MKeCL: Medical Knowledge-Enhanced Contrastive Learning for Few-shot Disease Diagnosis

Yutian Zhao*, **Huimin Wang***, **Xian Wu†**, **Yefeng Zheng**
Jarvis Research Center, Tencent YouTu Lab
Shenzhen, China
{yutianzhao, hmmmwang, kevinxwu, yefengzheng}@tencent.com

Abstract

Artificial intelligence (AI)-aided disease prediction has gained extensive research interest due to its capability in supporting clinical decision-making. Existing works mainly formulate disease prediction as a multi-label classification problem and use historical Electronic Medical Records (EMR) to train supervised models. However, in real-world clinics, such purely data-driven approaches pose two main challenges: 1) long tail problem: there are excessive EMRs for common diseases and insufficient EMRs for rare diseases, thus training over an imbalanced data set could result in a biased model that ignores rare diseases in diagnosis; 2) easily misdiagnosed diseases: some diseases can be easily distinguished while others sharing analogous conditions are much more difficult. General classification models without emphasizing easily misdiagnosed diseases may generate incorrect predictions. To tackle these two problems, we propose a Medical Knowledge-Enhanced Contrastive Learning (**MKeCL**) approach to disease diagnosis in this paper. MKeCL incorporates medical knowledge graphs and medical licensing exams in modeling in order to compensate for the insufficient information on rare diseases; To handle hard-to-diagnose diseases, MKeCL introduces a contrastive learning strategy to separate diseases that are easily misdiagnosed. Moreover, we establish a new benchmark, named **Jarvis-D**, which contains clinical EMRs collected from various hospitals. Experiments on real clinical EMRs show that the proposed MKeCL outperforms existing disease prediction approaches, especially in the setting of few-shot and zero-shot scenarios.

Keywords: Disease Diagnosis, Contrastive Learning, Knowledge Augmentation

1. Introduction

In recent years, we have witnessed a rapid development of artificial intelligence technologies (especially deep neural networks) in disease diagnosis and clinical decision support systems (Berner, 2007). Given the medical notes of a patient (e.g., Electronic Medical Record (EMR)), automatic disease prediction aims to predict the most possible diseases which can help doctors to make correct clinical decisions. In-time and accurate disease prediction can also assist in early intervention, leading to optimized disease management and efficient allocation of healthcare resources. Previously rule-based and statistic knowledge-based approaches were widely adopted in disease classification (Shortliffe, 2012; Kohn et al., 2014); however, they suffer from inflexibility and heavy labor costs. Recent advances in deep learning have gained great success in clinical disease modeling (Lipton et al., 2015; Rasmy et al., 2021).

One of the biggest challenges in applying deep learning to disease prediction based on EMR is to learn feature patterns from a limited number of annotated samples. Deep learning approaches require a large quantity of annotated medical records, which are often impossible to acquire in real world due to privacy concerns. Moreover, the accessible samples can be highly imbalanced due to the heterogeneity of clinical presentations. Therefore, directly using historical EMR for training has two major drawbacks: 1) imbalanced dataset: in real-world clinics, common diseases have a rich collection of clinical cases while rare diseases often

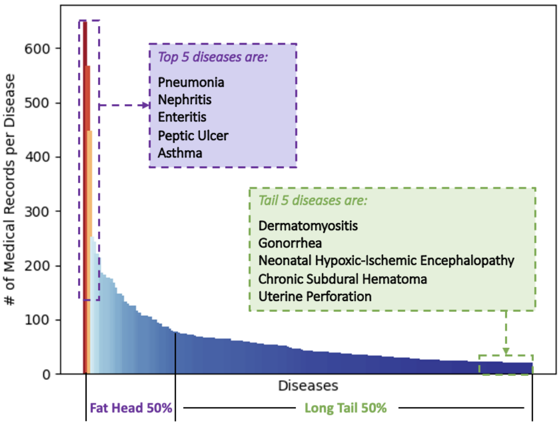
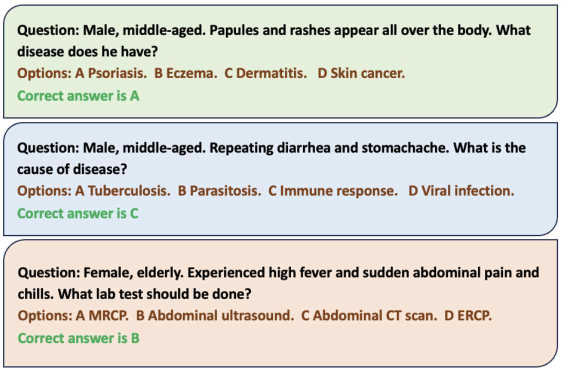


Figure 1: Distribution of 12,776 electronic medical records collected from five hospitals. Only half of the diseases have more than 40 associated records and less than 10% of diseases have more than 100 records.

* Equal Contribution
† Corresponding author

LREC-COLING 2024, pages 11394–11404
20-25 May, 2024.



Question: Male, middle-aged. Papules and rashes appear all over the body. What disease does he have?
Options: A Psoriasis. B Eczema. C Dermatitis. D Skin cancer.
Correct answer is A

Question: Male, middle-aged. Repeating diarrhea and stomachache. What is the cause of disease?
Options: A Tuberculosis. B Parasitosis. C Immune response. D Viral infection.
Correct answer is C

Question: Female, elderly. Experienced high fever and sudden abdominal pain and chills. What lab test should be done?
Options: A MRCP. B Abdominal ultrasound. C Abdominal CT scan. D ERCP.
Correct answer is B

Figure 2: Examples of multiple–choice questions in Medical Licensing Exam. Each question contains only one correct answer and the rest are distractors.

have a small number of medical records or even no records at all (Yoo et al., 2021). As shown in Figure 1, medical records of the top five diseases, including Pneumonia, account for more than 20% of cases, while the bottom five diseases such as Dermatomyositis are less than 1%. Training on such an imbalanced data set will result in a biased model that may ignore rare diseases during prediction (Chen et al., 2022b). This is unacceptable given that clinical practice requires high accuracy and misdiagnosis can lead to severe consequences; 2) diseases with analogous symptoms can be easily misdiagnosed: according to International Classification of Diseases (ICD10),[1] there are more than 68,000 diseases in total. For some diseases, like Psoriasis and Pneumonia, it is quite easy to separate them. While for others like Myocarditis and Myocardial Infarction, are both cardiovascular diseases and share very similar symptoms such as chest pain and abnormal heart rhythms. As a result, they can easily be misdiagnosed and improper treatments can lead to serious consequences; therefore, it is crucial for AI models to be able to distinguish between diseases with analogous conditions and provide doctors with valuable insights.

To handle the above two problems, we propose a Medical Knowledge-Enhanced Contrastive Learning(MKeCL) approach to disease diagnosis. For the imbalanced dataset problem, MKeCL introduces prior medical knowledge in the form of the knowledge graph. Such a knowledge graph can compensate for the lack of information on rare diseases in EMRs; As to the hard-to-diagnose diseases, we find that the diagnosis questions in the Medical Licensing Examination can help. As shown in Figure 2, there are many disease diagnosis-related multiple-choice questions in the examination. Given the symptoms, the question asks to choose the most proper disease from all five options. Usually, the distracting options are quite similar to the correct answer and are often misdiagnosed in clinics. Therefore, we can mine the differences between diseases with analogous symptoms from these multiple-choice questions in examinations.

To incorporate medical knowledge graphs and exam questions in modeling, we propose a medical knowledge-enhanced contrastive learning framework. In particular, upon the backbone classification neural network, we introduce two objectives for further optimization: 1) the separation between correct and incorrect knowledge graph triples; and 2) the separation between correct answers and distracting options. In this manner, data scarcity can be alleviated and the difference between similar diseases can be emphasized. In summary, we make the following contributions in this paper:

- We identify and target two challenges in disease diagnosis: data scarcity for rare diseases and commonly misdiagnosed diseases.

- We introduce two auxiliary data sources: 1) medical knowledge graph to compensate for the lack of information on rare diseases and 2) medical exam questions to distinguish between diseases with analogous symptoms. We also introduce a contrastive learning strategy to incorporate these two data sources in modeling without altering the network structure.

- We introduce a novel disease diagnosis benchmark, called JARVIS-D[2], from large-scale real-world clinical EMRs with labeled diagnosis results by professional clinicians.

- Extensive experiments are conducted on a real-world EMR dataset and experimental results demonstrate the superior performance of our proposed model, even compared with ChatGPT.

2. Related Work

Disease Diagnosis. Traditional machine-learning models were first used in single-label disease classification in the late 1990s. Prince (1996) employed the Bayes network to identify Alzheimer's and dementia. Recently, deep neural networks gradually became the dominating method in disease diagnosis. Green et al. (2006) studied both neural network and logistic regression in the prediction of acute coronary syndrome and

[1] https://icd.who.int/browse10/2019/en

[2] For the sake of privacy, we are only permitted by hospitals to release Jarvis-D after manual desensitization.

achieved promising results. To further improve diagnostic accuracy, more features were taken into consideration, e.g., Atkov et al. (2012) investigated to add genetic factors to detect coronary disease. However, the above approaches only focused on one or a few specific diseases. More recently, large language models have shown the potential of improving disease diagnosis's accuracy by pre-training on large electronic medical records (EMR) (Liu et al., 2021; Li et al., 2020; Rasmy et al., 2021). Modeled as a multi-label classification problem and taking the historical EMR as input, Li et al. (2020) proposed to use the Transformer-based model to predict a patient's possible diseases in the future, which encoded the 10th revision of the ICD-10. Rasmy et al. (2021) introduced Med-BERT, which adapted the BERT framework to the structured electronic health record (EHR). It fed BERT with three types of data involving diagnosis codes, the order of codes within each visit, and the position and name of each visit. The studies by Wang et al. (2023) and Chen et al. (2022a) explored how a diagnosis is formulated by leveraging all symptoms, utilizing sequence generation methodologies. The limitation of the above methods is they neglect the real-world data imbalance and sparsity problem in disease prediction. Only a few works in recent years have started to address this problem: Yang et al. (2022b) introduced a prototypical networks-based few-shot learning approach for dermatological disease diagnosis; Yang et al. (2022b) attempted to inject synonyms of medical terms in ICD-10 coding to relieve the data insufficiency for rare diseases.

Contrastive Learning. Contrastive learning aims to distinguish pairs of similar data points and has shown excellent performance on supervised contrastive Natural Language Processing (NLP) pretraining (Rethmeier and Augenstein, 2023). Supervised contrastive pretraining methods utilize human-annotated corpora such as parallel sentences, textual labels, or text summarizations to define text data augmentations for contrastive training, while self-supervised contrastive methods aim to scale pretraining by contrasting automatically expanded input texts or output pseudo-labels. Recently, Gao et al. (2021) proposed SimCSE, which applied two dropout masks to an input sentence to create two slightly different sentence embeddings that were used as a pair of positive (matching) sentence embeddings for self-supervised pretraining; Xu et al. (2022) introduced LaPraDoR, a pretrained dual-tower dense retriever that iteratively trained the query and document encoders with a cache mechanism, and it further integrated lexicon-enhanced dense retrieval to enhance dense retrieval with lexical matching; Ma et al. (2022) presented a novel contrastive span prediction task to pretrain the encoder alone, but still retain the bottleneck ability of autoencoder; Muennighoff (2022) applied in-batch negatives to train a Transformer decoder for generating sentence embeddings. After being pretrained using contrastive learning, models are believed to be efficient for downstream tasks or zero-shot transfer, e.g., medical image analysis (Zhang et al., 2022a), clinical events forecasting (Zhang et al., 2022b), and disease prediction (Chen et al., 2022c; Wu et al., 2022).

3. Methodology

In this section, we present the proposed Medical Knowledge-Enhanced contrastive learning (MKeCL) model. Firstly, we introduce two types of external knowledge and simple prompts to convert them to question-answer pairs; Secondly, we describe the detailed process of how MKeCL uses Contrastive Learning for disease diagnosis.

3.1. Incorporating Medical Knowledge

One of the key characteristics that distinguishes our method from other systems (Yang et al., 2022b) is that we integrate external knowledge from multiple sources in modeling. Here we collect structured data from Knowledge Graph (KG) and text data from the Medical Licensing Exam.

- Medical Knowledge Graph: to compensate for the insufficient training data for rare diseases, we introduce the pre-built knowledge graph to enrich the connections between rare diseases and common diseases, between rare diseases and symptoms, etc.
- Medical Licensing Exam: to distinguish easily misdiagnosed diseases, we introduce multi-choice questions in the Medical Licensing Exam. As shown in Figure 2, for each question, there is one correct answer and multiple disturbing options. Since these disturbing options are carefully designed by medical experts, they are easily mistaken for the correct answer. We can improve the model by separating correct answers from these distractors.

To accommodate medical knowledge graphs and medical licensing exam data in modeling, we convert them into question-answer pairs and use them for pre-training via contrastive learning. In the medical knowledge graph, every triple consists of two entities and a relationship in the form of (entity$_a$, relation, entity$_b$). We transform these triples into question-answer pairs by using a simple prompt to combine entity$_a$ and relation into a question. For instance, we convert (Heart Failure, Symptom, Chest

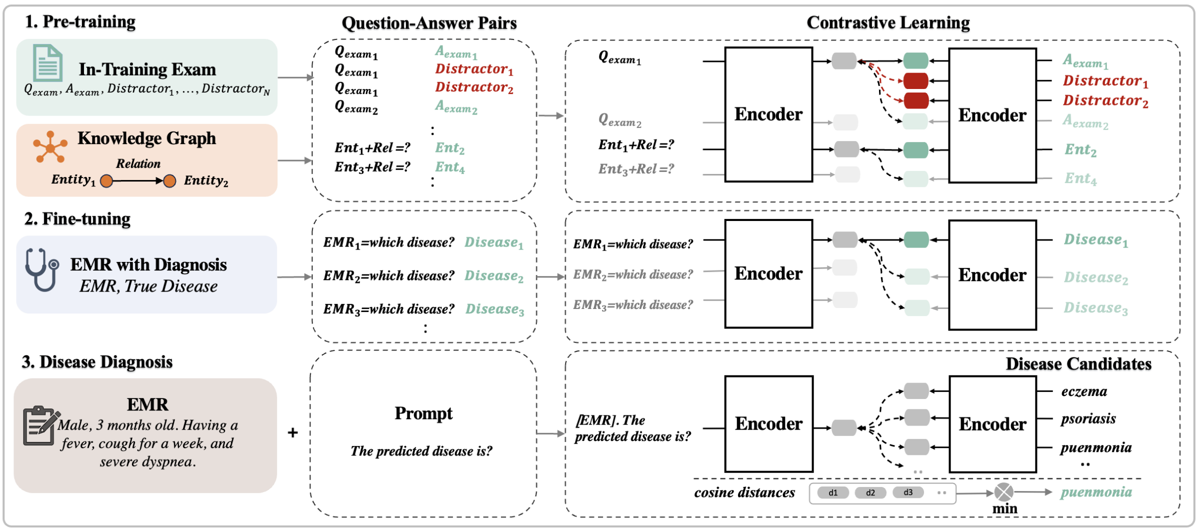


Figure 3: The workflow of Medical Knowledge-Enhanced Contrastive Representation Learning (MKeCL). Data collected from the Medical Licensing Exam and knowledge graph during pretraining are converted into question-answer pairs for contrastive learning. – between embeddings represents positive instances, - - means in-batch negatives and - - denote hard negatives. The Transformer encoder is used to generate embeddings for each question-answer pair. During the prediction stage, embeddings of all disease candidates are compared with that of EMR, and the one with the smallest cosine distance is the predicted disease.

Pain) into "What is the common symptom of Heart Failure?" - "Chest Pain"; As to the Medical Licensing Exam, we convert the pair (question, option) as a QA pair. For the correct answer, we create a positive instance, while for each distractor, we generate a negative instance.

3.2. Contrastive Learning for Diagnosis

Contrastive learning aims to learn effective representations by pulling semantically close sentences together and pushing unrelated sentences apart (Hadsell et al.). It has been proven effective in language model training and benefits various downstream NLP tasks. In this paper, we propose a training framework that applies contrastive learning to both pretraining and finetuning. As shown in Figure 3, the proposed framework consists of three major steps:

- Pretraining: we convert data collected from the knowledge graph and Medical Licensing Exam into question and answer pairs. Both questions and answers are sent to the same Transformer-based encoder. Then contrastive learning is conducted over the representation of questions and answers to optimize the parameters of the encoder. In this step, we try to overcome the challenges posed by insufficient training data for rare diseases and easily misdiagnosed diseases.
- Fine-tuning: we use the data of the downstreaming task: diagnosis prediction to finetune the parameters of the encoder. The input data is also in the format of question-and-answer pairs. We convert each EMR into a question by adding a simple prompt and regard the doctor's diagnosis as the answer;
- Disease Diagnosis: this step is the inference step that predicts the diagnosis given the input EMR.

3.2.1. Positive and Negative Instances

In this subsection, we compose the positive and negative cases for contrastive learning. We use (x_i, x_i^+) to denote a positive pair of instances, where x_i and x_i^+ are semantically related. Similarly, we use (x_i,x_i^-) to denote a negative pair of instance, where x_i and x_i^- have unrelated or even contradictory meanings. These pairs are used to train our encoder to better capture their underlying semantic meaning; therefore, it is critical to construct insightful positive (x_i, x_i^+) and negative (x_i, x_i^-) pairs. Common approaches in NLP include data augmentation techniques such as synonym replacement, random word insertion, and back translation (Su et al., 2021; Meng et al., 2021). However, given that medical terms are generally specific and precise, traditional data augmentation techniques may introduce extra noise when dealing with medical text. Therefore, we propose to generate positive and negative pairs as the following:

- Positive Pairs: Following the ideas in (Henderson et al., 2017; Gillick et al., 2019; Karpukhin et al., 2020), we construct positive pairs from the Medical Licensing Exam and knowledge graph. As mentioned in Section 3.1, we extract question and answer pairs from the Medical Licensing Exam and knowledge graph. We refer the representation of the question and answer (acquired from the encoder) as pairs of positive cases.
- Negative Pairs: we introduce two types of negative pairs: 1) In batch negatives: the questions with unpaired answers in the same training batch (Chen et al., 2020); 2) hard negatives: adding hard negatives has been proved to be effective in contrastive learning (Gao et al., 2021). As shown in Figure 2, since each multiple-choice question has several misleading options that are very similar to the correct answer, we include all incorrect options as hard negatives into our training dataset.

3.2.2. Encoding Electronic Medical Records

After pretraining MKeCL on external knowledge, we use it to generate embeddings for electronic medical records. Each EMR is a short paragraph with descriptions of the symptoms experienced by the patient and lab results if available. Firstly, we extend EMR with a prompt and convert it into a sequence of tokens and pass them through a Transformer encoder to produce a sequence of hidden vectors $b \in \mathbb{R}^{N_t \times H_d}$, where N_t is the length of tokens in the question and H_d is the dimension of the hidden states. A mean pooling layer further converts token embeddings into a final representation $\mathbf{h} \in \mathbb{R}^{H_d}$. A similar process is used to generate embeddings for the disease associated with the EMR.

3.2.3. Optimization and Inference

We follow the works in (Chen et al., 2020; Gao et al., 2021) and take a cross-entropy training objective ℓ_i with mini-batch of size N:

$$\ell_i = -log \frac{e^{sim(\mathbf{h}_i,\mathbf{h}_i^+)}}{e^{sim(\mathbf{h}_i,\mathbf{h}_i^+)} + \sum_{j=1,j\neq i}^{N} ce^{sim(\mathbf{h}_i,\mathbf{h}_j^-)}}, \quad (1)$$

where $\mathbf{h}_i, \mathbf{h}_i^+, \mathbf{h}_i^-$ denote the representations of (x_i, x_i^+, x_i^-), respectively. We use cosine similarity $\frac{\mathbf{h}_i^T \mathbf{h}_j}{||\mathbf{h}_i|| \cdot ||\mathbf{h}_j||}$ as the similarity function sim($\mathbf{h}_i$,$\mathbf{h}_j$).

During the disease prediction stage, we use MKeCL to encode EMR and each disease candidate respectively. Following the input data format used during the training stage of MKeCL, all EMRs are appended with a prompt question asking for the predicted disease. Each disease candidate's representation $\mathbf{h}_{d_i}$ is compared with the "EMR+prompt" representation $\mathbf{h}_{EMR}$, and the one that has the largest cosine similarity $(1 - \frac{\mathbf{h}_{EMR}^T \mathbf{h}_{d_i}}{||\mathbf{h}_{EMR}|| \cdot ||\mathbf{h}_{d_i}||})$ is selected as the most possible disease.

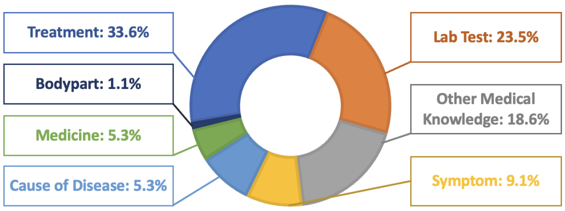


Figure 4: Distribution of different types of medical knowledge used in pretraining MKeCL.

4. Experiments

4.1. Datasets

External Source of Knowledge. We utilized 2,585 disease-related triples from a pre-built medical knowledge graph, which was extracted from professional medical textbooks by medical experts[3]. Therefore, the quality of these triples is assured. We also collected 41,626 multiple-choice questions from previous Medical Licensing Exams. As shown in Figure 4, our data mainly covers six types of medical knowledge, in which treatment and lab tests are the most prevalent. We apply simple prompts to convert the data into question-answer pairs and use them to pretrain MKeCL.

Electronic Medical Records. After the pretraining stage, our model is further finetuned on a total of 12,776 electronic medical records that cover 193 diseases. These medical records were obtained from five hospitals with signed approval from patients. Due to the anonymity requirement, we did not disclose the names of these five hospitals in this submission. Each EMR contains basic information about a patient such as sex and age, and also the patient's chief complaint and lab test results. An example is given in Figure 3. Training and testing datasets are randomly divided. In order to evaluate MKeCL's performance on few-shot and zero-shot settings, 0%, 1%, 3%, 5%, and 10% of training data are sampled respectively.

4.2. Baselines

We compared MKeCL with **BERT** (Devlin et al., 2018), which is a Transformer-based model and

[3] https://jarvislab.tencent.com/kg-intro.html

Model	0%		1%		3%		5%		10%		100%	
	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma
ALBERT (Lan et al., 2019)	-	-	30.9	31.8	36.1	43.2	48.3	44.7	61.6	57.6	87.7	85.9
BERT (Devlin et al., 2018)	-	-	42.6	44.1	43.8	50.7	55.4	52.1	68.6	63.2	90.3	**89.6**
MedBERT (Ting et al., 2020)	-	-	42.2	43.5	44.2	50.9	54.9	52.4	67.0	61.9	90.2	89.1
GP (Yang et al., 2022a)	-	-	43.5	52.8	49.2	59.0	48.7	55.8	54.8	58.5	82.7	85.6
KEPT (Yang et al., 2022b)	-	-	47.4	45.1	51.9	54.6	60.2	53.8	68.7	63.2	89.8	87.9
ChatGPT	40.9	41.2	41.1	40.5	43.0	44.6	45.2	43.5	45.3	44.8	-	-
GPT-4	41.7	34.4	40.6	32.4	44.1	36.2	46.3	37.7	46.2	37.3	-	-
MKeCL$_{mlm\ pretrain}$	-	-	50.6	43.56	51.2	49.4	58.1	50.2	68.4	60.5	89.7	86.0
MKeCL$_{w/o\ pretrain}$	-	-	45.2	39.9	49.2	49.2	55.2	45.3	66.5	59.1	88.3	84.1
MKeCL$_{w/o\ exam}$	23.7	16.6	48.3	44.3	52.4	51.6	56.8	47.1	67.2	59.2	88.9	85.5
MKeCL$_{w/o\ kg}$	49.5	46.1	59.3	54.4	60.4	58.3	65.4	57.5	71.5	64.0	90.2	87.8
MKeCL	**50.5**	**46.1**	**60.7**	**55.7**	**63.7**	**60.5**	**68.1**	**61.6**	**73.0**	**67.0**	**90.5**	87.4

Table 1: The micro F1 (Mi) and macro F1 (Ma) on EMR dataset. We sample 0%, 1%, 3%, 5%, and 10% of the dataset to train models respectively and evaluate their performance at zero-shot and few-shot settings. For ChatGPT and GPT-4, we adopt in-context learning to simulate the few-shot setting. This involves demonstrating 1, 2, 3, and 5 examples in the prompt to mimic the few-shot setup.

is further finetuned on disease classification tasks using EMR data; **ALBERT** (Lan et al., 2019), which is a Transformer architecture based on BERT but with fewer parameters; **MedBERT** (Rasmy et al., 2021), which takes BERT as the backbone and is finetuned with mask language modeling on an EMR dataset of 28,490,650 patients; **GP** (Yang et al., 2022a), which is a generative model pretrained with auto-regression and predicts the text description of each ICD code; **KEPT** (Yang et al., 2022b), which pretrains Longformer using domain-specific knowledge such as disease synonyms; and **ChatGPT**[4] as well as **GPT-4**,[5], which are two best sibling models to InstructGPT (Ouyang et al., 2022) trained to follow the instruction in a prompt and provide a detailed response.

To further examine the performance of pretraining MKeCL using external knowledge, we also finetune MKeCL on EMR directly. Then to examine the genuine advantages of MKeCL, we compare with its variants as outlined below: 1) **MKeCL**$_{mlm\ pretrain}$, a variant of MKeCL that undergoes pre-training using the masked language modeling task (MLM) instead of contrastive learning; 2) **MKeCL**$_{w/o\ pretrain}$, a variant of MKeCL that does not involve pretraining; 3) **MKeCL**$_{w/o\ exam}$, a variant of MKeCL that does not utilize medical exam knowledge; 4) **MKeCL**$_{w/o\ kg}$, a variant of MKeCL that does not incorporate data from knowledge graph.

4.3. Implementation Details

MKeCL and its variants take BERT$_{Base}$ as the backbone. All models are trained with a maximum sequence length of 256, a learning rate of 1×10^{-4} with 100 warm-up steps and a maximum of 100 epochs in both pretraining and finetuning stages. Since we take in-batch negatives during the training of MKeCL, the performance is believed to be sensitive to the training batch size (Henderson et al., 2017). We experiment with different batch sizes and choose the optimal batch size of 16 to train all versions of MKeCL.

4.4. Metrics

The F1 score, a measure of a classification model's performance, is the harmonic mean of precision and recall. We use Macro F1, which is the arithmetic mean of per-class F1 scores, and Micro F1, which also takes into account the support for each class as the first two metrics to evaluate the disease diagnosis results.

Contrastive learning-based methods predict the disease by differentiating the distance between the representations of EMR and disease. We also use *alignment* and *uniformity* (Wang and Isola, 2020) to assess the robustness of the produced representations. Alignment calculates the expected distance between the embeddings of positive pairs:

$$\ell_{align} \triangleq \underset{(x,x^{+})\sim p_{pos}}{\mathbb{E}} ||f(x) - f(x^{+})||^{2}. \qquad (2)$$

On the other hand, uniformity measures how far the negative instances are scattered over the hypersphere:

$$\ell_{uniform} \triangleq log \underset{(x,y) \overset{i.i.d}{\sim} p_{data}}{\mathbb{E}} e^{-2||f(x)-f(y)||^{2}}, \qquad (3)$$

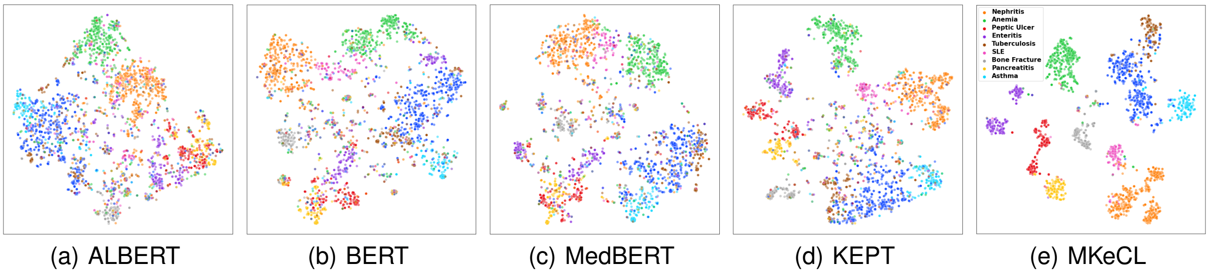
where p_{data} denotes the data distribution. For both ℓ_{align} and $\ell_{uniform}$, lower numbers are better.

4.5. Main Results

As illustrated in Table 1, MKeCL significantly outperforms all baseline models. When trained with

[4] https://openai.com/blog/chatgpt

[5] https://openai.com/gpt-4



(a) ALBERT (b) BERT (c) MedBERT (d) KEPT (e) MKeCL

Figure 5: Visualization of medical record representations generated by ALBERT, BERT, MedBERT, KEPT, and MKeCL using t-SNE. Models are trained on 1% of the training dataset and medical records from 10 diseases are sampled randomly for visualization.

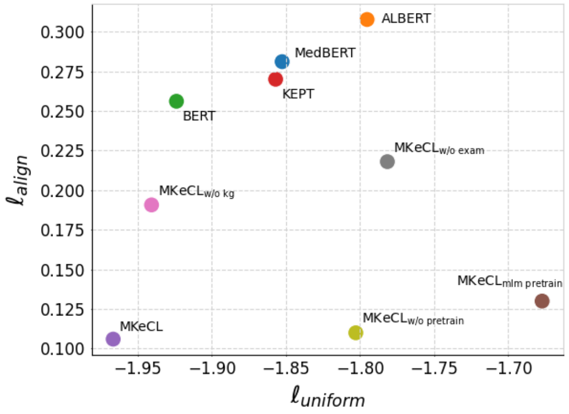
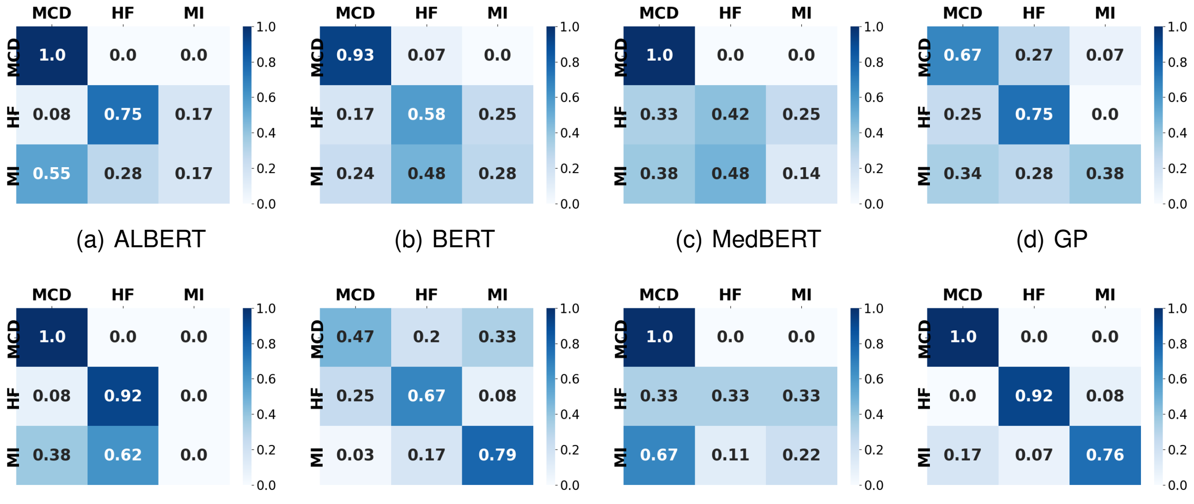


Figure 6: ℓ_{align}-$\ell_{uniform}$ plot of all baseline models and all variants of MKeCL.

just 1% of the dataset, MKeCL achieves a micro F1 score of 60.7% (an improvement of 13.3%) and a macro F1 score of 55.7% (an improvement of 2.9%). These figures in parentheses represent the improvements over the best-performing baseline models, KEPT and GP. MKeCL also surpasses classification models (BERT, ALBERT, MedBERT, and KEPT) and generative models (GP, ChatGPT and GPT-4) in terms of both micro and macro F1 scores across all datasets. This underscores the value of using contrastive learning to generate superior representations of EMR and diseases. Despite MedBERT being pretrained on approximately 28,490,650 health records (around 7000 times the data used to pretrain MKeCL), it scores about 20% lower on F1 scores in few-shot settings. This highlights the effectiveness of integrating external medical knowledge from multiple sources in MKeCL. To compare MKeCL's performance with ChatGPT and GPT-4, we simulated 1%, 3%, 5%, and 10% few-shot settings by providing ChatGPT and GPT-4 with 1, 2, 3, and 5 examples per EMR in the prompt, respectively. Due to ChatGPT's input token limitation, we couldn't test its full-shot capability. MKeCL outperforms both ChatGPT and GPT-4 in all settings, demonstrating its effectiveness. Interestingly, ChatGPT exhibits better performance in distinguishing dissimilar diseases, as indicated by its higher Macro F1 scores compared to GPT-4.

Visualization of Disease Representations. To examine the quality of representations generated by MKeCL, we randomly sampled EMRs from 10 diseases and visualized their embeddings generated by BERT, ALBERT, MedBERT, KEPT, and MKeCL using t-SNE respectively. As shown in Figure 5, when trained using 1% data, MKeCL is able to produce embedding clusters that are internally tighter and well-separated from each other compared to other baseline models. We also calculated ℓ_{align} and $\ell_{uniform}$ following Equation 2 and 3 and plot the result in Figure 6. MKeCL has the lowest loss for both alignment and uniformity and by comparing MKeCL and its variants, we can conclude that pretraining with both knowledge graph and medical exam questions leads to better representations.

Comparison of Easily Misdiagnosed Diseases. We also evaluate MKeCL's ability to distinguish diseases with analogous conditions. We use a heatmap to visualize each model's performance in identifying the correct disease based on EMR from a set of disease candidates with analogous symptoms. Myocarditis, Heart Failure, and Myocardial Infarction are cardiovascular diseases and they all present with similar symptoms such as chest pain, shortness of breath, fatigue, and palpitations. However, they have different causes and require different treatments, and therefore misdiagnosis would lead to serious health complications. One key difference between Heart Failure, Myocardial Infarction, and Myocarditis is their symptom duration. Heart failure typically involves long-term conditions that persist over months, whereas myocardial infarction presents with more severe and persistent symptoms over a shorter period of time. As shown in Figure 7, MKeCL performs the best in distinguishing between these three diseases, while other baseline models are only able to classify one or two diseases correctly. Examples of



(a) ALBERT (b) BERT (c) MedBERT (d) GP

(e) KEPT (f) ChatGPT (g) GPT-4 (h) MKeCL

Figure 7: Confusion matrix on easily misdiagnosed diseases. We use a heatmap to visualize each model's accuracy in classifying between Myocarditis (MCD), Heart Failure (HF), and Myocardial Infarction (MI). The x-axis represents the correct disease and the y-axis represents the predicted disease.

EMR	GT	BERT	ALBERT	MedBERT	GP	KEPT	ChatGPT	GPT-4	MKeCL
Case 1	HF	HF	HF	MCD	MCD	HF	HF	HF	HF
Case 2	MCD	HF	MCD	MCD	HF	MI	MI	MI	MCD

Table 2: Disease prediction results of different algorithms on two real cases. MCD, HF, and MI are Myocarditis, Heart Failure, and Myocardial Infarction, respectively, while GT is the ground truth. The detailed EMR content for Case 1 is: “Male, elderly, with a 3-year history of exertional angina. Over the past 2 weeks, the frequency of angina episodes has increased, and blood pressure has risen to 166/94 mmHg. The patient also experiences paroxysmal nocturnal dyspnea and is currently unable to lie flat.” Besides, the EMR for Case 2 states: “Female, middle-aged, with persistent chest pain for 6 hours. Physical examination: BP 110/70 mmHg. No crackles or wheezes were detected in both lungs. Heart rate is 125 beats/minute with a regular rhythm. No murmurs are heard in any cardiac valve areas. The electrocardiogram reveals partial ST-T elevation. Laboratory tests indicate elevated blood troponin levels.”.

each model's classification results are shown in Table 2. The first patient has had exertional angina for over three years, indicating a high possibility of heart failure, while the second patient experienced six hours of persistent chest pain, indicating a high possibility of myocardial infarction.

4.6. Ablation Studies

To investigate the contribution of each module in MKeCL, we compare it with its variants. Firstly, as shown in Table 1, information from medical license exam or medical knowledge graph is beneficial because dropping one of them (“MKeCL w/o exam” or “MKeCL w/o kg”) degrades the performance of MKeCL. Using the contrastive pretraining procedure is advantageous because using a masked language modeling task (“MKeCL mlm pretrain”) or erasing the contrastive pretraining (“MKeCL w/o pretrain”) for MKeCL impairs its performance. Masked language modeling is also effective, indicated by the superior performance of (“MKeCL mlm pretrain”) compared to without using pretraining task (“MKeCL w/o pretrain”). Furthermore, as illustrated in Figure 6, contrastive pretraining is crucial for enhancing uniformity. Removing it or substituting it with a masked language modeling task diminishes the uniformity of MKeCL. Simultaneously, the knowledge graph leads to improved alignment outcomes, and the medical license exam knowledge contributes positively to both uniformity and alignment.

5. Conclusions

Rare diseases impact over 300 million people globally, yet each has limited clinical records, and similar symptoms can lead to misdiagnosis. We propose the Medical Knowledge-enhanced Contrastive Learning framework (MKeCL) to address these issues. MKeCL integrates knowledge from a medical graph to supplement rare disease data and uses medical exam questions to differentiate similar diseases. A contrastive learning framework combines these data sources. Our experiments confirm MKeCL's effectiveness in diagnosing diseases in few-shot and zero-shot settings, potentially reducing diagnostic delays for rare disease patients.

Future research should focus on improving MKeCL's interpretability, crucial for clinician adoption, and examining its scalability to larger, more diverse datasets. As dataset complexity increases, so may computational and corpus requirements, potentially limiting MKeCL's applicability in resource-constrained settings.

Ethics Statement

In this work, we underscore the substantial risks that may arise from the improper application of the proposed models within the medical domain. The primary objective of our research is to explore more efficient and effective approaches to disease diagnosis. However, it is crucial to note that the proposed models are not yet ready for deployment in real-world medical settings. The potential for these models to mislead users about the underlying reasons for their predictions is a significant concern. Misinterpretations could lead to incorrect decisions, with potentially serious implications for patient care and outcomes. Moreover, the ethical considerations of our work extend beyond the accuracy and reliability of the models. The privacy and security of sensitive medical data are of paramount importance. During the process of data collection and utilization, we have implemented stringent measures to ensure the protection of this sensitive information. Our method adheres to all relevant national and international data protection regulations, demonstrating our commitment to ethical data practices. In addition to regulatory compliance, we have employed robust data anonymization and encryption techniques to safeguard patient confidentiality. These techniques ensure that individual patient identities cannot be linked to the data used in our models, thereby minimizing the risk of privacy breaches. We recognize that the trust of patients and healthcare providers in our work hinges on our ability to protect this sensitive information effectively. In conclusion, while our work holds promise for improving disease diagnosis, it is essential to approach its application with caution. We must continue to prioritize the ethical considerations of accuracy, transparency, data privacy, and security as we further develop and refine these models.

Oleg Yu Atkov, Svetlana G Gorokhova, Alexandr G Sboev, Eduard V Generozov, Elena V Muraseyeva, Svetlana Y Moroshkina, and Nadezhda N Cherniy. 2012. Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. *Journal of Cardiology*, 59(2):190–194.

Eta S Berner. 2007. *Clinical decision support systems*, volume 233. Springer.

Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. 2022a. Diaformer: Automatic diagnosis via symptoms sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4432–4440.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR.

Xiaocong Chen, Yun Li, Lina Yao, Ehsan Adeli, Yu Zhang, and Xianzhi Wang. 2022b. Generative adversarial U-Net for domain-free few-shot medical diagnosis. *Pattern Recognition Letters*, 157:112–118.

Yuhao Chen, Yanshi Hu, Xiaotian Hu, Cong Feng, and Ming Chen. 2022c. CoGO: a contrastive learning framework to predict disease similarity based on gene network and ontology structure. *Bioinformatics*, 38(18):4380–4386.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional Transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. *arXiv preprint arXiv:1909.10506*.

Michael Green, Jonas Björk, Jakob Forberg, Ulf Ekelund, Lars Edenbrandt, and Mattias Ohlsson. 2006. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. *Artificial Intelligence in Medicine*, 38(3):305–318.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

M Sun Kohn, J Sun, S Knoop, A Shabo, B Carmeli, D Sow, T Syed-Mahmood, and W Rapp. 2014. IBM's health analytics and clinical decision support. *Yearbook of Medical Informatics*, 23(01):154–162.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-Supervised Learning of Language representations. *arXiv preprint arXiv:1909.11942*.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(1):1–12.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. 2015. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677*.

Ning Liu, Qian Hu, Huayun Xu, Xing Xu, and Mengxin Chen. 2021. Med-BERT: A pretraining framework for medical records named entity recognition. *IEEE Transactions on Industrial Informatics*, 18(8):5600–5608.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Pre-train a discriminative text encoder for dense retrieval via contrastive span prediction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 848–858.

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. COCO-LM: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34:23102–23114.

Niklas Muennighoff. 2022. SGPT: GPT sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Martin J Prince. 1996. Predicting the onset of Alzheimer's disease using Bayes' theorem. *American Journal of Epidemiology*, 143(3):301–308.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4(1):86.

Nils Rethmeier and Isabelle Augenstein. 2023. A primer on contrastive pretraining in language processing: Methods, lessons learned, and perspectives. *ACM Computing Surveys*, 55(10):1–17.

Edward Shortliffe. 2012. *Computer-based medical consultations: MYCIN*, volume 2. Elsevier.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Liu Ting, Qin Bing, Liu Ming, Xu Ruifeng, Tang Buzhou, and Chen Qingcai. 2020. Pre-training model for Chinese medical text processing PCL MedBERT. *PCL blog*.

Huimin Wang, Wai-Chung Kwan, Kam-Fai Wong, and Yefeng Zheng. 2023. Coad: Automatic diagnosis through symptom and disease collaborative generation. *arXiv preprint arXiv:2307.08290*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Yawen Wu, Dewen Zeng, Zhepeng Wang, Yi Sheng, Lei Yang, Alaina J James, Yiyu Shi, and Jingtong Hu. 2022. Federated self-supervised contrastive learning and masked autoencoder for dermatological disease diagnosis. *arXiv preprint arXiv:2208.11278*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2022. LaPraDoR: Unsupervised pretrained dense retriever for zero-shot text retrieval. *arXiv preprint arXiv:2203.06169*.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hong Yu. 2022a. Multi-label Few-shot ICD Coding as Autoregressive Generation with Prompt. *arXiv preprint arXiv:2211.13813*.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022b. Knowledge Injected Prompt Based Fine-tuning for Multi-label Few-shot ICD Coding. *arXiv preprint arXiv:2210.03304*.

Tae Keun Yoo, Joon Yul Choi, and Hong Kyu Kim. 2021. Feasibility study to improve deep learning in OCT diagnosis of rare retinal diseases with few-shot classification. *Medical & Biological Engineering & Computing*, 59:401–415.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022a. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR.

Ziqi Zhang, Chao Yan, Xinmeng Zhang, Steve L Nyemba, and Bradley A Malin. 2022b. Forecasting the future clinical events of a patient through contrastive learning. *Journal of the American Medical Informatics Association*, 29(9):1584–1592.