BMC Medical Informatics and Decision Making

RESEARCH

Open Access

# Knowledge-point classification using simple LSTM-based and siamese-based networks for virtual patient simulation

Yih-Lon Lin[1], Yu-Min Chiang[2*], Tsuen-Chiuan Tsai[3] and Sheng-Gui Su[1]

## Abstract

**Background** In medical education, enhancing thinking skills is vital. The Virtual Diagnosis and Treatment Platform (VP) refines medical students' diagnostic abilities through interactive patient interviews (simulated patient interactions). By analyzing the questions asked during these interviews, the VP evaluates students' aptitude in medical history inquiries, offering insights into their thinking capabilities. This study aimed to extract insights from case summaries and patient interviews to improve evaluation and feedback in medical education.

**Methods** This study employs a systematic approach to knowledge-point classification by utilizing both simple long short-term memory (LSTM)-based and Siamese-based networks, coupled with cross-validation techniques. The dataset under scrutiny originates from the "Clinical Diagnosis and Treatment Skills Competitions" spanning the first to third years in Taiwan. The methodology involves generating knowledge points from sequential questions posed during case summaries and patient interviews. These knowledge points are then subjected to classification using the designated neural network architectures.

**Results** The experimental findings reveal promising outcomes, particularly when the Siamese-based network is used for knowledge-point classification. Through repeated (stratified) 10-fold cross validation, the accuracies achieved consistently exceeded 93%, with a standard deviation less than 0.007. These results underscore the efficacy of the proposed methodologies in enhancing virtual clinical diagnosis systems.

**Conclusions** This study underscores the viability of leveraging advanced neural network architectures, particularly the Siamese-based network, for knowledge-point classification within virtual clinical diagnosis systems. By effectively discerning and classifying knowledge points derived from case summaries and patient interviews, these systems offer invaluable insights into students' thinking capabilities in medical education. The robust accuracies attained through cross-validation affirm the feasibility and efficacy of the proposed methodologies, thus paving the way for enhanced virtual clinical training platforms.

**Keywords** Knowledge points, Siamese networks, Convolutional neural network (CNN), Long short-term memory (LSTM), K-fold cross-validation

*Correspondence:
Yu-Min Chiang
ymchiang@nfu.edu.tw
[1]Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Yunlin, Taiwan
[2]Department of Automation Engineering, National Formosa University, No.64, Wunhua Rd., 632 Huwei Township, Yunlin, Taiwan
[3]Landseed International Hospital, Taoyuan City, Taiwan

## Background

Medical education training allows physicians to conduct patient interviews based on case summaries, in which medical history inquiry and evaluation are indicators that can be used to evaluate medical decision-making. Mistakes in medical decision-making will lead to medical disputes and waste of medical resources. Improving medical history inquiry skills helps to improve physicians' professional ability [1].

In recent years, applications of virtual patient simulation have been proposed [2–4]. Furlan et al. [2] proposed the Hepius Virtual Patient Simulators (VPSs), which use the Siamese Long Short-Term Memory (SLSTM) model. Hepius compares student questions with database questions, ranking them based on similarity using the SLSTM model's output probabilities. The top three ranked questions' responses are determined using a threshold, with the SLSTM model employing two identical LSTM networks to process the input questions. The similarity between question features is measured using the Manhattan distance. Persad et al. [3] developed a VP simulator capable of identifying a predefined set of free-text questions using concept recognition and natural language processing (NLP). They emphasized that the natural-language question-answering component was capable of providing an appropriate response. Campillos-Llanos et al. [4] examined the development of a dialogue system in Spanish that was specifically designed for medical students to enhance their communication and anamnesis abilities. The system effectively replicates interactions with virtual patients and offers a detailed overview of its characteristics, with a particular emphasis on the Spanish components.

Similar to virtual patient simulations in medical education, question-answering systems have been widely applied in various fields to process large amounts of text data and automate responses. These systems not only enhance students' medical history inquiry skills but also strengthen interactions with virtual patients through natural language processing techniques. As demonstrated in [5] and [6], question-answering systems have shown their potential in question classification and answer retrieval when dealing with large-scale datasets during disasters, and these techniques can also be applied to medical decision-making training in virtual patient simulations. The authors of [5] introduced a question-answering (QA) system framework based on a Twitter dataset of more than nine million tweets during the Osaka North Earthquake in June 2018. Their research delved into the question structure and developed techniques to classify and retrieve answers from the dataset by leveraging ontology, word similarity, keyword frequency, and natural language processing. Kolomiyets and Moens [6] offered a comprehensive and comparative review of question-answering technology, highlighting the significance of retrieval models in the question-answering process. Their models encompass query and information document representations, along with retrieval functions that assess the relevance between a query and a potential answer.

In the field of ICD coding, numerous models have been explored to address the challenges of automatic coding and knowledge-point classification in medical training. Yuan et al. [7] proposed a multiple synonyms matching network that uses a shared LSTM to encode electronic medical record (EMR) texts and ICD code synonyms, emphasizing the importance of synonym matching. Similarly, Yang et al. [8] adopted a Transformer-based model specifically designed for processing long documents. They introduced a knowledge-injected prompt-based fine-tuning method for multi-label few-shot ICD coding, effectively leveraging domain-specific knowledge to enhance model performance in rare coding scenarios. In [9–12], several deep learning approaches were proposed for keyword and keyword extraction. Onan et al. [9] examined of the predictive performance of five statistical keyword extraction methods in the context of classification algorithms and ensemble methods for categorizing scientific text documents. Wu et al. [10] created a new probabilistic keyword extraction algorithm inspired by visual attention. By using an unsupervised neural network based pre-training method, their algorithm efficiently extracts keywords with rich contextual information from documents. Hu et al. [11] introduced the patent keyword extraction algorithm (PKEA), which is based on the distributed Skip-gram model for patent classification. Additionally, they proposed quantitative performance measures for evaluating keyword extraction, employing information gain, cross-validation, and SVM classification. Unlu and Cetin [12] introduced a model that tackles the keyword extraction problem by framing it as a sequence labeling task. They trained various classification algorithms, including Naive Bayes, Support Vector Machine, Polynomial Regression, Multi-Layer Perceptron, and Random Forest, individually within the Token Classification module of the model.

Text classification is a vital component in NLP and QA systems. Its primary role is to categorize questions based on their semantics and context, enabling the system to better understand and address different types of queries. Khilji et al. [13] developed a cooking QA system that employs deep learning techniques to contextually classify recipe questions into specific categories. The identified question class is then utilized to extract the relevant details from the recipe obtained through a rule-based approach, enabling the system to provide precise answers. In Han et al. [14], a dataset was created through an answer-driven approach, and a deep learning model with Transformer, Bi-GRU, and attention structures was

proposed. The model enhances coding for long sentences by incorporating Bi-GRU for segment interaction and emphasizes key semantics using the attention mechanism. Xia et al. [15] presented an integrated method for question classification in a Chinese cuisine question-answering system. They utilized domain knowledge to enhance question preprocessing and extracted classification features based on domain attributes to construct a rule-based classifier. Sarrouti and Alaoui [16] proposed a machine learning-based method for classifying biomedical question types, automatically assigning a category to each question. They extracted features from biomedical questions using handcrafted lexico-syntactic patterns and utilized these features in machine learning algorithms. Liu et al. [17] identified a limitation in the standard kernel function used for question classification, as it overlooks the question structure. To overcome this limitation, they proposed a question property kernel function that integrates syntactic dependency relationships and part of speech (POS) information. By introducing this kernel function, they were able to effectively leverage the question structure, resulting in improved accuracy in the classification process. A single biomedical answer or entity can be associated with multiple biomedical categories or semantic types. In the biomedical domain, a single biomedical answer or entity can be linked to multiple biomedical categories or semantic types. As highlighted by the authors [18], this inherent characteristic of biomedical entities transforms question classification in the biomedical field into a multi-label classification problem. This means that a question may anticipate answers that fall under multiple semantic types. Mallikarjuna and Sivanesan [19] proposed a question classification approach utilizing limited labeled data. To address the scarcity of labeled data, they employed data augmentation as a technique to generate additional training instances for question classification. To enhance the effectiveness of medical query intent classification and named entity recognition tasks. Tohti et al. [20] introduced a multi-task learning model, based on ALBERT-BILSTM, for the dual purposes of named entity recognition and intent classification in the context of Chinese online medical questions.

ChatGPT was released by OpenAI in 2021 (https://chat.openai.com/). It is a sibling model of InstructGPT, which is trained to efficiently respond to prompts by following instructions and providing detailed information. To examine the impact of ChatGPT on teaching and learning, it is necessary to gain insights into students' perceptions and evaluate its potential and challenges. Shoufan [21] sheds light on the capabilities and limitations of ChatGPT in an educational context, serving as a valuable resource for future research and development. Maddigan and Susnjak [22] explored the use of advanced pre-trained language models such as ChatGPT and GPT-3 to convert natural language into code for visualizations. The proposed system, Chat2VIS, leverages these models and showcases the benefits of prompt engineering in improving language understanding for more efficient and accurate end-to-end solutions. In this study, the simple LSTM-based and Siamese-based networks were proposed for knowledge-point classification. Knowledge-point classification helps design more targeted training case scenarios in virtual patient systems, ensuring that students acquire essential skills and knowledge.

Schmidgall [23] introduced a multimodal benchmarking tool designed to evaluate the performance of language models in simulated clinical environments. This tool enhances the realism of simulated clinical scenarios but also highlights the challenges and limitations that current language models face in real-world medical applications. Mehandru [24] explored the potential of large language models (LLMs) as intelligent agents in clinical settings. However, these challenges indicate that applying LLMs in clinical environments requires addressing multiple issues, particularly concerning the reliability, safety, and accuracy of the models.

## VP system and problem statement

In this study, we focused on interactive computer simulations of natural language processing (NLP) on virtual reality standardized patients.

### Virtual patient system

A virtual patient (VP) system for clinical training is usually designed to simulate interactively the real clinical processes, which can be used in health care education to train students. Medical students can use the VP system to practice diagnoses and treatments, as shown in Fig. 1. In some VP systems, e.g., the virtual standardized patients (https://www.ennovamed.com/), the main tool employed is rule-based keyword matching. However, exact matching usually yields worse prediction accuracy. This motivates us to consider knowledge-point classification using deep learning techniques. In general, an ideal VP should encompass (1) a case summary, (2) a patient interview, (3) a physical/neurological examination, (4) a quick test, (5) an initial diagnosis, (6) a laboratory test, (7) radiology and (8) a final diagnosis. This is illustrated in Table 1. Among these, the case summary and patient interview are the most pivotal elements for medical students' analytical skills. Incorrect patient interviews can lead to erroneous medical decisions and impact healthcare quality. The main goal of this study is to generate appropriate knowledge points based on case summaries and patient interviews. By predicting potential knowledge point categories based on the case summary and user question sentences, the VP system can map these knowledge points


**Fig. 1** VP system (https://www.ennovamed.com/)

**Table 1** The workflow of the diagnosis system from the virtual patient simulation system

| Input simulation system | Outcome of report |
|---|---|
| 1. **Case summary** | 1. Diagnosis result |
| 2. **Patient interview(QA)** | 2. Learning score |
| 3. Physical/neurological examination | 3. Learning record |
| 4. Quick test | 4. Health insurance points |
| 5. Initial diagnosis | 5. Feedback in learning |
| 6. Laboratory test | |
| 7. Radiology | |
| 8. Final diagnosis (Submit) | |

to a scoring mechanism for evaluating the user's medical thinking capabilities, as shown in Fig. 2.

### Dataset description

A case example for the VP system is shown in Table 2. In Table 2, each case includes structured and unstructured data. The dataset used in this study was based on the 2018 and 2019 datasets of the "National Clinical Diagnosis and Treatment Skills Competitions (NCDTSC)" in Taiwan. The establishment of the dataset occurs annually within each competition, involving annotated simulated cases by medical education experts and the collection of authentic dialogues from participants, who are medical students and also candidates for becoming a doctor. The collection of data is a process that requires gradual accumulation over time. Our dataset was provided by Innova Medical Technology (IMT) Corporation (https://www.ennovamed.com/). Currently, we have a total of 16 virtual patient cases and 184 knowledge points together with their corresponding response sentences. Please note that the original interview (QA) material is in Traditional

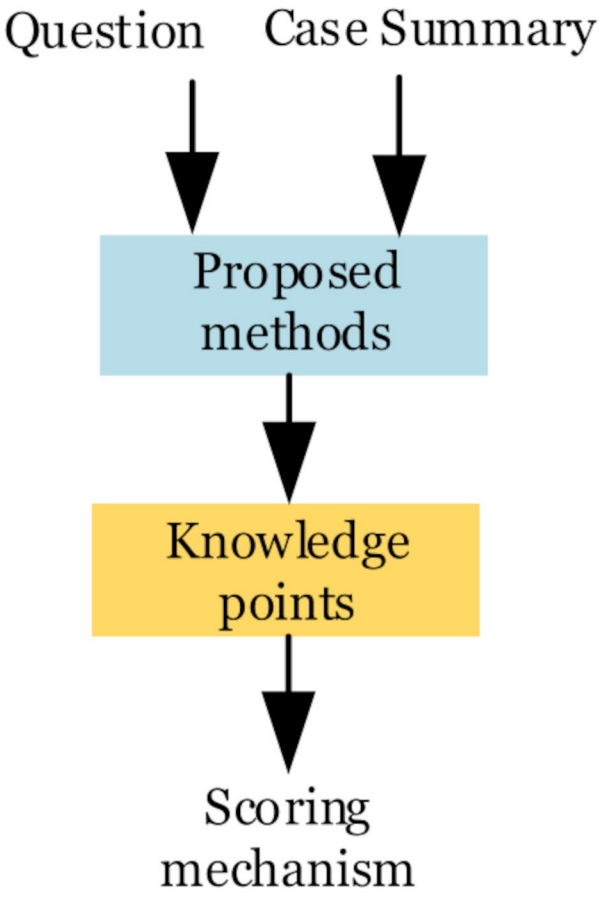


**Fig. 2** The purpose of the proposed methods

**Table 2** A case example for VP system

| **Input Pattern** | | | |
|---|---|---|---|
| Item | Case Summary | | Patient Interview |
| | Structured data | Unstructured data | QA sentences (Q: from participant; A: from virtual patient) |
| Statement | "Age": 58<br>"Sex": female<br>"Height":158<br>"Weight": 54<br>"Blood Oxygen": 98<br>"Blood Pressure": 138 – 85 "Respiratory Rate": 18<br>"SpO2": 98<br>"Temperature": 38.3 | Ms. Chen began to feel very uncomfortable in her stomach three days ago. Yesterday, she first went to the outpatient department of An-Kang Hospital, and the doctor asked her to come over for an examination. | Q: Hello.<br>A: Hello doctor.<br>Q: What's the matter with you?<br>A: I have a bit of stomachache.<br>… |
| **Output Indicator** | | | |
| **Knowledge- point classification** (Labeling from medical experts) | Knowledge point (1 ~ 184):<br>• Onset of the illness<br>• Duration of the illness<br>• ….<br>• The allergy history<br>• The medication history | | |

**Table 3** An illustrative knowledge point on "seeking healthcare"

| **Knowledge point** | **Possible questions** |
|---|---|
| Seeking for healthcare | • Have you been to other hospitals for treatment?<br>• Have you ever sought medical care at different healthcare facilities?<br>• Have you previously sought treatment at different hospitals?<br>• Did you receive medical services at other hospitals?<br>• Do you have any experience with medical care at other healthcare centers?<br>• Have you consulted doctors at different hospitals before?<br>• Before seeking the current medical treatment, have you been to other hospitals for medical care?<br>• Have you sought medical advice at other healthcare institutions before?<br>• Have you received treatment at other hospitals prior to this?"<br>• Have you ever had any medical experiences at other hospitals?<br>• … |

Chinese. The English version of the QA, along with the labeling of knowledge points, is presented in File 1. An illustrative knowledge point on "seeking healthcare" is shown in Table 3. The symptoms and final diagnoses of the 16 virtual patient cases are presented in Appendix 1. This study specifically concentrates on two evaluation indicators of the participants, "case summary" and "patient interview", within the dialogues between users (students) and system responses.

### Model description

LSTM-based classification leverages Long Short-Term Memory (LSTM) networks to process sequential data, making it well-suited for tasks involving time series or natural language. In our study, this aligns closely with the requirements of the simple QA system, where learning patterns from sequential inputs is critical for accurate knowledge point classification. On the other hand, Siamese-based networks consist of two neural networks with shared weights, designed to compare two inputs and primarily used for similarity measurement tasks, such as image or text matching. This approach directly supports our goal of classifying knowledge points by emphasizing the similarity between inputs, such as identifying whether two questions belong to the same knowledge point, rather than relying solely on direct categorization.

For knowledge-point classification, the two proposed models are simple LSTM-based and Siamese-based networks. These models are constructed by combining three primary layers: dense layers, convolutional layers, and LSTM layers. In this paper, the simple LSTM-based network integrates some dense layers and LSTM layers, whereas the Siamese-based network contains, in addition, a Siamese network. They are briefly reviewed in the following.

(a) Convolutional layer: The convolutional layer involves convolution operations of input patches with masking filters (also called kernels). The input features of a layer are transformed to output features via a convolutional layer. Many convolutional layers can be cascaded within a deep neural network.
(b) LSTM layer: The LSTM layer comprises a series of recursively connected blocks, known as memory blocks. The question answering problem involves time-ordered data.
(c) Siamese network: The standard Siamese neural network is used to measure the similarity between two inputs. The two inputs are processed through two separate neural networks. Each network transforms its inputs into a new space, resulting in new features that can be viewed as the representations of the inputs within this new space. Then, an energy function can be used to measure the similarity of the new features.

### Proposed methods

It is well known that correctly mapped questions lead to accurate knowledge points, which in turn contribute to effective questioning skills. The proposed model is trained using data from case summaries and patient interviews.

In this study, we use simple LSTM-based and Siamese-based networks to implement knowledge-point classification. The steps of the proposed method are listed as follows:

**Step 1** Defining knowledge points.
The definition of knowledge points was manually marked by the physician team of the second author. The compilation of each case was evaluated and confirmed by more than three doctors, especially the weights of knowledge points.

**Step 2** Collecting the corpus and building the vocabularies.
All the datasets were obtained from the "National Clinical Diagnosis and Treatment Skills Competitions" in Taiwan for the first three years. The candidates (users) for each competition include interns or nurses. In the process of education learning and examining, candidates showed pleasant attitudes in the education learning process and were very cautious in asking and answering questions on the exam. Therefore, the reliability of the corpus data is high, reflecting realistic questioning situations. We collected all vocabularies from non-structured data of case summaries and questions for QA dialogue.

**Step 3** Building and training the proposed models.
Because the QA dialogue is a type of time-ordered data, we built a simple LSTM-based network to implement knowledge-point classification. Because the data size is small, we also constructed a Siamese network to address this problem. Then, the proposed models were trained using the data from case summaries and patient interviews.

**Step 4** Testing the model performance.
The performances of the simple LSTM-based and Siamese-based network models are evaluated and compared.

Medical school students learn from patients who act as their educators or teachers, and they also acquire knowledge by studying patient cases. In this study, we collected 16 individual cases. In the following, we describe the details of our proposed simple LSTM-based and Siamese-based networks.

#### Simple LSTM-based network

The simple LSTM-based network, shown in Fig. 3, includes a word embedding, a concatenate, two LSTM, and some dense layers. The outputs (1) of model A and (2) of the model B are extracted from the "case no" and "question" sentences inputs, respectively. In model-1, the input is $1 \times 16$, representing 16 case no., and includes a single fully connected layer. In contrast, model-2 has a $1 \times 20$ input, representing the maximum token length, and comprises an embedding layer, a dropout layer, two LSTM layers, and a fully connected layer. The transformation from $20 \times 512$ to $1 \times 1024$ is performed by an LSTM layer. This LSTM layer processes the $20 \times 512$ input over 20 time steps (corresponding to the sentence length of 20) and produces a $1 \times 1024$ output. The LSTM captures the contextual information from the entire sequence and converts it into a fixed-length 1024-dimensional vector representation.

#### Siamese-based network

The proposed Siamese-based network, shown in Fig. 4, includes a word embedding, a concatenation, two LSTM, and some dense layers. The two leftmost building blocks (pretrained model-1 and pretrained model-2) constitute the pretrained models. Model-1 is extracted from a pretrained autoencoder model (not shown in Fig. 4). Model-2 is extracted from two leftmost portions, i.e., the "case" model and "question" model, of the pretrained simple LSTM-based network in Fig. 3. The Manhattan distance is adopted as the energy function of the Siamese network to calculate the similarity of the features produced by the Siamese network. The output of the whole network is the similarity between the output features of model-1 and model-2. The activation function of the last dense layer is the sigmoid function, which produces

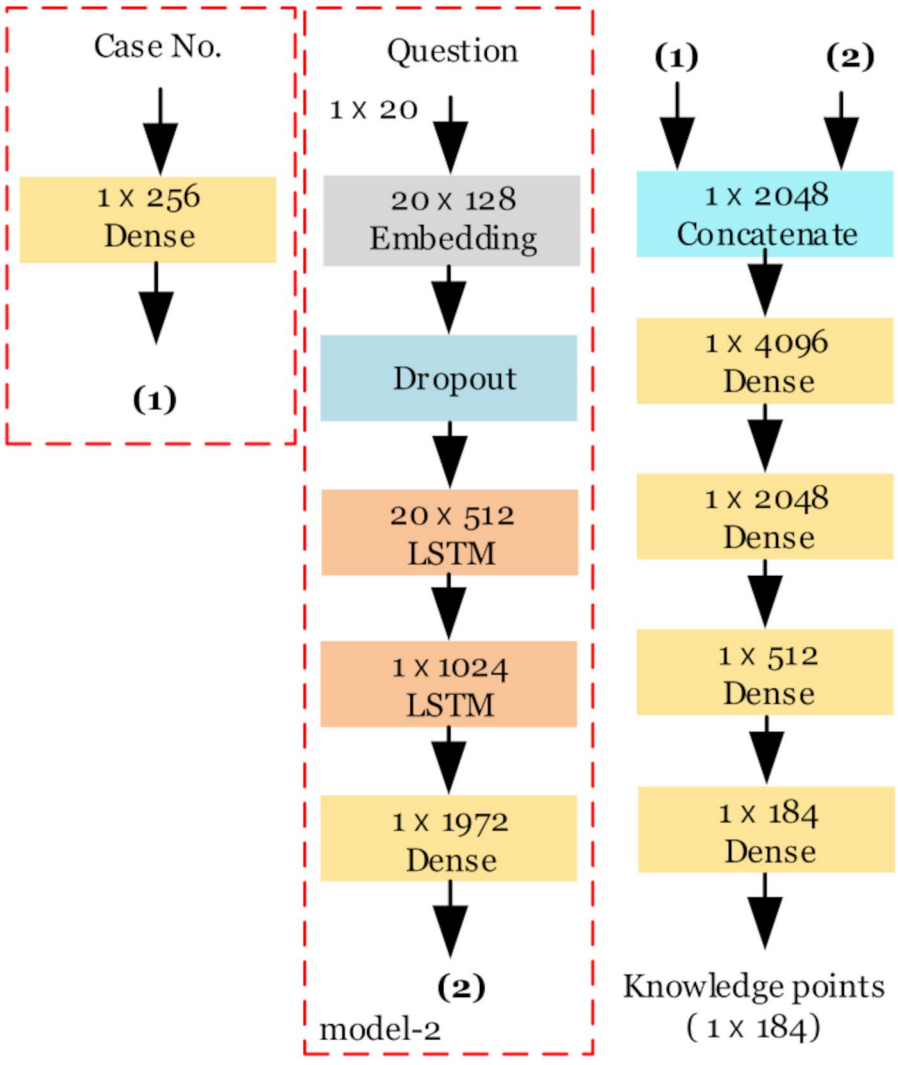


**Fig. 3** Simple LSTM-based network

values between 0 and 1. If the similarity value is greater than or equal to the threshold of 0.5, the output label is 1, indicating the matching of the case and the underlying knowledge point. On the other hand, if the similarity value is less than the threshold of 0.5, the output label is 0, indicating the mismatching of the case and the underlying knowledge point.

## Results

The description of the original datasets is listed in Table 4. The training dataset was collected from the NCDTSC in Taiwan for the first and second years, and the testing dataset was from the third year. The simple LSTM-based network was trained on the training dataset. According to the mechanism of the Siamese network, we have to randomly generate the positive and negative cases. The positive cases are data pairs from the same category, whereas the negative cases are data pairs from different categories. The description of the datasets used for the Siamese network is listed in Table 5.

The learning parameters of the proposed simple LSTM-based and Siamese-based networks are listed in Table 6. In the third column, the trainable parameters and non-trainable parameters of the Siamese network are 52,589,441 and 10,008,832, respectively.

All the text features were extract was from the NCDTSC data for the first three years. According to Table 7, the vocabulary size and number of keywords are 184 and 1,410, respectively.

For a fair evaluation of the trained models, we employed 10-fold cross validation. The cost to be minimized is the categorical cross-entropy for multi-class classification problems, and the performance metric for cross validation is the accuracy of the testing data.

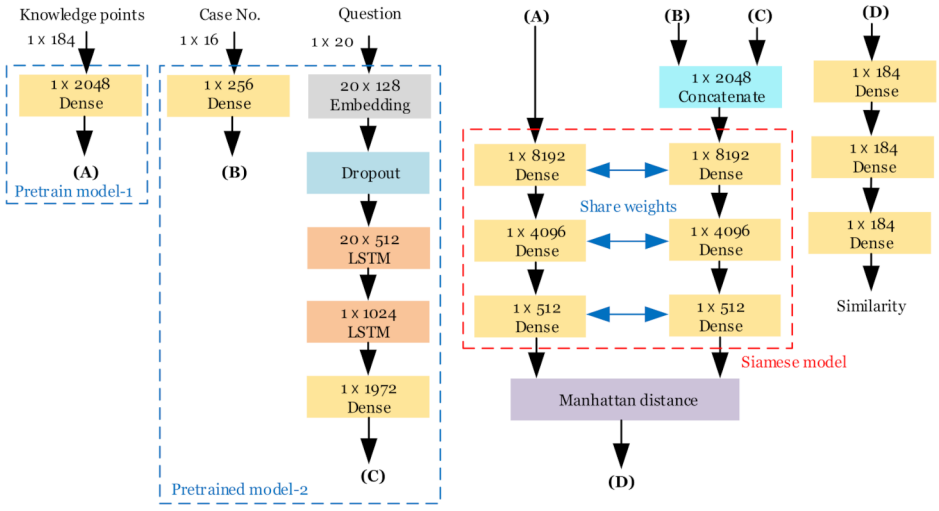


**Fig. 4** Siamese-based network

**Table 4** Description of the original datasets

| | Training dataset (The first and second sessions) | Testing dataset (The third session) |
|---|---|---|
| # of instances | 3247 | 217 |

**Table 5** Description of the datasets for siamese network

| | Training dataset | Testing dataset |
|---|---|---|
| # of instances | (3247, 6494)* | (217, 434) |

*(No. of positive cases, No. of negative cases)

**Table 6** The learning parameters of the simple LSTM-based and siamese-based networks

| Parameters | Simple LSTM-based network | Siamese-based network |
|---|---|---|
| Trainable | 27,556,792 | 52,589,441 |
| Non-trainable | 0 | 10,008,832 |
| Total | 27,556,792 | 62,598,273 |

**Table 7** Description of the parameters of text feature extraction

| Item | Value |
|---|---|
| Input length | 20 |
| Length of word vector | 128 |
| Output shape of word embedding | 20×128 |
| Number of cases | 16 |
| Number of keywords | 184 |
| Vocabulary size | 1,410 |

The simulation results for simple LSTM-based and Siamese-based models are listed in Table 8, with metrics including accuracy, precision, recall (sensitivity), specificity, F1-score, and AUC-ROC. The AUC-ROC metric evaluates a classification model's ability to distinguish between positive and negative classes by measuring the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate (also known as sensitivity or recall) against the false positive rate at various threshold settings. To provide a comprehensive assessment across all categories, we have now calculated the average AUC-ROC as an overall performance indicator. This approach enables us to present a single, integrated AUC-ROC, offering a clear and concise visualization of the model's classification capability across all categories. Table 8 shows that the LSTM-based model has a much higher training accuracy (0.9550) compared to testing accuracy (0.3917), indicating significant overfitting. In contrast, the Siamese-based model shows a smaller difference between training (0.9968) and testing accuracy (0.8464), suggesting only slight overfitting. In this study, the Siamese model shows a smaller gap between training and test accuracies, indicating effective reduction in overfitting when handling the virtual patient dataset and maintaining stable performance, thereby demonstrating better generalization. In comparison, the LSTM model exhibits a larger gap between training and test accuracies, suggesting that it performs well on training data but relies heavily on it, lacking robustness and generalizability. Additionally, the Siamese-based network consistently performs well across all metrics on both datasets, demonstrating greater robustness and better

**Table 8** The simulation results of the proposed models

| | Simple LSTM-based network | | Siamese-based network | |
|---|---|---|---|---|
| | Training dataset | Testing dataset | Training dataset | Testing dataset |
| Accuracy | 0.9550 | 0.3917 | 0.9968 | 0.8464 |
| Precision | 0.7824 | 0.1075 | 0.9513 | 0.8265 |
| Recall | 0.7674 | 0.0963 | 0.9393 | 0.7103 |
| F1-score | 0.7550 | 0.0933 | 0.9413 | 0.7137 |
| Specificity | 0.9622 | 0.4435 | 0.9638 | 0.9157 |
| AUC-ROC | 0.9629 | 0.2708 | 0.9959 | 0.8288 |

**Table 9** The validating accuracies of the simple LSTM-based and siamese-based networks

| **Models** | **Simple LSTM-based network with 10-fold cross validation** | **Siamese-based network with 10-fold cross validation** |
|---|---|---|
| Mean | 0.5005 | 0.9217 |
| Standard deviation | 0.1689 | 0.1081 |
| **Models** | **Simple LSTM-based network with stratified 10-fold cross validation** | **Siamese-based network with stratified 10-fold cross validation** |
| Mean | 0.4749 | 0.9372 |
| Standard deviation | 0.1452 | 0.0337 |
| **Models** | **Simple LSTM-based network with repeated 10-fold cross validation** | **Siamese-based network with repeated 10-fold cross validation** |
| Mean | 0.5477 | 0.9661 |
| Standard deviation | 0.0696 | 0.0053 |
| **Models** | **Simple LSTM-based network with repeated stratified 10-fold cross validation** | **Siamese-based network with repeated stratified 10-fold cross validation** |
| Mean | 0.5306 | 0.9638 |
| Standard deviation | 0.1126 | 0.0063 |

generalization to unseen data. This clearly shows that the Siamese-based model is superior and more reliable for the knowledge-point classification problem under consideration.

It is interesting to know the cross-validated estimates of the accuracy for both proposed models. Cross validation usually provides a more a reliable estimate of the metric under consideration. This study employs 10-fold cross-validation using data from the first two years, with 9/10 allocated for training and 1/10 for validation. To assess the model's generalization ability, data collected from the third year is utilized as an independent testing dataset. Here, accuracy estimates for both proposed models are obtained through four popular kinds of cross validation, namely, 10-fold cross validation, repeated 10-fold cross validation, stratified 10-fold cross validation, and repeated stratified 10-fold cross validation. The repetition number is 4. The simulation results are shown in Table 9. From this table, it can be seen that the estimated validation accuracies (respectively, standard deviation) using the Siamese-based model are much larger (respectively, smaller) than those using the simple LSTM-based model. This shows that the Siamese-based model is more robust and accurate than the simple LSTM-based model.

### Limitations

This study aims to improve the accuracy of virtual patient dialogue systems. However, several limitations remain: First, the training dialogue corpus used in this study is primarily sourced from the annual "Clinical Diagnosis and Treatment Skills Competitions," and compiling a comprehensive corpus will require additional time. Second, the collection, organization, and annotation of the corpus demand significant investments of manpower and time. Third, the platform's effectiveness depends heavily on the user's professional background. Finally, the dialogue data in this study is limited to applications within Chinese-language contexts.

## Conclusion

The VP system is designed to enhance thinking and reasoning skills for students in medical schools. The skill of posing pertinent questions is an important part of medical education. This paper emphasizes the importance of mapping from QA dialogues conducted to knowledge points. In this study, the simple LSTM-based and Siamese-based networks were proposed to address the knowledge-point classification problem. The accuracy estimates were obtained utilizing cross-validation method. The simulation results demonstrated that the Siamese-based network provides more robust and accurate estimates of accuracy than does the simple LSTM-based network. This means that the Siamese-based network is better at leveraging knowledge point matching as a classification mechanism. Furthermore, the overfitting to training data due to the many trainable parameters of the model is effectively mitigated when using the Siamese-based network.

The main contribution of this paper lies in addressing the challenge of knowledge-point classification in medical education. The challenge originates from the construction of a mapping from the QA dialogues conducted to knowledge points. It is worth of mentioning that the collection of conversational transcripts and expert annotations during the annual competition requires a substantial investment of time and money for data gathering. In the future, we expect to expand the scale of our dataset, which will significantly improve its ability to predict knowledge points. Additionally, we plan to further explore and compare alternative classification models, such as prototypical networks or matching networks.

**Abbreviations**

| | |
|---|---|
| AUC | Area under curve |
| AUC-ROC | Area under the ROC curve |
| CNN | Convolutional neural network |
| IMT | Innova medical technology |
| KP | Knowledge points |
| LSTM | Long short-term memory |
| NCDTSC | National Clinical Diagnosis and Treatment Skills Competitions |
| NLP | Natural language processing |
| QA | Question-answering |
| ROC | Receiver operating characteristic |
| VP | Virtual patient |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02866-3.

Supplementary Material 1

Supplementary Material 2

**Acknowledgements**
Note applicable.

**Author contributions**
YLL and YMC conceived of the presented idea and designed the neural network model and the computational framework. YLL designed and supervised the project. TCT contributed to sample preparation, and SGS carried out the implementation. YLL wrote the manuscript with support from YMC. All authors reviewed the manuscript.

**Funding**
The research has been supported by the National Science and Technology Council, Taiwan, under grant no. MOST 110-2622-E-224-004 and the "Intelligent Recognition Industry Service Center" from the Featured Areas Research Center-Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. The funder had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Data availability**
The datasets used and analyzed during the current study are not publicly available because IMT allows only authorized persons to access data. However, data are available from the corresponding author (YMC) upon reasonable request.

## Declarations

**Ethics approval and consent to participate**
This study complies with the principles outlined in the Declaration of Helsinki. The requirement for ethical approval has been waived by the review board of National Science and Technology Council, Taiwan. All data collection and processing adhere to the regulations of Taiwan Personal Information Protection Act. Strict measures are implemented during data collection to ensure the privacy and confidentiality of participants. All participants sign an informed consent form prior to data collection, and all personal identifying information is anonymized during data processing. The term "patient interviews" in our study refers to the interactive patient consultation simulations conducted within the Virtual Diagnosis and Treatment Platform (VP). These interviews are a built-in functionality of the VP system and were developed as part of the platform to facilitate medical training. The interview data collected were used only for this study. They are original to the platform and not based on any previously published materials, making external references inapplicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 9 July 2024 / Accepted: 13 January 2025
Published online: 24 January 2025

**References**

1. Wang SY, Chen CH, Tsai TC. Learning clinical reasoning with virtual patients. Med Educ. 2024;54(5):523–56.
2. Furlan R, Gatti M, Menè R, et al. A natural language processing-based virtual patient simulator and intelligent tutoring system for the clinical diagnostic process: simulator development and case study. IMIR Med Inf. 2021;9(4). https://doi.org/10.2196/24073.
3. Persad A, Stroulia E, Forgie S. A novel approach to virtual patient simulation using natural language processing. Med Educ. 2016;50(11):1162–3. https://doi.org/10.1111/medu.13197.
4. Campillos-Llanos L, Thomas C, Bilinski E, et al. Terminological and language resources for developing a virtual patient dialogue system in Spanish. Procesamiento del Lenguaje Nat. 2019;63:205–8.
5. Kemavuthanon K, Uchida O. Integrated question-answering system for natural disaster domains based on social media messages posted at the time of disaster. Information. 2020;11(9):456. https://doi.org/10.3390/info11090456.
6. Kolomiyets O, Moens MF. A survey on question answering technology from an information retrieval perspective. Inform Sci. 2011;181(24):5412–34.
7. Yuan Z, Tan C, Huang S. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022; pp. 808–814, Dublin, Ireland.
8. Yang Z, Wang S, Rawat BPS, Mitra A, Yu H. Knowledge injected prompt-based fine-tuning for multi-label few-shot ICD coding. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on empirical methods in natural language processing. 2022: 1767, NIH Public Access.
9. Onan A, Korukoglu S, Bulut H. Ensemble of keyword extraction methods and classifiers in text classification. Expert Syst Appl. 2016;57:232–47.
10. Wu X, Du ZK, Guo YK. A visual attention-based keyword extraction for document classification. Multimedia Tools Appl. 2018;77(19):25355–67.
11. Hu J, Li SB, Yao Y, et al. Patent keyword extraction algorithm based on distributed representation for patent classification. Entropy. 2018;20(2):104. https://doi.org/10.3390/e20020104.
12. Unlu HK, Cetin A. Keyword extraction as sequence labeling with classification algorithms. Neural Comput Appl. 2022;35(4):3413–22.
13. Khilji AFUR, Manna R, Laskar SR, et al. Question classification and answer extraction for developing a cooking QA system. Comput Y Sistemas. 2020;24(2):921–7.
14. Han DF, Tohti T, Hamdulla A. Attention-based transformer-BiGRU for question classification. Information. 2022;13(5). https://doi.org/10.3390/info13050214.
15. Xia L, Teng Z, Ren FJ. Question classification for Chinese cuisine question answering system. IEEJ Trans Electr Electron Eng. 2009;4(6):689–95.
16. Sarrouti M, El Alaoui SO. A machine learning-based method for question type classification in biomedical question answering. Methods Inf Med. 2017;56(3):209–16.
17. Liu L, Yu ZT, Guo JY, et al. Chinese question classification based on question property kernel. Int J Mach Learn Cybernet. 2014;5(5):713–20.
18. Wasim M, Mahmood W, Asim MN, Ghani MU. Multi-label question classification for factoid and list type questions in biomedical question answering. IEEE Access. 2019;7:3882–96.
19. Mallikarjuna C, Sivanesan S. Question classification using limited labelled data. Inf Process Manage. 2022;59(6). https://doi.org/10.1016/j.ipm.2022.103094.
20. Tohti T, Abdurxit M, Hamdulla A. Medical QA oriented multi-task learning model for question intent classification and named entity recognition. Information. 2022;13(12):581. https://doi.org/10.3390/info13120581.
21. Shoufan A. Exploring students's perceptions of ChatGPT: thematic analysis and follow-up survey. IEEE Access. 2023;11. https://doi.org/10.1109/ACCESS.2023.3268224.
22. Maddigan P, Susnjak T. Chat2VIS: Generating data visualizations via natural language using ChatGPT, codex and GPT-3 large language models. IEEE Access. 2023;11. https://doi.org/10.1109/ACCESS.2023.3274199.
23. Schmidgall S, Ziaei R, Harris C, Reis E, Jopling J, Moor M. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments, arXiv preprint arXiv:2405.07960.

24. Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. Npj Digit Med, 7(84): 2024.

## Publisher's note