Assayed *et al. Journal of Engineering and Applied Science* (2025) 72:212
https://doi.org/10.1186/s44147-025-00800-y

Journal of Engineering and Applied Science

RESEARCH Open Access

Engineering intelligent healthcare systems: understanding medical queries with AI and NLP

Suha Khalil Assayed[1,2*], Chin-Shiuh Shieh[2] and Shashi Kant Gupta[2,3]

*Correspondence:
Suha Khalil Assayed
sassayed@gmail.com
[1]Faculty of Engineering and IT, Department of Computer Science, The British University in Dubai, Dubai, UAE
[2]Research Institute of IoT Cybersecurity, Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan
[3]Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Pujab, India

Abstract

Industry 5.0 introduces a human-centered approach where engineering and applied science are combined to create smarter systems that directly improve human well-being. In healthcare, this approach is realized through Healthcare 5.0, which uses Artificial Intelligence (AI) and Natural Language Processing (NLP) to design intelligent platforms that can interpret patient questions and provide accurate responses. This study addresses the engineering challenge of intent classification in medical question-answering systems, an essential step in developing reliable healthcare chatbots and decision-support tools. Using the MedQuad dataset of 14,979 labeled medical questions, we evaluate classical machine learning models such as Logistic Regression, Naive Bayes, Support Vector Machines (SVM), and Random Forest, along with the transformer-based BERT model. Nonetheless, to improve classification under imbalanced data, the Synthetic Minority Oversampling Technique (SMOTE) was applied. In the training phase, the Random Forest model attained 100% accuracy, whereas its inference accuracy on unseen data (without SMOTE) was 80%, demonstrating its effectiveness in generalizing beyond the training set, while other models performed moderately, and BERT required more domain-specific tuning. The findings highlight the contribution of computational engineering methods to healthcare applications and demonstrate how applied AI models can support human-centered solutions at the intersection of engineering and medical sciences.

Keywords Artificial intelligence, BERT, Healthcare, NLP, NLU, LLM, Machine learning, Medical, SMOTE

Introduction

As we transition into the era of Industry 5.0, the focus of technological advancement is shifting from automation toward more human-centric solutions [1]. This new industrial vision emphasizes the integration of Artificial Intelligence (AI) and Machine Learning (ML) with human values such as empathy and well-being. In this context, Healthcare 5.0 has emerged as a paradigm where engineering and applied sciences are employed to design technologies that not only support but also enhance human competence, enabling the collaboration between doctors, patients, and intelligent systems [2]. From an engineering perspective, such advancements require strong computational models, suitable algorithms, and efficient human–machine interfaces to ensure accuracy and usability. As

Springer Open

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

a result, patients increasingly expect immediate and precise responses to their medical inquiries, making intelligent, human-centered healthcare systems a fundamental component of delivering high-quality care and support. Patients often seek instant answers to a different health related enquiries ranging from symptoms to medication details. Therefore, several authors have contributed to developing various AI-powered virtual assistants that provide instant healthcare support, enabling patients to access timely and accurate information at any time [3].

In fact, the accuracy of any AI model mostly depends on its ability to understand the context and intent behind each query [4]. This process, known as Natural Language Understanding (NLU), enables the AI to interpret the meaning of questions and enquiries by analyzing the context and the intents, so it accomplishes both the intent classification and entity extraction [5]. When the patient asks a question, the NLU system understands the meaning and the intention behind this question [6], helping to identify what the patient is truly seeking. This process is known as intent recognition.

Intent recognition plays a critical role in the development of conversational AI systems, including chatbots, question-answering platforms, and dialogue systems [7, 8]. Consequently, research interest in this area has expanded across various domains. Although intent recognition has been applied to healthcare and the medical field, the volume of research in these domains is still relatively limited compared to others. Despite significant progress in general-purpose intent recognition, its application to healthcare and medical communication remains relatively limited. Recent research has focused on improving intent recognition in the medical domain using deep learning and transformer models across different datasets and domains, though studies often face limitations in model techniques, dataset types, or dataset size.

Therefore, this study aims to address these limitations by engineering intelligent healthcare systems that leverage AI and Natural Language Processing (NLP) to better understand and classify medical queries. Specifically, it compares the performance of traditional machine learning models and transformer-based NLP models in medical intent recognition. By evaluating these approaches, the study seeks to enhance the accuracy, adaptability, and human-centered functionality of AI-driven healthcare systems.

This work is organized as follows: The Related Work section reviews previous studies and highlights the research gap. The Methodology section describes the dataset preparation, preprocessing, and model implementation. The Results and Discussion section presents and interprets the experimental findings, followed by sections on Limitations, Conclusions, and Future Research Directions.

Related work

Comparative developments in medical query classification models

Recent studies have explored various deep learning and transformer-based models for medical question answering and intent recognition, each contributing distinct advancements. Tohti et al. [9] evaluated multiple architectures on a Chinese medical dataset and found that the ALBERT-BiLSTM model achieved the highest F1-score of 92.3%, although its focus only on Chinese data. Similarly, Lee et al. [11] introduced BioBERT, a domain-adapted version of BERT pretrained on biomedical texts, which showed notable improvements over the vanilla BERT in biomedical named entity recognition tasks. Turchin et al. [10] extended this work by comparing regular BERT, BioBERT, and

ClinicalBERT across only electronic health record (EHR) descriptions. Other studies proposed hybrid and enhanced approaches for instance, Guan and Tezuka [13] combined BERT with TextCNN, knowledge graphs, and rule templates to achieve 85.6% accuracy in question linking, while Shi et al. [14] introduced prompt-based learning to boost intent recognition accuracy to 89.4%. Tresner-Kirsch et al. [12] explored multilingual intent classification using traditional and neural models, finding that LSTM yielded the best average F1-score of 81%, though cultural and linguistic diversity posed challenges. Lastly, Luo et al. [15] proposed TAI-FLB, which integrated attention and label boosting, reaching a 90.2% F1-score and surpassing previous baselines.

Overall, these studies collectively underscore the growing effectiveness of deep learning and hybrid models in healthcare NLP, while also revealing persistent challenges related to data diversity, domain adaptation, and scalability. Table 1 summarizes the progression of research efforts within the medical domain, highlighting the specific limitations identified in each study.

Challenges and limitations in medical intent recognition models

While previous studies have shown promising results, they also reveal persistent challenges related to data diversity, domain adaptation, and scalability as shown in Table 1. These limitations highlight the need for more generalizable and adaptable models for medical intent recognition. For instance, a multi-task learning model was introduced in [9] based on ALBERT-BiLSTM for intent classification and named entity recognition in Chinese medical questions. Although the results demonstrated that combining pre-trained language models with multi-task learning improves the understanding of medical texts, the model was trained exclusively on Chinese data, limiting its applicability to other languages. Similarly, the study in [10] applied intent recognition models to user messages in Hausa, Hindi, Swahili, and English through the askNivi chatbot, which focuses on sexual and reproductive health education and access. However, because the chatbot deals with culturally sensitive topics, the models faced challenges in understanding variations in how users from different linguistic backgrounds express their queries.

Another challenge is reflected in the work of Guan and Tezuka [13], who proposed a model combining intent recognition and named entity recognition (NER) to link user questions to a medical knowledge graph using rule-based templates formed during slot filling. While effective for structured queries, this method lacks flexibility in handling complex or unstructured questions. Furthermore, Shi et al. [14] explored the use of pre-trained transformer models enhanced with prompt learning methods for intent recognition and question answering in the medical domain. Although this approach improved performance, it relies heavily on domain-specific prompt engineering, which is time-consuming, lacks standardization, and limits transferability across medical domains and tasks.

Additional challenge is highlighted in the work of Luo et al. [15], who incorporated a text-to-label attention interaction mechanism based on label embeddings to better interpret ambiguous or incomplete queries, which is particularly valuable in the medical domain where queries can vary widely in complexity and specificity. However, the model's performance heavily depends on the quality of the label embeddings, and poor or insufficient embeddings may limit its ability to generalize effectively. Moreover, Turchin et al. [10] compared several BERT variants such as classic BERT, BioBERT, and

Table 1 Overview of recent intent recognition studies in medical QA systems with goals and limitations

Author(s)	Model/Algorithm	Goals	Metrics Used	Performance Results	Limitation
Tohti et al. (2022) [9]	ALBERT-BiLSTM	A comparative experiment of different models on a Chinese medical question dataset	Accuracy, Precision, F1-score	ALBERT-BiLSTM achieved the best performance with an F1-score of 92.3%	Focus only on Chinese medical data; may not generalize well to other languages.
Turchin et al. (2023) [10]	Regular BERT, BioBERT, ClinicalBERT	Compare accuracy in identifying complex medical concepts in EHR narrative notes	Macro-F1, Precision, Recall; also precisio	ClinicalBERT achieved highest mean Macro-F1 (0.761), BioBERT 0.735, regular BERT 0.699.	Only one implementation of ClinicalBERT tested;
Lee et al. (2019) [11]	BioBERT	Adapt BERT to biomedical domain; improve performance on biomedical NER	F1	Over vanilla BERT: +0.62% F1 in NER	Focused on biomedical literature (not clinical notes
Guan & Tezuka (2022) [13]	BERT-TextCNN + Knowledge Graph + Rule Templates	Linked user questions to a medical knowledge graph using rule templates formed during slot filling	Accuracy, Recall	Achieved 85.6% accuracy in linking questions correctly using the hybrid model	Relies on rule-based templates and slot filling; limited flexibility for unstructured queries.
Shi et al. (2024) [14]	Pre-trained Transformer + Prompt Learning	Introduced prompt-based learning to enhance intent recognition in medical QA	Accuracy, F1-score, Precision	Prompt learning outperformed fine-tuned BERT baselines, achieving up to 89.4% accuracy	Domain-specific prompt engineering required; lacks scalability and standardization.
Tresner-Kirsch et al. (2023) [12]	SVM, Random Forests, LSTM	Explored intent recognition for chatbot queries on sexual/reproductive health using supervised ML over a multilingual corpus	Accuracy, F1-score, Cross-validation	LSTM model achieved best average F1-score of 81% across multiple languages	Intent classification may suffer due to cultural and language variations.
Luo et al. (2023) [15]	TAI-FLB (Text-Label Attention Interaction + Focal Loss Boost)	Proposed a new model for intent recognition in medical queries using attention and label boosting	Precision, Recall, F1-score	TAI-FLB achieved 90.2% F1-score, outperforming traditional and transformer baselines	Relies heavily on label embedding quality; performance may degrade with sparse or noisy label representations.

ClinicalBERT to evaluate their ability to identify complex medical concepts in electronic health record (EHR) narratives. Their results showed that ClinicalBERT achieved the highest mean Macro-F1 score (0.761), outperforming BioBERT (0.735) and regular BERT (0.699). However, the study was limited to a single implementation of ClinicalBERT, which restricted broader generalization. Similarly, Lee et al. (2019) introduced

BioBERT, an adaptation of BERT for the biomedical domain, which improved performance on biomedical named entity recognition (NER) tasks by 0.62% F1 over the vanilla BERT model. Nevertheless, this work primarily focused on biomedical literature rather than clinical or patient-generated text.

While prior work has explored a variety of neural and transformer-based techniques, most studies have focused on specific models or languages. There remains a lack of comprehensive comparative analyses assessing both classical machine learning models such as Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Random Forest as well as other transformer-based approaches within particularly the medical question-answering domain.

The growing use of AI and Big Data is transforming healthcare, making it more efficient, insightful, and focused on improving patient care [16, 17]. However, there's still a gap in creating NLP models that can easily adapt, use data efficiently, and truly understand different types of medical questions.

This study addresses these gaps by evaluating and comparing machine learning and fine-tuning transformer-based NLP models for intent classification within English medical question-answering datasets. Unlike prior studies that focused on specific languages or highly specialized models, this work emphasizes models that are less complex, more efficient, and easier to integrate into real-world healthcare systems. The study aims to develop intelligent healthcare solutions that use AI and NLP to better understand and classify medical queries, thereby improving healthcare delivery and supporting more human-centered, accessible, and responsive care.

Methodology

In this study, we developed a model to classify user intent in medical question-answering (QA) tasks, using Python to evaluate the performance of multiple machine learning algorithms in comparison with language models. Specifically, we assessed the accuracy and effectiveness of Bidirectional Encoder Representations from Transformers (BERT), a fine-tuned large language model introduced by Devlin et al. [18]. The methodology involved preprocessing and balancing the dataset, followed by training traditional machine learning classifiers and fine-tuning the BERT model. The aim was to compare their performance in terms of accuracy and effectiveness for intent recognition in medical queries. Fig. 1 provides a detailed overview of the experimental design and methodology that used in this study.

Dataset

The dataset used in this study was obtained from the MedQuad collection, a comprehensive repository of medical question-answer pairs designed for natural language processing tasks, and is publicly available on Kaggle [19]. It contains 14,979 patient-generated medical questions and 16,407 expert-provided answers. Each question is classified into 15 categories, as detailed in Table 2 below.

The majority of the answers are concentrated in particular classes. For example, the "Information" category has the highest number of answers, with 4,535, followed by "Symptoms" and "Treatment" with 2,748 and 2,442 answers, respectively. On the other hand, categories such as "Stages" and "Complications" are significantly underrepresented, with only 77 and 46 answers, respectively. Consequently, this dataset is

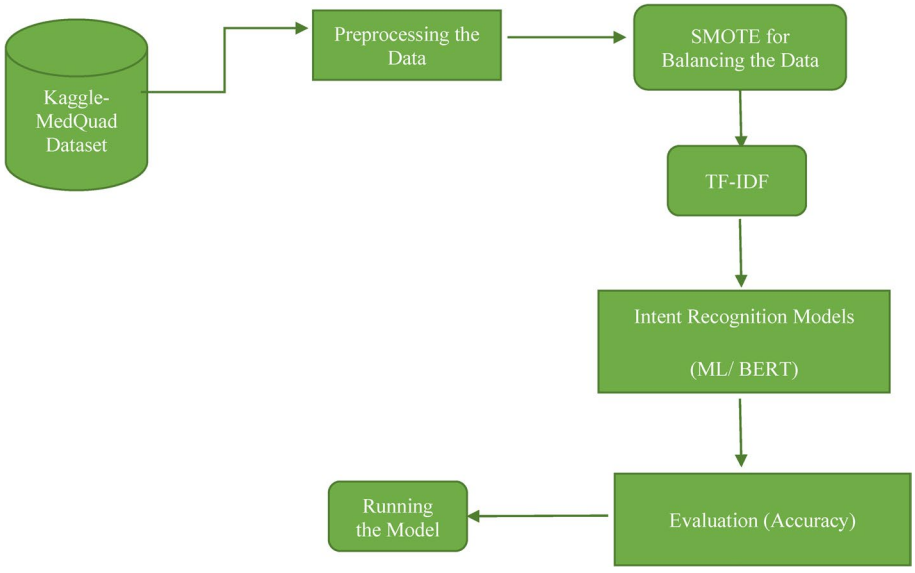

Fig. 1 Experimental workflow comparing ML models algorithms for intent recognition in medical QA tasks

Table 2 MedQuad dataset summary, showing question count, answers, and category labels

Class	# of provided answers.
Information	4535
Symptoms	2748
Treatment	2442
Inheritance	1446
Frequency	1120
Genetic Changes	1087
Causes	727
Exams and Tests	653
Research	395
Outlook	361
Susceptibility	324
Considerations	235
Prevention	210
Stages	77
Complications	46
	16,406

imbalanced, which could negatively impact the model's performance. The Table 3 shows the percentage for the answers per class.

Balancing the dataset

An imbalanced class is a common challenge in many real-world applications [20], since it could lead to a model bias toward the majority class, as there are insufficient training samples from the minority class to accurately represent their patterns. This imbalance can delay the model's ability to learn from underrepresented categories resulting in poor generalization and reduced performance, especially for minority class predictions.

In this study, we applied the SMOTE technique to balance the dataset. The resulting class distribution after applying SMOTE is presented in Table 4.

Table 3 MedQuad dataset summary showing question count, answers, and category labels

Class	Answers	% of Total
Information	4535	27.6%
Symptoms	2748	16.7%
Treatment	2442	14.9%
Inheritance	1446	8.8%
Frequency	1120	6.8%
Genetic Changes	1087	6.6%
Causes	727	4.4%
Exams and Tests	653	4.0%
Research	395	2.4%
Outlook	361	2.2%
Susceptibility	324	2.0%
Considerations	235	1.4%
Prevention	210	1.3%
Stages	77	0.5%
Complications	46	0.3%

Table 4 Class distribution after applying SMOTE

Class	Sample Count
susceptibility	4535
symptoms	4535
exams and tests	4535
treatment	4535
prevention	4535
information	4535
frequency	4535
complications	4535
causes	4535
research	4535
outlook	4535
considerations	4535
inheritance	4535
stages	4535
genetic changes	4535

After applying SMOTE, the dataset was balanced by ensuring that each class contained an equal number of samples, which is expected to enhance the model's performance. The Synthetic Minority Oversampling Technique (SMOTE), as described by Pradipta et al. [21], is a widely used method for addressing class imbalance [22]. It works by generating synthetic data points in the feature space using an instance and its K-nearest neighbors. This technique helps mitigate overfitting and enables the classifier to better learn and define decision boundaries between classes.

Models for intent recognition

Naive Bayes

The Naive Bayes algorithm is a classification technique based on Bayes' Theorem. It is primarily used to determine the most likely hypothesis given certain evidence, relying heavily on probability calculations [23]. It is widely used in classification and data analysis because it is simple and fast [24].

Random forest

Random Forest (RF) constructs multiple decision trees and combines their outputs, which helps minimize the effect of errors and noise. This makes it well-suited for intent recognition, even when user queries contain typos, filler words, or inconsistent phrasing. As a flexible, non-parametric classification method, it performs effectively on complex, irregular, or unknown data patterns [25].

Logistic regression

Logistic regression is a simple, interpretable and fast used to classify text into two or more intent categories. It estimates the probability that a user query belongs to a specific intent based on input features [26].

Support vector machine (SVM)

Support Vector Machines (SVMs) are effective supervised learning techniques often used for classification and regression. They work by finding the most suitable boundary called a hyperplane, that creates the widest possible separation between different categories of data [27].

Performance evaluation & results

After preprocessing the dataset and applying TF-IDF for feature extraction, we trained and evaluated the models in two different scenarios: (1) without applying the SMOTE technique, and (2) after applying SMOTE to balance the class distribution. This comparison aimed to assess how well each model performs in both imbalanced and balanced data settings.

To begin the evaluation process, the dataset was split into training and testing sets by allocating 80% of the data for training the model and reserved 20% for testing. using the following code.

X_train, X_test, y_train, y_test = train_test_split(X_vect, y, test_size = 0.2, random_state = 42)

The results show that Random Forest works very well with imbalanced data, reaching almost perfect accuracy even without using SMOTE, this result means it can handle unbalanced datasets better than the other models, as shown in Table 5.

On the other hand, Naive Bayes performs poorly with imbalanced data, so SMOTE is more helpful for that model. Fig 2 illustrates how each model's accuracy changed before and after balancing the dataset. It shows the Comparison of model accuracy before and after applying the (SMOTE) across four classification algorithms: Naive Bayes, Random Forest, Logistic Regression, and Support Vector Machine (SVM). The x-axis represents the machine learning models, while the y-axis shows accuracy scores ranging from 0.85 to 1.00.

Table 5 Comparison of model accuracy before and after applying SMOTE for class balancing

Model	Accuracy Before SMOTE	Accuracy after SMOTE
Naive Bayes	0.89	0.9529
Random Forest	1	0.9996
Logistic Regression	1	0.9924
Support Vector Machines (SVM)	0.997	0.9998

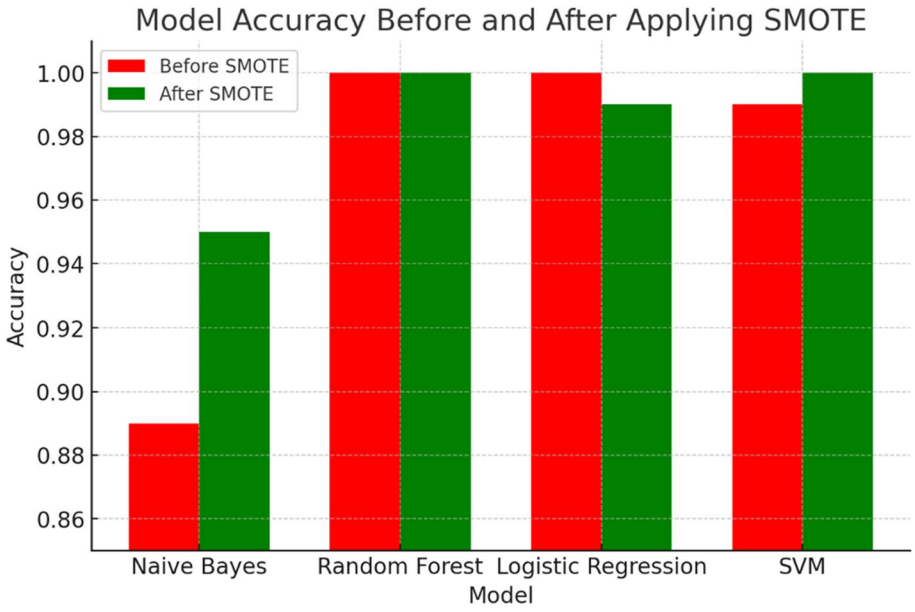

Fig. 2 The model's accuracy before and after balancing the dataset

Table 6 The model accuracy after applying 5-fold cross-validation

Model	Average Accuracy for 5-fold cross-validation	Standard Deviation
Naive Bayes	0.8798	0.0021
Random Forest	0.9936	0.0005
Logistic Regression	0.9931	0.0005
Support Vector Machines (SVM)	0.9929	0.0004

In our study, Random Forest performed very well even before applying SMOTE, showing that it handles this type of data robustly. We made sure that there was no data leakage. To check the reliability of our results, we used 5-fold cross-validation: the dataset was split into five parts, and each model was trained and tested five times on different splits. The accuracy scores were very similar across all folds, which shows that the models are stable and consistent. We also calculated the mean accuracy and standard deviation for each model, as shown in Table 6. These results come from training and validation, but we expect the model to perform similarly on new, unseen data.

Fine-tuned BERT model

Bidirectional Encoder Representations from Transformer) BERT (processes user queries by breaking them into tokens (words or sub-words) and looks at the full sentence from both directions to understand the meaning. Its self-attention mechanism helps the model focus on the most important words related to the user's intent. This makes BERT very effective for intent recognition, as it captures the context and meaning of each word in the query [28]. However, BERT has inspired numerous researchers to fine-tune the model for various domain-specific applications, such as banking system [29], education [30] and healthcare [31] due to its strong contextual understanding and transfer learning capabilities.

In general, fine-tuning BERT does not strictly require a balanced dataset because it leverages powerful pretrained language representations that effectively capture

contextual meaning. In this study, we used a small learning rate (0.00002) to carefully adjust the model's weights during training, processing 8 samples per training step. Table 7 lists the key parameters used for fine-tuning the BERT model. Notably, the model was initially trained on only 20% of the full dataset (equivalent to 0.2 epochs), which allowed for faster testing but resulted in moderate accuracy. When the fine-tuning was increased to one full epoch, the accuracy improved from 50% to 60%, demonstrating that longer training enables the model to learn better representations despite the limited dataset size.

For fine-tuning BERT, we set a few key training parameters:

1 Learning_rate: Set to 2e-5 (0.00002) to make small, careful updates during training.
2 Batch size: We used 8 samples per batch during training and evaluation to balance memory use and speed.
3 Num_train_epochs: Initially set to 0.2 epochs, increasing this to 1 epoch improved accuracy from 50% to 60%. More epochs help the model learn better but require more computing power.
4 Evaluation_strategy and save_strategy: The model was evaluated and saved after each epoch to monitor progress and keep the best version.
5 Load_best_model_at_end: Automatically loads the best-performing model at the end of training for final evaluation.

We measured performance using accuracy, comparing predicted labels to true labels. This setup allowed effective fine-tuning of BERT within our hardware limits and showed improvement when training duration was increased.

Running the model

The trained models, both traditional machine learning and the fine-tuned BERT, process user-input questions by first converting them into suitable numerical features (e.g., TF-IDF vectors for ML models or token embeddings for BERT). These inputs are then passed to the model, which instantly predicts the corresponding class label. This enables real-time and accurate classification of new and unseen medical questions.

Sample Inputs and Predicted Outputs:

1. *Input*: What is the best medicine for controlling high blood pressure?

   Predicted Label: ['information']

2. *Input*: How can I protect myself from high blood pressure?

   Predicted Label: ['prevention']

3. *Input*: How many tablets from Panadol should I take for my stomach pain?

Table 7 Lists the parameters used for fine-tuning the BERT model

Parameter	Value	Description
eval_loss	0.2614	The loss on the evaluation set
eval_accuracy	98.39%	Model accuracy on training the data
eval_runtime	740.65 s	Total time taken to fine tune the data
eval_samples_per_second	4.43	The number of samples were evaluated per second
eval_steps_per_second	0.555	The number of steps ran per second
epoch	0.2, then increased to 1	Training ran first for 20% of one full epoch, then full epoch

Table 8 Comparison of average processing time per question across various models

Model	SVM	RF	Naive Bayes	Logistic	BERT (0.2)	BERT(1 epoch)
Time/Seconds	10	40	1	3	3,480	11,609

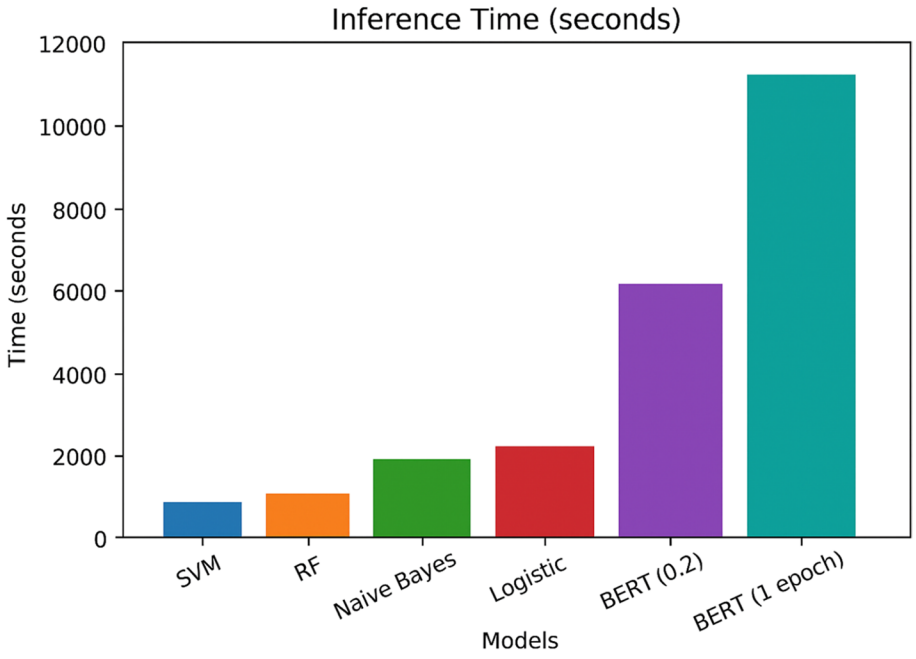

Fig. 3 Inference time (seconds) per question for five models: Naive Bayes, Logistic Regression, SVM, Random Forest, and BERT

Predicted Label: ['frequency']

Results and discussions

All model training and evaluation were conducted on a personal computer equipped with an Intel(R) Core(TM) i7-1065G7 CPU @ 1.30 GHz and no dedicated GPU. Due to these hardware limitations, fine-tuning deep learning models such as BERT was restricted to one (1) epoch to avoid memory overflow errors. However, all models demonstrated satisfactory accuracy during evaluation and, more importantly, showed good generalization on entirely new questions. While the processing time varied across models, Table 8 presents the time each model required to handle individual questions.

The bar chart below in Fig. 3 compares the inference time (in seconds) required by five different machine learning models to process a single question. The models include Naive Bayes, Logistic Regression, Support vector Machine (SVM), Random Forest, and BERT. The x-axis represents the machine learning models, while the y-axis shows the time in seconds.

Traditional models like Naive Bayes and Logistic Regression are highly efficient in terms of speed, making them well-suited for real-time or low-resource environments. In contrast, BERT, while offering superior accuracy and deep contextual understanding, has a significantly higher inference time. This makes it less practical for fast-response applications unless it is optimized or deployed on high-performance hardware. However, performance gains are clearly evident when increasing the number of epochs during fine-tuning, as shown in Table 9.

However, after testing 10 medical questions, the results show that Random Forest handles imbalanced data effectively, achieving high accuracy without the need for SMOTE, while Naive Bayes benefits significantly from data balancing techniques. The trained models, both traditional machine learning and the fine-tuned BERT, process user-input questions by first converting them into suitable numerical features (e.g., TF-IDF vectors for ML models or token embeddings for BERT).

All models achieved over 60% accuracy on a 10-question medical test set. Random Forest performed best with 80% accuracy, followed by Support Vector Machine (SVM) and Logistic Regression at 70%. BERT, by contrast, achieved only 60% accuracy, misclassifying several key queries as depicted in Table 10. This lower performance is likely due to limited fine-tuning.

Furthermore, the result indicates that the models did not show signs of overfitting, which occurs when a machine learning model learns both true patterns and noise from the training data, resulting in poor performance on new and unseen inputs. Since, avoiding overfitting enhances the model's ability to generalize effectively [32]. However, increasing BERT's fine-tuning to one (1) epoch led to an improvement in accuracy, reaching 60%, suggesting that the model has the potential to perform better with additional training. Fig. 4 illustrates the comparative performance of these models. Unfortunately, further fine-tuning was not possible due to hardware constraints, specifically, limited RAM and processing power, which caused memory overflow errors when training exceeded one epoch.

Limitations

While the results of this study are promising, there are a few limitations. For instance, fine-tuning beyond a single epoch was not possible due to hardware constraints (limited RAM and processing power), which led to memory overflow errors. In addition, although the dataset was enough for our experiments, its size may limit how well the results apply to other situations or larger datasets.

Conclusion and future work

This study demonstrates how technological principles and intelligent computational models can be applied to advance healthcare technologies. Recognizing user intent is a critical component in developing effective health-related question-answering (QA) systems, as it enables the system to accurately interpret queries and deliver relevant and meaningful responses. The success of AI models in healthcare depends on understanding the purpose behind each query, a process known as Natural Language Understanding (NLU). In this study, we developed a computational model to classify user intent in medical QA tasks, evaluating multiple classical machine learning algorithms alongside the transformer-based BERT model using Python.

The MedQuad dataset, comprising 14,979 patient-generated medical questions and 16,407 expert-provided answers across 15 categories, was used for training and

Table 9 BERT model performance comparison by epoch

Epochs	Training Loss	Validation Loss	Accuracy (%)
0.2	0.357800	0.302671	0.967087
1	0.0106	0.0079	99.9

Table 10 Summary of the performance of traditional machine learning models and the BERT model

	Medical Question	Correct Label	SVM	RF	Naive Bayes	Logistic	BERT (0.2 Epochs)	BERT (1 Epochs)
1	Why do I have stomach pain?	Information	Information	Information	Information	Information	Exams and tests	information
2	How long does Panadol take to work?	Exams and Tests	Exams and tests	Exams and tests	Exams and tests	Exams and tests	Frequency	frequency
3	Can I give Panadol to babies?	Considerations (or Prevention)	Prevention	Prevention	Prevention	Prevention	Exams and tests	exams and tests
4	Is cold medicine safe when pregnant?	Considerations	Information	Inheritance	Information	Information	Inheritance	inheritance
5	Are flu complications dangerous?	Complications	Complications	Complications	Complications	Complications	Inheritance	inheritance
6	What causes iron deficiency?	Causes	Causes	Causes	Causes	Causes	Causes	Causes
7	What are signs of low iron?	Symptoms	Information	Information	Stages	Information	Symptoms	symptoms
8	Is low iron genetic?	Inheritance	Information	Inheritance	Information	Information	Inheritance	inheritance
9	How often to check iron levels?	Frequency/ Exams and tests	Exams and tests	Exams and tests	Prevention	Exams and tests	Frequency	Frequency
10	How often check blood pressure?	Frequency/ Exams and tests	Exams and tests	Exams and tests	Prevention	Exams and tests	Frequency	Frequency
	Average of Correct Answers		7/10	8/10	5/10	7/0	5/10	6/10

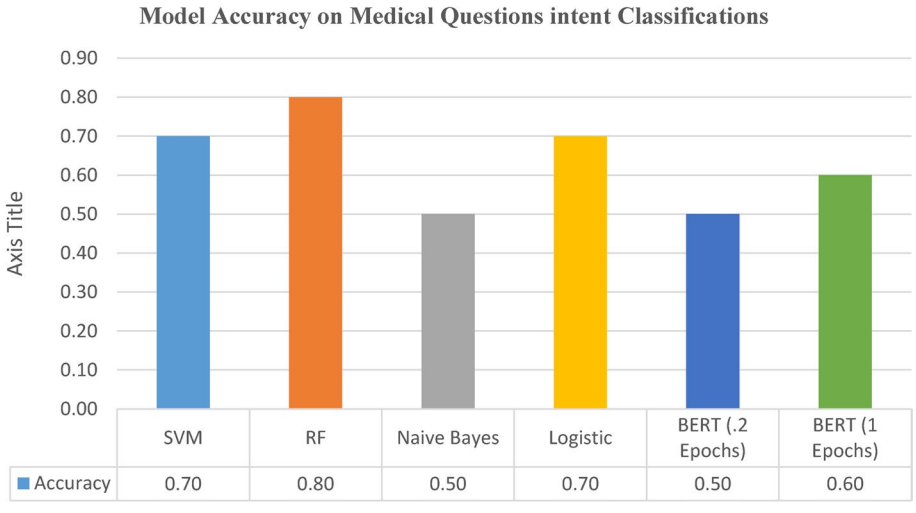

Fig. 4 Intent classification accuracy for medical questions using the model

evaluation. To address class imbalance, the SMOTE technique was applied, improving the model's ability to learn from underrepresented categories and enhancing generalization.

Evaluation results indicate that all models generalized well to unseen questions. Random Forest achieved the highest accuracy (80%) without data balancing, while Naive Bayes improved significantly with SMOTE. Traditional models like Naive Bayes and Logistic Regression also demonstrated fast inference times, making them suitable for real-time and resource-constrained healthcare environments. BERT, although initially undertrained, improved from 50% to 60% accuracy after one training epoch but remains computationally expensive, highlighting the trade-offs between model complexity, performance, and resource requirements. From an engineering and AI science perspective, these findings underscore the importance of computational modeling, algorithm optimization, and efficient implementation in designing human-centered healthcare AI systems.

In future will focus on fine-tuning BERT with additional training epochs to capture more complex medical language patterns, supported by enhanced computational infrastructure, including increased GPU capacity. Furthermore, we plan to evaluate all models not only for accuracy but also for sustainability metrics such as energy consumption and computational efficiency. Assessing these factors will provide a more comprehensive understanding of each model's practicality, environmental impact, and suitability for deployment in real-world healthcare systems, aligning with the broader goals of engineering innovation and applied science in addressing global healthcare challenges.

Abbreviations

AI	Artificial intelligence
BERT	Bidirectional encoder representations from transformers
BiLSTM	Bidirectional long short-term memory
GPT	Generative pre-trained transformer
GPU	Graphics processing unit
ML	Machine learning
NLP	Natural language processing
NLU	Natural language understanding
RF	Random forest
SMOTE	Synthetic minority oversampling technique
SVM	Support vector machine
TF-IDF	Term frequency–inverse document frequency

Acknowledgements
I gratefully acknowledge the guidance and support of my research supervisors, Chin-Shiuh Shieh and Shashi Kant Gupta, throughout the preparation of this review article.

Authors' contributions
S. A. wrote the main manuscript and led the ML algorithms. C. S. and S. G. reviewed and analyzed the results. S. A. and C. S reviewed the paper.
They have all read and approved the final study.

Funding
The authors did not receive financial support from any organization for the submitted work.

Data availability
The data used in this study is available upon request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 30 August 2025 / Accepted: 4 November 2025
Published online: 18 November 2025

References

1. Nahavandi S (2019) Industry 5.0—A human-centric solution. Sustainability 11(16):4371. https://doi.org/10.3390/su11164371
2. Basulo-Ribeiro J, Teixeira L (2024) The future of healthcare with industry 5.0: preliminary interview-based qualitative analysis. Future Internet 16(3):68. https://doi.org/10.3390/fi16030068
3. Sarella PNK, Mangam VT (2024) AI-driven natural Language processing in healthcare: transforming patient-provider communication. Indian J Pharm Pract, *17*(1) https://doi.org/10.5530/ijopp.17.1.4.
4. Assayed SK, Alkhatib M, Shaalan K (2024) Enhancing Student Services: Machine Learning Chatbot Intent Recognition for High School Inquiries. In: Al Marri, K., Mir, F.A., David, S.A., Al-Emran, M. (eds) BUiD Doctoral Research Conference 2023. Lecture Notes in Civil Engineering, vol 473. Springer, Cham. https://doi.org/10.1007/978-3-031-56121-4_24
5. Chandrakala CB, Bhardwaj R, Pujari C (2024) An intent recognition pipeline for conversational AI. Int J Inform Technol 16(2):731–743
6. Assayed SK, Shaalan K, Alkhatib MA, Chatbot Intent Classifier for Supporting High School Students (2022) Dec. EAI Endorsed Scal Inf Syst [Internet]. 21 [cited 2025 Aug. 28];10(3):e1. Available from: https://publications.eai.eu/index.php/sis/article/view/2948
7. Bouguelia S (2023) *Dialogue Patterns and Composite Intents Recognition in Task-oriented Human-Chatbot Conversations* (Doctoral dissertation, Université Claude Bernard-Lyon I) https://theses.hal.science/tel-04586537v1.
8. Jbene M, Chehri A, Saadane R, Tigani S, Jeon G (2024) Intent detection for task-oriented conversational agents: A comparative study of recurrent neural networks and transformer models. Expert Syst. Advance online publication https://doi.org/10.1111/exsy.13712
9. Tohti T, Abdurxit M, Hamdulla A (2022) Medical QA oriented multi-task learning model for question intent classification and named entity recognition. Information 13(12):581. https://doi.org/10.3390/info13120581
10. Turchin A, Masharsky S, Zitnik M (2023) Comparison of BERT implementations for natural Language processing of narrative medical documents. Inf Med Unlocked 36:101139
11. Lee LH, Lu Y, Chen PH, Lee PL, Shyu KK (2019), August NCUEE at MEDIQA 2019: medical text inference using ensemble BERT-BiLSTM-attention model. In *Proceedings of the 18th BioNLP workshop and shared task* (pp. 528–532) https://doi.org/10.18653/v1/W19-5058.
12. Tresner-Kirsch D, Mikkelson AA, Yinka-Banjo C, Akinyemi M, Goyal S (2023), June Intent Recognition on Low-Resource Language Messages in a Health Marketplace Chatbot. In 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI) (pp. 457–459). IEEE. https://doi.org/10.1109/ICHI57859.2023.00066
13. Guan F, Tezuka T (2022), December A medical Q&A system with entity linking and intent recognition. In *2022 IEEE symposium series on computational intelligence (SSCI)* (pp. 820–829). IEEE https://doi.org/10.1109/SSCI51031.2022.10022093.
14. Shi C, Chu C, Su J A Study on Question and Answer Intent Recognition in Medical Domain Based on Prompt Learning, *2024 Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC)*, Macao, China, 2024, pp. 1–3. https://doi.org/10.1109/CSRSWTC64338.2024.10811650
15. Luo Y, Xie Y, Yan E, Lee LK, Wang FL, Hao T (2023), July A User Intent Recognition Model for Medical Queries Based on Attentional Interaction and Focal Loss Boost. In *International Conference on Neural Computing for Advanced Applications* (pp. 245–259). Singapore: Springer Nature Singapore https://doi.org/10.1007/978-981-99-5847-4_18.
16. Gomathi L, Mishra AK, Tyagi AK (2023), April Industry 5.0 for healthcare 5.0: Opportunities, challenges and future research possibilities. In *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 204–213). IEEE https://doi.org/10.1109/ICOEI56765.2023.10125660.
17. Karatas M, Eriskin L, Deveci M, Pamucar D, Garg H (2022) Big data for healthcare industry 4.0: Applications, challenges and future perspectives. Expert Syst Appl 200:116912. https://doi.org/10.1016/j.eswa.2022.116912
18. Devlin J, Chang M-W, Lee K, Toutanova K (2019) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT), (pp. 4171–4186). https://arxiv.org/abs/1810.0480
19. The Devastator (n.d.). *Comprehensive Medical Q&A Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/thedevastator/comprehensive-medical-q-a-dataset
20. Susan S, Kumar A (2021) The balancing trick: optimized sampling of imbalanced datasets—A brief survey of the recent state of the Art. Eng Rep 3(4):e12298. https://doi.org/10.1002/eng2.12298
21. Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya INH, Ismail M (2021), November SMOTE for handling imbalanced data problem: A review. In *2021 sixth international conference on informatics and computing (ICIC)* (pp. 1–8). IEEE https://doi.org/10.1109/ICIC54025.2021.9632912.
22. Ghosh K, Bellinger C, Corizzo R, Branco P, Krawczyk B, Japkowicz N (2024) The class imbalance problem in deep learning. Mach Learn 113(7):4845–4901. https://doi.org/10.1007/s10994-024-06542-1
23. 20, Reddy EMK, Gurrala A, Hasitha VB, Kumar KVR (2022) Introduction to Naive Bayes and a review on its subtypes with applications. Bayesian Reasoning Gaussian Processes Mach Learn Appl, 1–14 https://doi.org/10.1201/9781003164265-1.
24. Wickramasinghe I, Kalutarage H (2021) Naive bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. Soft Comput 25(3):2277–2293. https://doi.org/10.1007/s00500-020-05277-5
25. Magidi J, Nhamo L, Mpandeli S, Mabhaudhi T (2021) Application of the random forest classifier to map irrigated areas using Google Earth engine. Remote Sens 13(5):876. https://doi.org/10.3390/rs13050876
26. Schober P, Vetter TR (2021) Logistic regression in medical research. Anesth Analgesia 132(2):365–366
27. Jabardi M (2025) Support vector machines: Theory, Algorithms, and applications. Infocommunications J, *17*(1) https://doi.org/10.36244/ICJ.2025.1.8.
28. Fernández-Martínez F, Luna-Jiménez C, Kleinlein R, Griol D, Callejas Z, Montero JM (2022) Fine-tuning BERT models for intent recognition using a frequency cut-off strategy for domain-specific vocabulary extension. Appl Sci 12(3):1610. https://doi.org/10.3390/app12031610
29. Wang X, Cao J (2024), January A BERT-Based Knowledge Selection Model for Bank Regulatory Reporting in Conversational Systems. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering* (pp. 436–440) https://doi.org/10.1145/3672758.3672829.
30. Assayed SK, Alkhatib M, Shaalan K (2024) A Transformer-Based generative AI model in education: Fine-Tuning BERT for Domain-Specific in student advising. In: Basiouni A, Frasson C (eds) Breaking barriers with generative Intelligence. Using

GI to improve human education and Well-Being. BBGI 2024. Communications in Computer and Information Science, vol 2162. Springer, Cham. https://doi.org/10.1007/978-3-031-65996-6_14
31. Babu A, Boddu SB (2024) BERT-based medical chatbot: enhancing healthcare communication through natural Language Understanding. Exploratory Res Clin Social Pharm 13:100419. https://doi.org/10.1016/j.rcsop.2024.100419
32. Montesinos López OA, Montesinos López A, Crossa J (2022) Overfitting, model Tuning, and evaluation of prediction performance. Multivariate statistical machine learning methods for genomic prediction. Springer, Cham. https://doi.org/10.1007/978-3-030-89010-0_4

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Suha Khalil Assayed is a researcher specializing in artificial intelligence, natural language processing, and deep learning, with a particular focus on the development of intelligent chatbots for education and healthcare. She has extensive professional experience in higher education, institutional research, and academic advising, bringing more than 20 years of expertise in guiding and supporting student success initiatives. Her research bridges advanced machine learning techniques with real-world applications, aiming to create human-centered, technology-driven solutions that enhance learning and well-being.

Chin-Shiuh Shieh received the M.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1991, and the Ph.D. degree from the Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, in 2009. He joined as a Faculty Member of the Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, in August 1991, where he is currently a Professor. His research interests include wireless networks and handover techniques.

Shashi Kant Gupta is a Post-Doctoral Fellow and Researcher in Computer Science and Engineering at Eudoxia Research University, USA, and also serves in honorary academic roles at institutions in Nigeria, Uzbekistan, and Taiwan. He earned his Ph.D. in CSE from Integral University, India, and has held various teaching positions including Assistant and Associate Professor in reputed universities. Currently, he is the Founder and CEO of CREP Pvt. Ltd., Lucknow, and actively contributes as Editor-in-Chief and Senior Editor in multiple international journals. His research interests include cloud computing, big data analytics, IoT, and computational intelligence in education. With over 12 years of teaching, 2 years of industry, and extensive editorial and publishing experience, he has authored numerous papers, books, and patents while receiving recognition from global organizations.