IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 May 2024, accepted 8 July 2024, date of publication 16 July 2024, date of current version 6 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3428918

RESEARCH ARTICLE

MediGPT: Exploring Potentials of Conventional and Large Language Models on Medical Data

MOHAMMAD ABU TAREQ RONY[1], MOHAMMAD SHARIFUL ISLAM[2], TIPU SULTAN[3], SAMAH ALSHATHRI[4], AND WALID EL-SHAFAI[5,6], (Senior Member, IEEE)

[1]Department of Statistics, Noakhali Science & Technology University, Noakhali 3814, Bangladesh
[2]Department of Computer Science and Telecommunication Engineering, Noakhali Science & Technology University, Noakhali 3814, Bangladesh
[3]Department of Computer Science, College of Computer and Information Sciences, Fordham University, Bronx, NY 10458, USA
[4]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P. O. Box 84428, Riyadh 11671, Saudi Arabia
[5]Security Engineering Laboratory, Computer Science Department, Prince Sultan University, Riyadh 11586, Saudi Arabia
[6]Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt

Corresponding authors: Mohammad Shariful Islam (shariful.43cse@gmail.com), Mohammad Abu Tareq Rony (abutareqrony@gmail.com), Samah Alshathri (sealshathry@pnu.edu.sa), and Walid El-Shafai (eng.waled.elshafai@gmail.com)

This work was supported by Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, through the Researchers Supporting Project PNURSP2024R197.

ABSTRACT Medical text classification organizes medical documents into categories to streamline information retrieval and support clinical decision-making. Traditional machine learning techniques, including pre-trained language models, are effective but require extensive domain-specific training data, often underperform across languages, and are costly and complex to deploy on a large scale. In this study, we employed four datasets: Clinical trials on cancer, encompassing 6 million statements from interventional cancer clinical trial protocols; the Illness-dataset, consisting of 22,660 categorized tweets from 2018 and 2019; the Multi-View active learning for short medical text classification in user-generated data, an extended version of the Illness-dataset including 22,660 documents from the same period; and the Symptom2Disease dataset, containing 1,200 data points used to predict diseases based on symptom descriptions. This study uses ChatGPT, particularly its ChatGPT-3.5 and ChatGPT-4 versions, as a viable alternative for classifying medical texts. We investigate essential aspects, including the construction of prompts, the parsing of responses, and the various strategic use of GPT models to optimize outcomes. Through comparative analysis with established methods like pre-trained language model fine-tuning and prompt-tuning, our findings indicate that ChatGPT addresses these challenges efficiently and matches the performance of traditional methods. Furthermore, the enhanced capabilities of the proposed MediGPT (Medical Generative Pre-Trained Transformers) have led to performance improvements of 14.3%, 22.3%, 13.6%, and 13.7% across the datasets, highlighting its adaptability and robustness in diverse medical text scenarios without the need for specialized domain adjustments. This research underscores the capability of ChatGPT to facilitate a versatile AI framework in medical text processing, which could revolutionize medical informatics practices.

INDEX TERMS Medical text, natural language processing, ChatGPT, classification, large language model.

I. INTRODUCTION

The exponential growth of digital medical data has created an urgent need for efficient information management and retrieval systems. Large Language Models (LLMs) and Natural Language Processing (NLP) play a crucial role in medical text classification by enabling the automated categorization of medical documents. LLMs, such as ChatGPT-4, leverage vast amounts of data to understand and generate human-like text, while NLP techniques process and analyze this text to identify relevant medical information.

Given the vast amount of unstructured medical content available online, including electronic health records and

The associate editor coordinating the review of this manuscript and approving it for publication was Antonio J. R. Neves.

research publications, accurate classification and organization of this data are crucial for healthcare professionals, researchers, and patients. Artificial intelligence, specifically medical text classification, has emerged as a powerful tool to automate categorizing and indexing this wealth of information.

Text classification has long been a pivotal technique in information retrieval and data mining, demonstrating its value across diverse sectors such as healthcare diagnostics, targeted marketing, entertainment, and data filtering. Recent advancements in data mining and NLP have sparked global research interest, leading to sophisticated automated text classification systems. These systems have revolutionized document categorization, enabling efficient organization and analysis of vast textual content from various sources, including substantial volumes of user-generated data from social media platforms [1].

The processing of medical documents has utilized Deep Learning(DL) methods, particularly pre-trained language models(PLMs) such as BERT (Bidirectional Encoder Representations from Transformers), BART (Bidirectional and Auto-Regressive Transformer), and T5 (Text-to-Text Transfer Transformer) [2], [3], [4]. These models have effectively predicted disease trends, evaluated patient sentiments, and extracted information from medical documents. However, deploying these PLMs poses significant challenges, including the scarcity of high-quality training data and the need for substantial computational resources, like GPUs and TPUs, due to the large size of the models [5].

Despite their potential, PLMs often need help with fine-tuning, development, and deployment due to the scarcity and low quality of training data, which can degrade model performance [6], [7]. Obtaining high-quality annotated datasets requires significant time and labor investment. Even with sufficient data, supervised learning models often struggle with generalization and maintaining robustness across diverse scenarios, including cross-linguistic applications. Additionally, the extensive parameter count in PLMs complicates deployment and necessitates high-performance computing resources.

These challenges highlight the limitations of conventional PLM-based methods in medical text classification, underscoring their shortcomings in achieving broader goals of General-Purpose Artificial Intelligence. Recently, innovations like OpenAI's ChatGPT have led to breakthroughs in NLP, recognized for their ability to deliver detailed responses to complex inquiries and perform tasks like multilingual translation, poetry creation, and code generation [5], [8]. The comprehensive language understanding and generation capabilities of ChatGPT have also been leveraged in interdisciplinary research, including radiology interpretation and sentiment analysis in healthcare settings [9], [10].

Given the capabilities of ChatGPT, it is compelling to explore its potential to enhance applications in digital healthcare. This study examines how ChatGPT can be strategically employed in medical text classification, assessing the capability of ChatGPT-3.5 and its extension ChatGPT-4 in classifying medical-related documents. According to ChatGPT-3.5, it can contribute to medical classification tasks, including disease identification, treatment recommendation, and medication classification, among other applications in Figure 1. Accompanying the proposed MediGPT framework, this paper introduces a distinctive paradigm that sets it apart from existing methodologies. By conducting a series of comparative experiments involving several mainstream text classification models, we systematically assess and showcase the superior performance of ChatGPT in these tasks. This marks a notable departure from conventional approaches.

In our experiments, we primarily evaluate the efficacy of ChatGPT-3.5 [11] and its extension, ChatGPT-4 [9], in classifying medical-related documents. Alongside the proposed MediGPT framework, this paper introduces a unique paradigm. We systematically assess and demonstrate the superior performance of ChatGPT in medical text classification tasks through a series of comparative experiments involving various mainstream text classification models, including traditional fine-tuned PLMs [1] and prompt-learning based on auto-regressive generative PLMs [12], [13]. This represents a significant departure from existing methodologies.

Our comprehensive literature review on ChatGPT-based question answering (QA) [14] and the prompt learning scheme [13] indicates that most language understanding tasks using ChatGPT can be interpreted as a novel form of prompt learning based on PLMs. This paper highlights the significant similarities and distinctions between the ChatGPT-based NLP paradigm and traditional methods through detailed examples and illustrations, as depicted in Figure 2.

Figure 2 offers a paradigmatic contrast between NLP solutions powered by ChatGPT and current methods of prompt learning, with medical data classification as an illustrative case. Part (a) of the diagram delineates the prototype for medical data classification, visually capturing the initial setup for applying these models to real-world medical data. Part (b) outlines the typical workflow for approaches based on ChatGPT and explains how ChatGPT processes and classifies medical texts. On the other hand, part (c) compares this workflow with masked language model (LM) prompt-tuning methods, showing the traditional approach and its differences in handling the same classification tasks.

Through this comprehensive exploration of ChatGPT's involvement in medical text classification, we aim to deepen our grasp of its capabilities and methodically evaluate its effectiveness using empirical evidence. By showcasing how ChatGPT can contribute to various facets of medical management, this research underscores the potential of ChatGPT to revolutionize healthcare practices, thereby fostering a more efficient and sustainable digital healthcare infrastructure.

The key contributions of this research are outlined as follows:

ChatGPT Prompt:

Tell us what you can do for medical application?

ChatGPT

I can assist in various ways in medical applications:

1. Information Retrieval: I can provide information on medical conditions, symptoms, treatments, medications, and procedures from reputable sources.
2. Patient Education: I can explain complex medical concepts in understandable language, helping patients better understand their conditions and treatment options.
3. Symptom Checker: I can help users assess their symptoms and provide guidance on whether they should seek medical attention.
4. Medication Information: I can offer details about medications, including dosages, side effects, interactions, and contraindications.
5. Health Monitoring: I can assist in tracking health metrics, such as blood pressure, blood glucose levels, and weight, and offer insights based on the data provided.
6. Appointment Scheduling: I can help users schedule appointments with healthcare providers, remind them of upcoming appointments, and provide directions to medical facilities.
7. Language Translation: I can translate medical information into different languages, facilitating communication between healthcare providers and patients with limited English proficiency.
8. Medical Research: I can aid researchers in literature review, hypothesis generation, and data analysis for medical studies.
9. Mental Health Support: I can provide resources and coping strategies for managing stress, anxiety, depression, and other mental health concerns.
10. Medical Training: I can serve as a tool for medical students and professionals to enhance their knowledge, practice clinical scenarios, and prepare for exams.

FIGURE 1. Recommendations from ChatGPT to improve medical governance (Query Date: 2024.4.14).

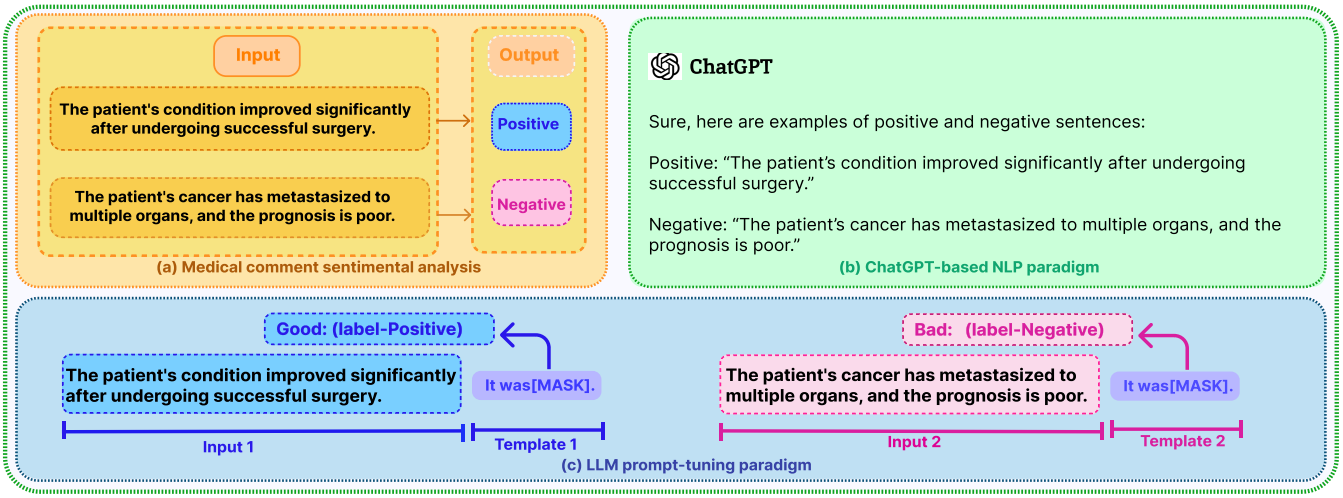

FIGURE 2. Comparing ChatGPT-based NLP solutions with established prompt learning approaches: A case study in medical text classification Part (a): Task prototype Part (b): Standard workflow of ChatGPT-based approaches Part (c): Standard workflow of masked language model prompt-tuning methods.

- We initiated a study on medical text classification motivated by the capabilities of advanced PLMs like ChatGPT. This led to MediGPT, a novel model powered by ChatGPT's architecture tailored for the medical domain.
- In comparative evaluations, MediGPT outperformed conventional approaches, showcasing enhanced semantic comprehension and sophisticated reasoning skills in particular medical contexts.
- Our few-shot, one-shot, and zero-shot learning trials highlighted MediGPT's ability to operate without needing supervised training data, manual labeling, or extensive medical expertise, showcasing its potential as a scalable solution in healthcare AI.
- As a cost-effective alternative to complex PLM frameworks, MediGPT offers a viable approach for integrating Artificial Intelligence technologies in healthcare settings.

The organization of this paper is as follows: Section II dives into a review of recent studies relevant to the classification of medical data. In Section III, provide detailed description of the MediGPT framework, explaining its algorithmic foundations. Sections IV offer an extensive analysis of experimental comparisons between MediGPT and other prevalent methods based on PLMs, including a series of ablation studies. Finally, Section V summarizes the findings of this research study.

II. RELATED WORK

This section explores pertinent literature concerning cross-linguistic medical text classification, recent breakthroughs with ChatGPT and its variants, and practical strategies employing fine-tuning based on PLMs and prompt-tuning to address classification challenges effectively.

A. CLASSIFICATION OF MEDICAL TEXT

Over the past decade, conventional machine learning models, including decision trees Convolutional Neural Networks (CNNs) [15], Long Short-Term Memory Networks (LSTMs), and Gated Recurrent Units (GRUs), have played pivotal roles in the classification of medical documents [16]. For instance, prior research has successfully employed Support Vector Machines (SVMs) and decision tree classifiers for regional medical data classification [17]. Additionally, Bi-LSTM models with attention mechanisms have been utilized to enhance semantic feature extraction in medical text classification [18]. The emergence of large-scale models like BERT [2] and ChatGPT [16] has revolutionized this field, often surpassing traditional methods by effectively handling complex scenarios [19]. For example, recent work has applied BERT to extract relevant information from unlabeled medical news, thereby improving corpus construction and classification accuracy [1].

B. CONVENTIONAL MACHINE LEARNING TECHNIQUES, PLM-BASED FINE-TUNING

Traditional machine-learning methods were once the standard for medical text processing. These methods often struggled to understand the complex nature of natural language, especially in detailed contexts [20]. Pre-trained language models (PLMs) have changed the field of NLP by consistently outperforming older methods through fine-tuning, which involves adjusting additional network parameters and focusing on specific tasks.

C. CHATGPT

Deep learning has significantly impacted various sectors, enhancing productivity and improving medical diagnostics. ChatGPT, particularly in its latest iterations like ChatGPT-3.5 [8] and ChatGPT-4, has transformed conversational AI, demonstrating exceptional medical data processing capabilities. These advancements highlight the ongoing evolution of AI technologies and underscore the critical need for responsible AI practices to maximize benefits and mitigate risks [8]. Millions of users utilizing language models have led to a diverse array of applications, revealing the capabilities of ChatGPT. Research findings indicate that ChatGPT performs exceptionally well in translating multiple languages, especially those with abundant resources [21].

III. MEDIGPT: CHATGPT-BASED MEDICAL TEXT CLASSIFICATION

A. METHODOLOGY OVERVIEW

This paper presents MediGPT, a pioneering study investigating the application of ChatGPT for medical text classification. This study is among the first to systematically adapt ChatGPT technology for healthcare applications, employing a systematic experimental analysis to explore its viability. There was a noticeable absence of systematic research utilizing ChatGPT for medical text classification before MediGP. To bridge this gap, this study outlines a structured workflow for implementing ChatGPT in this context, drawing heavily from the latest scholarly discussions.

The workflow of MediGPT illustrated in Figure 3 employs various prompting strategies to formulate prompts. These prompts are combined with the original sentence to create the ChatGPT question [32] ChatGPT generates a response based on these inputs. Answer alignment strategies are then implemented to categorize the response into predefined categories, ensuring the reactions are relevant and appropriately classified. The deployment of ChatGPT in medical text classification involves three main phases:

- **Prompt construction**: Developing specialized prompts for medical data to guide ChatGPT in generating responses.
- **Q&A Inference**: Processing input prompts and generating responses using ChatGPT as a closed system.
- **Answer normalization**: Translating ChatGPT's responses into categorized data based on a predefined medical taxonomy.

While the Q&A process with ChatGPT follows a fixed procedure, optimizing prompt construction and response alignment offers significant improvement opportunities. MediGPT operates as a pipeline where prompt quality, ChatGPT version, and response mapping strategies collectively enhance classification effectiveness.

This structured approach optimizes ChatGPT's performance in medical text classification by meticulously analyzing and refining each phase. MediGPT aims to establish a robust framework tailored to the unique challenges of this domain.

B. PROMPT QUESTION CONSTRUCTION

Prompt engineering is acknowledged as a complex skill requiring expertise and iterative refinement [13]. To harness ChatGPT effectively for sentence classification, we conducted thorough research to optimize prompt construction [7]. Figure 3 illustrates the strategies employed in this study including:

1) Manual prompt definition,
2) Prompt generation triggered by ChatGPT responses,
3) Prompts derived from zero-shot similarity comparisons,
4) Chain-of-Thought (CoT) prompting.

These strategies, detailed in subsequent sections, aim to enhance ChatGPT's input to produce accurate and relevant responses for medical classification tasks.

In our study, we have developed a set of manually designed prompts, detailed in Table 1, to evaluate MediGPT's effectiveness in classifying medical texts across different languages. These prompts are integral to our methodology for assessing MediGPT's ability to accurately interpret and categorize medical statements into specific medical specialties or categories. Each prompt is structured to provide MediGPT with a clear objective in determining the most suitable medical category for a given sentence. These prompts are designed to simulate real-world clinical queries

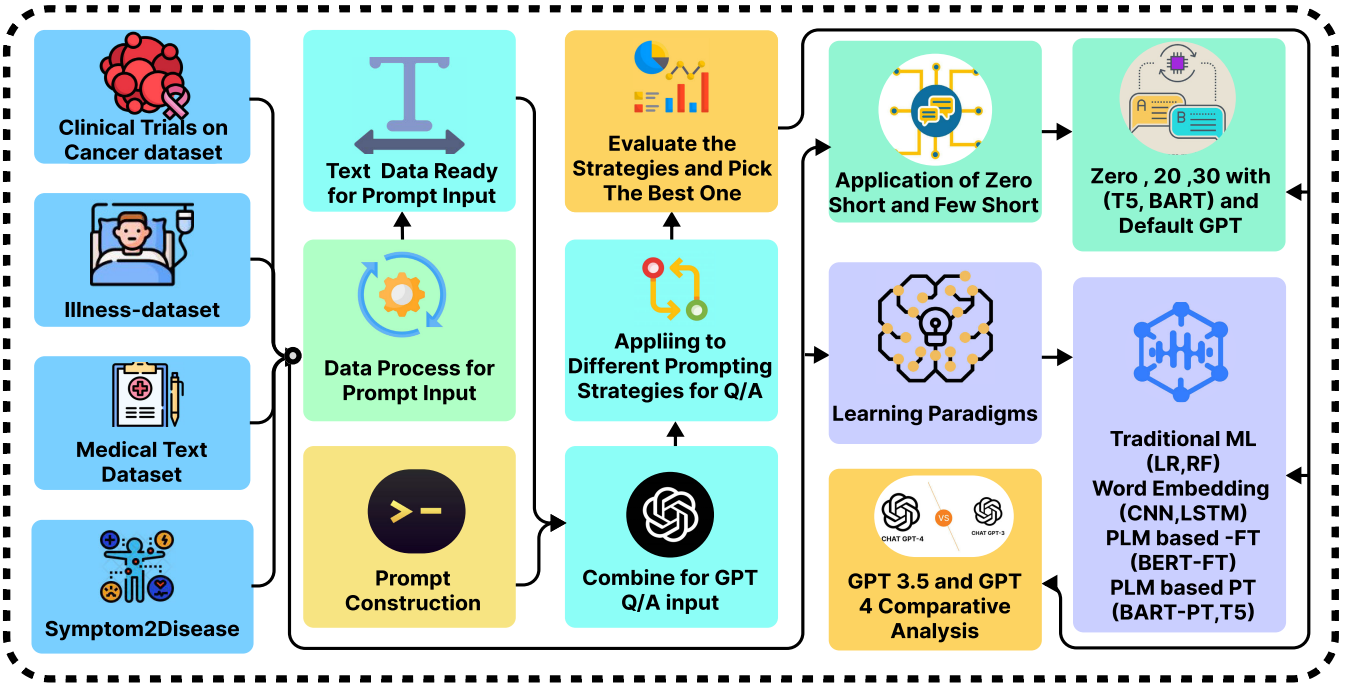

FIGURE 3. The workflow of proposed MediGPT method.

TABLE 1. Manually devised prompts for MediGPT.

No.	Prompting Template
1	Determine the medical category for the following sentence: [SENTENCE] Options: [CATEGORIES] Response: [Res]
2	Analyze the sentence below and assign the most relevant medical specialty: Sentence: [SENTENCE] Choose from: [CATEGORIES] Result: [Res]
3	Assess the following clinical statement and categorize it into the appropriate field: Text: [SENTENCE] Possible fields: [CATEGORIES] Output: [Res]

and emphasize precision and contextual understanding. The structure of each prompt includes the following:

1) Direct instruction to classify or analyze the sentence, ensuring clarity of the task.
2) Placeholder for the sentence ([SENTENCE]), accommodating the insertion of the medical text to be classified.
3) Options for potential categories ([CATEGORIES]), guiding MediGPT towards the expected type of response.
4) Response section ([Res]), prompting MediGPT to provide its classification results.

This setup adheres to established communication practices in medical consultations while incorporating specific adjustments to optimize MediGPT's performance:

1) Clarity and directness: Each prompt explicitly defines the task, reducing ambiguity and focusing MediGPT's response mechanism on classification rather than open-ended discussion.
2) Contextual relevance: By embedding medical context directly into the prompts, we ensure that MediGPT's responses are evaluated within the appropriate clinical context, enhancing the relevance and applicability of its classifications.
3) Simplicity in design: The prompts are designed to straightforwardly evaluate MediGPT's ability to classify texts under standardized conditions. This simplicity also aids in reducing the cognitive load on the model, focusing its capabilities on the classification task.

Our evaluation approach employs a sampling method to assess the effectiveness of these prompts. We select a fixed number of samples from a cross-linguistic medical text dataset and test the accuracy of each prompt. This systematic evaluation helps identify the most effective prompt configurations, ensuring that MediGPT is tuned to provide the most accurate and clinically relevant responses. A critical refinement in our approach is guiding MediGPT to focus exclusively on category classification, minimizing extraneous information. This guidance is essential to maintain clarity and precision in classification outcomes. Through rigorous testing of these prompts, we aim to enhance MediGPT's reliability as a tool for cross-linguistic medical text classification. This advancement aims to assist medical professionals in swiftly and accurately categorizing clinical statements.

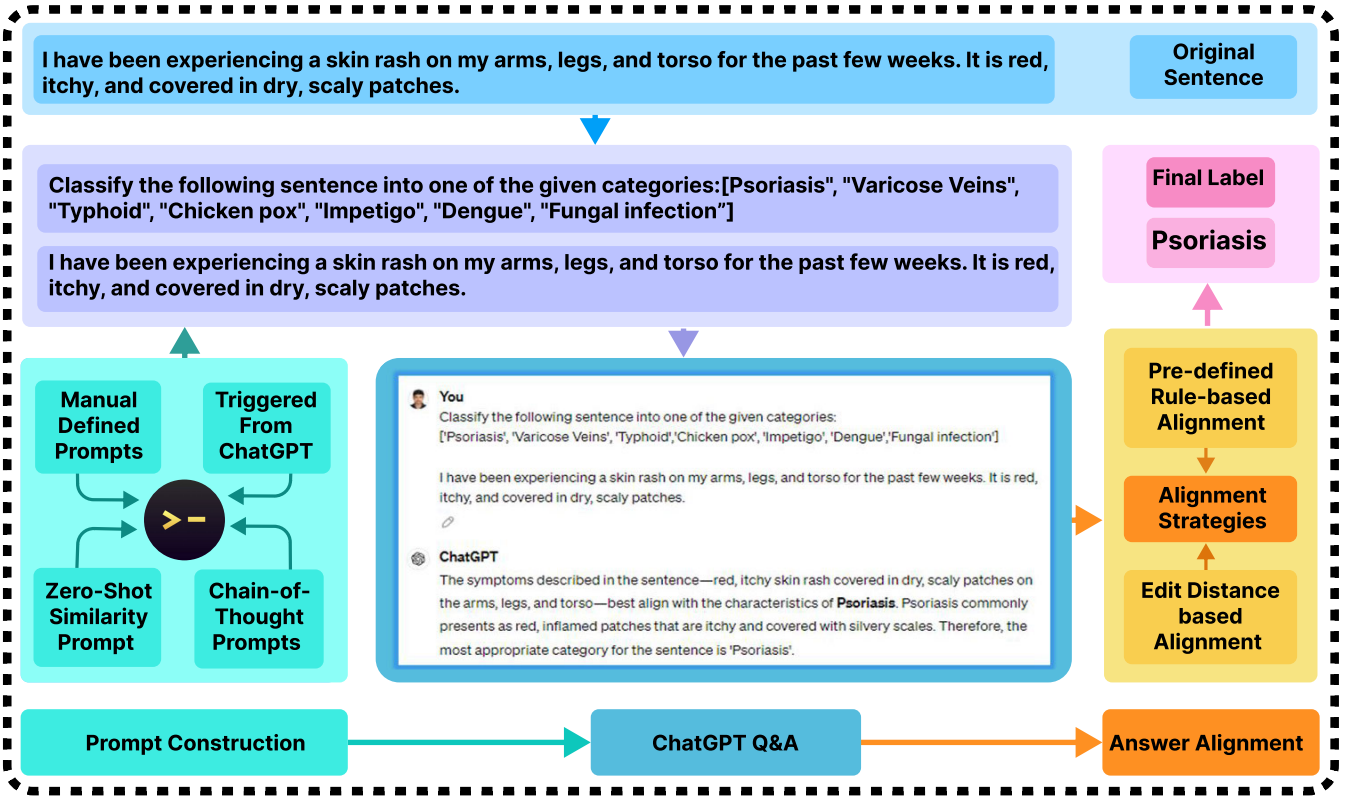

FIGURE 4. Illustration of the MediGPT framework using a typical example from the Medical Symptom dataset. (a) Various strategies are employed to construct prompts, integrating them with original sentences to form ChatGPT questions. (b) ChatGPT generates responses based on these inputs. (c) Strategies for answer alignment categorize responses into predefined categories.

C. MEDIGPT TRIGGERED PROMPTS

Based on prior studies [22], it was hypothesized that MediGPT could effectively generate high-quality prompt templates. Consequently, we utilized MediGPT to propose recommendations for template creation. Previous research demonstrated that task-specific prompts could be efficiently generated by directly engaging with human input. For instance, they recommended utilizing the query:

- Provide five succinct prompts or templates to address the [x] task.

Here, [x] represents different types of tasks. Our experiments confirmed this method's effectiveness across various settings.

MediGPT provided several potential prompts that were semantically consistent and exhibited unique stylistic variations. As shown in Figure 5; we asked MediGPT a similar question:

- Provide five succinct prompts or templates that can address the task of medical text classification.

We used a sampling-based evaluation method to select the most effective prompt from the generated set for use in subsequent comparative experiments. The selected prompt for medical text classification is:

- Classify the medical text: [SENT] based on its primary condition [CATE].

This method of using MediGPT-triggered prompts helps us understand how automated systems can improve task-specific templates in medical text classification. The insights gained are crucial for refining the interaction between human operators and AI systems in clinical informatics. Building on previous studies in few-shot and zero-shot learning using a meta-learning framework [23], we developed new prompting strategies called similarity-driven prompting. Traditional few-shot object classification often uses examples and classifiers from similar categories using distance measures like cosine similarity and squared Euclidean distance. In a few-shot learning scenario for image classification, an image to be categorized is presented alongside a representative image from each category. These images are embedded into a shared feature space using Siamese, prototypical, and matching networks. A similarity threshold is then set to help classify the image by comparing it to the representative images from different categories. Adapting this approach to medical text classification, our Zero Shot similarity-based prompting strategy involves:

1) **Embedding generation**: Medical texts are embedded into a shared low-dimensional space using language models fine-tuned for medical contexts.
2) **Template creation**: We generate initial prompt templates that capture the essence of potential medical text queries.
3) **Similarity assessment**: Using the embedded representations, we evaluate the semantic similarity between the prompt templates and an extensive repository

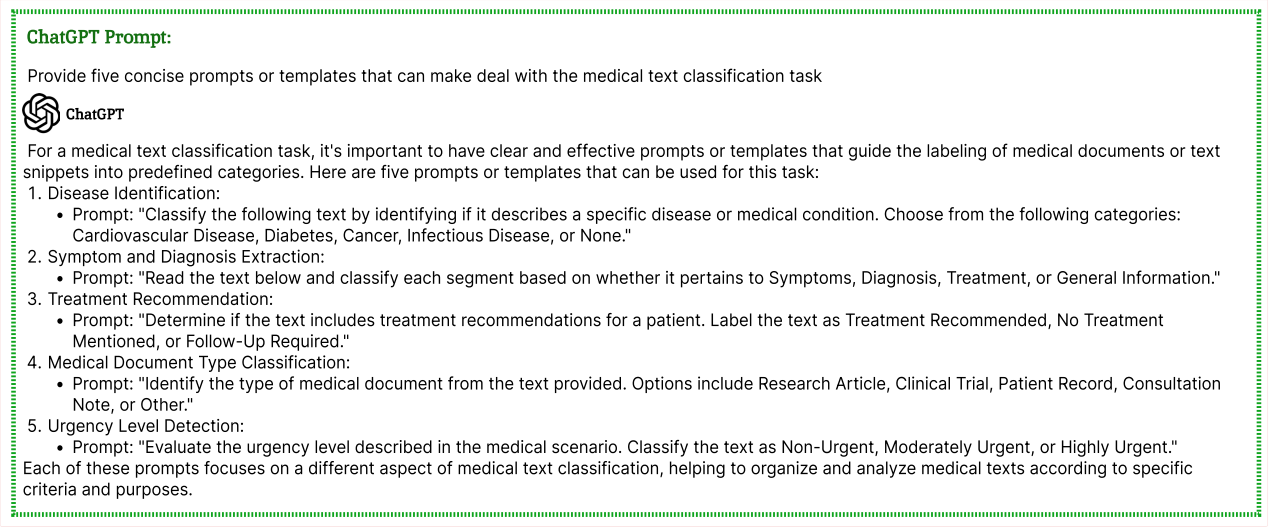

FIGURE 5. Templates generated upon requests to ChatGPT (Version: ChatGPT-3.5, Date of Query: April 26, 2024).

of unlabeled medical texts using cosine similarity measures.

4) **Prompt refinement**: Based on the similarity scores, prompts are refined to more accurately match the nuances of the medical texts they are designed to classify.

This methodology allows us to use Zero-Shot learning to generate effective and contextually appropriate prompts without needing labeled examples for every possible medical scenario. This approach aims to enhance MediGPT's ability to process and classify diverse medical texts with high accuracy and minimal supervision.

D. QA-BASED SIMILARITY EVALUATION IN MEDICAL TEXT CLASSIFICATIONS

We present a novel method using the ChatGPT interface to measure text similarity, which is essential for medical text classification. This approach implements two distinct QA modes for evaluating sentence similarity:

- **Direct End-to-End QA-Based Classification**: This method employs direct questions to classify texts. For example:
  - Given sentence S: [SENT1], which option, A: [SENT2], B: [SENT3], etc., do you think is most similar to sentence S? Please choose A, B, etc., or C.
- This approach allows for quick classification based on a single QA interaction, aligning the target sentence with the most similar category, as shown in (Figure 6).
- **Progressive Comparison QA-Based Classification**: This approach systematically compares pairs of sentences, improving accuracy through incremental assessments. Inspired by the bubble sort method. The prompt used is:
  - Given sentence S: [SENT0], which of the following sentences, A: [SENT1] or B: [SENT2], do you think is more similar to sentence S? Please respond with A or B.
- This method is particularly effective for complex datasets like medical texts, where subtle differences between categories are crucial.

These methods leverage QA interactions to enhance the classification of medical texts by focusing on sentence similarity.

E. CHAIN-OF-THOUGHT TRIGGERED PROMPTS

We adopt a Chain-of-Thought (CoT) prompting strategy where ChatGPT provides final classification outcomes accompanied by detailed reasoning. The CoT approach initiates with a directive that extends the original QA prompt:

- Explain the semantics and keywords, elucidating the corresponding classification rationale.

Figures 6, 7, and 8 demonstrate these methods, highlighting progressive similarity measurement and the efficacy of CoT strategies in text classification. In practice, both manually crafted and ChatGPT-triggered prompts serve as foundational references. Our research indicates that while CoT-triggered prompts excel in datasets with numerous classification categories, their effectiveness may diminish in more straightforward datasets with fewer categories. This investigation into QA-based similarity assessment and CoT-triggered prompts underscores the intricacies of prompt design in medical text classification, emphasizing how dataset characteristics significantly influence the efficacy of prompting strategies.

F. CHATGPT Q&A INFERENCE

ChatGPT is built on the Generative Pre-trained Transformer architecture, specifically using ChatGPT-3.5 for this study. This architecture uses a transformer framework with parallel data processing and multi-headed attention mechanisms. These features help the model handle and generate language sequences based on the probabilities of different continuations. Using an autoregressive inference process, ChatGPT creates text by progressively building responses from a given

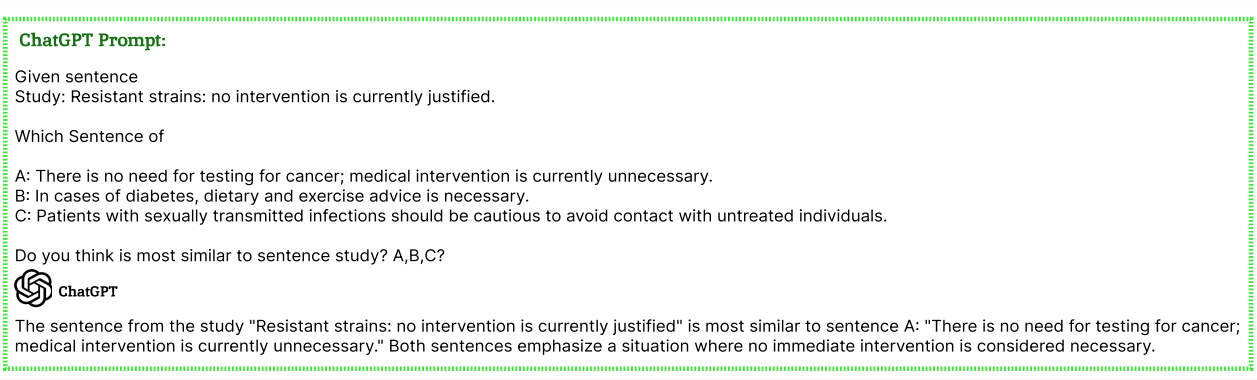

FIGURE 6. Illustrating QA-Based prompting for text classification through end-to-end direct similarity measurement.

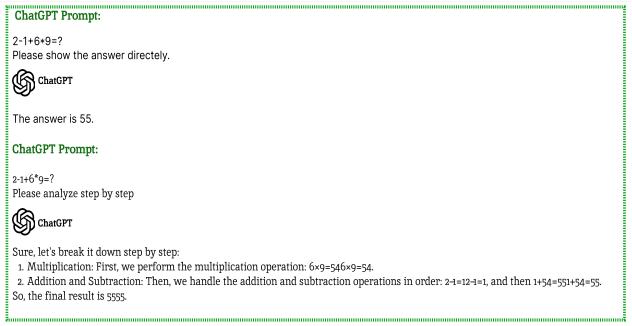

FIGURE 7. Advancing text classification with progressive similarity measurement: A QA-Based prompting approach.

prompt and selecting subsequent words using a probabilistic model. Trained on extensive text corpora, the model learns complex patterns and dependencies in language, allowing strong performance over long sequences.

ChatGPT excels in nuanced language comprehension and generation capabilities, particularly its 175 billion parameter configuration (ChatGPT-3.5). Further enhancements are achieved through supervised fine-tuning and reinforcement learning mechanisms that refine responses based on user interactions. The underlying mathematical foundation for the question-answering process can be encapsulated in the following formula:

$$p(y|x) = \prod_{t=1}^{T} p(y_t|y_1, \ldots, y_{t-1}, x) \tag{1}$$

where $\prod$ denotes the product of probabilities, y_t is the token at time t, and T is the sequence length. This formulation ensures that each generated token considers the contextual information provided by preceding tokens and the input prompt. Our investigation posits ChatGPT as a potent Zero-Shot (ZS) text classification tool, especially in interactive contexts. Each user query initiates a new conversational context, ensuring responses remain consistent and contextually independent from prior interactions. We have also integrated ChatGPT-4 into our evaluations [8], using it on three proposed datasets as shown in Figure 9. The figure provides an example of prompt input and its response. Which, as the latest iteration by OpenAI, demonstrated enhanced capabilities in handling complex semantic classifications. These findings will be further elaborated in the evaluation section of this paper.

TABLE 2. Medical text classification datasets categories for MediGPT.

Dataset	Categories
Clinical trials on cancer	Eligible, Not Eligible
Illness-dataset	Alzheimer's, Parkinson's, Cancer, Diabetes
Medical text	Thyroid Cancer, Colon Cancer, Lung Cancer
Symptom2Disease	Psoriasis, Varicose Veins and more

G. DATASET OVERVIEW

In exploring ChatGPT's capabilities in medical text classification, we employed four extensive datasets encompassing diverse medical domains and languages. These datasets form the empirical foundation for rigorously evaluating the robustness and adaptability of MediGPT. Table 2 summarizes the categories of our selected datasets, outlined as follows:

1) Clinical trials on cancer: This dataset is drawn from 18 years of interventional cancer clinical trial protocols to aid in understanding eligibility criteria. It comprises 6 million free-text statements annotated for eligibility in clinical trials, enabling nuanced semantic analysis and drug-treatment associations. [1]
2) Illness dataset: This dataset is extended from the one introduced in the EMNLP Findings 2022 paper by Karisani. It encompasses 22,660 tweets from 2018 and 2019, categorized into Alzheimer's, Parkinson's, Cancer, and Diabetes domains. A positive label denotes a tweet mentioning a diagnosed individual. [2]
3) Extended illness dataset: This dataset includes 22,660 biomedical documents collected between 2018 and 2019, covering Alzheimer's, Parkinson's, Cancer, and Diabetes. It provides valuable resources for research in medical text classification with user-generated data, distinguishing

[1] **Dataset available at** https://www.kaggle.com/datasets/auriml/eligibilityforcancerclinicaltrials.

[2] https://github.com/p-karisani/illness-dataset/blob/main/data.txt

ChatGPT Prompt:

Given sentence
Study: Resistant strains: no intervention is currently justified.

Which Sentence of

A: There is no need for testing for cancer; medical intervention is currently unnecessary.
B: In cases of diabetes, dietary and exercise advice is necessary.
C: Patients with sexually transmitted infections should be cautious to avoid contact with untreated individuals.

Do you think is most similar to sentence sstudy? A,B,C?
Please provide a step-by-step analysis towards the semantic and keywords, and present the corresponding classification reasoning process.

ChatGPT

Let's analyze each option in comparison to the study sentence:

1. Study Sentence: "Resistant strains: no intervention is currently justified."
   - Keywords: "resistant strains," "intervention," "justified"
   - Semantic Meaning: No action is deemed appropriate for dealing with resistant strains at the moment.
2. Option A: "There is no need for testing for cancer; medical intervention is currently unnecessary."
   - Keywords: "testing for cancer," "medical intervention," "unnecessary"
   - Semantic Meaning: No necessity for medical intervention in cancer testing currently.
3. Option B: "In cases of diabetes, dietary and exercise advice is necessary."
   - Keywords: "diabetes," "dietary and exercise advice," "necessary"
   - Semantic Meaning: Dietary and exercise advice is essential for managing diabetes.
4. Option C: "Patients with sexually transmitted infections should be cautious to avoid contact with untreated individuals."
   - Keywords: "sexually transmitted infections," "cautious," "avoid contact," "untreated individuals"
   - Semantic Meaning: Caution is advised for individuals with STIs to avoid contact with untreated persons.

Analysis:

- The study sentence emphasizes the lack of justification for intervention regarding resistant strains.
- Option A echoes a similar sentiment, discussing the unnecessary nature of medical intervention but in the context of cancer testing.
- Option B addresses the necessity of advice for managing diabetes, which is unrelated to the study sentence.
- Option C discusses caution for individuals with STIs, not directly related to the study sentence's theme.

Conclusion:
Option A is the most similar to the study sentence because it shares the concept of no current intervention being justified, albeit in a different context.
Options B and C diverge from the study sentence's focus.

FIGURE 8. The CoT prompting strategy: Leveraging simple and direct QA approach. (Model: ChatGPT-3.5, Query Date: 2024.5.1).

between negative and positive instances (mentions of diagnosed individuals). [3]

4) Symptom2Disease dataset: Comprising 1200 data points, this dataset is essential for developing models that predict diseases from symptom descriptions in natural language. It encompasses 24 diseases, each described by 50 symptom profiles, facilitating early diagnosis and remote consultation application [4]

Incorporating these diverse datasets, our evaluations of MediGPT are comprehensive, covering various textual formats and medical specialties. The model demonstrates its adaptability to different document lengths, complexities, and medical terminologies across these datasets, showcasing its advanced linguistic and domain-specific understanding.

H. EXPERIMENTAL SETUP

Conventional ML methods and LLMs were constructed using an experimental framework with libraries such as Sci-kit-learn, Pytorch, TensorFlow, and additional language modeling tools. The programming language selected for these implementations was Python 3.10. The testing environment for these experiments was a computer fitted with an Intel(R) Core(TM) i5-10300H CPU, featuring 32 GB of RAM and a processor speed of 2.50 GHz.

[3]https://www.kaggle.com/datasets/falgunipatel19/biomedical-text-publication-classification

[4]https://www.kaggle.com/datasets/niyarrbarman/symptom2disease

I. BASELINES

In contemporary text classification research, the field is characterized by five primary approaches to training models. Each approach offers distinct methods to tackle the challenges of understanding and processing natural language.

- Traditional ML Models: This approach includes Logistics Regression and Random Forest methods. These techniques rely on manually crafted features to distinguish between different classes of text data, providing a baseline for understanding ML applications in text classification [24].
- DL Models: Dominant architectures like CNN and LSTM represent this paradigm. They utilize word embeddings to represent textual information in a continuous vector space, capturing complex semantic relationships within the data.
- Fine-tuning of PLMs: Models such as BERT, BART, and T5 fall under this category. These PLMs adapt to text classification tasks by leveraging large-scale pre-trained language representations, requiring minimal additional training to achieve high performance [2].
- Prompt Learning with PLMs: This innovative approach involves generating specific prompts that guide the model's responses, harnessing the rich internal knowledge of PLMs for more precise and contextually relevant text classification.
- Zero-shot(ZS) Learning: This cutting-edge paradigm, exemplified by applications like ChatIE, ChatEventExtract, and our MediGPT, enables text

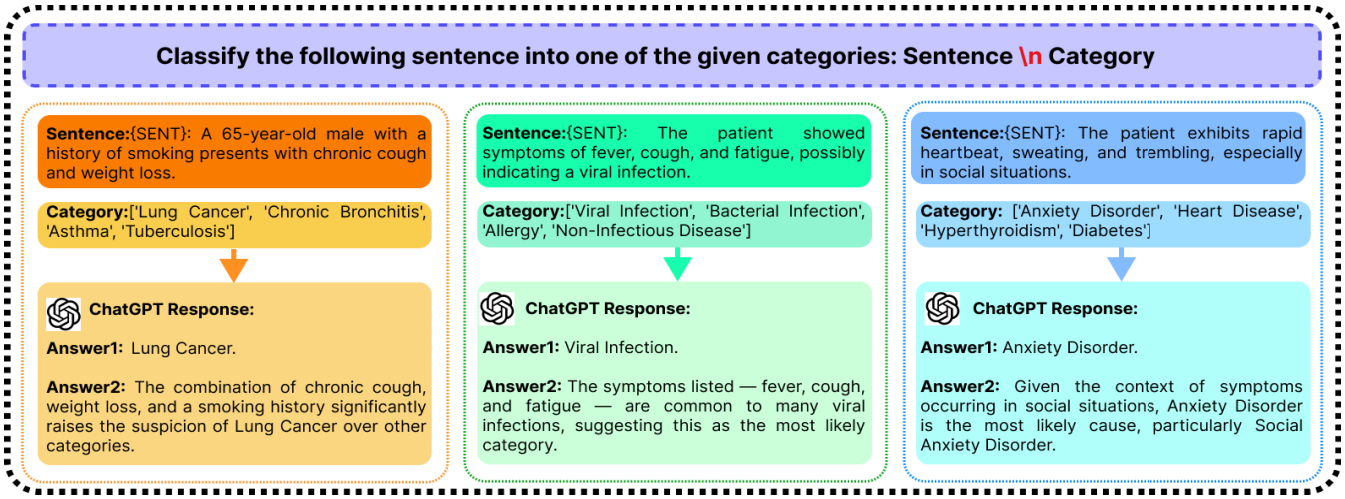

FIGURE 9. The intermediate responses generated by ChatGPT across three medical classification datasets showcase its remarkable proficiency in delivering precise answers and elucidating the rationale behind them.

classification without explicit training on labeled examples. It utilizes the model's understanding of language and context to perform classification tasks.

J. EXPERIMENTAL BENCHMARKS

To comprehensively evaluate these paradigms, our research includes benchmarks against each method within their respective contexts:

- Logistic Regression: Logistic Regression(LR) is a statistical model used to estimate the probability of a binary outcome, utilizing one or more predictor variables. [25]. It uses the logistic function to model a binary dependent variable [26]. The formula for LR can be expressed as:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \tag{2}$$

  where p represents the probability of the dependent variable equaling a case (usually 1), $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients, and $x_1, \ldots, x_n$ are the predictor variables.
- Random Forest: Random Forest(RF) is an ensemble learning technique that constructs multiple decision trees during training and outputs the most common class in classification tasks or the average prediction in regression tasks. It aims to enhance prediction accuracy and robustness by averaging multiple trees, thereby reducing overfitting [27]. The formula for prediction using RF is typically given by:

$$Y = \text{mode}(\{y_1, y_2, y_3, \ldots, y_n\}) \tag{3}$$

  where Y is the predicted outcome, and $y_1, y_2, y_3 \ldots, y_n$ are the outputs from the individual trees.
- Convolutional Neural Networks (CNNs): CNNs for text classification apply convolutional layers to extract features from the text, transformed into a matrix via embeddings [28]. Critical operations include convolution with filters, ReLU activation, max pooling to reduce dimensions, and a softmax function for classification [29]. The classification operation can be formulated as follows:

$$y = \text{softmax}(W_c \cdot \text{maxpool}(\text{ReLU}(W_f * x + b)) + b_c) \tag{4}$$

  where W_f, W_c, b, and b_c are model parameters.
- Long Short-Term Memory (LSTM): LSTM networks are a type of recurrent neural network (RNN) ideal for processing sequences. They are designed to address the vanishing gradient problem in traditional RNNs by incorporating gates that regulate the flow of information. These gates include the input, forget, and output gates, each playing a role in updating the cell and hidden state [30]. The LSTM update equations are:

$$\begin{aligned} i_t &= \sigma(W_{xi} x_t + W_{hi} h_{t-1} + b_i), \\ f_t &= \sigma(W_{xf} x_t + W_{hf} h_{t-1} + b_f), \\ o_t &= \sigma(W_{xo} x_t + W_{ho} h_{t-1} + b_o), \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c), \\ h_t &= o_t \circ \tanh(c_t), \end{aligned} \tag{5}$$

  where σ denotes the sigmoid function, $\circ$ denotes the Hadamard product, and W, b represent weights and biases, respectively.
- BERT-based Fine-tuning: Adjusts BERT's pre-trained embeddings to capture detailed semantic and syntactic information tailored to specific text processing tasks with limited data [2].
- T5-based Prompt-tuning: Activates T5's internal knowledge through targeted Text-to-Text tasks, using prompts to generate appropriate responses [4].
- BART-based Prompt-tuning: Combines bidirectional context modeling and auto-regressive transformers to enhance the effectiveness of prompt learning [3].

K. EVALUATION METRICS

This study used two evaluation metrics to assess the effectiveness of the different technique models, including our proposed MediGPT. These two metrics provide a comprehensive view of model performance, accounting for various aspects of prediction accuracy and reliability [31].

- Accuracy: Accuracy is a key metric used to assess the overall correctness of a model across all prediction tasks. It is the ratio of correct predictions to the total number of predictions made. The formula expresses this ratio:

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions} \quad (6)$$

- F1 score: The F1 score is a metric used to assess the performance of a classification model, combining precision and recall into a single value to provide a balanced measure of both metrics. The F1 score is calculated using the following formula:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

IV. EXPERIMENTAL RESULTS AND ANALYSES

We started our study with comparison experiments to build a foundation. We conducted ablation experiments to understand the role of different factors affecting MediGPT's performance in medical text classification tasks. We also explored how different prompting strategies impact text classification accuracy. Additionally, we integrated ChatGPT-4 into our analysis and compared its performance to the base version, ChatGPT-3.5. Our extensive findings highlight the significant potential, feasibility, and wide-ranging applicability of ChatGPT in medical text classification tasks. We also validated the competitiveness and effectiveness of MediGPT compared to other state-of-the-art models.

A. ENHANCING CHATGPT THROUGH ADVANCED PROMPTING TECHNIQUES

To comprehensively evaluate the enhancements facilitated by advanced prompting strategies in generative pre-trained transformer models within medical domains, Table 3 presents a detailed analysis based on experimental comparisons involving the base configuration of MediGPT (referred to as MediGPT-base) and its variants augmented by diverse prompting mechanisms. The study focuses on four critical areas: clinical trials on cancer, general illness, medical texts, and symptom-to-disease mappings, with performance metrics, quantified through accuracy (Acc) and weighted F1-score (W-F1), as reported in the table.

The baseline model MediGPT-base utilizes manually defined prompts, establishing initial benchmarks of 83.2% accuracy and 82.7% W-F1 for clinical trials on cancer, 82.8% accuracy, 82.3% W-F1 for illness categorization, 90.0% accuracy and 88.8% W-F1 for medical texts, and achieves near-perfect performance with 99.0% for both metrics in symptom-to-disease mappings. These figures are foundational metrics used to assess the effectiveness of more nuanced prompting strategies. The first variant employs ChatGPT Triggered Prompts, where the model dynamically generates prompts based on the immediate context of the query. This approach yields marginal yet consistent improvements across most tasks, enhancing accuracy by 0.8% and W-F1 by 1.0% for cancer trials, 2.1% and 1.4% for illness, and 0.5% and 0.9% for medical texts respectively, with no change observed in the symptom-to-disease task. These increments suggest that context-aware prompt generation can refine the model's response quality, particularly in complex scenarios requiring nuanced understanding. The Zero-Shot similarity prompts strategy, which utilizes the model's capability to generate suitable responses under a ZS learning paradigm without explicit task-specific training, demonstrates more significant improvements. This approach enhances accuracy and W-F1 in illness categorization by 3.0% and 2.2%, respectively, and by 1.2% and 1.4% in cancer trials, indicating a robust capability of the model to generalize from limited inputs effectively. However, a slight decrement of 0.3% in both metrics for symptom-to-disease mappings underscores potential limitations in ZS applicability to tasks requiring deep domain-specific knowledge. Significant progress is demonstrated through the implementation of CoT Triggered Prompts. This method encourages the model to process and articulate intermediate cognitive steps before reaching conclusions. It aligns with the intricate decision-making processes in medical diagnostics and significantly improves outcomes. It elevates performance in illness categorization by 3.8% in accuracy and 3.0% in W-F1 and cancer trials by 1.8% and 2.1%, respectively. This method also slightly improves symptom-to-disease task scores by 0.3% in both metrics, reinforcing the value of explicit reasoning in enhancing diagnostic accuracy and reliability.

B. EXPLORING FEW-SHOT PROMPT-TUNING AND ZERO-SHOT(ZS) CAPABILITIES OF MEDIGPT

Table 4 outlines two primary methodologies under investigation: Transformer-based T5 and BART models in ZS and few-shot learning scenarios. The ZS approach, where models generate outputs without specific training on task-related data, serves as a baseline. For instance, the T5 model achieves an accuracy of 55% and a weighted F1-score of 55.7% in clinical trials on cancer, reflecting the difficulty of generating accurate medical content without targeted learning. Similarly, BART performs slightly better with an accuracy of 57.9% and a w-F1 of 57.3% in the same ZS scenario. Both models show poor performance in illness-related tasks, underscoring the challenge of modeling medical data complexities without fine-tuned datasets.

As the model exposure increases to 10-shot learning, where each model has been fine-tuned with ten examples from the task-specific data, there is a noticeable improvement in performance across all tasks. For instance, the T5 model's performance in clinical trials on cancer jumps to 63.9% accuracy and 62.9% w-F1, demonstrating how even a

TABLE 3. Comparative analysis of MediGPT variants utilizing advanced prompts: Evaluating the performance of MediGPT-base as the baseline (Query Date: 2023.3.24.)

Prompting Strategies	Clinical Trials on Cancer		Illness		Medical Text		Symptom2Disease	
	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1
MediGPT-base Prompts(Manually)	0.832	0.827	0.828	0.823	0.900	0.887	0.990	0.990
ChatGPT(Triggered Prompts)	0.840	0.837	0.849	0.846	0.905	0.896	0.990	0.990
	+0.96%	*+1.21%*	*+2.54%*	*+2.80%*	*+0.56%*	*+1.01%*	*+0.00%*	*+0.00%*
ZS Similarity Prompts	0.844	0.841	0.858	0.855	0.908	0.900	0.987	0.987
	+1.44%	*+1.70%*	*+3.62%*	*+3.89%*	*+0.89%*	*+1.47%*	*-0.30%*	*-0.30%*
CoT Triggered Prompts	0.850	0.848	0.866	0.863	0.908	0.900	0.993	0.993
	+2.17%	*+2.54%*	*+4.59%*	*+4.86%*	*+0.89%*	*+1.47%*	*+0.30%*	*+0.30%*

minimal amount of targeted training can significantly boost model capability. BART shows a similar trend, achieving 66.1% accuracy and 64.3% w-F1 in the same category. The trend continues with further improvement as the exposure increases to 30 examples (30-shot learning), where T5 and BART register their best performances, with BART reaching an accuracy of 72.8% and a w-F1 of 72.2% in clinical trials on cancer, underscoring the effectiveness of incremental learning in enhancing predictive accuracy and model reliability in specialized domains. We observe a significant improvement when comparing these results with the specialized Zero-Shot performance of ChatGPT, which benefits from extensive and nuanced pre-training along with advanced prompt engineering. ChatGPT's ZS model scores extraordinarily high across all categories with 83.2% accuracy and 82.7% w-F1 in clinical trials on cancer and near-perfect scores in the symptom-to-disease category. This illustrates the profound impact of advanced pre-training techniques and model architectures optimized for generative tasks, even without task-specific tuning. The improvement percentages listed at the bottom of the table quantify the relative gains offered by ChatGPT over the best-performing few-shot models (30-shot BART), which range from 14.3% in clinical trials on cancer to a notable 13.7% in both accuracy and w-F1 in symptom-to-disease mappings. These metrics underscore the advanced capabilities of ChatGPT in managing intricate medical queries and highlight the potential of generative AI to transform information synthesis and decision support in healthcare settings. This capability is achieved without requiring extensive task-specific data, thus lowering the implementation barriers across diverse medical scenarios.

C. COMPARISION BETWEEN GPT3.5 AND CHATGPT-4

In comparing ChatGPT-3.5 and ChatGPT-4 across various medical and health-related datasets, a noticeable improvement in performance metrics is observed when transitioning from ChatGPT-3.5 to ChatGPT-4. As illustrated in Figure 10, both models were evaluated based on their accuracy and weighted F1-score (W-F1) across four distinct datasets: Clinical Trials on Cancer, Illness, Medical Text, and Symptom2Disease. For the Clinical Trials on Cancer dataset, ChatGPT-4 demonstrates an accuracy of 83.2%, compared to 81.0% for ChatGPT-3.5, reflecting a 2.2% increase. The W-F1 score also improves, with ChatGPT-4 achieving 82.7%, 2.3 points higher than ChatGPT-3.5's score of 80.4%. This improvement indicates that ChatGPT-4's refined understanding and processing capabilities better capture the nuances and complexities inherent in clinical trial data related to oncology. In the Illness dataset, both models performed closely, with ChatGPT-4 slightly outperforming ChatGPT-3.5. ChatGPT-4 achieved an accuracy of 82.8% compared to ChatGPT-3.5's 81.3% and a W-F1 of 82.3%, surpassing ChatGPT-3.5's score of 81.0%. These more minor improvements highlight ChatGPT-4's enhanced ability to interpret and classify data about various illnesses accurately. Significant improvements are evident in the Medical Text dataset, where ChatGPT-4's accuracy increased to 90.0% from 86.6%, and its W-F1 score rose to 88.8 from 85.6. This dataset likely benefits from ChatGPT-4's advanced language models, which more effectively grasp the specialized vocabulary and complex sentence structures typical of medical texts. The Symptom2Disease dataset showed notable gains, with ChatGPT-4 reaching an accuracy of 86.2%, compared to 82.7% for ChatGPT-3.5, and a W-F1 of 85.8, up from 83.3%. These results indicate that ChatGPT-4 is better equipped to link symptoms with potential diseases, a critical capability in medical diagnostics. The comparative analysis shows that ChatGPT-4 consistently achieves higher accuracy and weighted F1 scores across all analyzed datasets. This improvement can be attributed to ChatGPT-4's enhanced model architecture, which likely includes more advanced training algorithms and a larger, more diverse training dataset. These enhancements enable ChatGPT-4 to comprehend better and handle complex, specialized content in the medical and healthcare domain, resulting in more precise predictions and classifications.

D. PERFORMANCE COMPARISON OF DIFFERENT MODELS AND TECHNIQUES

Table 5 presents an in-depth performance comparison of different models and techniques. This comparison spans traditional ML approaches, word embedding techniques, pre-trained language models (PLMs), both Fine-Tuned (FT) and Prompt-Tuned (PT), and our proposed MediGPT. The traditional ML models, LR and RF, demonstrated moderate performance across all datasets. LR achieved accuracy scores ranging from 0.648 to 0.784 and w-F1 scores from 0.645 to 0.763. RF showed slightly better results with accuracy

TABLE 4. Performance metrics of different methods.

Few-shot learning	Methods	Clinical Trials on Cancer		Illness		Medical Text		Symptom2Disease	
		Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1
ZS	T5	0.555	0.557	0.508	0.500	0.596	0.579	0.631	0.631
	BART	0.579	0.573	0.473	0.465	0.607	0.600	0.673	0.673
	Improvement	*4.3%*	*2.9%*	*-6.9%*	*-7.0%*	*1.8%*	*3.6%*	*6.7%*	*6.7%*
10-Shot	T5	0.639	0.629	0.619	0.612	0.708	0.685	0.791	0.791
	BART	0.661	0.643	0.597	0.588	0.677	0.660	0.795	0.795
	Improvement	*3.4%*	*2.2%*	*-3.6%*	*-3.9%*	*-4.4%*	*-3.6%*	*0.5%*	*0.5%*
30-Shot	T5	0.713	0.708	0.690	0.681	0.766	0.753	0.865	0.865
	BART	0.728	0.722	0.677	0.663	0.792	0.780	0.888	0.888
	Improvement	*2.1%*	*2.0%*	*-1.9%*	*-2.6%*	*3.4%*	*3.6%*	*2.7%*	*2.7%*
ZS(Default)	ChatGPT	0.832	0.827	0.828	0.823	0.900	0.887	0.992	0.993
	Improvement	*14.3%*	*14.5%*	*22.3%*	*24.2%*	*13.6%*	*13.7%*	*13.7%*	*13.7%*

TABLE 5. Performance statistics of baselines and MediGPT on adopted datasets. We underline the best values.

Learning Paradigms	Baselines	Clinical Trials on Cancer		Illness		Medical Text		Symptom2Disease	
		Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1
Traditional ML	LR	0.648	0.645	0.693	0.676	0.784	0.763	0.544	0.543
	RF	0.669	0.664	0.685	0.673	0.808	0.776	0.574	0.555
Word Embedding	CNN	0.769	0.763	0.736	0.725	0.855	0.837	0.813	0.806
	LSTM	0.748	0.746	0.728	0.718	0.866	0.848	0.833	0.822
PLM-based FT	BERT-FT	0.788	0.785	0.757	0.735	0.890	0.860	0.847	0.840
PLM-based PT	T5-PT	0.826	0.820	0.785	0.774	0.895	0.878	0.880	0.875
	BART-PT	0.821	0.816	0.778	0.788	0.896	0.886	0.888	0.883
ChatGPT-based QA	MediGPT	0.832	0.827	0.828	0.823	0.901	0.887	0.992	0.993
	Improvement	*+1.34%*	*+1.35%*	*+6.43%*	*+4.44%*	*+0.56%*	*+0.11%*	*+11.7%*	*+11.0%*

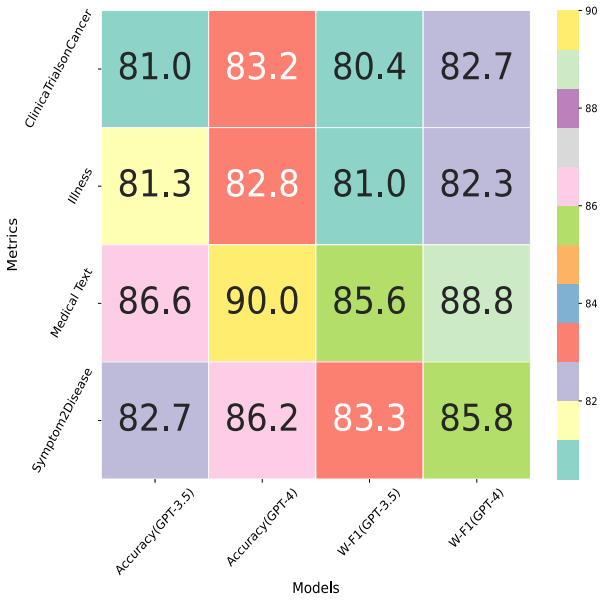

FIGURE 10. Comparision between ChatGPT 3.5 and GPT 4.

between 0.669 and 0.808 and w-F1 scores from 0.664 to 0.776. These models performed exceptionally well on the Medical Text dataset but lagged on the Symptom2Disease dataset. Word embedding-based models, such as CNN and LSTM, improved upon the traditional models. CNN achieved accuracy scores between 0.736 and 0.855 and w-F1 scores from 0.725 to 0.837. LSTM showed similar performance with accuracy ranging from 0.728 to 0.866 and w-F1 scores from 0.718 to 0.848. These models excelled in handling complex medical texts, indicating the advantage of DL techniques in capturing semantic information. Pre-trained language models fine-tuned on specific tasks, like BERT-FT, enhanced performance. BERT-FT achieved accuracy scores from 0.757 to 0.890 and w-F1 from 0.735 to 0.860. The improved performance across all datasets highlighted the effectiveness of leveraging large-scale pre-trained models for medical text classification tasks. The fine-tuned PLM models, such as T5-PT and BART-PT, demonstrated even higher performance. T5-PT achieved accuracy scores ranging from 0.785 to 0.895 and w-F1 scores from 0.774 to 0.878. BART-PT showed similar results with accuracy from 0.778 to 0.896 and w-F1 scores from 0.788 to 0.886. These models exhibited superior performance on more challenging datasets, such as Symptom2Disease, underscoring the benefits of task-specific fine-tuning. Our proposed model, MediGPT, based on a ChatGPT-based QA framework, outperformed all the baseline models across all datasets. MediGPT achieved the highest accuracy and w-F1 scores, with notable performance on the Symptom2Disease dataset, scoring 0.992 for both metrics. Table 5 also illustrates the improvement percentages of MediGPT over BART-PT. MediGPT showed a 1.34% improvement in accuracy and a 1.35% improvement in w-F1 for the Clinical Trials on Cancer dataset, a 6.43% increase in accuracy and a 4.44% increase in w-F1 for the Illness dataset, a 0.56% improvement in accuracy and a 0.11% improvement in w-F1 for the Medical Text dataset, and an impressive 11.7% increase in accuracy and an 11.0% increase in w-F1 for the Symptom2Disease dataset. These improvements highlight MediGPT's enhanced ability to handle diverse and

complex medical datasets effectively. MediGPT's superior performance can be attributed to its ability to understand and generate contextually relevant responses, thus enhancing its classification capabilities. The performance comparison highlights the progressive improvements from traditional ML models to advanced pre-trained and fine-tuned language models. MediGPT is the best-performing model, showcasing its potential for advancing medical text classification tasks. These results demonstrate the significant impact of leveraging state-of-the-art natural language processing techniques in medical informatics.

V. CONCLUSION AND FUTURE WORK

Medical text classification is crucial in organizing a vast and expanding volume of medical information. Current PLM based models face challenges such as dependency on annotated data, limited transferability across languages, and deployment complexities. The introduction of ChatGPT has provided new avenues to address these challenges, particularly in enhancing the sustainable management of medical information through text classification. In this study, we explored the capabilities of ChatGPT in medical text classification and introduced MediGPT, a novel framework designed for this purpose. MediGPT represents an initial qualitative assessment of ChatGPT's application in healthcare text classification. Our research compared MediGPT against conventional machine learning methods, PLM-based fine-tuning approaches, and prompt-based learning techniques. We conducted extensive evaluations across diverse datasets and devised strategies for generating prompts to enhance the quality of ChatGPT's outputs. Furthermore, we evaluated the performance of ChatGPT-4 through comparative experiments. MediGPT demonstrated significant performance improvements across the four selected datasets, with accuracy increases of 14.3%, 22.3%, 13.6%, and 13.7%, respectively. These enhancements underscore its efficacy in handling diverse medical texts compared to traditional models. Our findings affirm ChatGPT's superiority in medical text classification, representing substantial progress in leveraging AI for managing medical information. This research sets the stage for future applications in advancing sustainable healthcare practices, promoting digital innovation, and enhancing operational efficiency. The strides made by MediGPT contribute to the evolution of medical text classification, emphasizing the integration of advanced AI models in healthcare to optimize the management and utilization of medical information.

ACKNOWLEDGMENT

The authors would like to acknowledge Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R197), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would like to thank Prince Sultan University for their support.

REFERENCES

[1] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. A. Almazroi, "A fine-tuned BERT-based transfer learning approach for text classification," *J. Healthcare Eng.*, vol. 2022, pp. 1–17, Jan. 2022.
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.
[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
[5] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "GPTs are GPTs: An early look at the labor market impact potential of large language models," 2023, *arXiv:2303.10130*.
[6] N. Xia, H. Yu, Y. Wang, J. Xuan, and X. Luo, "DAFS: A domain aware few shot generative model for event detection," *Mach. Learn.*, vol. 112, no. 3, pp. 1011–1031, Mar. 2023.
[7] J. Gao, H. Yu, and S. Zhang, "Joint event causality extraction using dual-channel enhanced neural network," *Knowl.-Based Syst.*, vol. 258, Dec. 2022, Art. no. 109935.
[8] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
[9] Q. Lyu, J. Tan, M. E. Zapadka, J. Ponnatapura, C. Niu, K. J. Myers, G. Wang, and C. T. Whitlow, "Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: Results, limitations, and potential," *Vis. Comput. for Ind., Biomed., Art*, vol. 6, no. 1, p. 9, May 2023.
[10] T. Susnjak, "Applying BERT and ChatGPT for sentiment analysis of Lyme disease in scientific literature," in *Borrelia Burgdorferi: Methods and Protocols*. New York, NY, USA: Springer, 2024, pp. 173–183.
[11] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.
[12] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023.
[13] Y. Shen, X. Ma, Z. Tan, S. Zhang, W. Wang, and W. Lu, "Locate and label: A two-stage identifier for nested named entity recognition," 2021, *arXiv:2105.06804*.
[14] M. U. Haque, I. Dharmadasa, Z. T. Sworna, R. N. Rajapakse, and H. Ahmad, "'I think this is the most disruptive technology': Exploring sentiments of ChatGPT early adopters using Twitter data," 2022, *arXiv:2212.05856*.
[15] M. S. Islam, M. A. T. Rony, and T. Sultan, "GastroVRG: Enhancing early screening in gastrointestinal health via advanced transfer features," *Intell. Syst. Appl.*, vol. 23, Sep. 2024, Art. no. 200399.
[16] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Med. Res. Methodol.*, vol. 22, no. 1, p. 181, Dec. 2022.
[17] A. K. Mohanty, M. R. Senapati, S. Beberta, and S. K. Lenka, "Texture-based features for classification of mammograms using decision tree," *Neural Comput. Appl.*, vol. 23, nos. 3–4, pp. 1011–1017, Sep. 2013.
[18] Y. Li, S. Zhang, and C. Lai, "Agricultural text classification method based on dynamic fusion of multiple features," *IEEE Access*, vol. 11, pp. 27034–27042, 2023.
[19] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun, "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," 2023, *arXiv:2302.09419*.
[20] I. Spasic and G. Nenadic, "Clinical text data in machine learning: Systematic review," *JMIR Med. Informat.*, vol. 8, no. 3, Mar. 2020, Art. no. e17984.
[21] J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou, "Is ChatGPT a good NLG evaluator? A preliminary study," 2023, *arXiv:2303.04048*.
[22] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT," 2023, *arXiv:2302.10198*.
[23] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, May 2021.

[24] J. O. Bappi, M. A. T. Rony, and M. S. Islam, "BNVGLENET: Hypercomplex Bangla handwriting character recognition with hierarchical class expansion using convolutional neural networks," *Natural Lang. Process. J.*, vol. 7, Jun. 2024, Art. no. 100068.
[25] A. Das, "Logistic regression," in *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht, The Netherlands: Springer, 2024, pp. 3985–3986.
[26] M. S. Sayed, M. A. T. Rony, M. S. Islam, A. Raza, S. Tabassum, M. S. Daoud, H. Migdady, and L. Abualigah, "A novel deep learning approach for forecasting myocardial infarction occurrences with time series patient data," *J. Med. Syst.*, vol. 48, no. 1, p. 53, May 2024.
[27] M. R. Ali, S. M. A. Nipu, and S. A. Khan, "A decision support system for classifying supplier selection criteria using machine learning and random forest approach," *Decis. Anal. J.*, vol. 7, Jun. 2023, Art. no. 100238.
[28] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," 2015, *arXiv:1510.03820*.
[29] J. O. Bappi, M. A. T. Rony, M. S. Islam, S. Alshathri, and W. El-Shafai, "A novel deep learning approach for accurate cancer type and subtype identification," *IEEE Access*, vol. 12, pp. 94116–94134, 2024.
[30] M. Alizamir, J. Shiri, A. F. Fard, S. Kim, A. D. Gorgij, S. Heddam, and V. P. Singh, "Improving the accuracy of daily solar radiation prediction by climatic data using an efficient hybrid deep learning model: Long short-term memory (LSTM) network coupled with wavelet transform," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106199.
[31] Z. Lian, Y. Ma, M. Li, W. Lu, and W. Zhou, "Discovery precision: An effective metric for evaluating performance of machine learning model for explorative materials discovery," *Comput. Mater. Sci.*, vol. 233, Jan. 2024, Art. no. 112738.
[32] B. Zhao, W. Jin, J. Del Ser, and G. Yang, "ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification," *Neurocomputing*, vol. 557, 2023, Art. no. 126708, doi: 10.1016/j.neucom.2023.126708. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223008317

MOHAMMAD ABU TAREQ RONY received the Bachelor of Science degree in statistics from Noakhali Science & Technology University, Noakhali, Bangladesh. Additionally, he possesses expertise in devising advanced analytics strategies using data. His diverse professional experience includes three years of research in artificial intelligence. He is working as a part-time Research Data Scientist at AiQuest Intelligence. Dhaka, Bangladesh. He has published articles in refereed journals and conference proceedings, such as IEEE Access, *Data in Brief* (Elsevier), *Children* (MDPI), IEEE, and Springer international conferences. Moreover, he actively engages in partnerships with international researchers, recognizing that research is indispensable in fostering innovation. Overall, he is hardworking and has taught himself various skills, such as data analysis, statistics, ML, and DL.

MOHAMMAD SHARIFUL ISLAM received the B.Sc. degree in computer science and telecommunication engineering from Noakhali Science & Technology University, Bangladesh, in 2023, brings a deep passion for cutting-edge technologies to the research community. His academic journey, rooted in computer science and telecommunications, has evolved into a genuine pursuit of specialized areas, including data science, ML, natural language processing, and image processing. His work in these fields is driven by a quest to uncover hidden insights within data, develop intelligent learning algorithms, bridge the communication gap between humans and machines, and artistically enhance digital imagery. As a Researcher, his approach is characterized by a blend of technical proficiency and creative problem-solving, aiming to contribute significantly to the frontiers of technology and its application in understanding and improving our digital world.

TIPU SULTAN received the B.Sc. degree (Hons.) in mechatronics engineering from Kyungsung University, South Korea, and the M.Sc. degree in data science from Fordham University, Lincoln Center, in April 2024. Currently, interning as a Data Analyst with the Authentic Brands Group, he specializes in ML, Tableau dashboard creation, Power BI, and SQL database management. With over five years of experience in various industries as a Data Analyst and a Machine Learning Researcher, he brings a wealth of expertise to his roles. Proficient in SQL, R, Python, Tableau, and Excel Power BI, he possesses a solid foundation in machine learning. With a genuine passion for data analysis, he is well-equipped to excel in data science.

SAMAH ALSHATHRI received the Bachelor of Computer Science and Master of Computer Engineering degrees from King Saud University, Riyadh, Saudi Arabia, and the Ph.D. degree from the Department of Computer and Mathematics, Plymouth University, Plymouth, U.K. She is currently an Assistant Professor with the Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University (PNU), Riyadh, Saudi Arabia. Her research interests include wireless networks, cloud computing, fog computing, the IoT, data mining, machine learning, text analytics, image classification, and deep learning. She has authored or co-authored many articles published in well-known journals in the research field. She was the Chair of the Network and Communication Department and participated in organizing many international conferences.

WALID EL-SHAFAI (Senior Member, IEEE) was born in Alexandria, Egypt. He received the B.Sc. degree (Hons.) in electronics and electrical communication engineering from the Faculty of Electronic Engineering (FEE), Menoufia University, Menouf, Egypt, in 2008, the M.Sc. degree from Egypt-Japan University of Science and Technology (E-JUST), in 2012, and the Ph.D. degree from FEE, Menoufia University, in 2019. Since January 2021, he has been a Postdoctoral Research Fellow with the Security Engineering Laboratory (SEL), Prince Sultan University (PSU), Riyadh, Saudi Arabia. He is currently a Senior Cybersecurity Researcher with the SEL Laboratory and an Assistant Professor with the College of Computer Science and Information Systems. Also, he is an Associate Professor with the Department of Electronics and Communication Engineering (ECE), FEE, Menoufia University. His research interests include wireless mobile and multimedia communications systems, image and video signal processing, efficient 2D video/3D multi-view video coding, multi-view video plus depth coding, 3D multi-view video coding and transmission, quality of service and experience, digital communication techniques, cognitive radio networks, adaptive filters design, 3D video watermarking, steganography, encryption, error resilience and concealment algorithms for H.264/AVC, H.264/MVC, H.265/HEVC video codecs standards, cognitive cryptography, medical image processing, speech processing, security algorithms, software-defined networks, the Internet of Things, medical diagnoses applications, FPGA implementations for signal processing algorithms and communication systems, cancellable biometrics and pattern recognition, image and video magnification, artificial intelligence for signal processing algorithms and communication systems, modulation identification and classification, image and video super-resolution and denoising, cybersecurity applications, malware and ransomware detection and analysis, deep learning in signal processing, and communication systems applications. He also serves as a reviewer for several international journals.