 2019 Annual report of Korean emergency medicine. https://www.e-gen.or.kr/nemc/statistics_annual_report.do. Updated 2020. Accessed January 13, 2021.
2. Bernstein SL, Aronsky D, Duseja R, Epstein S, Handel D, Hwang U, et al. The effect of emergency department crowding on clinically oriented outcomes. *Acad Emerg Med* 2009;16(1):1-10.
   PUBMED | CROSSREF
3. Stang AS, Crotts J, Johnson DW, Hartling L, Guttmann A. Crowding measures associated with the quality of emergency department care: a systematic review. *Acad Emerg Med* 2015;22(6):643-56.
   PUBMED | CROSSREF
4. Park J, Lim T. Korean Triage and Acuity Scale (KTAS). *J Korean Soc Emerg Med* 2017;28(6):547-51.
5. Ryu JH, Min MK, Lee DS, Yeom SR, Lee SH, Wang IJ, et al. Changes in relative importance of the 5-level triage system, Korean Triage and Acuity Scale, for the disposition of emergency patients induced by forced reduction in its level number: a multi-center registry based retrospective cohort study. *J Korean Med Sci* 2019;34(14):e114.
   PUBMED | CROSSREF
6. Kwon H, Kim YJ, Jo YH, Lee JH, Lee JH, Kim J, et al. The Korean Triage and Acuity Scale: associations with admission, disposition, mortality and length of stay in the emergency department. *Int J Qual Health Care* 2019;31(6):449-55.
   PUBMED | CROSSREF
7. Park JB, Lee J, Kim YJ, Lee JH, Lim TH. Reliability of Korean Triage and Acuity Scale: interrater agreement between two experienced nurses by real-time triage and analysis of influencing factors to disagreement of triage levels. *J Korean Med Sci* 2019;34(28):e189.
   PUBMED | CROSSREF
8. Moon SH, Shim JL, Park KS, Park CS. Triage accuracy and causes of mistriage using the Korean Triage and Acuity Scale. *PLoS One* 2019;14(9):e0216972.
   PUBMED | CROSSREF
9. Kim JH, Kim JW, Kim SY, Hong DY, Park SO, Baek KJ, et al. Validation of the Korean Triage and Acuity Scale compare to triage by emergency severity index for emergency adult patient: preliminary study in a tertiary hospital emergency medical center. *J Korean Soc Emerg Med* 2016;27(5):436-41.
10. Kim JH, Hong JS, Park HJ. Prospects of deep learning for medical imaging. *Precis Future Med* 2018;2(2):37-52.
    CROSSREF
11. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One* 2018;13(7):e0201016.
    PUBMED | CROSSREF

12. Raita Y, Goto T, Faridi MK, Brown DF, Camargo CA Jr, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* 2019;23(1):64.
PUBMED | CROSSREF

13. 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.
PUBMED | CROSSREF

14. Patel SJ, Chamberlain DB, Chamberlain JM. A machine learning approach to predicting need for hospitalization for pediatric asthma exacerbation at the time of emergency department triage. *Acad Emerg Med* 2018;25(12):1463-70.
PUBMED | CROSSREF

15. Choi SW, Ko T, Hong KJ, Kim KH. Machine learning-based prediction of Korean Triage and Acuity Scale level in emergency department patients. *Healthc Inform Res* 2019;25(4):305-12.
PUBMED | CROSSREF

16. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One* 2017;12(4):e0174708.
PUBMED | CROSSREF

17. Lee KS, Song IS, Kim ES, Ahn KH. Determinants of spontaneous preterm labor and birth including gastroesophageal reflux disease and periodontitis. *J Korean Med Sci* 2020;35(14):e105.
PUBMED | CROSSREF

18. Lee KS, Ahn KH. Artificial neural network analysis of spontaneous preterm labor and birth and its major determinants. *J Korean Med Sci* 2019;34(16):e128.
PUBMED | CROSSREF

19. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res* 2019;20(177):1-81.

20. Alabi RO, Elmusrati M, Sawazaki-Calone I, Kowalski LP, Haglund C, Coletta RD, et al. Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. *Int J Med Inform* 2020;136:104068.
PUBMED | CROSSREF

21. Shew M, New J, Bur AM. Machine learning to predict delays in adjuvant radiation following surgery for head and neck cancer. *Otolaryngol Head Neck Surg* 2019;160(6):1058-64.
PUBMED | CROSSREF

22. Karadaghy OA, Shew M, New J, Bur AM. Development and assessment of a machine learning model to help predict survival among patients with oral squamous cell carcinoma. *JAMA Otolaryngol Head Neck Surg* 2019;145(12):1115-20.
PUBMED | CROSSREF

23. Rajpurkar P, Yang J, Dass N, Vale V, Keller AS, Irvin J, et al. Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: a prespecified secondary analysis of a randomized clinical trial. *JAMA Netw Open* 2020;3(6):e206653.
PUBMED | CROSSREF

24. Lee KS. Research about chief complaint and principal diagnosis of patients who visited the university hospital emergency room. *J Digit Converg* 2012;10(10):347-52.
CROSSREF

25. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 1998;10(7):1895-923.
PUBMED | CROSSREF

26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019, 4171-86.

27. Vaswani A, Noam S, Niki P, Uszkoreit J, Jone L, Gomez AN, et al. Ateention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing System*. 2017, 6000-10.

28. Moh'd A, Mesleh A. Chi square feature extraction based svms arabic language text categorization system. *J Comput Sci* 2007;3(6):430-5.
CROSSREF

29. Whissell JS, Clarke CL. Improving document clustering using Okapi BM25 feature weighting. *Inf Retr Boston* 2011;14(5):466-87.
CROSSREF

30. Lee YH, Lee SB. A research on enhancement of text categorization performance by using okapi BM25 word weight method. *J Korea Acad Ind Coop Soc* 2010;11(12):5089-96.
CROSSREF

31. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol* 2011;2(1):37-63.

32. Cortes C, Mohri M. Confidence intervals for the area under the ROC curve. In: Saul L, Weiss Y, Bottou L, editors. *Advances in Neural Information Processing Systems 17 (NIPS 2004)*. Cambridge, MA, USA: MIT Press; 2005, 305-12.

33. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44(3):837-45.
PUBMED | CROSSREF

34. Shapley LS. *A Value for n-Person Games. Contributions to the Theory of Games*. Princeton, NJ, USA: Princeton University Press; 1953.

35. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Cambridge, MA, USA: MIT Press; 2017, 4765-74.

36. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng* 2011;13(2):22-30.
CROSSREF

37. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825-30.

38. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. *arXiv*. Forthcoming 2019. https://arxiv.org/abs/1912.01703

39. Yoo HJ, Seo S, Im SW, Gim GY. The performance evaluation of continuous speech recognition based on Korean phonological rules of cloud-based speech recognition open API. *Int J Networked Distrib Comput* 2021;9(1):10-8.
CROSSREF

40. Hosmer JR, David W, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 2nd ed. Hoboken, NJ, USA: John Wiley & Sons; 2013.

41. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput* 2006;18(7):1527-54.
PUBMED | CROSSREF

42. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;15(1):1929-58.

43. Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P, Nasrin MS, et al. A state-of-the-art survey on deep learning theory and architectures. *Electronics (Basel)* 2019;8(3):292.
CROSSREF

44. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2016, 1715-25.