IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 August 2023, accepted 22 September 2023, date of publication 13 October 2023, date of current version 20 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324376

RESEARCH ARTICLE

Automatic Diagnosis of Medical Conditions Using Deep Learning With Symptom2Vec

MINJI KIM AND INWHEE JOE
Department of Computer Science, Hanyang University, Seongdong-gu, Seoul 04763, South Korea
Corresponding author: Inwhee Joe (iwjoe@hanyang.ac.kr)

ABSTRACT In this paper, a medical examination algorithm is proposed that can collect users' symptoms and automatically issue a diagnosis. The proposed algorithm makes use of ''Symptom2Vec'' and the ''analysis model of responses on self-diagnosis questions'' (AMoRSD) for real-time interviews with users. Symptom2Vec can learn about the relationship between terms related to the symptoms and disease, and establish questioning criteria to be used in patient health checkups, as well as general appropriate follow-up questions based on patient symptomology. AMoRSD analyzes the patient's emotional expressions and responses to self-diagnostic questions, classifying them into ''Sick,'' ''Not Sick,'' and ''Neutral'' categories based on patterns. Compared to traditional models, Symptom2Vec earned the highest mean symptom similarity score of 0.983. Furthermore, compared to other models that only learn from patient responses, AMoRSD demonstrates an area under curves (AUC) of 0.99%, indicating that jointly learning the relationship between emotions and patient responses improves the accuracy of user response classification. The combined algorithm of Symptom2Vec and AMoRSD enhances the efficiency and accuracy of user symptom collection and appropriate diagnosis generation. The data were collected from reliable medical sources such as WebMD Dictionary, NHS inform, Snomed Ct, and Cleveland Clinic, encompassing 526 disease names and 2078 symptoms. Additional data were obtained for AMoRSD, focusing on conversations within a hospital context, and effectively trained and evaluated the model using diverse and representative datasets. This research addresses the importance of medical history-taking and contributes to the field by providing a robust framework for real-time symptom-based diagnosis in clinical environments.

INDEX TERMS Artificial intelligence, BERT, disease prediction, healthcare, history taking, natural language processing, self-diagnosis, supervised learning, Symptom2Vec, Word2Vec.

I. INTRODUCTION

Acommon misunderstanding of medical diagnosis is that a doctor can accurately determine a patient's needs by performing tests. However, a wealth of studies has revealed that probing questions are the most important step in patient symptomology [1], [2], [3]. Taking note of a patient's medical history is an interview process which is deemed important in making a medical diagnosis in approximately 75% of cases. This is therefore seen as a sign of good practice and is often taught to medical students at medical school [2], [3], [4]. However, given the lack of a standardized checklist, a great emphasis must be put on a doctor's experience and meticulousness when in conversation with a patient [1], [4], [5], [6]. This unstructured approach isn't purely a matter of efficiency; it can also be cause problems and even be fatal if, for example, a resident and the doctor make different diagnoses for the same patient that conflicts with the patient's medical history. Boston Medical College in the United States analyzed a clinical case in which a patient died when the medical history questioning process was not performed properly [4]. Consequently, the importance of a doctor's evidence collection through an accounting of a patient's medical history is essential [4]. However, it is difficult to generalize the medical history taking process because the medical conditions of patients vary, and the degree of symptomology and stage of disease advancement are also highly diverse. For example, an eye related question

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

should be asked to a person who comes to the hospital with eye pain, and the disease might be correctly diagnosed through the process of asking the patient follow-up questions about other symptoms, including the frequency or severity of those symptoms. However, the method of taking a medical history, which requires interaction with many patients, is a time-consuming process.

As interest in medical care has soared since 2017, papers on developing diagnostic algorithms based on machine learning have increased rapidly [7]. In the oriental medicine symptom diagnosis system, Word2Vec was utilized to solve the synonym problem of symptoms, and studies to help diagnose by applying Word2Vec such as Disase2Vec and Patient2Vec are also being conducted steadily [8], [9], [10], [11]. However, deep learning-based medical chatbots face various challenges in actual clinical scenarios, such as bias issues caused by supervised learning and concerns regarding their reliability and trustworthiness [12]. Furthermore, the data used in the study are also conducted with patients' personal information which requires strict management [7], [13], [14], [15]. This is disadvantageous as it is difficult to apply in a clinical environment because prediction can be attempted only when a patient visits a hospital, and patient data must be collected and pre-processed which is very time-consuming. These restrictions limit the usefulness of deep learning-based medical chatbots in real-world clinical situations.

The following six reasons limit the usefulness of deep learning-based medical chatbots in real clinical settings.

1) The process of generating follow-up questions related to symptoms to patients is important because the diagnosis varies depending on the medical history taking, but there is no standardized way for medical history taking.
2) Doctors generate follow-up questions based on their medical experience, but medical chatbots have some set questions in the form of Rule-based.
3) Since the symptoms of the patient generate follow-up questions while predicting the disease, the weight of the disease increases, causing inference bias.
4) Since deep learning-based models are black box models, reliability problems arise.
5) Since the medical data used for model learning is sensitive data containing individual information, a lot of resources are consumed in the process of collecting and preprocessing data.
6) Chatbot users may not be aware of their symptoms clearly or may find it difficult to express.

It is difficult to use deep learning-based models in real-world clinical practice. Therefore, in this paper, the following three tasks were conducted to solve these problems.

1) Generate follow-up questions based on symptoms using our proposed Symptom2Vec, which expresses the relationship between a disease and symptoms on a vector.
2) Categorize user responses to self-diagnosis questions into three classes (sick, not sick, and neutral) using BERT-based AMoRSD
3) Establish the criteria for generating follow-up questions and using those criteria to standardize the previously unregulated history taking process.

In order to determine their relationships, only disease and symptom data were used, and patient data was omitted. Based on this relationship, an automatic diagnosis health checkup algorithm is proposed that asks patients about similar symptoms. For instance, if a patient presents a symptom such as a runny nose, the algorithm will include inquiries about related symptoms like sneezing or nasal congestion.

This paper consists of five sections. Section II reviews the related research. First, the history of medical information for patient symptom collection and related research are introduced. Next, the studies are introduced which uses deep learning and Word2Vec to predict diseases. Section III concerns methods and materials. This section proposes Symptom2Vec and follow-up question generation conditions for self-diagnosis, and illustrates the data and experimental environment required for learning. Section IV evaluates the performance of each model, and finally, Section V discusses the conclusions of this paper and future research.

II. RELATED WORK

A. HISTORY TAKING

Medical history taking is the process of obtaining information on a patient's disease, such as the time, symptoms, location, and intensity of occurrence. Until the 1850s, there was no record of history taking, and the diagnosis was focused on the immediate presentation of physical symptoms alone [16]. The first medical history taking tool was the Woodworth Personal Data Sheet (WPDS) developed in 1928 for largescale soldiers suffering from World War I-related mental illnesses. It was a questionnaire-type measurement tool to identify neurotic patients [1], [17]. From this point on, tracking a patient's medical history has proven to be an essential part of the diagnosis process, and research related to history taking instruments (HTI) has been actively conducted to effectively record patient medical histories [18], [19]. Since the 1980s, software based HTI studies have been conducted for physicians; however, most clinicians have been reluctant to use computers because they doubt the reliability of the software.

Despite this view, computerizing such data has proven to have many benefits. Collected data are more reliable and accurate than face-to-face interviews [1], [20], [21]. HTI also has the advantage of saving clinicians' time and being able to diagnose and treat patients regardless of location [1], [22]. This is because clinicians spend most of their time listening to patients' medical histories. Although these numerous potential uses, such as its advantageous role in combating the recent COVID-19 pandemic have been

demonstrated, utilizing it in the actual clinical setting has proven challenging due to significant drawbacks. In practice, history taking primarily relies on questionnaires rather than direct interaction with a doctor. As a result, there is a possibility of misinterpreting questions and doubt cast on the reliability of the answers. In addition, this can lead to inference bias because it performs inductive reasoning that queries the symptoms of a disease by increasing the weight for a specific disease based on patient symptom information, rather than deductive reasoning that diagnoses a disease after collecting patient information. This is because the algorithm of medical history listening mainly uses sequential Bayesian algorithms [20], [23], [24]. Bayesian algorithms have the advantage of dramatically reducing time compared to formalized questionnaires as they select the next question based on the previous answers of the patient [25], but these algorithms rely on access to a vast amount of medical procedure data and require some rule-based sequential fixation which may cause inference bias.

B. DISEASE PREDICTION USING DEEP LEARNING

Deep learning technology, which is at the center of the development of artificial intelligence, uses an artificial neural network that imitates the brain structure of living things. This innovation is being used across various fields including image and natural language processing. The medical industry has also adopted this technology, and it is playing a role in complementing the shortcomings of the existing medical system, including reducing inefficiencies and inaccuracies in the disease diagnosis process. For example, the European Respiratory Society announced a technology to diagnose COVID-19 by voice. Mel-spectrogram analysis technology was utilized with the average verification AUROC at 83.23%, and test AUROC also showed a high degree of accuracy at 78.3% [26]. However, deep learning model applied to the medical field also has some limitations. The first problem is that it is difficult to balance the data. As the importance of quality data has become clear, many hospitals around the world are conducting computerization to collect data from patients and continuing research to extract data that meets each condition. Examples of computerized data include EHR and MIMIC, which are big data with vast amounts of data [27], [28]. However, these data are difficult to access and limited compared to other data because of privacy policy issues [29]. Furthermore, some data are too dense or missing specific classes in collected data sets, so researchers must obtain additional data from hospitals or clinics or use a GAN model to generate additional data [7].

However, since the data thus obtained are likely to have the same problem, it is necessary to prove the validity of the deep learning model [7]. The second issue is the uncertainty surrounding the accuracy of the deep learning model, which may not be reflected in the figures published in current research. It is unclear whether a diagnosis made using real patient data with multiple variables will result in satisfactory performance, as opposed to using cases learned in the deep learning model. IBM's Watson, developed in the United States, had a consistency rate of 40% with doctors when using patient data from within the country. However, it was noted that the data Watson was trained on was not representative of certain patient groups such as Asian characteristics as it was trained using only American data [30]. The third issue is the lack of clarity around responsibility in the case of a human accident caused by a malfunction of the deep learning model. It is uncertain as to who should be held accountable for such unexpected incidents, including hospitals, doctors, and the developers of the technology. Currently, the deep learning model operates as a black box model during the process of inputting data and receiving results, making it unclear what information the deep learning model uses when making a diagnosis. This lack of transparency reduces the reliability of diagnoses. This result, various issues with artificial intelligence make it difficult to be widely adopted in the medical field, leading to ongoing research and development. However, it remains challenging to implement the deep learning model in actual professional settings. To address this gap, this paper focuses on representing the relationships between symptoms and disease using symptom and disease data, rather than limited medical data due to privacy and ethical constraints. Additionally, by explaining the collected symptoms in a method that enables secondary usage by both doctors and patients, the uncertainty of black-box models that could potentially affect trustworthiness was addressed.

C. DISEASE DIAGNOSIS USING Word2Vec METHOD

One of the techniques that came out to solve the disadvantage of not being able to calculate the similarity between words, a problem with one-hot encoding, is Word2Vec. This technique predicts the central word using the peripheral word by learning their relationship [31]. Word2Vec employs a distributed hypothesis; namely, a hypothesis where each word appearing in a sentence is semantically related to the surrounding word. For example, in the sentence ''Students study at school,'' the words ''students'' and ''school,'' ''students'' and ''study,'' and ''school'' and ''study'' are related. In this way, in an article with one theme, different words have a semantic relationship. Expressing this on a vector is called word embedding, and a representative learning method for word embedding is Word2Vec. Word2Vec, which learns each sentence, identifies the role and meaning of words in the context [31]. As medical texts are focused on a single topic with tightly related vocabulary, such approaches are particularly suitable for disease diagnosis. In 2018, a study vectorized text data in the BioNLP process using word2vec and compared the predictive model with logistic regression, revealing that type 2 diabetes outbreaks could be predicted with a degree of accuracy using the XGBoost model [8]. In addition to these studies, information on each patient was collected using the Electronic Health Record (EHR) system, and based on this, a study of Patient2Vec, dealing with

the in-depth representation of personal patient information, was conducted resulting in an area under the curve(AUC) of 0.799 [9]. Moreover, in 2020, a Disase2Vec study was conducted that could classify diseases based on microbial information by connecting microorganisms and diseases [10].

D. NATURAL LANGUAGE PROCESSING IN THE LANGUAGE PROCESSING IN THE MEDICAL FIELD

In recent years, there has been an increase in research focusing on utilizing digital biometric indicators in natural language processing (NLP) to analyze medical texts [32], [33]. This has enabled low-cost pathology diagnosis, classification, and monitoring. One study introduced the transformer for relation extraction (TRE) model, which demonstrated that using a pre-trained Transformer significantly improved sample efficiency [34]. By utilizing only 20% of the training data compared to training from scratch on the TACRED dataset, the pre-trained model achieved comparable performance [34]. Additionally, there has been a rapid increase in NLP research, particularly in the field of mental health [35]. This is due to the convenience of analyzing clinical documents written in free-form text and accessing behavioral and emotional reference documents for patient treatment [35], [36]. Ongoing research aims to efficiently extract relevant information from numerous medical literature sources, including rapid and accurate retrieval of documents related to specific diseases. In the medical field, there is a growing focus on fine-tuning pre-trained models like BERT to enhance model performance [37], [38], [39], [40]. For instance, a study analyzed the performance of NLP integrated with machine learning for accurately classifying medical subjects in text-based health counseling data used for disease counseling in medical chatbot systems [40]. The study compared five models, including LSTM and BERT, and found that the BERT model achieved the highest performance metrics, with approximately 75% accuracy in four evaluation indicators(precision, accuracy, recall, and F1-score) [40]. Precision, accuracy, recall, and F1-score are commonly used evaluation metrics in the medical field due to the importance of accurate information [36]. To improve the model accuracy, data quantity and quality are crucial factors, and activities such as changing pre-trained models or adjusting hyperparameters are performed. Related research has investigated the performance variations based on different pre-trained models and the stability of fine-tuning with limited medical resources [36]. A comparison of BERT-BASE, BERT-LARGE, and ELECTRA models identified specific techniques for performance enhancement, such as freezing lower layers for BERT-BASE and layerwise decay for BERT-LARGE and ELECTRA models [37]. The research also emphasized the importance of domain-specific vocabulary and pre-training for robust fine-tuning models [37]. AMoRSD utilizes a pre-trained model specifically designed for sentiment analysis, which sets it apart from previous studies that employed large-scale pre-trained models like BERT-BASE and BERT-LARGE with fine-tuning. While sentiment analysis may not have direct relevance to the healthcare domain, the representation of responses to self-diagnosis questions is closely linked to human emotional expression. As a result, AMoRSD possesses the capability to learn both human emotions and symptom expressions simultaneously, enabling accurate real-time classification of users' responses to self-diagnosis questions.

III. METHODS AND MATERIALS

A. DISTRIBUTION SEMANTIC

The Symptom2Vec proposed in the paper is an idea based on the concept of distributed semantics from Word2Vec. In Word2Vec, words in a sentence are embedded in a way that reflects the meaning of their neighboring words, allowing the algorithm to identify the context of the sentence. In contrast, Symptom2Vec embeds symptoms of a disease instead of words in a sentence because symptoms are essential in diagnosing a disease. However, the learning process of Symptom2Vec is structurally similar to that of Word2Vec, and it allows the algorithm to recognize the relationship between symptoms of a disease.

Fig. 1 compares the learning process of Word2Vec and Symptom2Vec under the assumption that the window size is 2, and it is confirmed that both have structural similarities in terms of distributed semantics. Among the criteria for judging a disease, symptoms are the most important clue. As the figure shows, colds are accompanied by symptoms such as nasal congestion, cough, fever, sore throat, and headache. Symptom2Vec understands them in the same context because they have a similarity to the symptoms of a cold.

In the case of Symptom2Vec, multiple symptoms are incorporated into one sentence. Each sentence represents a collection of symptoms associated with a particular disease, and the sentences shown in the figure represent the symptoms of a cold. This design allows Symptom2Vec to grasp the relationships and associations between various symptoms, aiding in more accurate disease diagnosis based on similar symptom patterns.

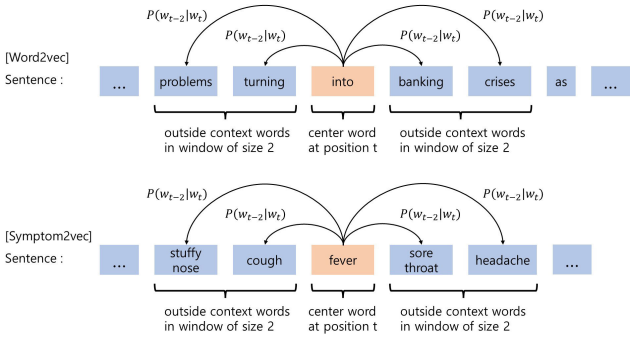

FIGURE 1. Comparison of the learning process between Word2Vec and Symptom2Vec.

B. Symptom2Vec

Similar symptoms are clustered in Symptom2Vec, just as similar words are clustered in Word2Vec. Symptom2Vec

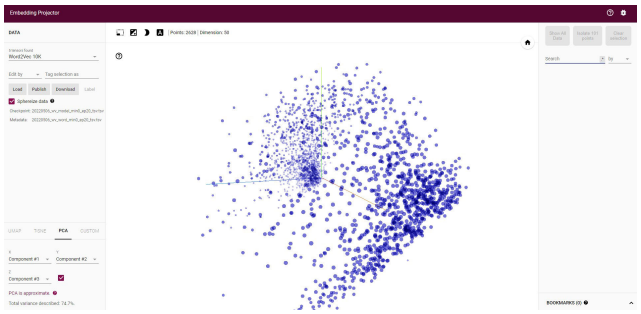

FIGURE 2. Visualization of Symptom2Vec in three dimensions using Embedding Projector of Google.

learns the symptoms of a single disease in the same context, so the cluster of symptoms has unique characteristics. This cluster contains information about the disease. Among the criteria for determining disease, symptoms are the most important clue. For example, colds are accompanied by symptoms such as nasal congestion, cough, fever, sore throat, and headache, which means that each symptom is similar to the other and in the same context by the standards of colds. In conclusion, Symptom2Vec understands the symptoms of each disease as a context and proceeds with learning.

Fig. 2 is a three-dimensional visualization of Symptom2Vec learned using Google's Embedding Projector. The symptoms were clustered by disease according to the learning characteristics of Symptom2Vec. The confirmation of cluster similarity for specific diseases is conducted using cosine similarity. By utilizing Symptom2Vec in self-diagnosis, it becomes possible to identify similar symptoms to the user's reported symptoms and generate subsequent questions. Detailed methods, along with an explanation of the variable "potentialValue," will be introduced in the following section.

C. CONDITIONS FOR GENERATING FOLLOW-UP QUESTIONS FOR SELF-DIAGNOSIS

If similar symptoms have been identified using Symptom2Vec, it is time to set a standard for when to proceed with additional interviews. The criteria can be defined using the cluster characteristics of Symptom2Vec. In this paper, there are two criteria for completing additional interviews. The first criterion is, "Is the number of symptoms mentioned by the user sufficient to predict the disease?" The data used for learning will be covered in more detail in E and F part, but the average number of symptoms for each disease in the data used for Symptom2Vec learning is 7.97. Therefore the first criterion is defined as 8 because a disease has an average of eight symptoms. This can be expressed as Equation 1. The second criterion is "Is the user's symptom cluster similar to the disease's symptom cluster?" Clusters should be judged based on density, not area. A large number of symptoms must be concentrated in a narrow area to be recognized as a cluster. Suppose the user complains of having symptoms such as a runny nose, sneezing, coughing, and headaches. An additional questionnaire asking appropriate follow-up questions for related illnesses, such as those relating to chest pain, weight loss, and hyperuria, can specify which disease cluster the user corresponds to. Criteria for determining the degree of integration of these symptoms can be obtained through the average of the similarity of symptoms. If you know all the coordinates of the points, let's say that the mean of d is the potential value. When the density is large, the average value becomes smaller, and when the density is small, the average value becomes larger, so the potential value and the density can be defined by Equation 2.

$$AVG(diseaseSymptomsCount) = 7.97$$
$$\therefore userSymptomsCountStandard = 8 = uSCS \quad (1)$$

$$Density \propto \frac{1}{potentialValue} \quad (2)$$

Equations 3 to 6 provide a formalization of the process for defining the potential standard based on the symptoms mentioned by a user, denoted as Sn for the nth mention. Among these equations, Equation 5 defines the *potentialValue* for each disease's symptom cluster, which is used to calculate the average similarity of the symptom clusters for a disease. By averaging the *potentialValues*, the average density of symptoms that define a disease can determined and it is represented as the *potentialStandard* in Equation 6. In this calculation, the cosine similarity method, considering only the direction of the vectors and accounting for the vector's magnitude, is employed. By computing the minimum, maximum, and average values through these equations, the following results are obtained: the minimum average symptom similarity value among the diseases is 0.77, and the maximum average symptom similarity value is 1. The average symptom similarity value across all diseases was 0.94. Consequently, the *potentialStandard*, which serves as the second criterion, is defined as 0.94.

Therefore, Equations 1 to 6 provide a framework for defining the two criteria for generating follow-up questions in self-diagnosis. Equation 7 summarizes these criteria, indicating that follow-up question generation is terminated when the number of symptom mentions by the user is sufficient to predict a disease and when the mentioned symptoms form clusters that are sufficiently indicative of a disease.

$$d_{cosine}(X, Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} \quad (3)$$

$$S = S_1 + S_2 + S_3 + \ldots + S_n \quad (4)$$

$$potentialValue(S) = \frac{AVG(d_{cosine}(S_0, S_{whole-0})}{n} + \frac{AVG(d_{cosine}(S_1, S_{whole-1})}{n} + \ldots + \frac{AVG(d_{cosine}(S_n, S_{whole-n}))}{n} \quad (5)$$

$$potentialStandard = AVG(potentialValue(diseaseSymptoms)) = 0.94 \quad (6)$$

$$If\,(n \geq uSCS) and (potentialValue \geq potentialStandard) : EndFollowUpQuestioning \quad (7)$$

D. AMoRSD

The Analysis Model of responses to self-diagnosis questions (AMoRSD) is a model used in the diagnostic process to determine the presence of symptoms by classifying patient responses into three categories: ''Sick,'' ''Not Sick,'' and ''Neutral.''

The core idea of AMoRSD is to fine-tune a pre-trained BERT model using symptom expression data for self-diagnosis questions by leveraging sentiment analysis data based on BERT. Although sentiment analysis itself is not commonly used in the medical domain, the expression of responses to self-diagnosis questions is closely tied to human emotions. Figure 3 illustrates example sentences of response expressions for each class and their association with human emotions.

For instance, a sentence like ''I have a headache'' could be classified as a negative sentiment and associated with the ''Sick'' class. Therefore, AMoRSD classifies response expressions based on human emotions, allowing for efficient learning with limited data.

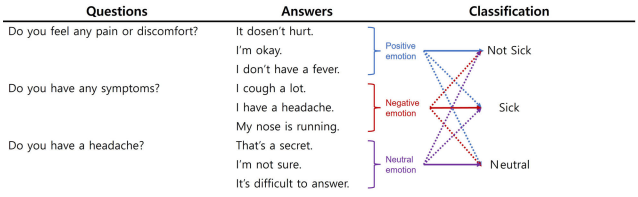

FIGURE 3. Association between response expressions to self-diagnosis question and human emotions.

E. OVERALL ALGORITHM

Fig. 4 represents an example of a medical examination algorithm's overall progress using Symptom2Vec and AMoRSD. This is a situation where a user complains of eye pain. First, keyword extraction and chunking rules are used to extract the symptoms from the text. This is the ''Find Symptom''. Second, AMoRSD analyzes the conversation to determine if the user has the symptoms. It also checks if the user has a neutral attitude and does not relate to the symptoms. Based on the results of the analysis, the user's symptom information is updated. The information remains until the last stage of the algorithm. Third, based on the updated symptom list, those classified as ''Sick'' or ''Not Sick'' are verified for additional questionnaire conditions to confirm whether to proceed with additional questions. The other one is ''Neutral'', so the additional questionnaire is used. Fourth, if the user is not satisfied with the conditions, the user needs to conduct an additional questionnaire to extract symptoms similar to those of the user through Symptom2Vec and ask the user questions. If the conditions related to the additional questionnaire are satisfied, the top three diseases inferred from the information gathered are checked along with the matching rate. Finally, if the match rate does not exceed 70%, the user is asked for one of the symptoms corresponding to the first-rank disease. However, if the concordance rate exceeds 70%, the predicted diseases are explained to the user.

F. DATA BASED ON MEDICAL KNOWLEDGE

Data were collected from websites such as WebMD Dictionary, NHS inform, Snomed Ct, and Cleveland Clinic, which are known for reliable medical data. Data collection was completed based on 526 disease names by referring to the papers related to medical questionnaires. Fig. 5 shows a partial excerpt of the data, which consists of 526 disease names and 2078 symptoms. The column consists of a total of 8 columns, each representing the disease name, description, prevention method, cause, symptom, accompanying disease, treatment department, and treatment method.

G. DATA FOR AMoRSD

Additional data was collected for AMoRSD as it requires conversation-related data to determine whether the user has symptoms. The collected dataset consisted of Google search results focusing on conversations within a hospital context. A corpus of 2078 symptoms was used to generate sentences related to being sick, not being sick, and neutral expressions, and these sentences were then categorized into three classes and labeled accordingly. The dataset was split into an 80:20 ratio for training and evaluation purposes. The training data comprised 1161 instances classified as Sick, 1216 instances classified as Not Sick, and 1008 instances classified as Neutral. The remaining 20% of the data served as the test set, which consisted of 307 instances classified as Sick, 292 instances classified as Not Sick, and 248 instances classified as Neutral. By employing this methodology, a diverse and representative dataset was utilized to train and evaluate the AMoRSD model, enabling effective classification of the purpose of user talks within the medical conversation context.

H. LEARNING

Table 1 summarizes the specifications of the hardware and software used in Symptom2Vec learning. The CPU used AMD Ryzen Threadripper 3960X 24-Core Processor, while the GPU, which plays an important role in model learning such as BERT, used NVIDIA TITAN RTX. The RAM capacity is 98304 MB, and CUDA Toolkit installed version 11.7.0 in consideration of GPU model. The version of Tensorflow (including Keras) used 2.7.0 and Transformers used 4.10.0. Finally, Python and Pytorch used versions 3.7.10 and 1.10.1, respectively, and for Sklearn, which serves as an auxiliary aid, such as dividing training data and test data, 0.24.2 was installed to finally complete the configuration.

Table 2 provides a summary of the learning parameters used for Symptom2Vec and AMoRSD. Symptom2Vec employed a Skip-gram model using the Gensim library. The vector size was set to 200, and the window size, as defined in

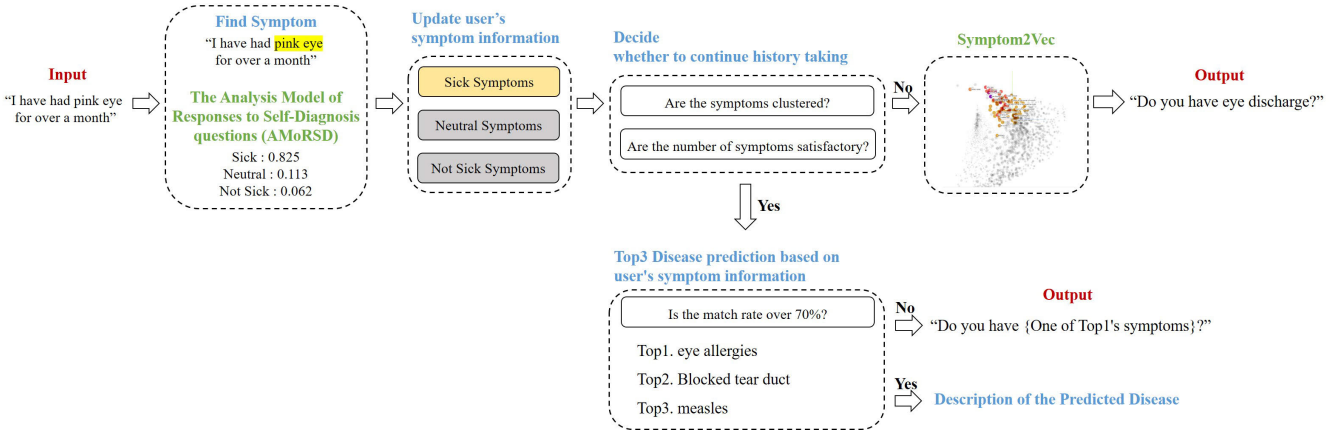

FIGURE 4. Process of medical examination algorithm using Symptom2Vec and AMoRSD.

	name	desc	prevent	cause	symptom	acompany	cure_department	cure_way
0	abdominal migraine	Abdominal migraines aren't headaches as their ...	Some of the same medicines used to treat migra...	Doctors dont know exactly what causes abdomina...	[headache, pale skin, loss of appetite, nausea...	[]	[internal medicine, neurology]	[medical treatment]
1	abdominal muscle deficiency syndrome	Abdominal muscle deficiency syndrome is a cond...	If your child has mild prune belly syndrome, h...	The underlying cause of pbs is unknown pbs can...	[club foot, poorly developed or absent abdomin...	[]	[internal medicine, digestive internal medicine]	[surgical treatment, supportive treatment]
...	...	...	...	...	...	...	...	...
523	xanthoma	A xanthoma is a skin lesion caused by the accu...	Reducing your blood lipids will improve your o...	Xanthomas are small skin blemishes that happen...	[firm, raised waxy-appearing papules or bumps,...	[leukemia]	[internal medicine, endocrinology]	[medical treatment, surgical treatment, suppor...
524	zellweger syndrome	Zellweger syndrome (zs) is a genetic disorder ...	Theres no way to prevent zs people with a fami...	Zs is the result of a mutation in any of the 1...	[weak, hepatorenal syndrome, face deformity, t...	[digestive tract bleeding]	[internal medicine, digestive internal medicine]	[generally use supportive therapy]
525	zinc poisoning	Zinc is an essential trace element in the huma...	In most cases, it is easy to avoid an overdose...	Poisoning is caused by accidental administrati...	[roughness of breath sounds, hepatomegaly, sou...	[shock]	[emergency department]	[medical treatment, supportive treatment]

FIGURE 5. Data used for Symptom2Vec learning and disease retrieval (data consists of 526 diseases, 2,078 symptoms, and 8 columns).

TABLE 1. Workstation specifications for learning.

Hardware Specification	
CPU	AMD Ryzen Threadripper 3960X 24-Core Processor (48 CPUs)
GPU	NVIDIA TITAN RTX
RAM	98304 MB
Software Specification	
CUDA Toolkit	11.7.0
Tensorflow	2.7.0
Transformers	4.10.0
Python	3.7.10
Pytorch	1.10.1
Sklearn	0.24.2

Equation 1, was determined to be 8. The min count parameter was set to zero to ensure all symptom-related words were learned, regardless of their frequency of occurrence. Additionally, the model's hyperparameters used four workers and 30 epochs. In this study, the yangheng/deberta-v3-base-absa-v1.1[1] model from Hugging Face was chosen as the pre-trained model for AMoRSD. This model was trained on over 30k ABSADatasets[2] and can classify human emotions into three classes (Negative, Positive, Neutral) when given a sentence, aligning well with the objectives of the proposed AMoRSD. The number of epochs was set to 30, and the maximum length of input sentences was limited to 64, and a batch size of 8 was chosen to strike a balance between training time and accuracy. The learning rate was set to 1e-5, and the Cross-Entropy Loss function, commonly used for multi-class classification in BERT, was employed. Lastly, the optimizer was set to AdamW, allowing for the identification of the global minimum even with a small amount of data and without reducing the learning rate. The training data was encoded using the tokenizer.encode_plus function. Data preprocessing involved tokenizing the input sentences, adding special tokens, limiting the maximum length, and performing padding. Subsequently, a DataLoader was created to fetch the encoded input sentences and attention_masks in batch units for training.

[1]https://huggingface.co/yangheng/deberta-v3-base-absa-v1.1
[2]https://github.com/yangheng95/ABSADatasets

TABLE 2. Hyperparameters for Symptom2Vec and AMoRSD.

Hyperparameters for Symptom2Vec	
Model	Skip-gram
Vector size	200
Window size	8
Min count	0
Workers	4
Epochs	30
Hyperparameters for AMoRSD	
Pretrained model name	deberta-v3-base-absa-v1.1
Epochs	30
Max Sequence length	64
Batch size	8
Learning rate	1e-5
Loss Function	CrossEntropyLoss
Optimizer	AdamW

IV. EVALUATION METHOD

Cosine similarity was employed as the evaluation metric for Symptom2Vec. This method calculates similarity by measuring the cosine angle between two vectors. Unlike Euclidean distance or Jacquard similarity, the length of the document does not influence the similarity score. In the context of Symptom2Vec, it is crucial to assess the semantic similarity between symptoms, irrespective of the number of symptoms associated with each disease. Hence, in this paper,

Symptom2Vec is evaluated using the Cosine similarity score function provided by the Gensim library.[3]

As mentioned in the section on ADDITIONAL QUESTIONNAIRE CONDITIONS, the higher average similarity implies that symptoms associated with a particular disease exhibit well-defined clustering, indicating a cohesive grouping. Therefore, when considering all diseases, the value of avg(AvgSimilarity) indicates that symptoms have effectively formed distinct clusters for each respective disease. Additionally, the min(AvgSimilarity) aims to identify the disease with the lowest average similarity among all diseases. If in cases where the list of symptoms for a specific disease was empty, the similarity score was considered as 1.

Symptom2Vec was evaluated by comparing it with three other models: Word2Vec, pre-trained Word2Vec, and pre-trained Sence2Vec. The first comparison model, Word2Vec, was trained using the same parameters as Symptom2Vec and utilized medical knowledge-based data descriptions for learning. Additional comparisons were performed using a pre-trained Word2Vec model sourced from the Google News corpus,[4] which was fine-tuned using the same approach as the first comparison model. The third comparison model, Sence2Vec, is a variation of Word2Vec specifically developed to address the issue of word polysemy [41]. This model was installed through pip install using code provided on GitHub,[5] imported as a library, and then fine-tuned using the same approach as the first comparison model.

AMoRSD was evaluated by comparing it with six models: five commonly used models for multi-class classification (Softmax Regression, SVM, LSTM, Random Forest, BERT (base)), and AMoRSD (large) as the comparison model. For AMoRSD (large), a similar pre-trained model, deberta-v3-large-absa-v1.1,[6] was used as the base, and fine-tuning was conducted following the same process as AMoRSD. The main difference between the two models lies in the training model: AMoRSD utilizes the FAST-LCF-BERT model based on microsoft/deberta-v3-base,[7] while AMoRSD (Large) uses the FAST-LCF-BERT model based on microsoft/deberta-v3-large.[8]

For the remaining five models, similar hyperparameters were employed as in AMoRSD. The LSTM model had a max_sequence_length set to 100, and Adam optimizer and categorical_crossentropy loss function were used. BERT (base) utilized the widely-used bert-base-uncased as the pre-trained model, while the other models had no further modifications.

To assess the model's performance, the model was set to evaluation mode by invoking model.eval(). Evaluation metrics such as classification_report, confusion_matrix, roc_curve, and auc were employed. Since the model is a multiple classification model with three classes, in addition to accuracy, class-specific F1-Score, precision, and recall were computed. Accuracy represents the ratio of correctly classified instances to the total number of instances. Precision measures the proportion of true positive predictions among instances predicted as positive. Recall calculates the ratio of true positive predictions among actual positive instances. The F1-Score provides a balanced measure of model performance, taking the harmonic mean of precision and recall. The following equations represent the four evaluation metrics:

$$Accuracy : \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision : \frac{TP}{TP + FP} \tag{9}$$

$$Recall : \frac{TP}{TP + FN} \tag{10}$$

$$F1 - Score : 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{11}$$

[3]https://radimrehurek.com/gensim/models/word2vec.html
[4]https://github.com/mmihaltz/word2vec-GoogleNews-vectors
[5]https://github.com/explosion/sense2vec
[6]https://huggingface.co/yangheng/deberta-v3-large-absa-v1.1
[7]https://huggingface.co/microsoft/deberta-v3-base
[8]https://huggingface.co/microsoft/deberta-v3-large

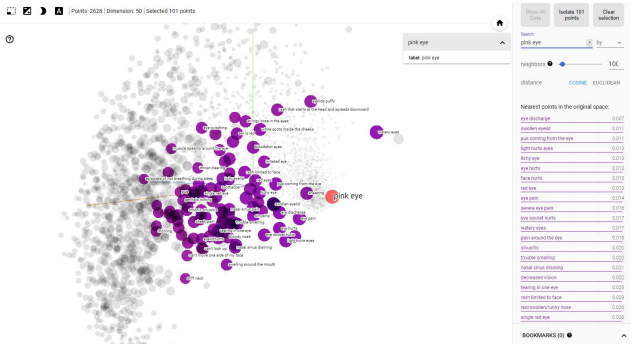

FIGURE 6. Visualize Symptom2Vec in three dimensions using Google's Embedding Projector (Pink eye).

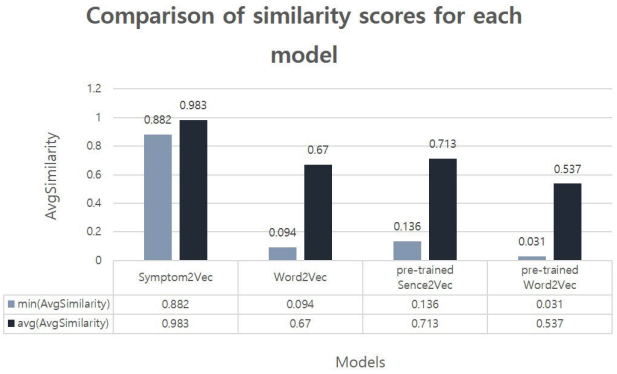

FIGURE 7. Comparison of similarity scores for each model.

V. PERFORMANCE EVALUATION

Fig. 6 shows what happens when "pink eye" is provided as an input to Symptom2Vec. The numerical values on the right side of the figure represent the cosine distance between the reference word and the comparison words. These distances are arranged in ascending order, indicating the proximity

Model name	Accuracy	F1-Score	Precision	Recall	Class 0 Precision	Class 0 Recall	Class 0 F1-Score	Class 1 Precision	Class 1 Recall	Class 1 F1-Score	Class 2 Precision	Class 2 Recall	Class 2 F1-Score
Softmax Regression	0.9456	0.9458	0.9452	0.9466	0.95	0.95	0.95	0.95	0.93	0.94	0.94	0.96	0.95
SVM	0.9693	0.9694	0.9688	0.9707	0.96	0.98	0.97	0.99	0.94	0.96	0.96	0.99	0.97
LSTM	0.97	0.97	0.97	0.97	0.96	0.98	0.97	0.97	0.95	0.96	0.97	0.97	0.97
Random Forest	0.9716	0.9715	0.971	0.9721	0.97	0.99	0.98	0.98	0.95	0.97	0.96	0.98	0.97
BERT (base)	0.9752	0.9752	0.9755	0.9752	0.97	0.99	0.98	0.99	0.96	0.97	0.96	0.98	0.97
AMoRSD (large)	0.9811	0.9811	0.9814	0.9811	0.98	0.99	0.98	1.00	0.96	0.98	0.97	0.99	0.98
AMoRSD (base)	**0.9823**	**0.9823**	**0.9825**	**0.9823**	**0.97**	**0.99**	**0.98**	**1.00**	**0.97**	**0.98**	**0.98**	**0.99**	**0.98**

FIGURE 8. Classification report for each model.

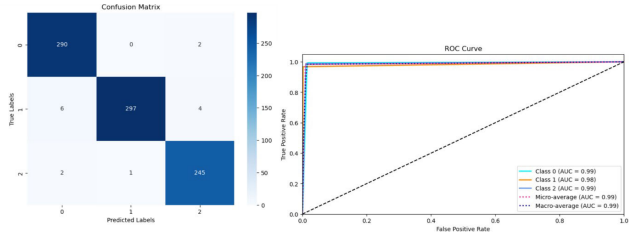

FIGURE 9. Confusion matrix and ROC curve graphs in AMoRSD (base).

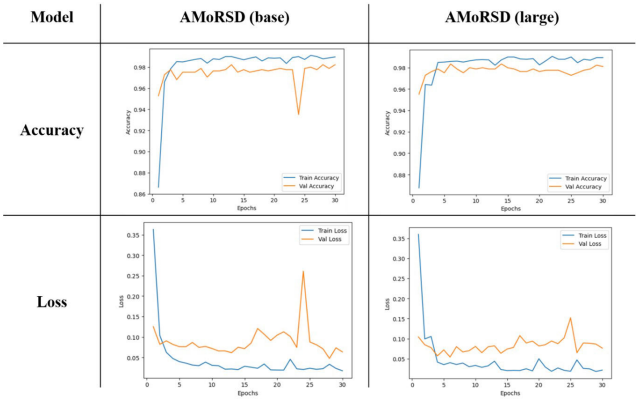

FIGURE 10. Learning graphs of AMoRSD (base) and AMoRSD (large).

between the words. The left side of the figure, which shows the vector visualization, reveals that various eye-related symptoms, including ''eye discharge'' and ''swollen eyelid,'' are positioned in close proximity to pink eye. This proximity suggests a concentration of similar symptoms. In essence, individuals exhibiting symptoms located nearby are more likely to have the same disease. Consequently, by calculating the symptom distances for each disease, it becomes possible to determine the average density that defines a particular disease.

Figure 7 illustrates a comparison of symptom similarity scores between Symptom2Vec and other models. Word2Vec encountered a substantial number of instances (1,755 cases) where symptoms were not adequately expressed, despite training with DESC. Therefore, additional comparisons were conducted using pre-trained Word2Vec. The model had a vector size of 300 and had been trained using Google News and encompassed approximately 3 million embedded words. However, this lacked the inclusion of symptom-related words. When comparing symptom similarities, it was observed that the avg(AvgSimilarity) was relatively low at 0.537. Another pre-trained model, Sense2Vec, exhibited slightly improved performance compared to other models, but the min(AvgSimilarity) remained low at 0.136. The proposed Symptom2Vec model in this study demonstrates superior performance compared to other models in exploring similar symptoms for generating appropriate follow-up questions based on a patient's symptoms. This model effectively captures the relationship between symptoms and diseases, thereby enhancing expressiveness and accuracy.

Figure 8 provides a summary of the Classification Report for the proposed AMoRSD model and the six alternative comparison models based on the information presented thus far in the paper. The AMoRSD model achieves the highest accuracy with a value of 0.9823. Additionally, when comparing precision, recall, and F1-score for each class, it consistently exhibits the best performance. Figure 9 depicts the Confusion Matrix and ROC Curve graph for AMoRSD (base). These results suggest that the responses to the self-diagnostic questions are closely associated with human emotions, and such association positively impact the model's performance.

A notable aspect is the comparison between AMoRSD (large) and AMoRSD (base). Figure 10 illustrates the training graphs for AMoRSD (base) and AMoRSD (large). These models differ in terms of their ''large'' and ''base'' configurations. The training graph indicates that the large model exhibits more stable learning. This can be attributed to the larger amount of pre-trained data available for the large model, which compensates for the challenge of finding direct associations in AMoRSD, where a sentiment analysis model is employed as a pre-trained model. For instance, a sentence like ''I have a headache'' might be classified as a negative emotion in the sentiment analysis model and linked to the ''pain'' class. However, when asked the question ''Do you have a headache?'' and the response is ''Yes,'' it becomes difficult to establish a direct association with the ''pain'' class, which is classified as a positive emotion in the sentiment analysis model. Considering all the previous evaluation metrics, it can be considered that the AMoRSD (base) model, which strikes a suitable balance between

human emotion data and self-diagnostic question responses, is the more appropriate choice.

VI. CONCLUSION

The objective of this study was to propose a deep learning-based self-diagnosis algorithm that can be used effectively in real clinical situations. To achieve this, Symptom2Vec and AMoRSD are introduced.

The research highlights the advantages of using disease and symptom data for prediction, as it significantly reduces the time and resources required for data collection and pre-processing compared to traditional methods. By relying on symptoms alone, unnecessary hospital visits can be minimized. Moreover, the use of patient symptoms for disease prediction allows for better communication with patients and potentially reduces treatment time [12].

Symptom2Vec, unlike conventional methods, learns from disease and symptom data, not patient-specific information, leading to cost-effective learning with reduced data and analysis expenses. The algorithmic criteria established in this study address inference bias issues observed in other symptom checker studies. This approach benefits patients who struggle to find appropriate medical care, doctors and residents with limited experience, or busy hospital settings with numerous patients. Collecting accurate patient data through questionnaires becomes the foundation for effective treatment.

The automatic questionnaire algorithm presented in this paper is based on a vector representation of disease symptoms, which enables the establishment of standardized criteria and procedures for medical history gathering. This opens up possibilities for various applications, including serving as an auxiliary indicator for deep learning-based disease diagnosis and supporting doctor's diagnoses.

Symptom2Vec demonstrates an average symptom similarity score of 0.983, and AMoRSD achieves an AUC of 0.99%. Therefore, it is expected that the symptom collection process using Symptom2Vec will increase the reliability of self-examination.

However, the data used in the study is limited to about 2,000 pieces of symptom information and 526 diseases. In addition, The disadvantage is that the nomenclature of the symptoms is inconsistent, which may cause some inconvenience in terms of services for users. Therefore, it is expected that a richer and more detailed symptom dimension would be created if the naming of symptoms is consistent with ICD CODE, currently known as the method to express symptoms globally. Therefore, future work needs to improve the model according to medical data standards and focus on individualized diagnosis by utilizing big data such as MIMIC utilizing ICD CODE.

REFERENCES

[1] H. G. R. Neto, M. T. Cavalcanti, and D. T. Correia, "Structured solutions for medical history taking: A historical review," *Int. J. Psychiatry*, vol. 7, no. 2, pp. 144–152, 2022.

[2] B. P. Schmitt, M. S. Kushner, and S. L. Wiener, "The diagnostic usefulness of the history of the patient with dyspnea," *J. Gen. Internal Med.*, vol. 1, no. 6, pp. 386–393, Nov. 1986.

[3] M. C. Peterson, J. M. Holbrook, D. V. Hales, N. L. Smith, and L. V. Staker, "Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses," *Western J. Med.*, vol. 156, no. 2, pp. 163–165, 1992.

[4] S. Ramani, "Promoting the art of history taking," *Med. Teacher*, vol. 26, no. 4, pp. 374–376, Jun. 2004.

[5] N. E. Biederwolf, "A proposed evidence-based shoulder special testing examination algorithm: Clinical utility based on a systematic review of the literature," *Int. J. Sports Phys. Therapy*, vol. 8, no. 4, pp. 427–440, Aug. 2013.

[6] J. H. Park and H. J. Lee, "Clinical nurses' knowledge and educational needs about dizziness," *J. Korean Biol. Nursing Sci.*, vol. 21, no. 4, pp. 259–265, Jan. 2017.

[7] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," *Healthcare*, vol. 10, no. 3, p. 541, Mar. 2022.

[8] M. Nagata, K. Takai, K. Yasuda, P. Heracleous, and A. Yoneyama, "Prediction models for risk of type-2 diabetes using health claims," in *Proc. BioNLP workshop*, 2018, pp. 172–176.

[9] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2Vec: A personalized interpretable deep representation of the longitudinal electronic health record," *IEEE Access*, vol. 6, pp. 65333–65346, 2018.

[10] V. Chandrasekhar, "Disease2Vec: A method determining disease from gut microbiome using neural embeddings, Ph.D. thesis, Dept. Extension Stud., Harvard Univ., Cambridge, MA, USA, 2020. [Online]. Available: https://www.proquest.com/docview/2562816787/abstract/B59989D2FDEF4CDEPQ/1?accountid=11283

[11] D. Gligorijevic, J. Stojanovic, and Z. Obradovic, "Disease types discovery from a large database of inpatient records: A sepsis study," *Methods*, vol. 111, pp. 45–55, Dec. 2016.

[12] S. Roy, T. Meena, and S.-J. Lim, "Demystifying supervised learning in Healthcare 4.0: A new reality of transforming diagnostic medicine," *Diagnostics*, vol. 12, no. 10, p. 2549, Oct. 2022.

[13] G. A. Beathard, "An algorithm for the physical examination of early fistula failure," in *Seminars in Dialysis*, vol. 18. Hoboken, NJ, USA: Wiley, 2005, pp. 331–335.

[14] M. Siebelt, D. Das, A. Van Den Moosdijk, T. Warren, P. Van Der Putten, and W. Van Der Weegen, "Machine learning algorithms trained with pre-hospital acquired history-taking data can accurately differentiate diagnoses in patients with hip complaints," *Acta Orthopaedica*, vol. 92, no. 3, pp. 254–257, May 2021.

[15] J. M. Fritz, G. P. Brennan, S. N. Clifford, S. J. Hunter, and A. Thackeray, "An examination of the reliability of a classification algorithm for subgrouping patients with low back pain," *Spine*, vol. 31, no. 1, pp. 77–82, Jan. 2006.

[16] J. Gillis, "The history of the patient history since 1850," *Bull. Hist. Med.*, vol. 80, no. 3, pp. 490–512, 2006.

[17] H. E. Garrett and M. R. Schneck, "A study of the discriminative value of the woodworth personal data sheet," *J. Gen. Psychol.*, vol. 1, nos. 3–4, pp. 459–471, Jul. 1928.

[18] L. E. Emerson, "A psychoanalytic study of a severe case of hysteria (concluded)," *J. Abnormal Psychol.*, vol. 8, no. 3, pp. 180–207, Aug. 1913.

[19] S. J. Reiser, *Medicine and the Reign of Technology*, Cambridge, U.K.: Cambridge Univ. Press, 1981.

[20] J. D. Stoeckle and J. A. Billings, "A history of history-taking," *J. Gen. Internal Med.*, vol. 2, no. 2, pp. 119–127, 1987.

[21] H. Kyrk and K. Lewin, "Resolving social conflicts: Selected papers on group dynamics," *Social Service Rev.*, vol. 22, no. 3, pp. 405–406, Sep. 1948.

[22] J. D. Stoeckle, "The role of academician as a teacher of patient care," *Bull. New York Acad. Med.*, vol. 61, no. 2, p. 144, 1985.

[23] W. Xue, Y. Sun, and Y. Lu, "Research and application of data mining in traditional Chinese medical clinic diagnosis," in *Proc. 8th Int. Conf. Signal Process.*, vol. 4. IEEE, 2006.

[24] F. Deutch and W. F. Murphy, *The Clinical Interview—Volume 1: Diagnosis, A Method Of Teaching Associative Exploratione*. New York, NY, USA: Guilford Press, 1954. [Online]. Available: https://www.proquest.com/docview/1308900199/citation/79E33A8E911F4E37PQ/1?accountid=11283

[25] A. Baker, Y. Perov, K. Middleton, J. Baxter, D. Mullarkey, D. Sangar, M. Butt, A. DoRosario, and S. Johri, "A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis," *Frontiers Artif. Intell.*, vol. 3, Nov. 2020, Art. no. 543405.
[26] S. Rao, V. Narayanaswamy, M. Esposito, J. Thiagarajan, and A. Spanias, "Deep learning with hyper-parameter tuning for COVID-19 cough detection," in *Proc. 12th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Jul. 2021, pp. 1–5.
[27] M. D. Brannock, R. F. Chew, A. J. Preiss, E. C. Hadley, S. Redfield, J. A. McMurry, P. J. Leese, A. T. Girvin, M. Crosskey, A. G. Zhou, R. A. Moffitt, M. J. Funk, E. R. Pfaff, M. A. Haendel, and C. G. Chute, "Long COVID risk and pre-COVID vaccination in an EHR-based cohort study from the RECOVER program," *Nature Commun.*, vol. 14, no. 1, p. 2914, May 2023.
[28] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-W.-H. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV—A freely accessible electronic health record dataset," *Sci. Data*, vol. 10, p. 1, Jan. 2023.
[29] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, Apr. 2023.
[30] Z. Jie, Z. Zhiying, and L. Li, "A meta-analysis of Watson for oncology in clinical application," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, Mar. 2021.
[31] Y. Goldberg and O. Levy, "Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method," 2014, *arXiv:1402.3722*.
[32] G. Gagliardi, "Natural language processing techniques for studying language in pathological ageing: A scoping review," *Int. J. Lang. Commun. Disorders*, Mar. 2023.
[33] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023.
[34] C. Alt, M. Hübner, and L. Hennig, "Improving relation extraction by pre-trained language representations," 2019, *arXiv:1906.03088*.
[35] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: A narrative review," *NPJ Digit. Med.*, vol. 5, no. 1, p. 46, Apr. 2022.
[36] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, A. D. Shah, K. Morley, D. Osborn, J. Hayes, R. Stewart, J. Downs, W. Chapman, and R. Dutta, "Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances," *J. Biomed. Informat.*, vol. 88, pp. 11–19, Dec. 2018.
[37] R. Tinn, H. Cheng, Y. Gu, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Fine-tuning large neural language models for biomedical natural language processing," *Patterns*, vol. 4, no. 4, Apr. 2023, Art. no. 100729.
[38] A. Turchin, S. Masharsky, and M. Zitnik, "Comparison of BERT implementations for natural language processing of narrative medical documents," *Informat. Med. Unlocked*, vol. 36, Jan. 2023, Art. no. 101139.
[39] 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, Aug. 2022, Art. no. 13847.
[40] Y. W. Sung, D. S. Park, and C. G. Kim, "A study of BERT-based classification performance of text-based health counseling data," *Comput. Model. Eng. Sci.*, vol. 135, no. 1, pp. 795–808, 2023.
[41] A. Trask, P. Michalak, and J. Liu, "sense2vec—A fast and accurate method for word sense disambiguation in neural word embeddings, 2015, *arXiv:1511.06388*.

MINJI KIM received the B.S. degree in smart system software engineering from Hyupsung University, South Korea, in 2021, and the M.S. degree in computer engineering from Hanyang University, Seoul, South Korea, in 2023. From 2017 to 2023, she participated in various computer vision projects, application development projects, and natural language processing projects. Her current research interests include machine learning, deep learning, natural language processing, sentiment analysis, data mining, and computer vision.

INWHEE JOE received the B.S. and M.S. degrees in electronics engineering from Hanyang University, Seoul, South Korea, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1998. Since 2002, he has been a Faculty Member of the Division of Computer Science and Engineering, Hanyang University. His current research interests include mobile internet, cellular systems, and PCS, wireless sensor networks, mobile ad-hoc networks, multimedia networking, and performance evaluation.