Journal of King Saud University - Computer and Information Sciences 36 (2024) 101933

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Full length article

MED-Prompt: A novel prompt engineering framework for medicine prediction on free-text clinical notes

Awais Ahmed [a], Xiaoyang Zeng [a], Rui Xi [a], Mengshu Hou [a,b,*], Syed Attique Shah [c]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China - UESTC, Sichuan, 611731, China
[b] School of Big Data and Artificial Intelligence, Chengdu Technological University, Sichuan, 611730, China
[c] School of Computing and Digital Technology, Birmingham City University, STEAMhouse, B4 7RQ, Birmingham, United Kingdom

ARTICLE INFO

Keywords:
AI-enabled healthcare decisions
Medical prompts
Pretrained models
Free-text clinical notes
And natural language processing

ABSTRACT

Existing AI-based medicine prediction systems require substantial training time, computing resources, and extensive labeled data, yet they often lack scalability. To bridge these gaps, this study introduces a novel MED-Prompt framework that employs pretrained models such as BERT, BioBERT, and ClinicalBERT. The core of our framework lies in developing specialized prompts, which act as guiding instructions for the models during the prediction process. MED-Prompt develops prompts that help models interpret and extract medical information from clinical corpus. The clinical text was derived from the widely known MIMIC-III [1] dataset. The study performs a comparative analysis and evaluates the performance of Manual-Prompt and GPT-Prompts. Further, a fine-tuned approach is developed within MED-Prompt, leveraging transfer learning to achieve prompt-guided medicine predictions. The proposed method achieved a maximum F1-score of 96.8%, which is more than 40% F1-score higher than the pretrained model. In addition, the fine-tuned also showed an average of 2.38 times better processing performance. These results revealed that MED-Prompt is scalable regarding the number of training records and input prompts. These results not only demonstrate the proficiency and effectiveness of the framework but also significantly reduce computational requirements. This also indicates that the proposed approach has the potential to significantly improve patient care, reduce resource requirements, and increase the overall effectiveness of AI-driven medical prediction systems.

1. Introduction

Accurate medicine prediction holds immense importance. In recent years, the healthcare domain has experienced a notable surge in the adoption of applied artificial intelligence applications (Feng et al., 2022; Javaid et al., 2022). Among these applications, medicine prediction is a pivotal aspect of developing healthcare decision support systems (Ahmad et al., 2022; Fernandes et al., 2020; Ahmed et al., 2023; Sylolypavan et al., 2023; Sajde et al., 2022). The emergence of applied AI in healthcare has introduced a new dimension to the field, with AI-based diagnosis systems gaining prominence (Gupta and Kumar, 2023; Ali et al., 2023). Clinicians now aspire to have their customized intelligent healthcare diagnosis systems to aid decision-making and expand the scope of services provided to society (Rasheed et al., 2022).

Accurate, effective, and precise medicine prediction outcomes are crucial to ensure the delivery of optimal healthcare decisions and treatments. Notably, AI-enabled models have the potential to anticipate patient outcomes, optimize treatment plans, and enhance overall healthcare quality. By harnessing the power of AI, these prediction models offer valuable insights that can revolutionize healthcare decision-making (Pandey et al., 2022; Karthikeyan et al., 2023; Firouzi et al., 2021; Zhang et al., 2023). However, conventional, AI-driven

* Corresponding author at: School of Computer Science and Engineering, University of Electronic Science and Technology of China - UESTC, Sichuan, 611731, China.
E-mail addresses: 202014080105@std.uestc.edu.cn (A. Ahmed), 202011081605@std.uestc.edu.cn (X. Zeng), ruix.ryan@gmail.com, ruix2022@uestc.edu.cn (R. Xi), mshou@uestc.edu.cn (M. Hou), syedattique.shah@bcu.ac.uk (S.A. Shah).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

[1] https://physionet.org/content/mimiciii/1.4/

https://doi.org/10.1016/j.jksuci.2024.101933
Received 10 October 2023; Received in revised form 13 January 2024; Accepted 15 January 2024
Available online 20 January 2024

healthcare systems based on machine learning and deep learning demand significant training time, computing resources, large amounts of labeled data, and scalability issues.

To address the aforementioned challenges, the current study takes advantage of recent technology of Large Language Models (LLMs) regarding domain-specific factual knowledge. This study harnesses the capabilities of pretrained models designed for transfer learning (TL), also known as fine-tuning approaches. The proliferation of language models such as GPT (Radford et al., 2018), BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), and Google BARD (Chang et al., 2023) has demonstrated remarkable efficacy across diverse applications, spanning language generation, machine translation, question answering, sentiment analysis, and text classification in various domains like social media, e-commerce, customer support, gaming, and healthcare. However, within healthcare, specifically in processing clinical notes authored by medical experts, the utilization of language models encounters inherent hurdles. Notably, most existing models lack training on medical domain-specific data, limiting their capacity to provide accurate and domain-specific information (Thirunavukarasu et al., 2023). This limitation regarding factual knowledge within LLMs remains a critical concern in extracting precise domain-specific insights from medical text. This study acknowledges this challenge and aims to explore methodologies that mitigate this gap, seeking ways to enhance the precision and relevance of LLMs in medical contexts.

When designing MED-Prompt framework and preparing this manuscript, we conducted a comprehensive literature review to identify relevant studies on medicine prediction using the prompts technique. However, to our knowledge, we did not find any recent studies specifically addressing resource-intensive, precise, and accurate medicine prediction. We have undertaken the following challenges to propose a state-of-the-art medicine prediction system to achieve the core objectives of why a prompt-based medicine prediction system is needed.

- Time and resource-intensive solutions
- Reliance on training-based techniques
- Extensive labeled data is needed
- Domain-specific challenges due to the evolving nature of medicine

This study introduces MED-Prompt, a novel prompt engineering framework for an accurate and precise medicine prediction framework. A "prompt" is considered an additional contextual natural language sentence or phrase fed to an AI model, which helps to activate models and provoke certain responses to the tasks (Ding et al., 2022). Prompts could play a crucial role in healthcare and help extract contextual guidance and explicit instructions to AI models, aiding in the accurate extraction of medical information and improving prediction precision.

The preliminary objective of the MED-Prompt framework is to leverage pretrained language models such as BERT, BioBERT, and ClinicalBERT to address issues associated with conventional machine learning classification-based methods. The main innovation is the development of specific prompts to guide models during prediction. The specifically designed prompts guide pretrained or fine-tuned models to understand and help the framework extract medical information from clinical text data. The MED-Prompt architecture utilizes pretrained models, indicating that they are not specifically retrained on medical data. Conversely, these models are meticulously adjusted and directed by specific prompts to accommodate medicine predictions. To further emphasize prompt engineering techniques adopted by humans and machines. We developed and evaluated our framework on two prompts: manual-designed and GPT-generated. Additionally, the study aims to optimize results through fine-tuning (transfer learning) approaches to enhance the predictive capabilities of the models. To evaluate the effectiveness of the proposed MED-Prompt framework, we derived a subset of data from MIMIC-III[2] dataset clinical text. By conducting multiple experiments that test different medicine predictions, including sets of medicines (m30, m50, m70, and mOrignal), the research aims to assess the framework's performance across various medical scenarios comprehensively. A core objective of this study is to develop a resource-efficient and scalable approach to overcome the challenges associated with medicine prediction from free-text clinical notes.

In summary, the main contributions of this work are as follows:

- **MED-Prompt Framework**: The study presents the development of the MED-Prompt framework, a novel approach to medicine prediction. This framework utilizes pretrained natural language processing models, incorporating the core architecture of transformers. By leveraging transfer learning techniques, MED-Prompt achieves accurate and efficient medicine prediction. The detailed design of the MED-Prompt framework is illustrated in Fig. 1.
- **Harnessed Pretrained Models:** The study harnesses the potential of large transformer-based pretrained models. By utilizing pretrained models, the framework benefits from their learned linguistic patterns, contextual linkages, and semantic representations, enhancing the accuracy and effectiveness of medicine prediction.
- **Experimentation and Evaluation:** The framework carries out multiple experiments using different combinations of dimensions, such as the set of medicines, prompts, and records. This rigorous experimentation evaluates the framework's performance and sheds light on the impact of different factors on the prediction results.

The rest of the paper is structured as follows. The next subsequent Section 2 highlights the related work, followed by the research methodology in Section 3, and Section 4 details the experimental setup. At the same time, the comparative and ablation experiments are presented in Section 5. Additionally, Section 6 briefly discusses the limitations and future work. The study concludes this work in Section 7.

2. Related work

The research landscape in text-based healthcare classification, particularly concerning prompt-based techniques for medicine prediction, remains relatively limited. However, the emergence of advanced language models has brought increased attention and progress to the medical domain in this regard. Significant works have contributed to effective medicine prediction frameworks. This study provides a comparative literature overview in Table 1, focusing on prompt-based methodologies to highlight MED-Prompt's potential. This detailed investigation identifies the research gap that MED-Prompt aims to address by understanding the strengths and limitations of existing approaches, emphasizing MED-Prompt's contributions to medicine prediction with prompt-based techniques and pretrained language models.

- ***Advancements in Text-based Healthcare:*** The BEHRT framework by Li et al. (2020) marks an early milestone in text-based healthcare. This framework harnesses the capabilities of language models to make accurate healthcare predictions efficiently. BEHRT, a deep neural sequence transduction model, was introduced for the simultaneous prediction of multiple diseases using electronic health records (EHR) data. Their experiments were conducted using the Clinical Practice Research Datalink (Herrett et al., 2015), a substantial public dataset. Impressively, their methodology achieved an 80% accuracy, setting a state-of-the-art benchmark at the time of publication. Building on this foundation, Luo et al. (2020) introduced the HiTANet framework, further elevating text-based healthcare classification. HiTANet is

[2] https://physionet.org/content/mimiciii/1.4/.

Table 1
A comprehensive list of related literature to prompt engineering in healthcare.

Ref.	Contribution	Shortcoming	Disadvantage	Research gap	Scope/Domains	Performance
MedKPL (Lu et al., 2023)	This study aims to fill the existing research gap by proposing medical knowledge-enhanced diagnoses based on integrating different sources of prompt	Noise propagation, Poor extensibility, and No relationship between diseases	Need to manually design the knowledge text format whenever applicable	Future plan is to design more templates and conduct integration learning of templates to evaluate methodology	Chinese EHR disease classification (6 diseases), Pediatric Patient EHR (PP-EHR)	82.53% accuracy.
AFKF (Li et al., 2023)	A novel ALBERT-based fusion Kalman, namely the AFKF model, was presented. Core concepts were based on a sliding window scheme and a fusion Kalman-filter model to deal with the coupling relationships of large sequences	There is no reduction in the differences between downstream and pre-training tasks. Also, the study lacks extensibility and huge resource consumption.	Lack of multi-granular information	Authors aims to fuse medical text features to work on multi-granularity visual feature representation	Surgical recommendation support system	The study recorded (76.6%, 81.4%, 75.6%, and 78.0%) accuracy, recall, precision, and F1 respectively.
CTC (Kambar et al., 2022)	The study focuses on implementing machine learning methods to analyze the mechanism of action of Alzheimer's drugs in COVID-19.	Cannot fully maintain sentence-level information and context.	Sensitive to hyperparameters. Exploration of different SVM kernels. Also, the dataset is limited in size and lacks generalization.	Authors intend to apply ensemble learning or other ways of combining models to improve results.	Clinical Text Classification of Alzheimer's Drugs	95% accuracy, 100% recall, and 92.0% F1-score.
Gao et al. (2021)	The key application presented by this study is an extension of the length limitation of transformer applications in medical NLP. Additionally, the authors introduce techniques such as hierarchical or multi-head attention mechanisms proposed to handle longer, better clinical texts and capture context effectively.	The study lacks implicit knowledge and high resource consumption	Splitting a long input sequence into several segments and feeding them into the model together.	Using word pieces tokenization method defined in general corpora and has not been adjusted or expanded to clinical terminology corpora.	EHRs disease classification (6 diseases). MIMIC-III and SEER cancer pathology report	76.15% accuracy.
TCM (Yao et al., 2019a)	Using a pretrained language model for disease classification based on Traditional Chinese Medicine clinical records	There is no reduction in the differences between downstream and pre-training tasks, and the study lacks extensibility.	They directly fine-tuned the model based on the pretrained model without reducing pretrained model base tasks, which resulted in resource consumption and extensive training time	The BERT model achieves state-of-the-art results, and they aim to improve by fine-tuning unlabeled clinical corpora. Further, the fine-tuning approach can also be enhanced by adding knowledge-based objectives.	TCM clinical records based disease classification (5 diseases), TCM clinical records data set	89.39% accuracy, 88.67% Marco F1, and 89.39% Micro F1.
CTC (Yao et al., 2019b)	The study employs rule-based features and knowledge-guided CNNs to classify clinical content. The authors suggest a CNN model with domain-specific knowledge characteristics to boost classification performance.	Lacks pre-training on huge datasets. Poor grasping of complicated clinical text data linkages and nuances	CNN-based structure failed to capture the long dependency and had difficulty in long text sequences and difficult text. Framework needs to define the knowledge organization rules well. Lack of pretrained knowledge	They plan to design more principled methods and evaluate their methods on more clinical text datasets.	Clinical text classification	Disease-wise, (80.14%, 97.60%) F1 macro and micro scores, respectively. Category-wise, (67.64%, 96.18%) F1 macro and micro scores, respectively.
HiSANs (Gao et al., 2019)	This study introduces HiSANs architecture to classify cancer pathology reports with hierarchical self-attention networks. Further, it extended the length limitation of CNN applications in medical NLP.	Their approach faces difficulties in classifying reports belonging to tumor IDs associated with multiple reports. Lacking implicit knowledge	Ground truth granularity at the tumor level rather than the report level (multi-site problem).	HiSANs aim to enhance performance on challenging tasks like histology, exploring uncertainty quantification, and leveraging contextual tumor information for improved prediction	Cancer pathology reports classification-based information extraction (site, laterality, behavior, histology, and grade)	Site (75.16% accuracy and 62.52% marco F1), Histology (70.41% accuracy and 27.71% marco F1)

(*continued on next page*)

Table 1 (*continued*).

Ref.	Contribution	Shortcoming	Disadvantage	Research gap	Scope/Domains	Performance
Hughes et al. (2017)	This study proposed novel sentence-level document classification on medical records using deep convolutional neural networks to represent complex features.	The proposed framework cannot capture linguistic patterns beyond a small window, usually 3–5 words, and lacks implicit knowledge.	No pretrained knowledge enhanced and CNN-based structure	Their future direction includes scaling the proposed network to handle more data and finer clinical classifications, further exploring medical and online resources to gather more data, and applying domain adaptation techniques and convolutional neural network features for high-level patient representations.	Sentence level clinical text classification with 26 medical categories and 4000 sentences	68% accuracy.

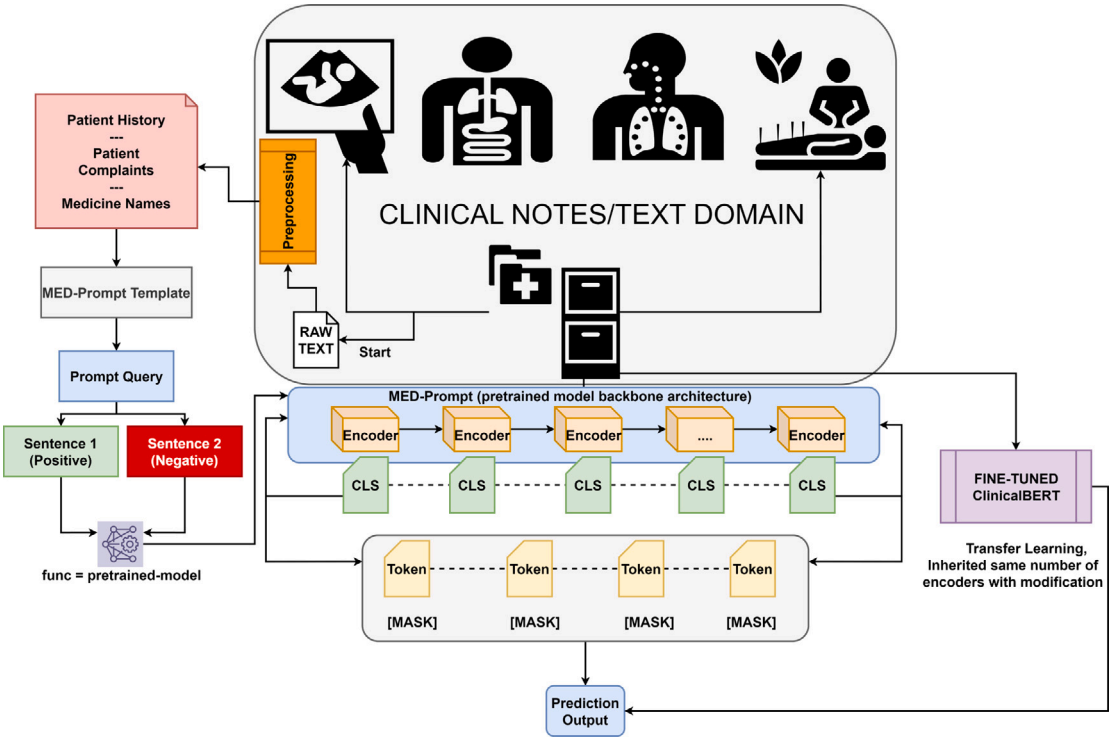

Fig. 1. Proposed framework with an in-depth representation of the interconnected modules and stages.

designed to capture temporal patterns and dependencies within EHR data. It incorporates patient- and visit-level attention mechanisms, enabling time-aware attention focusing on critical time points and capturing long-term dependencies in EHR sequential data. This attention-based technique emphasizes significant events while disregarding irrelevant ones in EHR data's temporal characteristics, ultimately facilitating risk prediction using EHR data.

- ***Prompt-based Techniques in Medicine Prediction:*** In Gururangan et al. (2020), authors stress the versatility of pretrained language models in various domains, setting a standard across computer science publications, news, reviews, and biomedical data. They fine-tune the ROBERTA-base model for eight classification tasks, comparing its performance in high-end and low-resource settings on various domain tasks. This study underscores the importance of domain adaptation and task-specific fine-tuning for improved pretrained model performance, providing a foundation for their application in specific contexts. Rasmy et al. (2021) introduce Med-BERT, tailored for the medical field, emphasizing its effectiveness in medicine prediction using pretrained language models. They assess Med-BERT on two diverse EHR datasets (Truven (Truven, 2023) and Cerner (CernerHealthFacts, 2023)), demonstrating its applicability in healthcare and superior performance over baseline models like BEHRT (Li et al., 2020) and G-BERT (Shang et al., 2019). In Sivarajkumar and Wang (2022), HealthPrompt is introduced, leveraging prompt-based techniques in healthcare classification. Zero-shot learning and task-specific templates are used, and the framework is evaluated with the MIMIC-III dataset, providing a competitive reference for the MED-Prompt framework. Lu et al. (2023) present the MedKPL framework in 2023, offering fresh insights and methodologies to enhance medicine prediction using language models. The study in Lee et al. (2023) assesses the effectiveness of an automated prompt system in electronic medical records, aiming to boost the response rate of patients with metabolic conditions. They achieve a 27% improvement and suggest the need for further research in this area, marking the first study of its kind for increasing screenings in obese children with metabolic conditions.
- ***Pretrained Model's Application Oriented work:*** The authors of Huang et al. (2019) introduced ClinicalBERT, the first model tailored specifically for clinical data. ClinicalBERT was designed to

Table 2
Set of positive and negative prompts for [GPT and manual] designed for MED-Prompt framework.

Prompt text	Predicted [MASK]	Type of prompt
Taking {medicine} has been [MASK] for your health	Important, Appropriate	gpt_prompt1p
The effectiveness of {medicine} as a treatment has been widely [MASK]	Maintained, Limited	gpt_prompt2p
Using {medicine} can [MASK] your recovery process	Help, Improve	gpt_prompt3p
Using {medicine} can expedite your recovery process and [MASK] you regain your health faster	Help, Should	gpt_prompt4p
Taking {medicine} at the recommended dosage can [MASK] your symptoms	Control, Improve	gpt_prompt5p
Research studies have shown that {medicine} is [MASK] in reducing pain	Effective, Interested	gpt_prompt6p
Taking {medicine} has been not [MASK] for your health	Important, Sufficient	gpt_prompt1n
The effectiveness of {medicine} as a treatment has not been widely [MASK]	Changed, Maintained	gpt_prompt2n
Using {medicine} cannot [MASK] your recovery process	Help, Improve	gpt_prompt3n
Using {medicine} can expedite your recovery process and not [MASK] you regain your health faster	Help, Until	gpt_prompt4n
Taking {medicine} at the recommended dosage cannot [MASK] your symptoms	Improve, Control	gpt_prompt5n
Research studies have shown that {medicine} is not [MASK] in reducing pain	Interested, Effective	gpt_prompt6n
The medicine you are taking is [MASK] for you	Good, Relevant	manual_prompt1p
You have been using medicine for a long and it [MASK] your disease and reduces the risk	Helps, Controls	manual_prompt2p
The medicine you have been prescribed has [MASK] your health	Changed, Improved	manual_prompt3p
The medicine has [MASK] your health	Changed, Helped	manual_prompt4p
The medicine you are taking is not [MASK] for you	Tested, Taking	manual_prompt1n
You have been using medicine for a long, and it is [MASK] your disease and reducing the risk	Concerning, Causing	manual_prompt2n
The medicine you have been prescribed has [MASK] your health	Changed, Improved	manual_prompt3n
The medicine has not [MASK] your health	Changed, Affected	manual_prompt4n

leverage the advantages of BERT, originally trained on extensive Wikipedia and Book-Corpus data. This development marked a significant advancement in applying BERT to the medical field, enhancing the analysis of clinical text data. Another noteworthy study by Müller et al. (2023) focuses on COVID-19 analysis using a pretrained BERT model applied to Twitter content. This work explores various tasks, including classification, question-answering, and chatbot-related tasks, all centered around COVID-19. The COVID-19-BERT study is relevant to our research, as it also examines the use of pretrained models like BERT for specific healthcare tasks specifically related to COVID-19 analysis on social media.

3. Research methodology

This section discusses the components within the methodology to develop the MED-Prompt framework for medicine prediction using pretrained language models and fine-tuned approach. Additionally, key components of the proposed study are visually depicted in Fig. 1, further (Algorithms 1, 2) elucidate structure and functioning.

3.1. Data preparation

This study involves extracting a pertinent subset from the MIMIC-III dataset, particularly the NOTEEVENTS file, through meticulous data engineering techniques. Adhering to official guidelines, we selected this dataset while fulfilling all legal prerequisites. It is imperative to abide by the no-distribution mandate and uphold ethical considerations following HIPAA regulations. We performed several preprocessing steps, encompassing:

(i) *Reading NOTEEVENTS file:* This file contains approximately 2.1M Rows and 11 columns of clinical notes in free-text format.
(ii) *Feature selection:* Relevant features were extracted from one of the elaborate columns carrying a total of 14 keywords (including demographics of patients [Admission, Discharge, Date of Birth, and Gender]. Another set of keywords includes Services provided to patients, Allergies, Major Complaints of the patient, Patient history, medication at admission and discharge and discharge notes, etc.).
(iii) *Keyword filtering:* After a rigorous review of technicalities and to achieve the objectives of this study, we finalized the four keywords including [“Chief Complaint”, “Medications on Admission”, “Discharge Medications”, and “History OF Present Illness”].

Table 3
Medicine list and their matched percentage employed in this manuscript.

Medicine list	Match percentage
m30 and m50	56.00%
m30 and m70	19.28%
m30 and mOriginal	0.00%
m50 and m70	43.37%
m50 and mOriginal	0.00%
m70 and mOriginal	0.00%

(iv) *Data filtering:* The raw text was converted to dictionary keys, and records were filtered to retain only those containing predefined keywords, resulting in 36 000 records.
(v) *Summarization:* Summaries of the “History OF Present Illness” feature was generated using the T5 pretrained model.
(vi) *Classical experiment:* Before delving with the designed MED-Prompt framework, separate data was maintained to conduct machine learning experiments as the preliminary study.
(vii) *Prompt generation:* Prompt preparation steps remained under consideration during experiments.
(viii) *Template design:* Template design is a crucial step in the pipeline, comprising multiple factors as shown in Eq. (2).

3.2. Dataset and medicine cohorts

Following rigorous preprocessing efforts as discussed in Section 3.1, a subset of data was prepared from an extensive clinical text dataset consisting of raw text, summarized patient history, chief complaint, medication on admission, discharge medications, and discharge diagnosis and condition at discharge. To facilitate the analysis, distinct medicine cohorts were strategically selected and organized into several subsets, such as a collection of 30 medicines as m30, for a group of 50 medicines as m50 cohort, followed by m70 for a set of 70 medicines, and lastly, a set of 100 medicines termed as mOrignal. Table 3 highlights the different sets of medicines and their match percentage concerning each other. Selecting a representative medicine cohort is a key component, allowing us to explore various medicine cohorts and their predictive outcomes thoroughly, and the impact of different sets of cohorts is also used to measure the scalability of the MED-Prompt framework.

3.3. Prompt engineering

Prompt Engineering is always challenging, and manually designing prompts for models is even more complex. It requires deep knowledge of the problem and a better understanding of the scientific problem for which the prompts are being designed (Wang et al., 2023). Working within the medical domain poses additional challenges due to the limited research. Consequently, researchers encounter difficulties in locating relevant examples. We took help from related HuggingFace[3] respective model for manually designed prompts. We tested before feeding to the model for real-time experiments and optimized the prompts wherever required. Table 2 records and presents the exact number of prompts. The table records prompt text while maintaining the top two [MASK] generated by the model. Table only records results generated from ClinicalBERT, but all models were tested accordingly. The final column represents the prompt type, such as "manual_generated_prompt" or "gpt_generated_prompt".

The prompt engineering step is crucial because prompt structure plays an important role. Several considerations may be taken, and a few are listed below. It is also suggested when using a language model for tasks like summarization, question answering, or text production. It might be helpful to provide the model with a well-prepared prompt to help it produce answers that are more accurate and relevant to the task. The following instructions should not be considered the final list as they may vary from task to task, but this study utilized the following suggestions for prompt engineering.

- Clear and specific task instructions
- Prompts contextual text
- Concise use of formatting and placeholders
- Length and grammatical structure of a sentence
- Multiple variations of prompts

3.4. Prompt design

Prompt design is a meticulous task focused on crafting concise, clear, informative, and task-specific prompts. The objective is to input the model with adequate information to generate the desired output without being overall specific or general. The essence of prompt design lies in enabling the model to exhibit creativity and produce text that transcends mere replication of the prompt.

Both prompt design and prompt engineering are integral and important modules within prompt-based predictions. The substantial difference between them lies in their respective focuses and scopes, each tailored to the specific domain and task.

3.5. Prompt-guided predictions - MED-Prompt

Prompt-guided predictions are robust tools to generate text for various tasks such as code generation, script writing, completing musical pieces, email and letter samples, etc. Further, they can also be used for tasks such as question answering and summarization (Liu et al., 2023).

This study investigates prompt-guided predictions as a distinctive approach for improving medical text classification on clinical notes. The study presents an in-depth approach that optimizes deep learning-based pretrained models for medicine prediction on free-text clinical notes through prompt engineering. We carefully design prompts to aid the model in interpreting and extracting medical information from the clinical corpus. This includes incorporating domain-specific terminology and context-specific information and ensuring the accuracy and precision of the predictions, which are crucial in medical data analysis. Prompt-guided prediction is a key component of MED-Prompt framework, which takes a prompt input template as a prompt query and forwards it to the model. Within MED-Prompt, the prompts are designed to help reduce the overhead of training and prediction time and computational resources. Eq. (1) generalizes the overall concept of medicine prediction using a prompt. Further, Eq. (2) defines the construction of input sequences for positive and negative prompts, denoted as *N_input_text* and *P_input_text*, respectively.

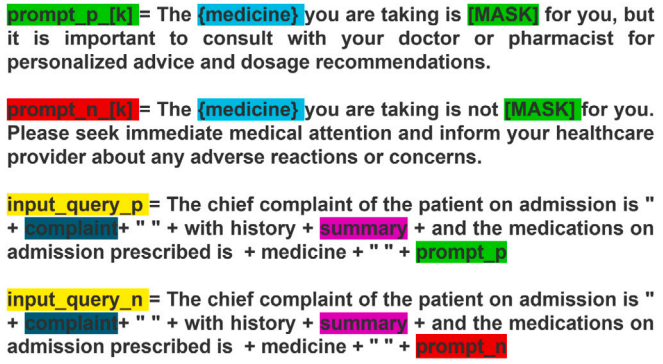

Fig. 2. MED-Prompt template.

$$\text{Prediction}(M, P) = \text{Model}(m, p) \quad \forall (m \in M, p \in P) \tag{1}$$

In this equation, "Prediction" represents the predicted outcome or probability of medicine m from a set of medicines M based on a given prompt p. The function "*Model*" represents the employed pretrained model, which takes the selected medicine m and prompt p as inputs to generate the prediction. The symbol ∀ denotes that the prediction is computed for every medicine m in the set of medicines M.

$$\begin{aligned} N\_input\_text &= \sum_{i=1}^{n} complnt_i + summary_i \\ &\quad + medicine_i + prompt\_list\_n\_[k] \\ P\_input\_text &= \sum_{i=1}^{n} complnt_i + summary_i \\ &\quad + medicine_i + prompt\_list\_p\_[k] \end{aligned} \tag{2}$$

For both sub-equations, i represents the index of the input sequence (ranging from 1 to n). $complnt_i$ and $summary_i$ is the patient's complaint and history text for ith input sequence, respectively. Further, $medicine_i$ presents the medicine information in the current input sequence. Lastly, $prompt\_list\_[k]$ shows the current prompt query being processed within MED-Prompt, whether positive or negative prompt text. The selection of prompt type depends on the specific query processed.

MED-Prompt: The core idea of MED-Prompt framework as presented in Fig. 1 lies in developing the specialized prompts, which act as guiding instructions for the models during the prediction process such that the first component of MED-Prompt is an input set of prompts. Prompts guide the generation of text or responses from a language model. In this context, prompts can be designed before or after prompt templates, which are used as input for baseline algorithms. Further, the patient's subset of data, such as medical history and current medications, is chosen and seamlessly integrated into designed prompts. This ensures contextually relevant information is fed to the model to interpret, as shown in the model architecture diagram within a gray-colored rectangle box, further detailed in Fig. 2. Feature extraction, involving identifying pertinent details from input data, is conducted based on positive and negative input prompts. Further, the "func = pre-trained-model" component integrates the pre-trained model, which generates a vector of predictions to which we compare the actual existence of medicines' names to finalize the output.

The Fig. 2 illustrates the prompt template within the MED-Prompt. The template has two main components: positive prompts (prompt_p)

[3] https://huggingface.co/.

and negative prompts (prompt_n). Each prompt follows a corresponding input query (input_query_p for positive and input_query_n for negative prompts). The positive prompts are designed to elicit contextual information that supports the prediction of a specific medicine. For instance, a positive prompt might ask, "What is the most likely medicine for a patient with a headache?" In contrast, negative prompts are designed to provide information that contradicts the prediction of a specific medicine. For example, a negative prompt might ask, "What is the least likely medicine for a patient with a headache?" Combining positive and negative prompts helps the model learn to distinguish between relevant and irrelevant information, improving its ability to predict medicines accurately. The number of prompts and input queries used during training determines the overall size and complexity of the prompt template. A larger prompt template can capture more contextual information and improve model performance. Overall, the prompt template plays a crucial role for both approaches of the proposed MED-Prompt, enabling it to effectively learn from both positive and negative examples and make accurate medicine predictions. One of the key strengths of the framework is its ability to handle any number of prompts, making it highly versatile and offering scalability. This independence is achieved through the framework's core design, which allows for the seamless integration of any number of positive and negative prompts as long as they maintain an equal balance. The framework is flexible to cope with various datasets. Further, it allows seamless integration with various pretrained models, leveraging their pre-existing knowledge and capabilities to enhance performance and reduce training time.

Further, the key details of with and without fine-tuned approaches are presented in Algorithms 1 and 2, which are responsible for computing predictions for provided sets of medicine cohorts m30, m50, m70, and mOrignal. Further, it depends on the number of records to be fed, the pretrained model, and the tokenizer type. We modified these algorithms per the need for different experiments conducted throughout the study.

In conclusion, Algorithm 1 presents the overall architecture in a high-level presentation, and we aim that future researchers can adopt it easily. Then, a modified version of our previous algorithm is shown in Algorithm 2. We have adopted the changes in this version and resumed the pretrained approach to leverage pretrained models for achieving a fine-tuned approach. By utilizing pretrained models, we aim to improve the performance of our algorithm, particularly in scenarios where the available training data is limited.

3.6. Classical model analysis and MED-Prompt architecture

Before delving into the MED-Prompt results, this study initially evaluated the classical machine learning methods for medicine prediction to highlight their inherent challenges, such as requiring substantial training time and computing resources, extensively labeled data, and lacking scalability. Table 4 presents the experimental results that underpin these challenges. The best performer among classical models, XGBClassifier (m30), exhibited the highest accuracy of 93.9% and an F1-score of 91.8%; however, time taken by this model recorded as 3150 s, approximately 52 min, which is underpinning the challenges, highlighting the urgent need for a system capable of addressing these issues effectively.

4. Experimental setup

This section describes the experimental setup and provides an overview of the baseline pretrained models and evaluation metrics used in the study.

4.1. System parameters

The study leveraged an NVIDIA TITAN RTX GPU for computational operations, equipped with driver version 470.161.03 and CUDA version 11.4. The PyTorch deep learning framework was employed to leverage the GPU's capabilities. During model training, a total of 24 217 MiB memory was utilized.

Table 4
Classical machine learning model evaluation results.

Model	mcount	Performance metrics			
		Acc	Pre	Rec	F1
XGBC	m30	**0.939**	**0.875**	**0.965**	**0.918**
	m50	0.900	0.820	0.951	0.880
	m70	0.900	0.820	0.951	0.880
MLPC	m30	**0.899**	**0.882**	0.820	**0.850**
	m50	0.719	0.881	0.114	0.202
	m70	0.724	0.837	0.126	0.219
RC	m30	**0.906**	**0.859**	**0.869**	**0.864**
	m50	0.774	0.864	0.391	0.538
	m70	0.774	0.864	0.391	0.538
RFC	m30	0.739	0.943	0.059	0.111
	m50	0.742	0.889	0.123	0.216
	m70	**0.749**	**0.845**	**0.185**	**0.303**
kNNC	m30	**0.710**	**0.436**	**0.131**	**0.201**
	m50	0.686	0.432	0.108	0.172
	m70	0.686	0.432	0.108	0.172
DTC	m30	**0.906**	**0.856**	**0.867**	**0.861**
	m50	0.875	0.814	0.837	0.825
	m70	0.868	0.812	0.837	0.824

Algorithm 1: MED-Prompt High-Level Algorithm - Using pretrained

```
input : modeltype, token, medicine_count, DATA_NUM
output: Vectors of Predicted and Actual Values
Initialize empty lists: results_list, matched_medicines, actual_result, predictions
for i ← 1 to DATA_NUM do
    Get inputs:
    chief complaint (complaint)
    medications (medications)
    summary for the current sample
    Concatenate inputs
    Clear the matched_medicines list
    for each medicine in medicine_count do
        if medicine.lower() is in medications.lower() then
            Append medicine to matched_medicines
        end
    end
    Join the elements of matched_medicines with '—' and assign it to matched_medicines_pattern
    Initialize lists:
    prompt_list_pos and prompt_list_neg
    for each medicine in medicine_count do
        for each prompt in prompt_list_pos and prompt_list_neg do
            Construct the input_text as in Equation (2)
            Initialize res_a and res_b
            Used the model to predict the masked token
            Compare the scores to determine the prediction
            if res_a > res_b then
                Set prediction to 1 (Positive)
            end
            else
                Set prediction to 0 (Negative)
            end
        end
    end
    Append a dictionary to results_list containing the 'Predicted Values' and 'Actual Values' lists
end
return Output Vectors
```

4.2. Baseline pretrained models

The baseline models used in this manuscript include BERT (Devlin et al., 2018), BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019; Huang et al., 2019). These models were selected after rigorous considerations as these models have been utilized for similar tasks in literature (Lu et al., 2023; Yao et al., 2019b). We aim to use them as a baseline for our proposed MED-Prompt framework and to evaluate the effectiveness and utilization of such models for medicine prediction tasks using prompt engineering. The BERT is the base model for both ClinicalBERT and BioBERT. The former two models were designed on top of BERT using core architecture.

4.2.1. BERT

BERT is a transformer-based model used for various general natural language-based tasks such as language generation, machine translation, Question Answering, sentiment analysis, named entity recognition, summarization, text classification, etc. It is pretrained on the large

Algorithm 2: MED-Prompt High-Level Algorithm -Transfer Learning

```
input : modeltype, token, medicine_count, DATA_NUM, Epoch_NUM
output: Vectors of Predicted and Actual Values
Initialize empty lists: results_list, matched_medicines, actual_result, predictions
for i ← 1 to Epoch_NUM do
    for i ← 1 to DATA_NUM do
        Get inputs:
        chief complaint (complaint)
        medications (medications)
        summary for the current sample (summary)
        Concatenate inputs
        Clear the matched_medicines list
        for each medicine in medicine_count do
            if medicine.lower() is in medications.lower() then
                Append medicine to matched_medicines
            end
        end
        Join the elements of matched_medicines with '—' and assign it to
          matched_medicines_pattern
        for each medicine in medicine_count do
            Initialize lists:
            prompt_list_pos and prompt_list_neg;
            for each prompt in prompt_list_pos and prompt_list_neg do
                Feed the input_text (pos and neg) separately to the Model for training
                Calculate Score using verbalize_score()
                Compare the scores to determine the prediction
                Calculate Loss using procedure loss()
                Adjusting train loss
                loss.backward()
                optim.step()
                if score_pos > score_neg then
                    Set prediction to 1 (Positive);
                end
                else
                    Set prediction to 0 (Negative);
                end
                actual_result;
            end
        end
        Append a dictionary to results_list containing the 'Predicted Values' and 'Actual
          Values' lists
    end
end
return Output Vectors;
```

corpus of English language data with case and uncased versions. Originally, this model was released in case and uncased variation with the English language. Further, this model has been pretrained with several large data corpora, including multilingual data. This study employed the "BERT Base Uncased" version of the model. At the time of writing, 24 versions have been released, including TinyBERT, MediumBERT, SmallBERT, MiniBERT, etc. (Devlin et al., 2018).

4.2.2. BioBERT

BioBERT is a specialized language representation model based on the BERT architecture, specifically designed for the biomedical domain. It is pretrained on a massive corpus of biomedical text data, including PubMed abstracts and full-text articles (Lee et al., 2020). It addresses the limitations of BERT by dealing with complexities of biomedical language specialized terminology, domain-specific abbreviations, and intricate relationships between entities.

4.2.3. ClinicalBERT

ClinicalBERT, another specialized language model based on BERT, is designed for the clinical domain. ClinicalBERT is pretrained on a massive corpus of clinical notes, electronic health records, and discharge summaries (Alsentzer et al., 2019). This pre-training allows ClinicalBERT to acquire a deep understanding and interpret clinical language, including medical terminologies, abbreviations, and relationships in the clinical context.

The experimental results show that the proposed MED-Prompt performed better with the ClinicalBERT model because it is pretrained on relevant data. This study performed several experiments on a subset of data extracted from the MIMIC-III dataset and then conducted various experiments to evaluate the proposed framework. MED-Prompt leverages pretrained natural language processing models based on deep learning architectures, including BERT, BioBERT, and ClinicalBERT, as baseline algorithms for experiments.

The models briefly discussed above are transformer-based, and their architecture consists of 12 encoder layers. The dimensionality of the hidden states is 768. Each transformer encoder block has two main components - a self-attention layer and a feed-forward network. This architecture ensures the model can process long text sequences and capture long-range dependencies specific to the task. As discussed above, BERT is a general-purpose model, while BioBERT and ClinicalBERT are designed for specific tasks. However, both models can be fine-tuned for domain-specific tasks, but the current study focuses on ClinicalBERT due to its wider application for clinical purposes. It is designed to capture specific domain knowledge and is suitable for healthcare tasks such as MED-Prompt fine-tuned for prompt guided medicine predictions. Further, additional technical details of individual models are beyond the scope of this study and are not discussed.

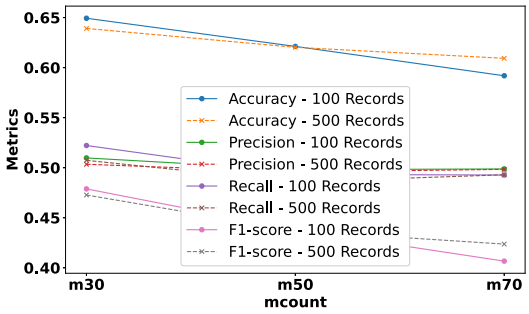

Fig. 3. Comparative analysis of evaluation matrices on BERT, BioBERT, and ClinicalBERT with Manual Prompt Set 1 with 100 and 500 records respectively.

Furthermore, additional experiments were conducted on different medicine cohorts for comparative analysis for a fine-tuned approach. The subsequent subsections present a comprehensive series of experiments to assess the effectiveness and efficiency of the proposed MED-Prompt framework. Additionally, Table 14 is prepared to summarize the key results obtained from these individual experiments.

4.3. Evaluation metrics

This study employs various metrics, including accuracy, precision, recall, and F_1 score as shown in (4), to evaluate the pretrained and fine-tuned models under different experimental setups. Amongst, F_1 is crucial to evaluate and compare the performance with existing studies. Achieving a higher F_1 indicates that the model minimizes false positives and negatives, showcasing a balanced predictive capability.

$$\text{tp} = \sum((\text{pred\_res}==1)\&(\text{test\_actual\_res}==1)) \tag{3a}$$

$$\text{fp} = \sum((\text{pred\_res}==1)\&(\text{test\_actual\_res}==0)) \tag{3b}$$

$$\text{fn} = \sum((\text{pred\_res}==0)\&(\text{test\_actual\_res}==1)) \tag{3c}$$

$$\text{tn} = \sum((\text{pred\_res}==0)\&(\text{test\_actual\_res}==0)) \tag{3d}$$

From Eq. (3), (tp, fp, fn, tn) are variables that help the model to evaluate predicted (true, false) and actual (true, false) and their combinations. Further, these four are used to calculate performance metrics of Eq. (4).

$$\begin{aligned} \text{Accuracy} &= \frac{\text{tp}+\text{tn}}{\text{tp}+\text{tn}+\text{fp}+\text{fn}} \\ \text{Precision} &= \frac{\text{tp}}{\text{tp}+\text{fp}} \\ \text{Recall} &= \frac{\text{tp}}{\text{tp}+\text{fn}} \\ \text{F1} &= 2 * \frac{\text{precision} * \text{recall}}{\text{precision}+\text{recall}} \end{aligned} \tag{4}$$

Table 5
Comparative results of manual vs. GPT prompts; The bold values are regarded as optimal results.

mCount	Model	Accuracy	Precision	Recall	F_1
Three pairs of (Positive and Negative) GPT prompts experiment with 500 records					
m30	BERT	0.626	0.501	0.502	0.462
	ClinicalBERT	**0.697**	**0.494**	**0.488**	**0.476**
	BioBERT	0.524	0.506	0.514	0.424
m50	BERT	0.639	0.502	0.504	0.448
	ClinicalBERT	**0.751**	**0.501**	**0.501**	**0.482**
	BioBERT	0.545	0.506	0.519	0.413
m70	BERT	0.649	0.500	0.501	0.431
	ClinicalBERT	**0.790**	**0.501**	**0.502**	**0.478**
	BioBERT	0.558	0.502	0.506	0.395
Three pair of (Positive and Negative) manual prompts experiment with 500 records					
m30	BERT	**0.752**	**0.498**	**0.496**	**0.489**
	ClinicalBERT	0.716	0.501	0.504	0.484
	BioBERT	0.263	0.508	0.514	0.255
m50	BERT	**0.770**	**0.500**	**0.499**	**0.482**
	ClinicalBERT	0.731	0.505	0.516	0.478
	BioBERT	0.255	0.506	0.517	0.238
m70	BERT	**0.787**	**0.502**	**0.511**	**0.473**
	ClinicalBERT	0.741	0.503	0.519	0.461
	BioBERT	0.245	0.502	0.509	0.217

Table 6
Comparative performance metrics for BERT_Large.

Prompt type	mCount	Accuracy	Precision	Recall	F_1
Manual Set1	m30	0.854	0.427	0.500	0.461
	m50	0.904	0.452	0.500	0.475
	m70	**0.960**	**0.480**	**0.500**	**0.490**
GPT Set1	m30	0.853	0.426	0.500	0.460
	m50	**0.902**	**0.452**	**0.498**	**0.474**
	m70	0.956	0.478	0.498	0.488
Manual ensemble	m30	**0.527**	**0.478**	**0.457**	**0.425**
	m50	0.516	0.484	0.453	0.398
	m70	0.525	0.499	0.492	0.376
GPT ensemble	m30	**0.358**	**0.498**	**0.497**	**0.342**
	m50	0.338	0.499	0.497	0.309
	m70	0.320	0.503	0.517	0.270

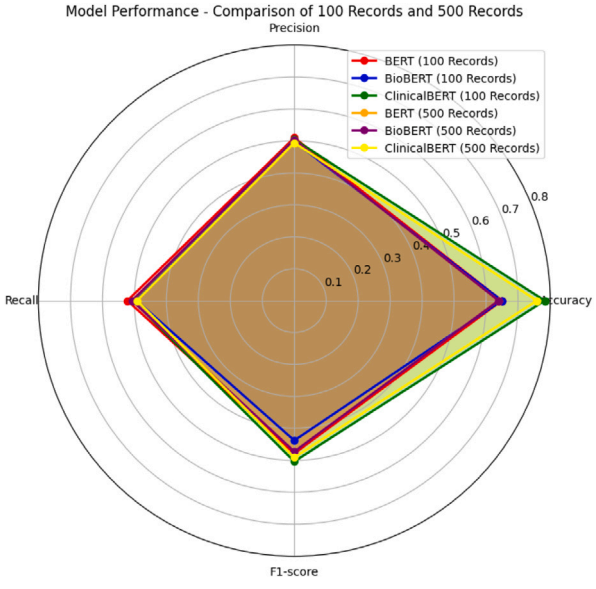

Fig. 4. Comparative analysis of evaluation matrices on BERT, BioBERT, and ClinicalBERT with Manual Prompt Set 1 with 100 and 500 records respectively (Radar plot analysis).

5. Results and discussion

This section provides a comprehensive analysis of the results obtained from the investigations undertaken to evaluate the effectiveness and efficiency of the MED-Prompt framework. Furthermore, this section provides detailed information about the exact outcomes achieved in each experiment and concludes with a summary of the noteworthy findings, accompanied by an analysis of their implications.

5.1. Comparative analysis

With this comparative analysis experimental setting, this study aims to record and analyze the performance of MED-Prompt framework with the baseline algorithms including but not limited to BERT, BioBERT, and ClinicalBERT using different experiments, such as Table 5 presenting comparative results between three sets of positive and negative prompts for Manual and GPT. Table 7 presents the comparative evaluations between 100 and 500 records set using all GPT prompts. Additionally, Table 9 shows an analysis between Manual Prompt Set 1 and 2 using 100 versus 500 records. Then, the study also presents detailed insights using a line plot analysis as shown in Fig. 5.

Table 5 presents the evaluations between the manually designed prompts set and GPT generated as listed in Table 2. We took three sets from each and experienced their performance with respective medicine counts starting from m30, m50, and m70. This is one of the most important experiments conducted as it highlights the performance of all pretrained models incorporated within MED-Prompt framework with a maximum number of records (500 records set as a selective cohort) and with no further deletion or addition to experimental settings. In this experiment, it is noticeable that MED-Prompt framework is highly dependent on the medicine set passed to the model as it achieved the maximum performance with a small medicine cohort (m30) irrespective of Manual or GPT prompts set. Further, it is also visible that Manual designed prompts are more effective in terms of medicine prediction tasks in MED-Prompt as we recorded a maximum score set of F1 (48.9%, 48.2%, and 47.3%) for all three medicine cohorts m30, m50, and m70 respectively with BERT remain the high achiever model in all cases. While with the GPT-Prompts set, BERT remains the top performer among the three models with an F1-score set of (46.2%, 44.8%, and 43.0%) for m30, m50, and m70, respectively.

Table 7 presents the individual evaluations of all GPT prompts as listed in Table 2. We fed each prompt as input to the model to compute ensemble results of medicine predictions using both 100 and 500

Table 7
Comparative results of GPT prompts (100 vs. 500 records); The bold values are regarded as optimal results.

Model	mCount	100 records				500 records			
		Accuracy	Precision	Recall	F_1	Accuracy	Precision	Recall	F_1
BERT	m30	**0.641**	**0.498**	**0.497**	**0.465**	0.883	0.510	0.503	0.495
	m50	0.655	0.500	0.500	0.452	0.914	0.511	0.504	0.502
	m70	0.666	0.500	0.498	0.435	**0.941**	**0.506**	**0.504**	**0.504**
BioBERT	m30	**0.632**	**0.502**	**0.506**	**0.465**	0.839	0.495	0.496	0.495
	m50	0.665	0.505	0.515	0.461	**0.850**	**0.654**	**0.536**	**0.534**
	m70	0.678	0.500	0.500	0.440	0.558	0.501	0.506	0.395
ClinicalBERT	m30	**0.634**	**0.501**	**0.500**	**0.465**	**0.833**	**0.516**	0.516	**0.516**
	m50	0.665	0.501	0.505	0.457	0.869	0.514	0.517	0.514
	m70	0.694	0.501	0.500	0.446	0.865	0.511	0.513	0.514

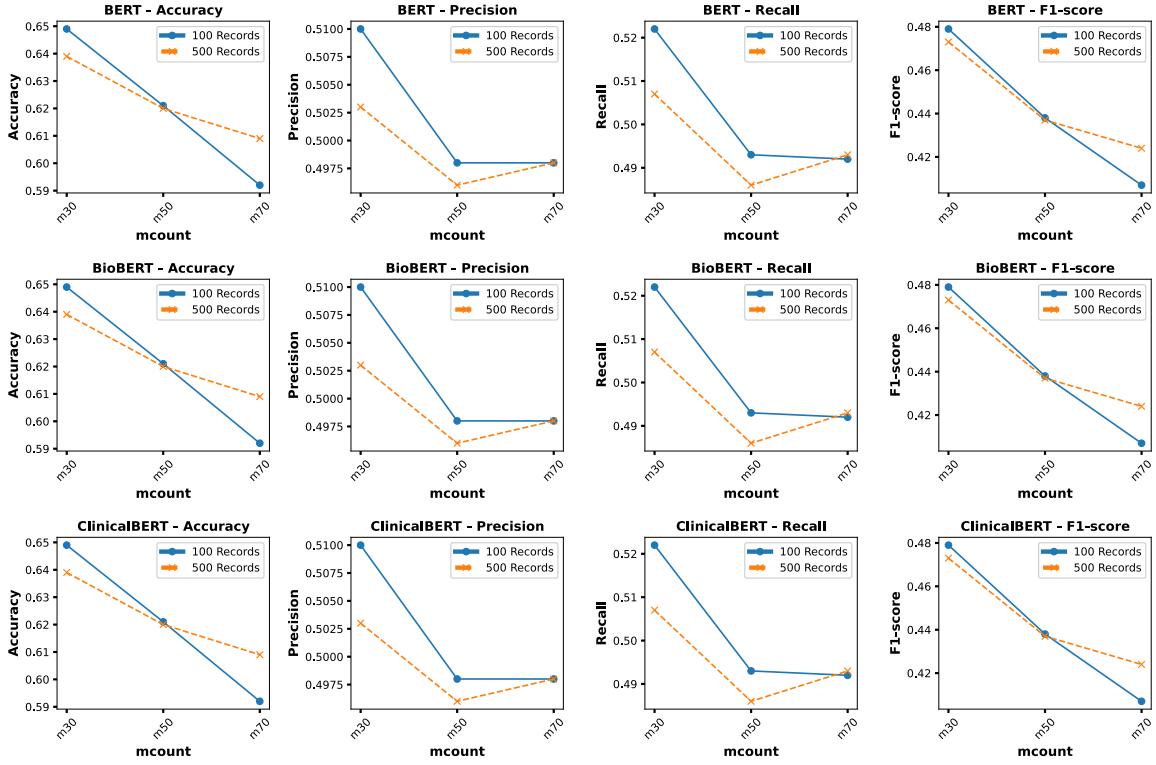

Fig. 5. Comparative analysis of evaluation matrices on BERT, BioBERT, and ClinicalBERT with Manual Prompt Set 1 with 100 and 500 records respectively (Line plot analysis).

Table 8
Comparative results of Manual vs. GPT prompts with mOrignal (Imbalance prompts); (Six pair of (Positive and Negative) GPT prompts experiment with 100 records) (Four pair of (Positive and Negative) manual prompts experiment with 100 records). The bold values are regarded as optimal results.

Model	Accuracy	Precision	Recall	F_1
BERT_Large	0.519	0.501	0.502	0.423
BERT	0.573	0.501	0.502	0.446
ClinicalBERT	0.641	0.495	0.490	0.465
BioBERT	**0.639**	**0.501**	**0.500**	**0.471**
BERT_Large	0.588	0.504	0.510	0.455
BERT	**0.708**	**0.508**	**0.513**	**0.493**
ClinicalBERT	0.638	0.502	0.504	0.472
BioBERT	0.331	0.500	0.500	0.313

records in each medicine cohort. The results indicate that more records improve performance across all evaluation metrics, including accuracy, precision, recall, and F1-score. In this experiment, the BioBERT model achieved the maximum scores (65.0%, 53.0%, and 53.0%) for precision, recall, and F1-score, respectively. While other models also achieve reasonable results compared to the previous experiment in Table 5. This evaluation analyzes the effectiveness of combining the maximum number of available prompts. As expected, this experiment achieved a high performance among all conducted experiments. This success can be attributed to the careful design of prompts. Despite being designed by GPT, we dedicated efforts to enhance their clarity and conciseness specifically tailored for medicine prediction tasks. Key findings/results are highlighted in bold.

In conclusion, the comparative experiment analysis setting investigated prompt-guided medicine prediction using several pretrained models and clinical note cohorts; further study assessed the effectiveness of medication counts m30, m50, and m70 with various prompt settings, as shown from Table 5 to 9. The evaluation results reveal the effectiveness of the proposed MED-Prompt framework, as demonstrated in Figs. 3 and 4. The analysis here supports the results.

5.2. Performance on key parameters

In this particular experimental setting, the study aims to observe the importance of key parameters of the framework, such as medicine count (mCount), the choice of the pretrained models, and the influence of varying numbers of data records. Through systematic evaluations, We observed and recorded all experiment outcomes. The experimental

Table 9
Comparative performance metrics for Manual Prompt Set 1 (manual_prompt1p and manual_prompt1n) and Set 2 (manual_prompt2p and manual_prompt2n) for 100 and 500 records, respectively; the bold values are recorded as optimal results.

Model		mCount	100 records				500 records			
			Accuracy	Precision	Recall	F_1	Accuracy	Precision	Recall	F_1
BERT	Set1	m30	**0.649**	**0.509**	**0.522**	**0.478**	**0.639**	**0.503**	**0.507**	**0.472**
		m50	0.621	0.498	0.493	0.438	0.620	0.495	0.485	0.437
		m70	0.592	0.498	0.492	0.406	0.609	0.498	0.493	0.423
	Set2	m30	0.850	0.466	0.484	0.472	0.655	0.505	0.512	0.470
		m50	0.878	0.475	0.482	0.478	0.630	0.497	0.489	0.435
		m70	**0.901**	**0.492**	**0.489**	**0.490**	**0.605**	**0.501**	**0.507**	**0.409**
BioBERT	Set1	m30	**0.610**	**0.510**	**0.503**	**0.464**	**0.640**	**0.504**	**0.507**	**0.472**
		m50	0.621	0.498	0.493	0.438	0.620	0.495	0.485	0.437
		m70	0.591	0.498	0.492	0.406	0.609	0.498	0.492	0.423
	Set2	m30	0.610	0.509	0.523	0.464	**0.593**	**0.514**	**0.540**	**0.455**
		m50	**0.657**	**0.512**	**0.538**	**0.465**	0.629	0.513	0.551	0.451
		m70	0.692	0.505	0.529	0.452	0.657	0.504	0.528	0.432
ClinicalBERT	Set1	m30	0.783	0.505	0.505	0.503	**0.760**	**0.493**	**0.491**	**0.491**
		m50	**0.788**	**0.511**	**0.523**	**0.503**	0.761	0.501	0.503	0.487
		m70	0.776	0.503	0.514	0.476	0.750	0.496	0.488	0.468
	Set2	m30	0.783	0.504	0.506	0.503	**0.774**	**0.495**	**0.492**	**0.489**
		m50	**0.788**	**0.511**	**0.523**	**0.504**	0.771	0.501	0.504	0.484
		m70	0.776	0.504	0.514	0.476	0.762	0.502	0.505	0.465

Table 10
Comparative analysis of ablation evaluation results between GPT and manual prompts.

mCount	Model	Accuracy	Precision	Recall	F_1	Ablation setting
GPT ablation with 100 records						
m30	BERT	**0.879**	**0.502**	**0.500**	**0.484**	GPT_AS1
	ClinicalBERT	0.225	0.517	0.521	0.225	
	BioBERT	0.688	0.499	0.499	0.480	
m30	BERT	0.249	0.504	0.506	0.245	GPT_AS2
	ClinicalBERT	0.884	0.446	0.495	0.469	
	BioBERT	**0.755**	**0.524**	**0.539**	**0.522**	
m30	BERT	**0.893**	**0.446**	**0.500**	**0.472**	GPT_AS3
	ClinicalBERT	0.582	0.506	**0.516**	0.451	
	BioBERT	0.892	0.446	0.499	0.471	
m50	BERT	0.209	0.498	0.496	0.203	GPT_AS2
	ClinicalBERT	0.922	0.463	0.497	0.480	
	BioBERT	**0.795**	**0.520**	**0.541**	**0.516**	
m70	BERT	**0.957**	**0.478**	**0.500**	**0.489**	GPT_AS3
	ClinicalBERT	0.564	0.497	0.485	0.394	
	BioBERT	0.956	0.478	0.500	0.489	
Manual ablation with 100 records						
m30	BERT	0.690	0.509	0.517	0.490	Manual_AS1
	ClinicalBERT	**0.803**	**0.500**	**0.500**	**0.500**	
	BioBERT	0.358	0.512	0.526	0.332	
m30	BERT	**0.735**	**0.473**	**0.459**	**0.463**	Manual_AS2
	ClinicalBERT	0.511	0.502	0.505	0.415	
	BioBERT	0.230	0.507	0.510	0.228	
m30	BERT	**0.686**	**0.501**	**0.502**	**0.481**	Manual_AS3
	ClinicalBERT	0.373	0.494	0.487	0.336	
	BioBERT	0.469	0.499	0.497	0.393	
m50	BERT	**0.879**	**0.488**	**0.492**	**0.489**	Manual_AS2
	ClinicalBERT	0.475	0.504	0.516	0.379	
	BioBERT	0.105	0.498	0.499	0.104	
m70	BERT	**0.725**	**0.502**	**0.508**	**0.458**	Manual_AS3
	ClinicalBERT	0.306	0.495	0.472	0.260	
	BioBERT	0.451	0.502	0.509	0.346	

results from Tables 5 to 11 demonstrate the significance and importance of key components and further emphasize their impact. Additionally, the study presents a concise overview in the following manner.

Impacts of mCount: The variation in the predicted medicines was analyzed in the mCount experiment. Four medicine cohorts were employed, such as m30, m50, m70 and mOriginal. There are various experiments to highlight the influence of mCount on MED-Prompt. Like, with the pretrained approach, BERT outperformed other models, especially achieving F1 scores of (48.9%, 48.2%, and 47.3%) for m30, m50, and m70, respectively, with ensemble prompt experiment as recorded in Table 5. Further, in this experiment, GPT-Prompts performed slightly lower, with F1-scores of (46.2%, 44.8%, and 43.1%) for m30, m50, and m70, respectively. The results highlight the dependency of MED-Prompt on specific medicine cohorts passed as input. Pretrained BERT consistently performed better, indicating its suitability for small medicine cohorts. The choice of medicines significantly influenced model performance.

Table 11
Comparative analysis between Normal setting vs. Ablation setting using Manual Prompt Set 3 and 4 with 100 records.

Model	mcount	Accuracy	Precision	Recall	F_1	Ablation setting
Prompt Set 3 with Normal setting						
BERT	m30	0.722	0.487	0.488	0.475	
BioBERT	m50	0.479	0.508	0.532	0.385	Normal
ClinicalBERT	m70	**0.956**	**0.478**	**0.500**	**0.488**	
Prompt Set 3 with Ablation setting						
BERT	m30	0.482	0.462	0.401	0.372	
BioBERT	m50	0.103	0.505	0.502	0.102	Manual_AS2*
ClinicalBERT	m70	**0.957**	**0.479**	**0.501**	**0.488**	
Prompt Set 4 with Normal setting						
BERT	m30	**0.578**	**0.527**	**0.571**	**0.466**	
BioBERT	m50	0.545	0.508	0.532	0.416	Normal
ClinicalBERT	m70	0.459	0.495	0.475	0.345	
Prompt Set 4 with Ablation setting						
BERT	m30	0.577	0.466	0.417	0.412	
BioBERT	m50	0.310	0.502	0.506	0.281	Manual_AS2**
ClinicalBERT	m70	**0.626**	**0.504**	**0.521**	**0.426**	

Impacts of pretrained model: The choice of the pretrained model significantly influences the performance of the proposed MED-Prompt framework. There are several experiments to discuss the influence of model selection as results presented from Tables 5 to 9. To present sample results to evaluate the influence of different pretrained models, this study kept the medicine count m50 cohort constant and training records constant. From Table 5, ClinicalBERT demonstrated the best performance, followed by BERT and BioBERT with F1-scores of (48.2%, 44.8%, and 41.3%). This model's advantage likely stems from its specialized training with clinical notes, allowing it to grasp medical nuances and contextual intricacies crucial for accurate medicine prediction. The model's architecture, training dataset, and domain knowledge help it to analyze and handle text data. ClinicalBERT is the most effective pretrained model in the MED-Prompt framework since it aligns with the clinical domain and extracts valuable insights from patient records. This emphasizes the necessity of choosing a task and domain-specific pretrained model.

Impacts of the number of training records: Lastly, to show the significance of data records for initiating any experiment, from the results, it is concluded that increasing the number of records for training significantly influences the model outcome. As evidenced by the experimental results, increasing the number of training records from 100 to 500 significantly improved model performance. However, it is essential to acknowledge that larger data size enhances performance and comes with trade-offs. Larger data sizes demand more training time and computational resources. The presented results emphasize the importance of the number of records for refining MED-Prompt framework predictions and achieving optimal performance.

5.3. Ablation study

The ablation experiments evaluated the effectiveness of particular core components or sub-modules within the MED-Prompt framework. Below are listed distinct settings designed for GPT and manually designed prompt ablations. All results are recorded in Tables 10 and 11.

- GPT_AS1: Tested only with P1 & P3 and deleted Major Complaint
- GPT_AS2: Tested only with P2 & P4 and deleted summary in existing architecture
- GPT_AS3: Tested only with P5 & P6 and deleted summary, but added modified MOD in existing architecture
- Manual_AS1: Tested only with P1 & P3 and deleted Major Complaint
- Manual_AS2: Tested only with P2 & P4 and deleted summary in existing architecture
- Manual_AS2*: Tested only with P3 and deleted summary in existing architecture
- Manual_AS2**: Tested only with P4 and deleted summary in existing architecture
- Manual_AS3:Tested only with P1 & P4 and deleted summary + but added modified MOD in existing architecture

Initially, three experiments were conducted by keeping the medicine count constant to m30 for the set of (AS1, AS2, and AS3) for GPT and Manual, respectively, as recorded in Table 10. Subsequently, we extended the experiments with AS2 and AS3 with medicine counts m50 and m70, respectively. Although several ablation experiments can be designed, we deliberately prioritized a limited number due to time constraints.

In the case of GPT_AS1, there was a notable decrease in the performance of the ClinicalBERT Model. This decline is attributed to the altered prompt text, which plays a significant role in the [MASK] prediction task. Additionally, deleting the Major_Complaint feature from the input has contributed to this decrease in results, alongside the limitation imposed by the medicine count. However, in subsequent ablation studies, ClinicalBERT exhibited improved performance compared to other models. On the other hand, both BERT and BioBERT showed consistent performance, with minimal differences observed in their results with and without ablation. Interestingly, in Manual_AS1, ClinicalBERT continued to perform well among the pretrained models with manual prompts. This indicates that the specific text used in the prompts played a role in the improved performance, whereas BioBERT's performance appeared to be considerably hindered.

In the case of GPT_AS2 and Manual_AS2, experiments were limited to prompt pair two and prompt pair four as listed in Table 2 and deleted the summary component to evaluate its impact. With GPT_AS2, it was noticed that the performance (except accuracy) remains approximately the same. From evaluations, it was observed that prompt text (gpt_prompt_2p and gpt_prompt_2n) and (gpt_prompt_4p and gpt_prompt_4n) were identified as the most effective among the six pairs of prompts. They possess the high potential to predict the most relevant mask [MASK]. Further, simultaneously with Manual_AS2, we noticed a significant drop in the F1-score for the BioBERT model, but precision and recall were approximately the same for all other pretrained models as per our expectations. Additionally, Table 10 details the comparative analysis between m30 and m50 for GPT_AS2 and Manual_AS2. In the GPT_AS2 setting, BioBERT achieved an F1-Score of 52.2% for m30 and 51.6% for m50. Meanwhile, with Manual_AS2, BERT achieved an F1-Score of 46.3% for m30 and outperformed the other two models with an F1-score of 48.9% for m50.

The GPT_AS3 ablation setting aimed to compare and assess the influence of prompt pairs P5 and P6. We intentionally excluded the summary feature, but Medication_On_Discharge was added in this ablation setting. Further, medicine count m70 was selected since medicine counts m30 and m50 were previously analyzed with GPT_AS1 and GPT_AS2, respectively. The results of GPT_AS3 showed that the BERT and BioBERT models achieved the same F1-score of 48.9%, indicating consistent performance with the GPT_AS3 ablation setting. However, the ClinicalBERT model exhibited the lowest performance among the evaluated models. This analysis sheds light on the effectiveness of prompt pairs P5 and P6 in medicine prediction for a larger cohort (m70). By comparing GPT_AS3 with GPT_AS2, we can further understand the influence of different prompt engineering strategies and features on the prediction performance of the MED-Prompt framework. At last, with Manual_AS3, prompt P1 and P4 pair were evaluated. In this ablation setting, we noticed BERT achieved high performance in terms of an F1-score of 45.8%, and at the same time, results remained compromised for both ClinicalBERT and BioBERT.

Table 12
Performance evaluation of individual medicines using a ClinicalBERT-based fine-tuned approach results. The results are shown for each medicine, with minimum, maximum, and mean values across the test set.

Medicine	Accuracy	Precision	Recall	F_1	Avg epoch time
Aspirin	0.800 0.800 0.800	0.800 0.800 0.800	1 1 1	0.888 0.888 0.888	1.98
Lisinopril	0 1 0.990	0 1 0.910	0 1 0.920	0 1 0.910	1.68
Atenolol	0.683 0.816 0.793	0.822 0.848 0.844	0.790 0.950 0.922	0.806 0.986 0.881	1.76
Spironolactone	0.600 0.929 0.729	0.750 1 0.911	0.750 0.911 0.732	0.750 0.953 0.812	2.48
Flowmax	0.167 1 0.583	0.167 1 0.583	1 1 1	0.286 1 0.737	1.92
Nitroglycerin	0.600 0.743 0.615	0.758 0.801 0.763	0.718 0.895 0.738	0.738 0.845 0.750	1.85
Levothyroxine	0.915 **0.938** 0.926	0.938 **0.946** 0.940	0.968 **1** 0.983	0.955 **0.968** 0.961	**1.21**
Average (Min)	**0.532**	**0.542**	**0.684**	**0.604**	N/A
Average (Max)	**0.812**	**0.835**	**0.971**	**0.879**	N/A
Average (Mean)	**0.722**	**0.779**	**0.895**	**0.822**	N/A

In continuation of the previous ablation evaluation, as in Table 10, we further present the comparative analysis between Normal experimental settings versus Ablation experimental settings to evaluate the effectiveness of our proposed model more specifically. This evaluation limits experiments to modified Manual_AS2* and Manual_AS2**. The choice of Manual_AS2 was due to its substantial influence on the overall results.

Table 11 presents a comparative analysis of normal and ablation tests performed using Prompt Sets 3 and 4. These particular prompts were not assessed in previous investigations. BioBERT displayed very poor outcomes. As a result, we were motivated to reassess our experimental methodology, which led us to discover that the prompt's design substantially impacted the model's prediction. When we switched to the modified Manual_AS2*, we saw that BioBERT had better F1 scores, whilst BERT's performance declined. The performance of ClinicalBERT was consistent across both the normal and modified Manual_AS2* ablation settings. With Manual_AS2**, the performance of BERT and BioBERT models was negatively affected, whereas there was little improvement observed in the case of the ClinicalBERT model.

In this ablations experimental setting, the analysis focuses on the effects of removing major components from the framework, pairing different prompt sets. Systematically modifying specific components allows for comparing the performance with the complete framework, revealing key factors influencing its performance. These findings can guide future improvements and optimizations in the MED-Prompt framework for enhanced medicine prediction in clinical notes.

5.4. Fine-tuned ClinicalBERT results discussion

For this particular experiment, the study evaluates a mOrignal set of medicines to highlight the strength of the proposed MED-Prompt framework. Experimenting with an original set of medicines illustrates and validates the effectiveness in actual data. The plots, illustrated in Figs. 6 and 7, depict the training, test, and validation results for prediction of a selected set of medicines using novel MED-Prompt fine-tuned ClinicalBERT as shown in Algorithm 2. The presented plots also represent training and testing values recorded in Tables 12 and 13. Wherein we maintain the record of training (minimum score) in Table 13 and for the testing portion, we present (minimum, maximum, and average) results respectively in Table 12.

Table 13
Performance evaluation of individual medicines using a ClinicalBERT-based fine-tuned approach results. The results are shown for each medicine, with minimum values across the train set.

Medicine	Accuracy	Precision	Recall	F_1
Aspirin	0.957	0.960	0.984	0.978
Lisinopril	0.875	0.750	0.750	0.750
Atenolol	0.850	0.906	0.906	0.918
Spironolactone	0.879	0.911	0.961	0.935
Flowmax	0.583	0.583	1	0.737
Nitroglycerin	0.830	0.878	0.939	0.907
Levothyroxine	0.922	0.928	0.988	0.959

Given the numerous experiments conducted on various individuals and medicine combinations, we showcase selective plots for their significance here. The plots, shown in Figs. 6 and 7, reveal promising training, validations, and testing patterns. The results demonstrate that the fine-tuned approach achieved an average F1 score of 87.9%, also recorded (60.4%, 87.9%, and 82.2%) for minimum average, maximum average, and average of average scores, respectively. These plots portray predictions of individual medicines (Aspirin, Lisinopril, and Metoprolol) and small sets of medicines, including combinations of medicines_sets (1–5, 25–31, and 65–69). In Fig. 6, a well-balanced train and validation loss for individual medicine experiments and the group of medicines. Further, the study also conducted experiments with a larger medicine group, such as a set of initial 15, random 15 medicines as well as the last 15 medicines, a further set of 30 medicines followed by a set of 50 medicines, and a set of 90 medicines as shown in Fig. 8. Extensive experimentation shows that the MED-Prompt improves with a large set of medicines. This suggests that adding more medicines improves the model's prediction power and efficacy. A bigger range of medicines gives the algorithm additional context and knowledge to make accurate predictions.

In conclusion, the find-tuned experiments were conducted with (500 records for the train set, 100 for validation, and 100 for the test set) using 50 epochs. After several attempts to select proportions for the training, validation, and testing sets, the allocation was optimal. The reason behind this allocation was determined based on a combination of scientific assumptions, resource constraints, and statistical considerations to ensure the robustness and generalizability of our study's findings. Experiments have shown that our MED-Prompt fine-tuned approach can achieve high performance even with fewer records. These experiments suggest that the model learns from the data effectively and generalizes well to new examples.

5.5. Summary of MED-Prompts Result

This study conducted a series of experiments utilizing the MED-Prompts framework. The experiments were structured into distinct categories, encompassing evaluating classical machine learning models, comparative results of Manual vs. GPT prompts, comparative performance metrics for GPT prompts with varying record numbers and medicine set, and the same with the rest of the nature of experiments.

The initial examination was conducted to evaluate the classical method for medicine prediction, and the results in Table 4 indicate that XGBClassifier (m30) performed the best among classical machine learning models with an accuracy of 93.9% and an F1-score of 91.8%. The total experiment time was recorded as 3150 s, approximately 52 min. ClinicalBERT (m50-GPT Prompt) and BERT (m30-Manual Prompt) showed approximately similar accuracy of 75.1% and 75.2%, respectively. Both had lower F1 scores of 48.2% and 48.8%. Using BioBERT, we examined how record count affects GPT prompts. Further, ClinicalBERT compared manual prompt sets 1 and 2 with 100 records.

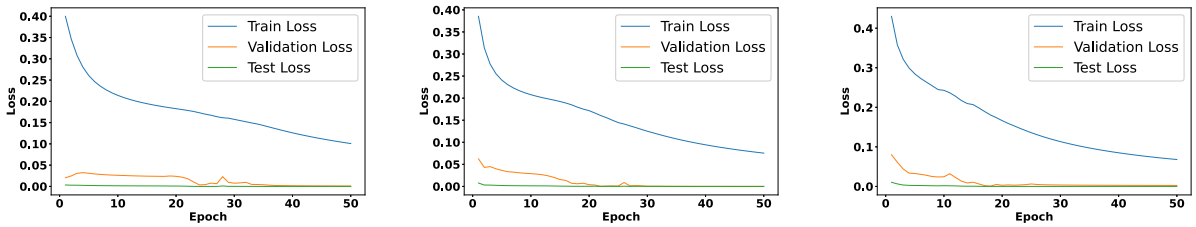

Fig. 6. Train, test, and validation plots of various individual medicines using fine-tuned ClinicalBERT.

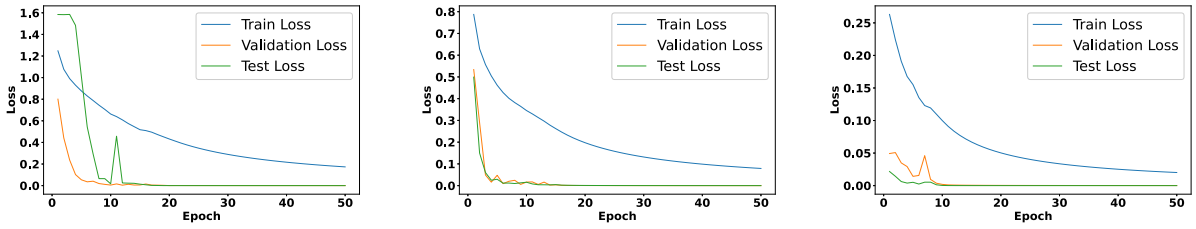

Fig. 7. Train, test, and validation plots of various small sets of medicines using fine-tuned ClinicalBERT.

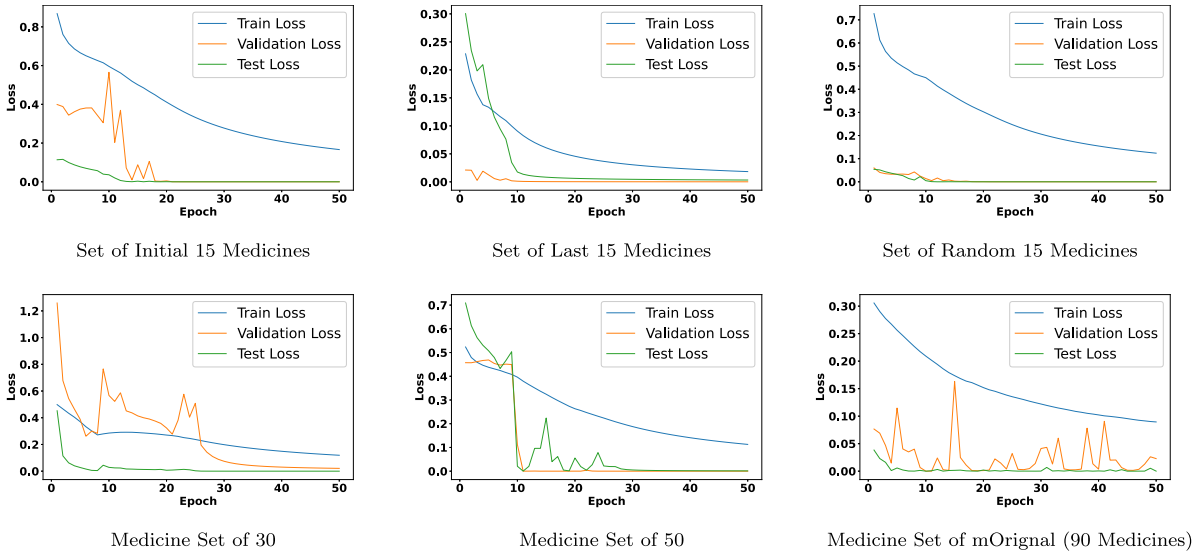

Fig. 8. Train, test, and validation plots of various large sets of medicines using fine-tuned ClinicalBERT.

Both sets had 78.8% accuracy and 50.3% F1 scores. BioBERT and ClinicalBERT compared GPT and manual ablation results. BioBERT (m30-GPT_AS2) scored 75.5% and ClinicalBERT (m30-Manual_AS1) 80.3%. Manual prompt sets 3 and 4, comprising 100 records each, were compared against standard and ablation configurations. In the case of ClinicalBERT (m70-Manual_AS2*-Set3), the achieved performance metric was 95.7%, whereas, for BERT (m30-Normal-Set4), the corresponding metric was 57.8%.

The experiment, as shown in Table 6, was conducted to evaluate the comparative performance of BERT_Large with different sets of prompts. The study also examined the consequences of an imbalance in prompts, especially assessing the implications of having an unbalanced number of ensemble prompts. The experiment utilized a BERT-large including three other pretrained models to evaluate ensembles of four and six prompt sets, as specified in Table 8. Notably, BioBERT exhibited superior performance. This also entails a pretrained model's relationship and efficacy in handling individual and ensemble prompts.

Table 14 provides a comprehensive summary of the MED-Prompts results across various experiments. It also records time metrics, illustrating comparative results of time evaluations between the proposed framework's approaches. For pretrained experiments, Time1 metric results are recorded as the average time taken by each record. Then, the Time2 metric is recorded by multiplying the number of records. For the fine-tuned approach, time is calculated as time taken by each epoch multiplied by the number of epochs. The pretrained models yielded a minimum average time of 2.89 s. In contrast, the proposed framework achieved an average time per epoch recorded at 1.21 s, which concluded that the MED-Prompt fine-tuned approach achieves 2.38% times better performance than the initial baseline approach.

To conclude, fine-tuned ClinicalBERT reached 93.8% accuracy and 96.8% F1-score. Further, it was noticed that the performance of GPT-Prompts and Manual-Prompts remain parallel, but Manual-Prompts achieves better results. This suggests that carefully designed prompts help the framework to achieve higher results.

Table 14
Summary of MED-Prompt results concerning nature of experiments — Score selected from the respected table is based on Max F1-Score and for details about each experiment refer to the Tables 4 to 13. Time1 is the average time of Epoch/Record, and Time2 is the total experiment time. Both time metrics are in seconds.

Nature of the experiment	Model	Accuracy	Precision	Recall	F_1	Time1	Time2
Classical machine learning model evaluation results	XGBClassifier (m30)	0.939	0.875	0.965	0.918	N/A	3150
Comparative results of set of three manual vs. GPT prompts	ClinicalBERT (m50-GPT Prompt)	0.751	0.500	0.501	0.482	5.15	2575
	BERT (m30-Manual prompt)	0.752	0.498	0.496	0.488	3.23	1615
Comparative results of GPT prompts (100 vs. 500 records)	BioBERT (m30-100records)	0.632	0.502	0.506	0.465	3.03	303
	BioBERT(m50-500records)	0.850	0.654	0.536	0.534	4.58	2290
Comparative performance metrics for Manual Prompt Set 1 and Set 2 for 100 and 500 records	ClinicalBERT (m50-Set1-100records)	0.788	0.511	0.523	0.503	2.89	289
	ClinicalBERT (m50-Set2-100records)	0.788	0.511	0.523	0.504	3.12	312
Comparative analysis of ablation evaluation results between GPT and manual prompts	BioBERT (m30-GPT_AS2)	0.755	0.524	0.539	0.522	2.94	294
	ClinicalBERT (m30-Manual_AS1)	0.803	0.500	0.500	0.500	3.17	317
Comparative analysis between Normal setting vs. Ablation setting using Manual Prompt Set 3 & 4 with 100 records	ClinicalBERT (m70-Manual_AS2*-Set3)	0.957	0.479	0.501	0.488	6.52	652
	BERT (m30-Normal-Set4)	0.578	0.527	0.571	0.466	2.95	295
Performance metrics for BERT_Large	(m70-Manual-Set 1)	0.960	0.480	0.500	0.490	7.27	727
	(m30-Manual-Ensemble of manual four sets)	0.527	0.478	0.457	0.425	4.19	419
Performance metrics for mOrignal	BERT (mOriginal)	0.708	0.508	0.513	0.493	7.39	3695
Performance metrics for transfer learning	ClinicalBERT (Levothyroxine)	0.938	0.946	1	0.968	1.21	60.5

The experimental results suggest the following conclusions for various parameters:

- Medicines Count: The number of medicines in the prediction task significantly influences the model's performance as each subfigure of Fig. 8 shows the influence of medicine count on the proposed model and their performance metrics. Experimental studies have shown that the prediction performance can vary greatly depending on the specific medicines being predicted (Kirchmair et al., 2015). Some drugs may exhibit easier relations, making them simpler to anticipate, while others might present more complex relationships and challenges for the model (Shehab et al., 2022). The medicine count is a critical factor to consider when evaluating the performance of the MED-Prompt's framework. However, it is essential to recognize that medicine count is not the sole determinant of optimized or improved results. Other key factors, such as the prompt engineering strategy, pretrained model selection, and dataset characteristics, also considerably impact the overall performance. Medicine prediction depends on the pretrained model (Clavié et al., 2023). Different models have different strengths and disadvantages, which affect accuracy, precision, memory, and F1 scores. Choose a pretrained model based on task requirements and compatibility for best results.
- Set of Prompts: The MED-Prompts framework's number of prompts can affect its capacity to extract meaningful clinical content. More prompts may improve the model's medicine prediction task capture and bring noise and redundant information. Accurate and efficient forecasts require a significant number of prompts.
- Ablation Setting: Ablation settings play a crucial role in understanding the contributions of different components in the MED-Prompt framework. Ablation experiments in the MED-Prompt elucidate the impact of modifying or removing specific model components. These analyses reveal prompt engineering's pivotal in enhancing model performance by assessing each prompt's significance. By isolating components, these experiments unveil influential prompts and offer insights for refining and optimizing the framework, guiding improvements for optimal results.

In summary, pretrained models and transfer learning make the MED-Prompts system flexible for medicine prediction. Medicines, models, prompts, ensemble tactics, and ablation settings affect framework performance. To find the best prediction configurations, careful experimentation and analysis are needed. The study also shows that transfer learning improves medication prediction accuracy and F1 scores, suggesting future research in healthcare applications. MED-Prompt elevates accuracy through domain-specific prompts, minimizing training time and enhancing computational efficiency, offering improved resource utilization and reduced costs for large-scale medical text classification.

6. Limitations and future works

Acknowledging the study's limitations is essential to ensure validity and transparency. Three major limitations are identified that warrant consideration for future work.

Generalization to different datasets: This work only uses MIMIC-III to evaluate the proposed model, which hinders comprehension of scenarios where symptoms associated with multiple other symptoms and diseases might arise as complications of various other conditions. However, exploring and evaluating models using other clinical note datasets would be valuable to assess their robustness and generalizability across different healthcare datasets. To address it, it is possible to access many clinical text datasets containing such information or necessitate modifications to the current framework, particularly prompt guided prediction component setting.

Domain-specific pretrained models: Another limitation lies in using domain-specific models. This study has experienced three models (BERT, BioBERT, and ClinicalBERT) among several available options. The selected models effectively handle the challenges and nuances clinical notes present. Considering limited model utilization, we recommend further exploitation of domain-specific pretrained models. Employing and validating several other models will enhance the credibility of the proposed MED-Prompt framework. Moreover, the transfer learning approach relied solely on ClinicalBERT, indicating a potential limitation. Future directions include comparative studies involving other models to enhance the proposed approach.

Generalization of prompt engineering techniques: The current study concentrated on prompt engineering for medicine prediction using pretrained models. Exploring additional techniques like zero-shot or few-shot learning would broaden the scope.

The landscape for future directions in prompt-based medicine predictions presents several noteworthy challenges that warrant attention

from researchers. The above discussed are a few from directly presented work. At the same time, some of the additional challenges include (i) multilingual medicine prediction by integrating multilingual datasets, (ii) achieving interpretability and reliability in medicine predictions, and (iii) integrating prompts with Medical Imaging. Addressing these challenges significantly enhances prompt-based medicine predictions, transforming healthcare facilities and improving patient outcomes and experiences.

7. Conclusion

In conclusion, this study introduces the MED-Prompt, a fine-tuned prompt-guided framework for medicine prediction in healthcare. The study utilized the MIMIC-III dataset and employed data engineering techniques to create a subset of relevant clinical text. Comparative experiments assessed baseline pretrained models like BERT, BioBERT, and ClinicalBERT, along with fine-tuned ClinicalBERT, introducing and analyzing Manual Prompts and GPT Prompts. The significant progress from 53% to 96% in terms of F_1 Score indeed demonstrates the effectiveness of the proposed MED-Prompt framework. Detailed investigations showed that the performance of "GPT Prompts" and "Manual Prompts" remained parallel for several experiments, but with certain scenarios, Manual-Prompts outperformed. This suggests that carefully designed prompts help the framework to achieve higher results. Importantly, MED-Prompt achieved substantial progress while utilizing fewer computational resources and time, showcasing robustness and efficiency. These findings hold implications for optimizing personalized care and improving healthcare outcomes, emphasizing the significance of prompt techniques in medicine prediction. Further exploration and refinement of this framework can lead to even greater healthcare decision-making and patient care.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and experiment source files supporting this study's findings are available from the first author upon reasonable request.

References

Ahmad, I., Asghar, Z., Kumar, T., Li, G., Manzoor, A., Mikhaylov, K., Shah, S.A., Höyhtyä, M., Reponen, J., Huusko, J., et al., 2022. Emerging technologies for next generation remote health care and assisted living. IEEE Access 10, 56094–56132.

Ahmed, A., Xi, R., Hou, M., Shah, S.A., Hameed, S., 2023. Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts. IEEE Access.

Ali, O., Abdelbaki, W., Shrestha, A., Elbasi, E., Alryalat, M.A.A., Dwivedi, Y.K., 2023. A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. J. Innov. Knowl. 8 (1), 100333.

Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., McDermott, M., 2019. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

CernerHealthFacts, 2023. Cerner health facts database. Accessed June 23, 2023. https://uthsc.edu/cbmi/data/cerner.php.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., et al., 2023. A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109.

Clavié, B., Ciceu, A., Naylor, F., Soulié, G., Brightwell, T., 2023. Large language models in the workplace: A case study on prompt engineering for job type classification. In: International Conference on Applications of Natural Language To Information Systems. Springer, pp. 3–17.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H., Sun, M., 2022. OpenPrompt: An open-source framework for prompt-learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 105–113.

Feng, J., Phillips, R.V., Malenica, I., Bishara, A., Hubbard, A.E., Celi, L.A., Pirracchio, R., 2022. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit. Med. 5 (1), 66.

Fernandes, M., Vieira, S.M., Leite, F., Palos, C., Finkelstein, S., Sousa, J.M., 2020. Clinical decision support systems for triage in the emergency department using intelligent systems: a review. Artif. Intell. Med. 102, 101762.

Firouzi, F., Farahani, B., Daneshmand, M., Grise, K., Song, J., Saracco, R., Wang, L.L., Lo, K., Angelov, P., Soares, E., et al., 2021. Harnessing the power of smart and connected health to tackle COVID-19: Iot, AI, robotics, and blockchain for a better world. IEEE Internet Things J. 8 (16), 12826–12846.

Gao, S., Alawad, M., Young, M.T., Gounley, J., Schaefferkoetter, N., Yoon, H.J., Wu, X.-C., Durbin, E.B., Doherty, J., Stroup, A., et al., 2021. Limitations of transformers on clinical text classification. IEEE J. Biomed. Health Inform. 25 (9), 3596–3607.

Gao, S., Qiu, J.X., Alawad, M., Hinkle, J.D., Schaefferkoetter, N., Yoon, H.-J., Christian, B., Fearn, P.A., Penberthy, L., Wu, X.-C., et al., 2019. Classifying cancer pathology reports with hierarchical self-attention networks. Artif. Intell. Med. 101, 101726.

Gupta, N.S., Kumar, P., 2023. Perspective of artificial intelligence in healthcare data management: A journey towards precision medicine. Comput. Biol. Med. 107051.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964.

Herrett, E., Gallagher, A.M., Bhaskaran, K., Forbes, H., Mathur, R., Van Staa, T., Smeeth, L., 2015. Data resource profile: clinical practice research datalink (CPRD). Int. J. Epidemiol. 44 (3), 827–836.

Huang, K., Altosaar, J., Ranganath, R., 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342.

Hughes, M., Li, I., Kotoulas, S., Suzumura, T., 2017. Medical text classification using convolutional neural networks. In: Informatics for Health: Connected Citizen-Led Wellness and Population Health. IOS Press, pp. 246–250.

Javaid, M., Haleem, A., Singh, R.P., Suman, R., Rab, S., 2022. Significance of machine learning in healthcare: Features, pillars and applications. Int. J. Intell. Netw. 3, 58–73.

Kambar, M.E.Z.N., Nahed, P., Cacho, J.R.F., Lee, G., Cummings, J., Taghva, K., 2022. Clinical text classification of alzheimer's drugs' mechanism of action. In: Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 1. Springer, pp. 513–521.

Karthikeyan, B., Nithya, K., Alkhayyat, A., Yousif, Y.K., 2023. Artificial intelligence enabled decision support system on E-healthcare environment.. Intell. Autom. Soft Comput. 36 (2).

Kirchmair, J., Göller, A.H., Lang, D., Kunze, J., Testa, B., Wilson, I.D., Glen, R.C., Schneider, G., 2015. Predicting drug metabolism: experiment and/or computation? Nat. Rev. Drug Discov. 14 (6), 387–404.

Lee, N.E., Parker, M.M., Concepcion, J.Q., 2023. An electronic medical record (EMR) prompt improves screening rates for metabolic conditions among children with obesity. Obesity.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 (4), 1234–1240.

Li, J., Huang, Q., Ren, S., Jiang, L., Deng, B., Qin, Y., 2023. A novel medical text classification model with Kalman filter for clinical decision making. Biomed. Signal Process. Control 82, 104503.

Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G., 2020. BEHRT: transformer for electronic health records. Sci. Rep. 10 (1), 1–12.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G., 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv. 55 (9), 1–35.

Lu, Y., Liu, X., Du, Z., Gao, Y., Wang, G., 2023. Medkpl: a heterogeneous knowledge enhanced prompt learning framework for transferable diagnosis. J. Biomed. Inform. 104417.

Luo, J., Ye, M., Xiao, C., Ma, F., 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 647–656.

Müller, M., Salathé, M., Kummervold, P.E., 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. Front. Artif. Intell. 6, 1023281.

Pandey, B., Pandey, D.K., Mishra, B.P., Rhmann, W., 2022. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. J. King Saud Univ.-Comput. Inf. Sci. 34 (8), 5083–5099.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training. Openai Blog.

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., Qadir, J., 2022. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. Comput. Biol. Med. 106043.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D., 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit. Med. 4 (1), 86.

Sajde, M., Malek, H., Mohsenzadeh, M., 2022. RecoMed: A knowledge-aware recommender system for hypertension medications. Inf. Med. Unlocked 30, 100950.

Shang, J., Ma, T., Xiao, C., Sun, J., 2019. Pre-training of graph augmented transformers for medication recommendation. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. IJCAI, 2019, pp. 5953–5959.

Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M.A., Shambour, M.K.Y., Alsalibi, A.I., Gandomi, A.H., 2022. Machine learning in medical applications: A review of state-of-the-art methods. Comput. Biol. Med. 145, 105458.

Sivarajkumar, S., Wang, Y., 2022. Healthprompt: A zero-shot learning paradigm for clinical natural language processing. In: AMIA Annual Symposium Proceedings. Vol. 2022, American Medical Informatics Association, p. 972.

Sylolypavan, A., Sleeman, D., Wu, H., Sim, M., 2023. The impact of inconsistent human annotations on AI driven clinical decision making. NPJ Digit. Med. 6 (1), 26.

Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W., 2023. Large language models in medicine. Nat. Med. 1–11.

Truven, 2023. Truven health MarketScan. Accessed June 23, 2023. https://marketscan.truvenhealth.com/marketscanportal/.

Wang, D.-Q., Feng, L.-Y., Ye, J.-G., Zou, J.-G., Zheng, Y.-F., 2023. Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare. MedComm–Future Med. 2 (2), e43.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. 32.

Yao, L., Jin, Z., Mao, C., Zhang, Y., Luo, Y., 2019a. Traditional Chinese medicine clinical records classification with bert and domain specific corpora. J. Am. Med. Inform. Assoc. 26 (12), 1632–1636.

Yao, L., Mao, C., Luo, Y., 2019b. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med. Inform. Decis. Mak. 19 (3), 31–39.

Zhang, J., Yu, X., Wang, Z., Zheng, X., 2023. GWBNER: A named entity recognition method based on character glyph and word boundary features for Chinese EHRs. J. King Saud Univ.-Comput. Inf. Sci. 35 (8), 101654.