IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 August 2024, accepted 19 September 2024, date of publication 25 September 2024, date of current version 8 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3467251

RESEARCH ARTICLE

BioBridge: Unified Bio-Embedding With Bridging Modality in Code-Switched EMR

JANGYEONG JEON[1], (Member, IEEE), SANGYEON CHO[1], (Graduate Student Member, IEEE), DONGJOON LEE[1], CHANGHEE LEE[2], (Member, IEEE), AND JUNYEONG KIM[1], (Member, IEEE)

[1]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea
[2]Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea

Corresponding author: Junyeong Kim (junyeongkim@cau.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) (Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics) under Grant 2022-0-00184, and in part by IITP Grant funded by the Korea Government (MSIT) (Artificial Intelligence Graduate School Program, Chung-Ang University) under Grant 2021-0-01341.

ABSTRACT Pediatric Emergency Department (PED) overcrowding presents a significant global challenge, prompting the need for efficient solutions. This paper introduces the BioBridge framework, a novel approach that applies Natural Language Processing (NLP) to Electronic Medical Records (EMRs) in written free-text form to enhance decision-making in PED. In non-English speaking countries, such as South Korea, EMR data is often written in a Code-Switching(CS) format that mixes the native language with English, with most code-switched English words having clinical significance. The BioBridge framework consists of two core modules: ''bridging modality in context'' and ''unified bio-embedding.'' The ''bridging modality in context'' module improves the contextual understanding of bilingual and code-switched EMRs. In the ''unified bio-embedding'' module, the knowledge of the model trained in the medical domain is injected into the encoder-based model to bridge the gap between the medical and general domains. Experimental results demonstrate that the proposed BioBridge significantly performance traditional machine learning and pre-trained encoder-based models on several metrics, including F1 score, area under the receiver operating characteristic curve (AUROC), area under the precision-recall Curve (AUPRC), and Brier score. Specifically, BioBridge-XLM achieved enhancements of 0.85% in F1 score, 0.75% in AUROC, and 0.76% in AUPRC, along with a notable 3.04% decrease in the Brier score, demonstrating marked improvements in accuracy, reliability, and prediction calibration over the baseline XLM model. The source code will be made publicly available at https://github.com/jjy961228/BioBridge.

INDEX TERMS Natural language processing, code-switching, electronic medical record, emergency department, pediatric emergency department.

I. INTRODUCTION

Overcrowding in Emergency Departments (EDs), including Pediatric Emergency Departments (PEDs), is an urgent problem that needs to be handled globally and immediately, including in South Korea. It is due to lead inefficient use of ED resources and the overworked healthcare professionals [1], [2], [3], [4]. Providing timely and accurate treatment is essential, especially for children who are more susceptible to infection than adults [5]. However, since children tend to visit the ED frequently, this further aggravates the overcrowding problem in PED [6].

As medical institutions worldwide adopt EMRs, vast amounts of healthcare data are stored digitally [7], [8], [9], [10]. EMR contains information about a patient's Present Illness (PI), past medical history, and textual data such as treatment and disposition information [7], [9], [10]. By using Natural Language Processing (NLP) techniques to EMR data,

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

clinicians can receive real-time decision support and increase the efficiency of emergency medical services by optimizing the assessment of patient severity and urgency in the ED [11], [12], [13]. Traditional approaches to applying NLP to EMR data have focused on Machine Learning (ML) [14], [15], [16] and Word2Vec-based approaches to extract context-free word embeddings [17], [18], [19]. Recently, the focus has shifted to improving predictive modeling performance by training EMRs on pre-trained encoder-based models [20], [21], [22] such as BERT [23]. Despite the potential benefits of EMRs, their practical application in healthcare decision support systems has been limited by challenges in applying NLP to extract key information, which still needs to be overcome [24], [25], [26].

First, EMRs consist of unstructured, free-text clinical notes written by physicians during patient examinations, often using varied wording to describe the same disease [27], [28], [29]. For example, medical symbols such as ''Cough/Sputum/Rhinorrhea'' are often used as ''C/S/R'' in EMRs, which are primarily used to record a patient's respiratory disease [30], [31]. Extensive text preprocessing is required to apply natural language processing techniques to these EMRs, and NLP toolkits such as the Unified Medical Language System (UMLS) [32], [33] have been widely adopted [32], [34], [35]. However, these standardized preprocessing methods incur high costs due to continuous resource maintenance and are often unavailable in languages other than English, including Korean. To mitigate these challenges, we have identified commonly used medical symbols and abbreviations in EMRs, proposed a straightforward preprocessing method to decode them, and integrated this preprocessed text with an ML-based [36], [37], [38], [39] and pre-trained encoder-based model [23], [40], [41], [42], [43].

Second, EMR data used in non-English speaking countries are recorded in Code-Switching (CS) form, where English and the native language are mixed [44], [45]. CS refers to using two languages in the sentence and is prevalent not only in EMRs but also in multicultural countries, social media, and online platforms [46], [47], [48]. Unfortunately, there is a lack of research on CS in non-English speaking countries, including languages such as Korean, Japanese, Chinese, and Arabic [48], [49], [50], [51]. In particular, few studies have explored the application of encoder-based models to code-switched EMR datasets from non-English-speaking countries [44], [45], [52], [53]. The EMR dataset used in our study is also code-switched to Korean-English. To bridge this research gap, we assume that language-specific modalities (Korean, English) exist in each CS sentence and propose a novel method of ''bridging modality in the context'' of learning each modality effectively by a pre-trained encoder. This approach is motivated by research in multimodal learning, where different inputs (visual, text, and audio input) are defined as modalities in a transformer and utilized for learning [54]. In the following Section II-B, we describe in detail the research related to multimodal learning.

Third, most of the terms that are code-switched have clinical significance. Using pre-trained encoder base models in general domains in NLP and applying them to clinical notes is challenging, and many attempts have been made to solve it [20], [21], [55]. This is due to most encoder-based models are not pre-trained with domain-specific data but are pre-trained using general-domain data such as English Wikipedia, BooksCorpus [56], and CommonCrawl [57]. Moreover, Korean is a very low-resourced language compared to English [41], [58], and more research needs to be done. Pre-trained encoder-based models on general domains have strengths in understanding general contexts but tend to perform poorly in medical domains [20], [21], [58]. To mitigate this gap, we propose a ''unified bio-embedding'' approach that effectively fine-tunes pre-trained encoder-based models [23], [40], [41], [42], [43], bridging the gap between general and medical domains.

Our goal is to alleviate overcrowding in PEDs. To achieve this, we propose to fine-tune the pre-trained encoder-based model to classify emergency and non-emergency cases [59] using EMR for the three to overcome the abovementioned challenges. We introduce ''unified-bio embedding with bridging modality in code-switched EMR (BioBridge),'' a framework comprising two training modules: (1) bridging modality in context, and (2) unified bio-embedding. We established Machine Learning (ML) and Transformer encoder-based models as baselines to validate BioBridge. The experimental results show that BioBridge has a surprisingly stable performance in predicting emergency cases. This is expected to alleviate PED overcrowding and is the first attempt to our knowledge. Our contributions are summarized as follows:

- Using a straightforward preprocessing method, we empirically identify and decode frequently used medical symbols and abbreviations in EMRs.
- We propose a method for effectively learning code-switched sentences.
- We propose a method to bridge the gap between general and medical domains, enhancing fine-tuning applicability.

II. RELATED WORK

A. NATURAL LANGUAGE PROCESSING FOR ELECTRONIC MEDICAL RECORD

The increasing availability of Electronic Medical Records (EMRs) has led to extensive research into utilizing EMRs in Natural Language Processing (NLP) [14], [21], [60], [61]. Traditional methods for applying EMR data to NLP have evolved from Machine Learning (ML) [36], [37], [38], [39] and Word2Vec-based approaches [17]. Word2Vec-based approaches have proven effective in identifying relationships between medical terms and symptoms by facilitating word encoding within a vector space where semantically similar words are closely aligned [17], [18]. However, Word2Vec is limited by its inability to account for context, and methodologies such as Sent2Vec [62] have attempted to

overcome this limitation by extending the Continuous Bag of Words (CBOW) approach of Word2Vec to increase the window size to whole sentences. In particular, BioSent2Vec [63], trained on over 30 million documents from academic articles in PubMed and clinical notes in the MIMMIC-III Clinical Database, has been able to generate useful word embeddings through biomedical text mining. In this paper, we use the pre-trained BioSent2Vec as a word-level feature extractor to derive English medical word embeddings.

Encoder-based models such as BERT [23] are gaining prominence in modern NLP due to their capacity to comprehend bidirectional context. Specifically, such as BioBERT [20] have been effectively fine-tuned on English EMR data for patient health prediction [21], [22]. However, the effectiveness of these encoder-based models is often constrained when applied to languages with limited resources, such as Korean [41], [45], [58], [64]. This limitation stems from the fact that English, an alphabet-based language, consists of only 26 characters and consequently faces fewer out-of-vocabulary (OOV) problems [41]. In contrast, Korean is a syllable-based language comprising 11,172 characters [41]. Korean medical texts, in particular, pose analytical challenges due to complex expressions and specialized jargon [44], [45], [58]. EMRs in Korea frequently involve CS between Korean and English and are typically maintained in a free-text format [27], [28], [29]. To address these challenges, approaches have been proposed to fine-tune multilingual-specific encoder base models [23], [42], [65] such as multilingual BERT (mBERT) [23] and Korean-specific encoder base models [40], [41] such as KR-BERT [41] with EMR data. Notable studies include one focused on extracting clinically significant information from free-text clinical notes [66], another on effectively triaging emergency patients [67], and another on de-identifying radiology reports [68]. The EMRs used for training in these studies cannot be a universal method for learning EMRs because the data cannot be released due to they contain personal patient information [66], [68], and [68]. Moreover, they only propose fine-tuning with EMR data without special techniques. In addition, previous studies have reported results using only one type of model, either multilingual-only or Korean-only encoder base models, which may not fully capture the nuances of EMRs documented in both Korean and English.

In this paper, we conduct extensive experiments with both model types to address these limitations. We also introduce an approach to effectively learn from universally code-switched EMRs without disclosing patient data.

B. MULTIMODAL LEARNING

Initial research in multimodal learning primarily focused on integrating text and images [69], [70]. Subsequently, video and text representations could provide complementary information, and research on combining video and text representations became a significant focus. For example, VideoBERT [71], [72] pursued richer information representation by combining video clips and text. Recently, advancements in Visual Question Answering (VQA) tasks in multimodal learning have centered on enhancing the capabilities of models to respond to complex questions across different modality types [54], [73], [74], [75]. These multimodal learning methods share a common goal of solving vision-and-language transfer tasks by separating inputs from different modalities (visual, text, audio) and integrating them into a transformer-based model.

Motivated by the methodology in multimodal learning, this paper hypothesizes the presence of distinct modalities for each language in also code-switched sentences involving two languages.

A notable example is the proposed method presented in [54], which introduced a straightforward yet effective method for tackling the Audio-Visual Scene-Aware Dialog (AVSD) task. The AVSD task is to analyze a video clip to identify information about its events, characters, and environment and to provide appropriate answers when the user asks questions about them. The AVSD task is given video, audio, caption, and dialog history as input. The study in [54] proposed an approach to concatenate video-audio, caption, and dialog history into one long sequence feature and add segment tokens for "[video]," "[caption]," "[user1]", and "[user2]" to distinguish each integrated modality to understand each modality better in AVSD task.

Inspired by the methodology in [54], this paper hypothesizes the presence of distinct modalities for each language in also code-switched sentences involving two languages.

III. PROPOSED METHOD

A. PROBLEM FORMULATION

Let $X = \{x_i\}_{i=0}^{N}$ denote PI, the text input contained in the EMR dataset, and let $X^{Eng} = \{x_i^{Eng}\}_{i=0}^{N}$ denote the English text input, $X^{Eng} \subset X$. The label $Y = \{y_i\}_{i=0}^{N}, y_i \in \{0, 1\}$ denote Non-emergency cases or Emergency cases.

Tokenized text input for X is defined as $X^{tok} = \{x_i^{tok}\}_{i=0}^{N}$. For instance, if a tokenized input x_i^{tok} consists of Korean followed by English, in that order, it is denoted as:

$$x_i^{tok} = \{[\text{CLS}], [\text{tokens}]^{kor}, [\text{tokens}]^{eng}, [\text{SEP}]\} \quad (1)$$

where, "[CLS]" signifies the first token of x_i^{tok} and "[SEP]" denotes the last token. $[\text{tokens}]^{kor}$ denotes the tokens for the tokenized Korean sentence within x_i^{tok}, and $[\text{tokens}]^{eng}$ denotes the tokens for the tokenized English sentence within x_i^{tok}.

We denote the batch size as b, the total number of tokens within a batch as n, and the number of English words as m. Additionally, we detail three components input to the pre-trained encoder $\mathcal{M}$.

Firstly, the token embedding layer of $\mathcal{M}$, denoted as f_θ^T, is defined as follows:

$$f_\theta^T(\{x_i^{tok}\}_{i=0}^{b}) \in \mathbb{R}^{n \times h_\mathcal{M}} \quad (2)$$

where, $h_\mathcal{M}$ is the hidden dimension of $\mathcal{M}$.

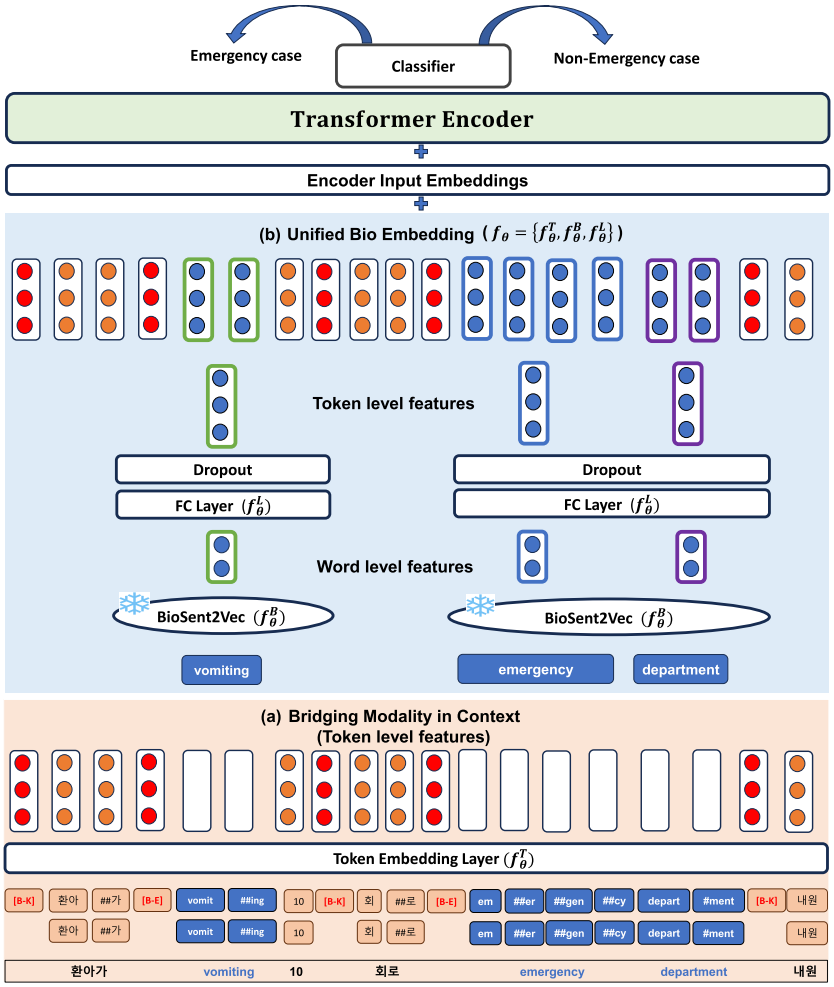


FIGURE 1. Overview of BioBridge. (a) In the bridging modality in context module, tokenized text input x_i^{tok} is processed to reconstruct $x_i^{bri} = \{[\text{CLS}], [\text{B-K}], [\text{tokens}]^{kor}, [\text{B-E}], [\text{tokens}]^{eng}, [\text{SEP}]\}$, where modalities for each language are separated using segment tokens. Where "[B-K]" and "[B-E]" are segment tokens that distinctly identify Korean and English. In the unified bio-embedding module, English tokens within x_i^{bri} are restructured at the word level to $\{x_i^{Eng}\}_{i=0}^b$ which then serve as input for $f_\theta^B(\{x_i^{Eng}\}_{i=0}^b) \in \mathbb{R}^{m \times h_B}$ to extract medical features. These features are subsequently input into f_θ^L, mapping them to the dimension space $h_\mathcal{M}$ of the pre-trained encoder $\mathcal{M}$.

Secondly, the medical feature extractor, BioSent2Vec f_θ^B [63], is defined as:

$$f_\theta^B(\{x_i^{Eng}\}_{i=0}^b) \in \mathbb{R}^{m \times h_B} \tag{3}$$

where, h_B is the hidden dimension of f_θ^B, and θ is a fixed parameter.

Thirdly, the Fully Connected (FC) Layer f_θ^L, designed to map dimensions, is defined as:

$$f_\theta^L : \mathbb{R}^{m \times h_B} \rightarrow \mathbb{R}^{m \times h_\mathcal{M}} \tag{4}$$

Challenges: In practice, applying natural language processing (NLP) to EMRs constructed in non-English-speaking regions presents two primary challenges: (1) Code-switched texts between the native language and English are complex and difficult to interpret. Specifically, it is particularly challenging to apply NLP to texts that are code-switched between English and resource-scarce languages, such as Korean. (2) Due to CS terms predominantly consisting of medical terminology, directly applying models pre-trained in general domains to EMR data proves difficult.

B. BIOBRIDGE FRAMEWORK

To address the challenges described above, we introduce a two-stage training approach for the BioBridge:

- **Bridging modality in context module**: We add a segment token before each modality within the text to enable the pre-trained encoder $\mathcal{M}$ [23], [40], [41], [42], [43] to effectively distinguish between the modalities (Korean, English) in the given text input X.
- **Unified bio-embedding module**: We extract features for English medical words using f_θ^B [63] and unify

these features into pre-trained encoder $\mathcal{M}$, enhancing the encoder's applicability to the medical domain.

We implement this methodology by training the pre-trained encoder $\mathcal{M}$ using the two training modules. Fig.1 shows an overview of the proposed BioBridge framework.

1) BRIDGING MODALITY IN CONTEXT

The pre-trained encoder $\mathcal{M}$ must learn to distinguish between Korean and English modalities in the text input X. To facilitate this, we prepend a segment token to each modality in the tokenized input X^{tok}. For instance, by adding segment tokens to x_i^{tok}, can be denoted as:

$$x_i^{bri} = \{[\text{CLS}], [\text{B-K}], [\text{tokens}]^{kor}, [\text{B-E}], [\text{tokens}]^{eng}, [\text{SEP}]\} \tag{5}$$

where "[B-K]" and "[B-E]" are segment tokens that differentiate between Korean and English, respectively. Additionally, with the inclusion of segment tokens, the dimension of token embedding layer f_θ^T is redefined as follows:

$$f_\theta^T(\{x_i^{bri}\}_{i=0}^b) \in \mathbb{R}^{(n+s)\times h_\mathcal{M}} \tag{6}$$

where s denotes the number of added segment tokens. The related framework is shown in Fig.1 (a).

2) UNIFIED BIO-EMBEDDING

In the unified bio-embedding module, we utilize f_θ^B to extract English medical features and integrate them into the pre-trained encoder $\mathcal{M}$. The tokenized inputs $\{x_i^{bri}\}_{i=0}^b$ constructed during the bridging modality in context module are segmented at the token level, while f_θ^B is optimized for word-level learning. To address this, we reassemble tokenized English tokens at the token level back into word-level constructs to form $\{x_i^{Eng}\}_{i=0}^b$. For example, tokens separated by the tokenizer, such as "vomit" and "##ing," are reconstructed into "vomiting" as illustrated in Fig.1 (b).

We then use the word-level reconstructed $\{x_i^{Eng}\}_{i=0}^b$ as input to f_θ^B, as specified in equation (3), to extract medical features. To integrate these extracted medical features into $\mathcal{M}$, the output dimension of f_θ^B must be mapped $\mathbb{R}^{h_B} \rightarrow \mathbb{R}^{h_\mathcal{W}}$ This mapping is facilitated using f_θ^L, as described in equation (4).

The output from f_θ^L produces token-level medical features, which are then integrated into the unified bio-embedding f_θ, which can be denoted as:

$$f_\theta(f_\theta^T, f_\theta^B, f_\theta^L) \in \mathbb{R}^{m\times h_\mathcal{M}} \tag{7}$$

where, $h_\mathcal{M}$ is the hidden dimension of $\mathcal{M}$. Ultimately, we integrate unified bio-embedding into the pre-trained encoder $\mathcal{M}$.

IV. EXPERIMENTS

In our experiments, we evaluate BioBridge's performance using the proposed EMR dataset. Additionally, all our experiments are designed for a classification task in which the predictive model, such as machine learning-based [36], [36], [37], [38], [39] and pre-trained encoder-based models [23], [40], [41], [42], [43], distinguishes between emergency and non-emergency cases [59].

Medical Symbols	Decoding Medical Symbols
C/S/R	Cough/Sputum/Rhinorrhea
N/V/D/C	Nausea/Vomiting/Diarrhea/Constipation
C.C or C/C	Chief Complaint
P.I or P/I	Present Illness
(+)	Positive
(-)	Negative
Medical Abbreviations	**Decoding Medical Abbreviations**
Hx	History
Fhx	Family History
Sx	Symptom
Dx	Diagnosis
Tx	Treatment
BT	Body Temperature
BW	Body Weight
LT	Left
RT	Right

FIGURE 2. Preprocessing example of present illness (PI) texts in an electronic medical record (EMR).

A. DATASET

The benchmark dataset used in this paper comprises electronic medical records (EMRs) from the Department of Pediatrics at the University Hospital of South Korea. The study population includes patients under 18 years of age who visited the Department of Pediatrics between January 1, 2012, and December 31, 2021. The EMR dataset includes Present Illness (PI) notes [76], urine and blood test results, and records of emergency and non-emergency cases [59]. Present Illness (PI) is an unstructured free-text clinical note written by a nurse or doctor about the main symptoms observed during the patient's visit to the PED. Additionally, Emergency and Non-emergency cases are data annotated according to whether the patient's visit to the PED was an emergency requiring treatment. Detailed dataset statistics are recorded in Table 1, and Section IV-B describes the preprocessing of the EMR data. SectionIV-B provides a detailed description of the emergency and non-emergency cases.

B. PREPROCESSING OF ELECTRONIC MEDICAL RECORD

We constructed an electronic medical record (EMR) dataset of patients who visited Pediatric Emergency Departments (PEDs) by excluding patients with no disposition record and those with missing information. Present Illness (PI) notes contain various medical symbols and abbreviations. For example, in Fig.2, the medical symbol "C/S/R" stands for "Cough/Sputum/Rhinorrhea," and the abbreviation "BT" stands for "Body Temperature". We decoded these medical symbols and abbreviations, as shown in Fig.2. Additionally,

TABLE 1. Statistical summary of character count in present illness (PI) texts of the electronic medical record (EMR). This table displays the character counts in the PI texts of EMR data, with 8,022,908 Korean and 14,717,723 english characters, indicating that most PI texts are code-switched between Korean and english.

-	Total number of data	Total number of characters	Number of Korean characters	Number of English characters	Number of Numerical characters	Number of special characters
Train	56,165	16,957,725	5,126,952	9,396,609	671,550	1,354,450
Dev	14,042	4,260,799	1,286,444	2,362,094	166,270	339,102
Test	17,552	5,333,652	1,609,512	2,959,020	211,526	426,016
total	**87,759**	**26,552,176**	**8,022,908**	**14,717,723**	**1,049,346**	**2,119,568**

we preprocessed the PI notes by systematically separating Korean, English, numeric values, and special characters when they are followed by a space and introducing a space. We split the preprocessed dataset into training, development, and test sets with ratios of 0.64, 0.16, and 0.20, respectively. In this paper, we integrated PI texts with simple preprocessing into a predictive model.

C. EMERGENCY AND NON-EMERGENCY CASES

The term ''Emergency'' refers to an abrupt and often unforeseen occurrence necessitating prompt intervention to reduce potential negative outcomes [59]. Cases classified as emergencies encompass patients who received interventions such as blood tests, urinalysis, intravenous hydration, nebulizer treatments, or immediate drug administration in the Pediatric Emergency Department (PED), as well as those who required hospital admission. Conversely, ''Non-emergency'' describes conditions that do not demand urgent medical care, including immediate evaluation, diagnosis, or treatment. Non-emergency cases typically involve patients who were released from the PED without undergoing urgent tests or receiving immediate medication or those who departed with only prescription medication at discharge.

D. BASELINES

1) MACHINE LEARNING BASED MODELS

For the Machine Learning (ML) Based Model, we constructed baseline models using one statistical method (Logistic Regression [37]) and three tree-based ensemble methods (Gradient Boosting [38], XGBoost [36], and Random Forest [39]).

a: LOGISTIC REGRESSION

Logistic regression [37] is a statistical model that estimates probabilities using a logistic function widely used for binary classification tasks. It models the probability of a default class (such as an emergency case) based on one or more predictor variables (features).

b: GRADIENT BOOSTING

Gradient Boosting [38] is an ensemble technique that builds models sequentially by correcting the previous models' errors. It uses decision trees as weak learners and focuses on minimizing a loss function, such as mean squared error for regression or log loss for classification, by optimizing the errors of successive models.

c: XGBOOST

XGBoost (Advanced Gradient Boosting) [36] enhances traditional gradient boosting by incorporating advanced features such as regularization to prevent overfitting and a more sophisticated tree pruning and splitting approach.

d: RANDOM FOREST

Random Forest [39] is an ensemble model consisting of multiple regression trees. It combines several classification trees and trains each on a slightly different set of dataset instances, splitting nodes in each tree considering a limited number of variables. The final predictions of the random forest are made by averaging the predictions of each tree.

2) ENCODER BASED MODELS

For the encoder base model, we established baselines using both the Korean-specific and the Multilingual-specific encoder-based models.

a: KOBERT

KoBERT [40] uses the SentencePiece tokenizer and character-level tokenization to reflect and handle the complexity of the Korean language. It is trained on Korean Wikidata, which comprises 5 million sentences and 54 million words, with a vocabulary of 8,002 tokens.

b: KR-BERT

KR-BERT [41] utilizes the WordPiece tokenizer to capture Korean morphological complexities. Unlike KoBERT, KR-BERT includes both sub-character and character-level tokenization. For our experiments, we used the KR-BERT-Medium variant. KR-BERT-Medium is an extension of the training data used in the original KR-BERT by adding Korean Wikipedia text, news articles, and Korean online comment datasets. The dataset used for training is 12.37 GB, and the vocabulary size is 20,000.

c: mBERT

Multilingual BERT [23] utilizes Masked Language Modeling (MLM), BERT's pre-training objective, to train on Wikipedia articles across 104 languages. It uses a shared WordPiece

vocabulary comprising 110,000 words to facilitate multilingual processing. This approach enables mBERT to process text in multilingual languages effectively.

d: XLM

XLM [42] utilizes a cross-lingual pre-training method on a translated parallel corpus and utilizes Byte-Pair Encoding (BPE) for multilingual languages to facilitate cross-lingual transfer learning. This approach has enhanced the performance of mBERT.

e: XLM-R

XLM-R [43] enhanced the XLM framework by training on 2.5 TB of CommonCrawl data, a significant increase over the dataset used for XLM. This model leverages the robust training approach of RoBERTa [77], an optimized version of BERT, which includes dynamic masking and removes the next sentence prediction objective. It also supports 100 languages.

E. PERFORMANCE METRICS

Our experiments evaluate model performance using several key metrics commonly used in classification tasks: F1 score, Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and Brier score. To ensure a balanced evaluation, we set the F1 score threshold at 0.595 to align with the label frequency ratio of the label 1 ratio. The F1 score, the harmonic mean of precision and recall, is particularly useful in scenarios with unbalanced datasets. The AUROC measures the model's ability to distinguish between the classes across all thresholds utilizing the False Positive Rate (FPR) and True Positive Rate (TPR). A high AUROC score indicates a strong capacity of the model to differentiate between positive and negative classes, meaning effective model performance. AUPRC is particularly useful in cases of significant class imbalance, which is the typical dataset. Unlike AUROC, the Area under the precision-recall curve focuses on the performance of the positive class only, making it more sensitive to detecting positive instances. It plots precision and recall for different threshold values, and a higher AUPRC value indicates both high recall and high precision, representing good retrieval of positive cases without many false positives. The Brier Score is used to measure the accuracy of probabilistic predictions. It is a mean squared error score that quantifies the difference between predicted probabilities and the actual outcomes. A lower Brier score indicates better accuracy, with 0 representing a perfect model and 1 indicating the worst model performance. This metric is particularly useful for evaluating the reliability of a model's probability estimates in our classification task, where predicting the correct probabilities for each class can be more informative than simply predicting the class.

TABLE 2. Hyperparameter used for training encoder-based model.

Model. Param	Batch Size	Learning Rate
Korean specific encoder based model		
krBERT	64	2e-5
*BioBridge-krBERT	64	5e-5
koBERT	64	6e-5
*BioBridge-koBERT	64	3e-5
Multilingual specific encoder based model		
XLM	22	1e-5
*BioBridge-XLM	22	5e-6
$\text{mBERT}_{\text{uncased}}$	64	4e-5
*BioBridge-$\text{mBERT}_{\text{uncased}}$	64	1e-5
$\text{mBERT}_{\text{cased}}$	64	5e-5
*BioBridge-$\text{mBERT}_{\text{cased}}$	64	6e-5
$\text{XLM-R}_{\text{base}}$	64	5e-5
*BioBridge-$\text{XLM-R}_{\text{base}}$	64	3e-5

F. TRAINING DETAILS

In our experiments, we developed a model using the PyTorch [78] and Transformer [79] libraries and utilized a pair of NVIDIA RTX A6000 GPUs for training.

1) TRAINING DETAILS ON MACHINE LEARNING BASED MODEL

We used the Term Frequency-Inverse Document Frequency (TF-IDF) to train a Machine Learning (ML) based model. TF-IDF is determined by multiplying two metrics: the Term Frequency (TF) score, which reflects the frequency of a word within a document, and the Inverse Document Frequency (IDF) score, which gauges the word's rarity across the entire document corpus. A higher TF-IDF score signifies greater importance of the word for prediction. We applied TF-IDF by treating each patient's Present Illnes (PI) as a single document.

2) TRAINING DETAILS ON ENCODER BASED MODEL

In our experiments, we initialized the pre-trained encoder using the checkpoints of KR-BERT, KoBERT, mBERT, XLM, and XLM-R. We also used the "[CLS]" token as the final sentence embedding. During training, we saved the checkpoint with the highest f1 score on the development set and evaluated its performance on the test set. Considering the extensive length of most Present Illness (PI) descriptions, we used the maximum sequence length of 512 across all experiments. Also, we used 5 epochs and gridsearch of learning rate $\in \{2e-6, 3e-6, 5e-6, 1e-5, 2e-5, 3e-5, 4e-5, 5e-5, 6e-5\}$ for all experiments. The optimal hyperparameters for the test set are shown in Table 2. Furthermore, all parameters of the medical feature extractor f_θ^B in BioBridge are fixed during training.

TABLE 3. Performance comparison of machine learning-based and encoder-based models in classifying emergency cases. *: results from the proposed BioBridge experiments. The best result within each row block is shown in bold, and The best results from the Machine learning-based model are underlined.

Model	F1 Score ↑	AUROC ↑	AUPRC ↑	Brier Score ↓
Machine learning based model				
Logistic Regression	66.60	69.58	76.24	21.43
XGBoost	66.46	69.77	76.43	21.38
Gradient Boosting	66.90	70.15	76.87	21.25
Random Forest	66.13	67.80	74.01	22.39
Korean specific encoder based model				
KR-BERT	78.30	78.45	83.20	19.87
*BioBridge-KR-BERT	**79.40**	**80.97**	**85.36**	**18.92**
koBERT	78.28	80.61	85.20	18.58
*BioBridge-koBERT	**78.95**	**80.92**	**85.27**	**18.12**
Multilingual specific encoder based model				
XLM	78.84	79.89	84.19	21.93
*BioBridge-XLM	**79.69**	**80.64**	**84.95**	**18.80**
$\text{mBERT}_{\text{uncased}}$	79.48	81.47	85.82	18.36
*BioBridge-$\text{mBERT}_{\text{uncased}}$	**79.50**	**81.98**	**86.22**	**17.43**
$\text{mBERT}_{\text{cased}}$	79.89	81.18	85.40	19.17
*BioBridge-$\text{mBERT}_{\text{cased}}$	**79.98**	**81.76**	**85.97**	**18.42**
$\text{XLM-R}_{\text{base}}$	79.55	81.05	84.07	**18.78**
*BioBridge-$\text{XLM-R}_{\text{base}}$	**79.99**	**81.32**	**85.57**	19.59

G. RESULTS

Table 3 shows the evaluation results on the test dataset. The comparison between machine learning-based models and encoder-based models (Korean and multilingual encoder-based models) clearly shows the superiority of encoder-based models. For the Machine Learning models, while Gradient Boosting achieved the highest F1 score of 66.90, the KR-BERT significantly improved, reaching an F1 score of 78.30. The improvement extends to AUROC, AUPRC, and Brier score metrics, revealing that Korean encoder-based models not only boost prediction accuracy but also refine the calibration of those predictions. The ML-based model was trained based on TF-IDF, so it would have performed worse than the encoder-based model that considered the context.

In the category of Korean-specific encoder-based models, *BioBridge-KR-BERT significantly outperforms the baseline KR-BERT across all evaluated metrics. The F1 score improves by 1.1%, AUROC by 2.52%, AUPRC by 2.16%, and the Brier score decreases by 0.95%. These results indicate that *BioBridge-KR-BERT enhances classification performance across various thresholds. Moreover, BioBridge demonstrates superiority in the accuracy of probabilistic predictions. This trend aligns with the performance gains observed between BioBridge-koBERT and the base koBERT model. Demonstrate the efficacy of BioBridge with Korean-based encoder models.

For the multilingual-specific encoder-based model, *BioBridge-XLM showed an increase of 0.85% in the F1 score, 0.75% in AUROC, and 0.76% in AUPRC compared to the baseline XLM. Notably, the Brier score decreased by 3.04%, signifying that BioBridge enhanced the prediction accuracy and the reliability and calibration of the model's outputs. Moreover, *BioBridge-$\text{mBERT}_{\text{uncased}}$ achieved the lowest Brier score.

BioBridge's consistent enhancement in performance across various encoder-based models attests to its robustness and adaptability to different encoder architectures and languages. These outcomes endorse the BioBridge method as a powerful strategy for classification tasks in the medical domain, notably in settings dealing with bilingual EMRs.

V. ABLATION STUDY

In this section, we perform extensive ablation studies to support our model design. For all experiments, we establish a multilingual-based model as our baseline. The BioBridge framework was designed in two modules. First the ''bridging modality in context'' module, which bridges modalities (Korean, English). This module is essential due to it connects each modality (Korean, English) in the CS context with a token. Excluding the token that bridges each modality in the CS sentence becomes a vanilla encoder-based model. The second module, ''unified bio-embedding,'' aims to infuse the encoder-based model, which lacks medical domain knowledge, with expertise in English medical terminology. The ''unified bio-embedding'' module extracts medical features from the medical feature extractor f_θ^B and incorporates them into the encoder base model. The vanilla encoder-based model is obtained by excluding the ''unified bio-embedding'' module. The optimal hyperparameters for the test set are shown in Table 4.

TABLE 4. Hyperparameter used for ablation studies on multilingual specific encoder based model.

Model. Param	Batch Size	Learning Rate
Multilingual specific encoder based model		
XLM		
w/ Bridging modality	22	1e-5
w/ Bio-embedding	22	5e-6
$mBERT_{uncased}$		
w/ Bridging modality	64	4e-5
w/ Bio-embedding	64	4e-5
$mBERT_{cased}$		
w/ Bridging modality	64	5e-5
w/ Bio-embedding	64	4e-5
$XLM\text{-}R_{base}$		
w/ Bridging modality	64	3e-5
w/ Bio-embedding	64	3e-5

TABLE 5. Ablation Study of a multilingual encoder-based model for emergency case classification. *: results from the proposed BioBridge experiments. The best result within each row block is shown in bold.

Model	F1 ↑	AUROC↑	AUPRC↑	Brier↓
Multilingual specific encoder based model				
XLM	78.84	79.89	84.19	21.93
w/ Bridging modality	79.10	80.19	84.52	21.76
w/ Bio-embedding	79.20	80.02	84.39	19.00
*BioBridge-XLM	**79.69**	**80.64**	**84.95**	**18.80**
$mBERT_{cased}$	79.89	81.18	85.40	19.17
w/ Bridging modality	**80.08**	81.36	85.75	19.36
w/ Bio-embedding	79.27	81.43	85.97	**17.83**
*BioBridge-$mBERT_{cased}$	79.98	**81.76**	**85.99**	18.42
$mBERT_{uncased}$	79.48	81.47	85.82	18.36
w/ Bridging modality	78.80	81.57	86.00	17.96
w/ Bio-embedding	78.69	81.34	85.66	17.49
*BioBridge-$mBERT_{uncased}$	**79.50**	**81.98**	**86.22**	**17.43**
$XLM\text{-}R_{base}$	79.55	81.05	84.07	18.78
w/ Bridging modality	79.27	81.32	85.48	**18.75**
w/ Bio-embedding	79.60	**81.67**	**85.99**	23.33
*BioBridge-$XLM\text{-}R_{base}$	**79.99**	81.32	85.57	19.59

A. RESULTS

We show insightful results for the BioBridge framework through extensive ablation studies. The results of the ablation study are shown in Table 5. Additionally, The optimal hyperparameters for the test set are shown in Table 4.

1) BRIDGING MODALITY IN CONTEXT

Integrating the ''bridging modality in context'' module into our baseline multilingual encoder-based model led to notable improvements across various performance metrics. In XLM, F1 score increased by 0.26%, AUROC increased by 0.3%, AUPRC increased by 0.33%, and Brier score decreased by 0.17%. These results highlight the importance of this module in enhancing comprehension of the context within code-switched scenarios.

2) UNINIFED BIO-EMBEDDING

Adding the ''unified bio-embedding'' module to our baseline multilingual encoder-based model also consistently enhanced performance on almost all metrics. In $mBERT_{cased}$, we observed a 0.62% decrease in the F1 score but a 0.25% increase in AUROC and a 0.57% increase in AUPRC, demonstrating robust performance at all thresholds. Notably, the Brier score decreased by 1.34%, affirming enhanced reliability and accuracy in probabilistic predictions. This improvement demonstrates that the ''unified bio-embedding'' module can effectively inject medical knowledge into a multilingual-based model trained on a general domain.

The results of these ablation studies highlight the importance of both modules in improving the encoder base model's ability to understand bilingual EMR accurately.

VI. LIMITATION

This paper proposes applying Natural Language Processing (NLP) to Electronic Medical Records (EMRs) to improve decision-making processes in pediatric emergency departments (PEDs). Our work has three primary limitations. Firstly, data privacy and security present challenges. EMRs contain sensitive personal information, necessitating strict compliance with privacy laws and regulations, which restricts the free sharing of data and limits its broader application. Secondly, the quality and consistency challenges. EMR data is written in free text, which makes the data inconsistent. Therefore, data input in different formats may hinder the model's ability to learn uniformly, potentially affecting the generalizability of the results. Thirdly, Model Interpretability challenges. The complexity of the Transformer model architecture used in this work reduces its interpretability. In a healthcare setting, where decisions can have significant implications for patient outcomes, the inability to fully explain how model predictions are made can severely hinder trust and adoption.

In future work, we plan to increase the generalization performance of our model in healthcare settings by obtaining EMR datasets from more hospitals while maintaining compliance. Furthermore, while this paper focuses on Korean-English code-switched EMRs, we intend to extend our approach to include code-switched EMRs from other languages with English to explore its broader applicability in multilingual healthcare environments. Additionally, to make the model more transparent and interpretable, we plan to use methods such as SHapley Additive exPlanations (SHAP) [80] to provide an interpretation of the model.

VII. CONCLUSION

In this paper, we proposed BioBridge, a framework that unified bio-embedding with bridging modality in code-switched EMR. This approach achieved state-of-the-art performance in classifying emergency cases in Korean-

English code-switched EMRs by fine-tuning a transformer encoder-based model. This approach addresses the critical challenges of overcrowding in PEDs and pioneers using code-switched EMRs with multilingual NLP techniques, setting a new standard for future applications.

REFERENCES

[1] M. Tanzariello, S. Marventano, S. Bucci, A. De Leva, W. Ricciardi, and A. de Belvis, "Emergency department crowding and access block: Is lean thinking a smart answer? A systematic review," *Eur. J. Public Health*, vol. 25, no. 3, pp. 19–4209, Oct. 2015.

[2] L. Zhou, P. Zhao, D. Wu, C. Cheng, and H. Huang, "Time series model for forecasting the number of new admission inpatients," *BMC Med. Informat. Decis. Making*, vol. 18, no. 1, pp. 1–11, Dec. 2018.

[3] H. R. Rasouli, A. A. Esfahani, M. Nobakht, M. Eskandari, S. Mahmoodi, H. Goodarzi, and M. A. Farajzadeh, "Outcomes of crowding in emergency departments; a systematic review," *Arch. Academic Emergency Med.*, vol. 7, no. 1, pp. 1–36, 2019.

[4] M. Sartini, A. Carbone, A. Demartini, L. Giribone, M. Oliva, A. M. Spagnolo, P. Cremonesi, F. Canale, and M. L. Cristina, "Overcrowding in emergency department: Causes, consequences, and solutions—A narrative review," *Healthcare*, vol. 10, no. 9, p. 1625, Aug. 2022.

[5] Committee on Pediatric Emergency Medicine, "Overcrowding crisis in our nation's emergency departments: Is our safety net unraveling?" *Pediatrics*, vol. 114, no. 3, pp. 878–888, 2004.

[6] B.-S. Kim, J.-Y. Kim, S.-H. Choi, and Y.-H. Yoon, "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*, vol. 8, no. 2, Feb. 2018, Art. no. e018208.

[7] F. Williams and S. Boren, "The role of the electronic medical record (EMR) in care delivery development in developing countries: A systematic review," *J. Innov. Health Informat.*, vol. 16, no. 2, pp. 139–145, Jul. 2008.

[8] C. A. Turner, A. D. Jacobs, C. K. Marques, J. C. Oates, D. L. Kamen, P. E. Anderson, and J. S. Obeid, "Word2Vec inversion and traditional text classifiers for phenotyping lupus," *BMC Med. Informat. Decis. Making*, vol. 17, no. 1, pp. 1–11, Dec. 2017.

[9] M. Hartswood, R. Procter, M. Rouncefield, and R. Slack, "Making a case in medical work: Implications for the electronic medical record," *Comput. Supported Cooperat. Work (CSCW)*, vol. 12, no. 3, pp. 241–266, Sep. 2003.

[10] N. Menachemi, "Benefits and drawbacks of electronic health record systems," *Risk Manage. Healthcare Policy*, p. 47, May 2011.

[11] N. W. Sterling, R. E. Patzer, M. Di, and J. D. Schrager, "Prediction of emergency department patient disposition based on natural language processing of triage notes," *Int. J. Med. Informat.*, vol. 129, pp. 184–188, Sep. 2019.

[12] X. Yang, "A large language model for electronic health records," *NPJ Digit. Med.*, vol. 5, p. 194, Dec. 2022, doi: 10.1038/s41746-022-00742-2.

[13] K. Mermin-Bunnell, Y. Zhu, A. Hornback, G. Damhorst, and T. Walker, "Use of natural language processing of patient-initiated electronic health record messages to identify patients with COVID-19 infection," *Jama Network Open*, vol. 6, Jun. 2023, Art. no. e2322299.

[14] I. Segura-Bedmar, C. Colón-Ruíz, M. Á. Tejedor-Alonso, and M. Moro-Moro, "Predicting of anaphylaxis in big data EMR by exploring machine learning approaches," *J. Biomed. Informat.*, vol. 87, pp. 50–59, Nov. 2018.

[15] R. Rivas, N. Montazeri, N. X. Le, and V. Hristidis, "Automatic classification of online doctor reviews: Evaluation of text classifier algorithms," *J. Med. Internet Res.*, vol. 20, no. 11, Nov. 2018, Art. no. e11141.

[16] H. Alzoubi, N. Ramzan, R. Alzubi, and E. Mesbahi, "An automated system for identifying alcohol use status from clinical text," in *Proc. Int. Conf. Comput., Electron. Commun. Eng. (iCCECE)*, Aug. 2018, pp. 41–46.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[18] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1495–1504.

[19] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "BioWordVec, improving biomedical word embeddings with subword information and MeSH," *Sci. Data*, vol. 6, no. 1, p. 52, May 2019.

[20] X. Yu, W. Hu, S. Lu, X. Sun, and Z. Yuan, "BioBERT based named entity recognition in electronic medical record," in *Proc. 10th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Aug. 2019, pp. 49–52.

[21] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *Npj Digit. Med.*, vol. 4, no. 1, p. 86, May 2021.

[22] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical BERT embeddings," 2019, *arXiv:1904.03323*.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[24] B. Polnaszek, A. Gilmore-Bykovskyi, M. Hovanes, R. Roiland, P. Ferguson, R. Brown, and A. J. H. Kind, "Overcoming the challenges of unstructured data in multisite, electronic medical record-based abstraction," *Med. Care*, vol. 54, no. 10, pp. e65–e72, 2016.

[25] S. Perera, A. Sheth, K. Thirunarayan, S. Nair, and N. Shah, "Challenges in understanding clinical notes: Why NLP engines fall short and where background knowledge can help," in *Proc. Int. Workshop Data Manage. Analytics Healthcare*, 2013, pp. 21–26.

[26] K. Adnan, R. Akbar, S. W. Khor, and A. B. A. Ali, "Role and challenges of unstructured big data in healthcare," in *Proc. ICDMAI*, vol. 1, 2019, pp. 301–323.

[27] C. J. Harrison and C. J. Sidey-Gibbons, "Machine learning in medicine: A practical introduction to natural language processing," *BMC Med. Res. Methodology*, vol. 21, no. 1, pp. 1–11, Dec. 2021.

[28] Y. Zhang et al., "High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP)," *Nature Protocols*, vol. 14, no. 12, pp. 3426–3444, Dec. 2019.

[29] V. M. Castro et al., "Validation of electronic health record phenotyping of bipolar disorder cases and controls," *Amer. J. Psychiatry*, vol. 172, no. 4, pp. 363–372, Apr. 2015.

[30] Y.-T. Park and D. Han, "Current status of electronic medical record systems in hospitals and clinics in Korea," *Healthcare Informat. Res.*, vol. 23, no. 3, p. 189, 2017.

[31] A. Shinozaki, "Electronic medical records and machine learning in approaches to drug development," in *Artificial Intelligence in Oncology Drug Discovery and Development*. London, U.K.: IntechOpen, 2020.

[32] J. J. Cimino, H. Min, and Y. Perl, "Consistency across the hierarchies of the UMLS semantic network and metathesaurus," *J. Biomed. Informat.*, vol. 36, no. 6, pp. 450–461, Dec. 2003.

[33] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. 1, pp. 267–270, Jan. 2004.

[34] B. L. Humphreys, A. T. McCray, and D. A. B. Lindberg, "The unified medical language system," *Yearbook Med. Informat.*, vol. 2, no. 1, pp. 41–51, 1993.

[35] C. P. Morrey, J. Geller, M. Halper, and Y. Perl, "The neighborhood auditing tool: A hybrid interface for auditing the UMLS," *J. Biomed. Informat.*, vol. 42, no. 3, pp. 468–489, Jun. 2009.

[36] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, vol. 11, Aug. 2016, pp. 785–794.

[37] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc. Ser. B, Stat. Methodology*, vol. 20, no. 2, pp. 215–232, Jul. 1958.

[38] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," in *Annals of Statistics*. USA: Institute of Mathematical Statistics, 2001, pp. 1189–1232.

[39] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[40] *Kobert*. Accessed: Sep. 23, 2019. [Online]. Available: https://github.com/SKTBrain/KoBERT

[41] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, "KR-BERT: A small-scale korean-specific language model," 2020, *arXiv:2008.03979*.

[42] A. Conneau and G. Lample, "Cross-lingual language model pretraining," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–20.

[43] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8440–8451.

[44] Y.-M. Kim and T.-H. Lee, "Korean clinical entity recognition from diagnosis text using BERT," *BMC Med. Informat. Decis. Making*, vol. 20, no. S7, pp. 1–9, Sep. 2020.
[45] Y. S. Bae, K. H. Kim, H. K. Kim, S. W. Choi, T. Ko, H. H. Seo, H.-Y. Lee, and H. Jeon, "Keyword extraction algorithm for classifying smoking status from unstructured bilingual electronic health records based on natural language processing," *Appl. Sci.*, vol. 11, no. 19, p. 8812, Sep. 2021.
[46] P. Auer, *Code-Switching in Conversation: Language, Interaction and Identity*. Evanston, IL, USA: Routledge, 2013.
[47] C. Baker, *Foundations of Bilingual Education and Bilingualism*. Bristol, U.K.: Multilingual Matters, 2011.
[48] J. Ahn, C. La Ferle, and D. Lee, "Language and advertising effectiveness: Code-switching in the Korean marketplace," *Int. J. Advertising*, vol. 36, no. 3, pp. 477–495, May 2017.
[49] J.-E. Park, "Korean/English intrasentential code-switching: Matrix language assignment and linguistic constraints," Univ. Illinois at Urbana, Champaign, IL, USA, Tech. Rep. 1, 1990.
[50] N. Miwa, "Intrasentential codeswitching in Japanese and English," Ph.D. dissertation, Dept. Linguistics, Univ. PA, Philadelphia, PA, USA, 1985.
[51] D. Amazouz, M. Adda-Decker, and L. Lamel, "Addressing code-switching in French/Algerian Arabic speech," in *Proc. Interspeech*, Aug. 2017, pp. 62–66.
[52] X. Li, H. Zhang, and X.-H. Zhou, "Chinese clinical named entity recognition with variant neural structures based on BERT methods," *J. Biomed. Informat.*, vol. 107, Jul. 2020, Art. no. 103422.
[53] H. Liu, Z. Zhang, Y. Xu, N. Wang, Y. Huang, Z. Yang, R. Jiang, and H. Chen, "Use of BERT (Bidirectional encoder representations from rransformers)-based deep learning method for extracting evidences in Chinese radiology reports: Development of a computer-aided liver cancer diagnosis framework," *J. Med. Internet Res.*, vol. 23, no. 1, Jan. 2021, Art. no. e19689.
[54] Z. Li, Z. Li, J. Zhang, Y. Feng, and J. Zhou, "Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, no. 1, pp. 2476–2483, May 2021.
[55] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," 2019, *arXiv:1903.10676*.
[56] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 19–27.
[57] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave, "CCNet: Extracting high quality monolingual datasets from Web crawl data," 2019, *arXiv:1911.00359*.
[58] Y. Kim, J.-H. Kim, J. M. Lee, M. J. Jang, Y. J. Yum, S. Kim, U. Shin, Y.-M. Kim, H. J. Joo, and S. Song, "A pre-trained BERT for Korean medical natural language processing," *Sci. Rep.*, vol. 12, no. 1, p. 13847, Aug. 2022.
[59] S. Fuchs, M. Terry, K. Adelgais, M. Bokholdt, J. Brice, K. M. Brown, A. Cooper, M. E. Fallat, K. E. Remick, K. Widmeier, W. Simon, and M. Marx, "Definitions and assessment approaches for emergency medical services for children," *Pediatrics*, vol. 138, no. 6, pp. 1–22, Dec. 2016.
[60] B. K. Reddy and D. Delen, "Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology," *Comput. Biol. Med.*, vol. 101, pp. 199–209, Oct. 2018.
[61] L.-H. Yao, K.-C. Leung, C.-L. Tsai, C.-H. Huang, and L.-C. Fu, "A novel deep learning-based system for triage in the emergency department using electronic medical records: Retrospective cohort study," *J. Med. Internet Res.*, vol. 23, no. 12, Dec. 2021, Art. no. e27008.
[62] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-Gram features," 2017, *arXiv:1703.02507*.
[63] Q. Chen, Y. Peng, and Z. Lu, "BioSentVec: Creating sentence embeddings for biomedical texts," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2019, pp. 1–5.
[64] W. Gao, X. Zheng, and S. Zhao, "Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF," *J. Phys. Conf. Ser.*, vol. 1848, no. 1, Apr. 2021, Art. no. 012083.
[65] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.
[66] H. K. Kim, Y. Park, Y. Park, E. Choi, S. Kim, H. You, and Y. S. Bae, "Identifying alcohol-related information from unstructured bilingual clinical notes with multilingual transformers," *IEEE Access*, vol. 11, pp. 16066–16075, 2023.
[67] S. Lee, J. Lee, J. Park, J. Park, D. Kim, J. Lee, and J. Oh, "Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage," *Amer. J. Emergency Med.*, vol. 77, pp. 29–38, Mar. 2024.
[68] J. An, J. Kim, L. Sunwoo, H. Baek, S. Yoo, and S. Lee, "De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained BERT," Seoul Nat. Univ., Seoul, Republic of Korea, Tech. Rep. 1, 2023.
[69] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–20.
[70] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," 2019, *arXiv:1908.07490*.
[71] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7463–7472.
[72] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13041–13049.
[73] P. Gao, Z. Jiang, H. You, P. Lu, S. C. H. Hoi, X. Wang, and H. Li, "Dynamic fusion with Intra- and inter-modality attention flow for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6632–6641.
[74] J. Kim, M. Ma, K. Kim, S. Kim, and C. D. Yoo, "Progressive attention memory network for movie story question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8329–8338.
[75] J. Kim, M. Ma, T. Pham, K. Kim, and C. D. Yoo, "Modality shifting attention network for multi-modal video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10103–10112.
[76] H. M. Adler, "The history of the present illness as treatment: Who's listening, and why does it matter?" *J. Amer. Board Family Pract.*, vol. 10, no. 1, pp. 28–35, 1997.
[77] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[78] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–25.
[79] T. Wolf et al., "HuggingFace's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*.
[80] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.



JANGYEONG JEON (Member, IEEE) received the B.S. degree from the Department of Big Data Statistics, Cheongju University, Cheongju, Republic of Korea, in 2022. He is currently pursuing the M.S. degree in artificial intelligence with Chung-Ang University, Seoul, South Korea. His main research interest includes natural and various multi-modal large language models.



SANGYEON CHO (Graduate Student Member, IEEE) received the B.S. degree in software engineering from Kookmin University, Seoul, Republic of Korea, in 2023. He is currently pursuing the M.S. degree with Chung-Ang University. His research interests include natural language processing, representation learning, and various multi modal learning.



DONGJOON LEE received the B.S. degree in IT media engineering from Kyonggi University, Suwon, South Korea, in 2022. He is currently pursuing the M.S. degree in artificial intelligence with Chung-Ang University, Seoul, South Korea. His main research interests include deep learning-based survival analysis and explainable AI for time series classification.



CHANGHEE LEE (Member, IEEE) received the B.S. and M.S. degrees from Korea University, Seoul, South Korea, in 2007 and 2011, respectively, and the Ph.D. degree from the University of California at Los Angeles, USA, in 2021. He has been an Assistant Professor with the Department of Artificial Intelligence, Chung-Ang University, Seoul, since 2021. He has also been an Assistant Professor with the Department of Artificial Intelligence, Korea University, since September 2024. His research interests include integrating multiple modalities, building advanced time-series models and causal inference models, discovering scientific knowledge from data, and interpreting "black-box" machine learning methods.



JUNYEONG KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from KAIST, Republic of Korea, in 2015, 2017, and 2021, respectively. He was a Postdoctoral Research Associate with the Artificial Intelligence and Machine Learning Laboratory, School of Electrical Engineering, KAIST. He is currently an Assistant Professor with the Department of AI, Chung-Ang University, Republic of Korea. His research interests include visual-language reasoning, visual question answering, and various video-based problems.