Medical Chatbot System Using Natural Language Processing

Anima Dahal
animadahal72@gmail.com
DOI : https://doi.org/10.3126/dmcj.v10i9.90605

Abstract

Access to timely and reliable healthcare is one of the primary issues in Nepal because of the challenging topography, lack of medical facilities, and the absence of healthcare professionals, especially in rural and remote areas. In order to overcome these issues, this research project is going to present a Medical Chatbot System, which is based on NLP and which is aimed at the provision of the initial disease diagnosis and situation-specific medical advice based on the symptoms described by the user. The system combines a fine-tuned BERT model to predict multiple labels of a disease with a Large Language Model (LLM) to be able to engage in natural dialogue and ask follow-ups. The system was trained and tested on a secondary medical dataset, which included descriptions of symptoms and disease names. The results of the experiment show that the proposed chatbot achieved an accuracy of 91.53%, as well as high precision and recall, which is evidence of reliable performance in terms of disease prediction. Safety mechanisms such as red-flag detection and fallback responses were incorporated to encourage responsible use and timely medical consultation. The system is launched as a web-based interface so that it is easy to use by people with different degrees of digital literacy, and it is supposed to be used during early health evaluation and not to substitute professional medical diagnosis. In general, the results imply the potential of smart conversational agents to enhance the accessibility of healthcare, lessen unneeded hospital admissions, and facilitate early medical care in resource-limited countries, like Nepal.

Keywords:

Medical Chatbot, BERT, Natural Language Processing, Large Language Model, Disease Prediction.

Introduction

Provision of quality and timely healthcare has been a challenge in Nepal, especially in the rural and other geographically remote areas where there is a shortage of medical facilities and professionals. Access to professional consultation is usually a challenge that compels the person to find alternative means to self-diagnose, like searching the internet or using home remedies. This enhances the chances of misinformation, delayed medical care, and unwarranted complications. Since Nepal is still experiencing gaps in health literacy and unequal distribution of healthcare facilities, the need to have easy-to-understand and accessible, and instant medical information has become increasingly urgent (PHC-Nepal, 2024).

To address these concerns, this study will suggest a Medical Chatbot System using a Natural Language Processing (NLP) algorithm that will help to perform a preliminary

assessment of the symptoms and to give credible information regarding health-related issues conversationally. This system uses the most updated language models to understand the symptoms reported by users, calculate the concerned patterns, and categorise the possible medical diagnoses. Then it provides the appropriate guidance, like self-care or advice to see a medical aid. The chatbot will decrease the number of unnecessary visits to the hospital and promote the timely health intervention through experience and exposure to various symptom descriptions and the generation of situation-specific responses.

The proposed solution is designed to be scalable and accessible: the system is operated via a convenient web interface that can be used by people of different levels of digital literacy, starting with urban inhabitants and including underserved rural groups. The automation of the preliminary assessment process will likely decrease the workload of Nepal's healthcare infrastructure, enhance the quality of the medical service, and make valuable medical information more accessible to the population.

In this paper, the design, development, and performance evaluation of the NLP-based Medical Chatbot System that combines fine-tuned BERT classification with an LLM-driven conversational module are presented. Section 3 shows the objectives of this research paper. Section 4 reviews existing AI-driven healthcare support tools and symptom-analysis systems. Section 5 explains the proposed methodology, including dataset preparation, preprocessing, and model architecture. Section 6 outlines the different equation. Section 7 presents experimental results and Section 8 presents the discussion. Finally, Section 9 concludes the study and provides possible directions for future improvements.

Research Objectives

The objectives of this study are:

1) To design and develop a highly accurate NLP-based medical chatbot for initial disease prediction using user-reported symptoms.
2) To fine-tune a BERT model for multi-label classification of medical conditions.
3) To integrate a Large Language Model (LLM) for generating contextually relevant follow-up questions and responses.
4) To assess the system's performance, reliability, and safety using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
5) To evaluate the suitability of the system in enhancing healthcare access in Nepal.

Literature Review

The development of Medical Chatbot Systems using Natural Language Processing has been explored through various approaches in recent literature. Multiple experiments have investigated how the combination of advanced AI models can enhance the accuracy of the diagnosis, the quality of conversation, and the accessibility of healthcare. This section reviews key research contributions that have informed the design and development of intelligent medical chatbots.

AI-Powered Healthcare Chatbots

Mhatre et al. (2024) explored the implementation of Large Language Models in healthcare chatbots aimed at addressing general illness inquiries. Their system used

LangChain architecture, MongoDB to handle data, and Retrieval-Augmented Generation (RAG) methods. The chatbot was based on text mining of medical records, smart chunking of data, storage using a vector database, and search algorithms based on the cosine similarity to align the query of the user with related medical knowledge. The data used in training came as the reference of reputable medical bodies such as the Indian Council of Medical Research and AIIMS. The system in question demonstrated a rate of accuracy that was over 70 percent when working with the MedMCQA test data that consisted of more than 194,000 questions related to a medical examination and covered 2,400 health topics. Although the authors have already pointed out such a potential in providing evidence-based and personalized healthcare advice, such issues as the reduction of algorithmic bias, the enhancement of the accuracy of advice in more complicated medical cases, and the ethics of AI-driven healthcare applications were highlighted.

Pharma LLM is an open-source Large Language Model prescriber chatbot created by Azam et al. to prescribe medicine (Azam et al., 2024). The study used the publicly available Kaggle EDA Medicine Dataset that had 11,000 records, but it was narrowed down to 7,515 records through an intensive data cleaning. The parameter-efficient method of Low-Rank Adaptation (LoRA) was used to fine-tune the Tiny Llama open-source model as the core system and achieve high performance with low computational costs. The training had been done in Google Colab GPU environment using optimized hyperparameters. The architecture also used improved transformer blocks such as RMS Norm, SwiGLU, RoPE, GQA, and Flash Attention to improve efficiency and contextual understanding. The uniqueness was the incorporation of speech-to-text and text-to-speech features, which ensured inclusivity to users with low levels of literacy or disability. The overall analysis produced some impressive results: 87.69 percent accuracy, 90.38 percent precision, 94.00 percent recall, and 92.16 percent F1-score. Regardless of these successes, the study recognized the limitations, such as bias in the data, limited scope of data, and the fact that further validation with the expert medical practice is required before the real-world application.

BERT-Based Medical Systems

Significant research has been published on Exploratory Research in Clinical and Social Pharmacy, introducing a BERT-based medical chatbot to improve healthcare communication by means of deep Natural Language Understanding (Babu & Boddu, 2024). The paper has discussed the weaknesses of existing medical chatbots, which often have limited contextual knowledge, fixed response structures, and flawed interpretation of medical terminologies. The solution aimed at the use of bidirectional context modelling through BERT to distinguish medical terminologies and handle multi-turn and complex conversation intonations. The training also used domain-specific datasets such as MIMIC-III, PubMed, BioASQ, and COVID-19 medical data, which allow subtleties of understanding healthcare-related queries. The deployment was conducted in several steps: extensive data cleaning with the help of tokenization, lemmatization, and vectorization; query interpretation with entity and intent recognition; and context management to ensure conversations remain on track. The Hugging Facebook Transformers library on Tensor Flow and spaCy was used as model fine-tuning and evaluated by using accuracy, precision, recall, F1-score, and AUC-ROC as performance metrics. The performance of the BERT-

based chatbot was quite impressive: 98% accuracy, 97% precision, and 97% AUC-ROC, which is far greater than that of the traditional models, such as LSTM, Bi-LSTM, and SVM. These findings revealed that BERT is more effective in interpreting the intention of the users and medical terminology to deliver more accurate and human-like responses.

Framework-Based Healthcare Chatbots

Ganguly (2023) introduced some practical implementations of a healthcare chatbot known as RISA based on the open-source RASA framework. This project showed how conversational AI systems can be utilized within healthcare settings to analyze symptoms and detect the location of a healthcare facility. Its implementation was based on the dual-component architecture of RASA, which includes Natural Language Understanding (NLU) and Dialogue Management components. The NLU component was used to process user inputs by tokenization, removal of stop-words, and elimination of punctuations, and then it was vectorized with Count Vectorizer. The similarity algorithms were based on cosine, and the user input was compared with medical data to get the related disease or medicine information. The Dialogue Management module made use of configuration files that contained intents, entities, response templates, and conversational flows. One of the strong points was the ability to integrate with external sources of data and APIs, as the Wikipedia API was used to obtain medical definitions, and Folium mapping was used with geolocation data to dynamically show nearby hospitals. HTML, CSS, and JavaScript were used to create the frontend, and Flask was used to communicate on the server-side. Nonetheless, the project admitted a range of weaknesses: the reliance on the quality of the symptom datasets to obtain the accuracy, the simplicity of the used vectorization strategy that prevents the comprehension of complex queries, the lack of evaluation of the standard performance metrics, and the lack of ethical and safety concerns when providing medical advice.

Synthesis and Research Gap

As the reviewed literature shows, there have been different methods for the development of medical chatbots, each with its own strengths and limitations. The systems based on LLM promise improvement in the quality of the conversation, but in most cases, they do not provide the accuracy required to diagnose the disease correctly. BERT-based systems are the best in terms of classification, but need to be combined with generative models to be able to speak to users naturally. Framework-based methods give more realistic implementations but can be less advanced in understanding language, like transformer-based models. This study fills these gaps by integrating the analytical accuracy of fine-tuned BERT classification with the converse intelligence of Large Language Models, forming a holistic system that is able to provide both reliable diagnosis and natural and occurrence-based dialogue. Moreover, the unique emphasis on the Nepal healthcare environment with references to the policies of diverse levels of digital literacy and resource-constrained environments is a major contribution to the accessibility of AI-driven healthcare support to developing areas.

Research Methodology

The medical chatbot system development was completed in a systematic way through integration of both Natural Language Processing (NLP) and deep learning to give accurate disease prediction and interactive health advice. The system is a combination of a BERT-

based symptom analysis classifier and a Large Language Model (LLM), which is Mistral, to produce conversational intelligence, wherein it can generate follow-up questions and self-care tips depending on the context.

Research Design

This study embraces the experimental research design to develop and test a Natural Language Processing (NLP)-based medical chatbot to predict diseases and provide healthcare advice in its initial version.

Research Data Source

In this research, secondary medical data were utilized. The dataset was retrieved from an openly available Kaggle source (Saleem, 2022), which contains the symptom descriptions, disease names, and corresponding medical responses. The dataset includes attributes such as short questions, short answers, medical tags, and disease labels, covering a variety of health conditions.

Data Preprocessing

The data was well preprocessed to ensure the quality of the data and model. This preprocessing included:

1. Text Cleaning: Removal of duplicate and irrelevant information, removing special characters and punctuations, converting all the text to lower case, and correcting spelling mistakes.
2. Tokenization: Breaking down text into smaller units (tokens) to make it understandable for the model.
3. Stop-word Removal: The NLTK library is used to delete common words (e.g., "is," "a," the) which do not have medical significance.
4. Lemmatization: he process of changing words into their base forms (e.g. running - run) to consider the variants of the word as one concept.
5. Label Encoding: This is the process of converting categorical disease labels to a numerical form to be used as a model training feature.

These preprocessing operations were performed using tools like Pandas, NumPy and NLTK.

Dataset Splitting

After preprocessing, stratified sampling was used to split the data into training (80%) and testing (20%) datasets to maintain balance across disease categories. The method allowed fair assessment of the model performance without the problem of class imbalance.

Model Development

The stage of model development involved making BERT (Bidirectional Encoder Representations from Transformers), a deep learning model that can comprehend contextual relationships in medical text, finer. BERT was used to train the multi-label disease classification, which offered the possibility to identify multiple possible conditions, given the same symptoms. The fine-tuning step was done on the Hugging Face Transformers library in a Google Colab GPU environment to bring about computational efficiency. Some of the main training parameters were:

- Number of training epochs: 12
- Batch size: 8

- Learning rate: 3e-5
- Optimizer: Adam

The BERT output layer was adapted to enable multi-label classification, which is predicting multiple diseases at the input of a single symptom. The adaptations included a sigmoid activation function.

Integration of Large Language Model

After training the BERT classifier, it was integrated with a Large Language Model-Mistral (LLM), which enhanced the capabilities of the chatbot to be more conversational. The LLM generates situational and human-like responses, such as follow-up queries and clarifications, thus making the chatbot more accurate and interactive in diagnosis. This BERT-LLM hybrid system enabled the system to integrate the analytical disease prediction and natural conversational flow.

Model Evaluation

The model performance was evaluated with common metrics computed from the scikit-learn library:

a) **Accuracy**: The overall correctness of predictions
b) **Precision**: The fraction of actual positives among those that we predicted as positive
c) **Recall**: The ratio of the true positives to all the relevant results
d) **F1-score**: The harmonic mean of precision and recall

Based on the results of the evaluation, fine-tuning was implemented to enhance performance. The red-flag detection was also part of the system, through which it identified severe symptoms (like chest pain or difficulty breathing) and prompted users toseek urgent medical help (Kavyasrirelangi, 2025).

System Deployment

After its validation, the chatbot was deployed in a web interface developed in Streamlit, where the user was allowed to enter the symptoms and receive instant predictions and personalized advice. It was deployed using the Iterative Model of Software. Development Life Cycle(SDLC), which enabled a gradual refining of it on successive cycles of requirement analysis, design, implementation, testing, deployment, review, and maintenance.

The functional requirements were based on disease prediction, follow-up question generation, and user-friendly interaction, whereas non-functional requirements were on real-time responsiveness, reliability, privacy, and scalability. This structured methodology ensured that the Medical Chatbot System effectively leveraged NLP and deep learning to provide intelligent, context-sensitive, and accessible medical support to users, where continuous implication is possible through periodical updates and user feedback.

Equations

The Medical Chatbot System applies some of the most vital mathematical and assessment equations that allow to give precise disease prediction and ensure the quality of performance. These equations are primarily used to evaluate the BERT-based classifier and its multi-label predictions, which are crucial for assessing the effectiveness of the system.

Accuracy

Accuracy is used to assess the general correctness of the model by assessing the percentage of all correct predictions (positive and negative) of the total predictions. It is an overall measure of model performance, but it can be deceptive when dealing with imbalanced data.

Equation 1

Accuracy

$$Accuracy = \frac{\mathbf{TP + TN}}{\mathbf{(TP + TN + FP + FN)}}$$

where:

- TP (True Positive): Correctly predicted presence of a disease.
- TN (True Negative): Correctly predicted absence of a disease.
- FP (False Positive): Incorrectly predicted disease when it was not present.
- FN (False Negative): Failed to predict a disease that was actually present.

a. **Precision**

Precision is used to measure the ratio of the number of correct cases of positive prediction to the number of cases that are predicted to be positive. The accuracy is high, so there will be fewer false alarms; this is very important in healthcare since it does not cause panic.

Equation 2: Precision

$$Precision = \frac{\mathbf{TP}}{\mathbf{TP + FP}}$$

Recall (Sensitivity)

Recall, also referred to as sensitivity, is a computational determination of the percentage of real cases that are correctly identified by the model. High recall is used to make sure that the majority of actual disease cases are identified, and there is a decreased probability of missing critical disease cases.

Equation 3

Recall

$$Recall = \frac{\mathbf{TP}}{\mathbf{(TP + FN)}}$$

F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure when both false positives and false negatives are important. It is particularly useful for imbalanced datasets, where some diseases occur less frequently.

Equation 4

F1-Score

$$F1\ score = 2 \times \frac{\mathbf{Precision \times Recall}}{\mathbf{Precision + Recall}}$$

Multi-Label Sigmoid Activation (BERT Output Layer)

For multi-label classification, the BERT model uses a sigmoid activation function to predict the probability of each disease independently. This allows the model to assign multiple disease labels for a single symptom input (geeksforgeeks, 2025).

Equation 5

Sigmoid Activation Function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Explanation:

- **x:** Input to sigmoid function (logits from BERT output layer)
- **σ(x):** Probability that a particular disease is present.
- The values range within [0, 1] and a threshold (usually 0.5) is adopted to determine if the disease is predicted.

Cross-Entropy Loss for Multi-Label Classification

During training, the BERT model is optimized using the binary cross-entropy loss function for multi-label disease prediction. It measures the difference between the predicted probabilities and actual labels (Bhartendu, 2021).

Equation 6

Binary Cross-entropy Loss Function

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Explanation:

- **N:** Total number of disease labels
- **y_i:** True label for disease i (0 or 1)
- **$\hat{y}_i$:** Predicted probability for disease i

This loss function penalizes incorrect predictions and guides the model to improve both precision and recall across all diseases.

Confidence Score Calculation

The system also provides a confidence score for each predicted disease, indicating the certainty of the prediction. This is directly derived from the sigmoid output of the model:

Equation 7

Confidence Score

$$\text{Confidence Score} = \sigma(x) \times 100\%$$

The larger the score of confidence, the larger the chances that the predicted disease will correspond to the symptoms of the user.

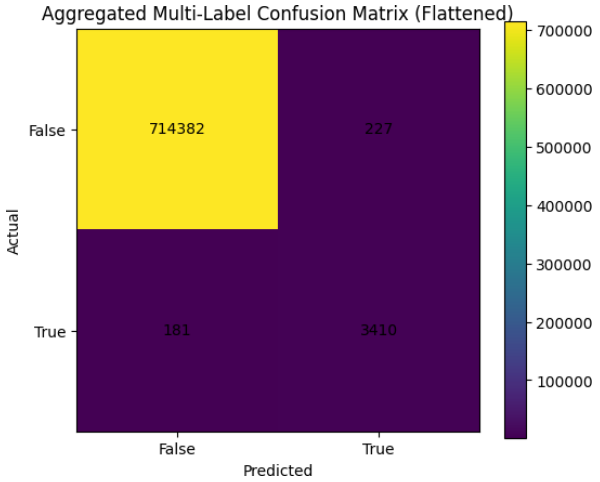
All these equations will be significant to guarantee the Medical Chatbot System is capable of classifying various diseases with a high probability, evaluating the success of the BERT classifier, and making predictions that are based on confidence. Quantitative evaluation metrics are the accuracy of the final outcome, precision, recall, and F1-score, whereas the sigmoid activation and cross-entropy loss determine the training and prediction probability of the model. The combination of these equations provides the system with both the accuracy and interpretability of analyses, which raises the confidence of its users and the reliability of the medical domain (Ultralytics, 2025).

Result

This section shows the experimental findings as far as the research objectives are concerned, including the accuracy of the disease predictions, reliability of the classification, and the safety of the system. The Medical Chatbot System uses a fine-tuned BERT model to diagnose the disease based on the symptoms entered by the user. The dataset comprises thousands of medical questions and labels, which allows the model to feel competitive on a variety of evaluation metrics such as accuracy, precision, recall, and F1-score. The pipeline was developed for not only predicting common diseases but also spotting emergencies and offering safe fallbacks in the decision-making when the prediction confidence is low.

Figure 1

Multi-Label Confusion Matrix (Flattened)



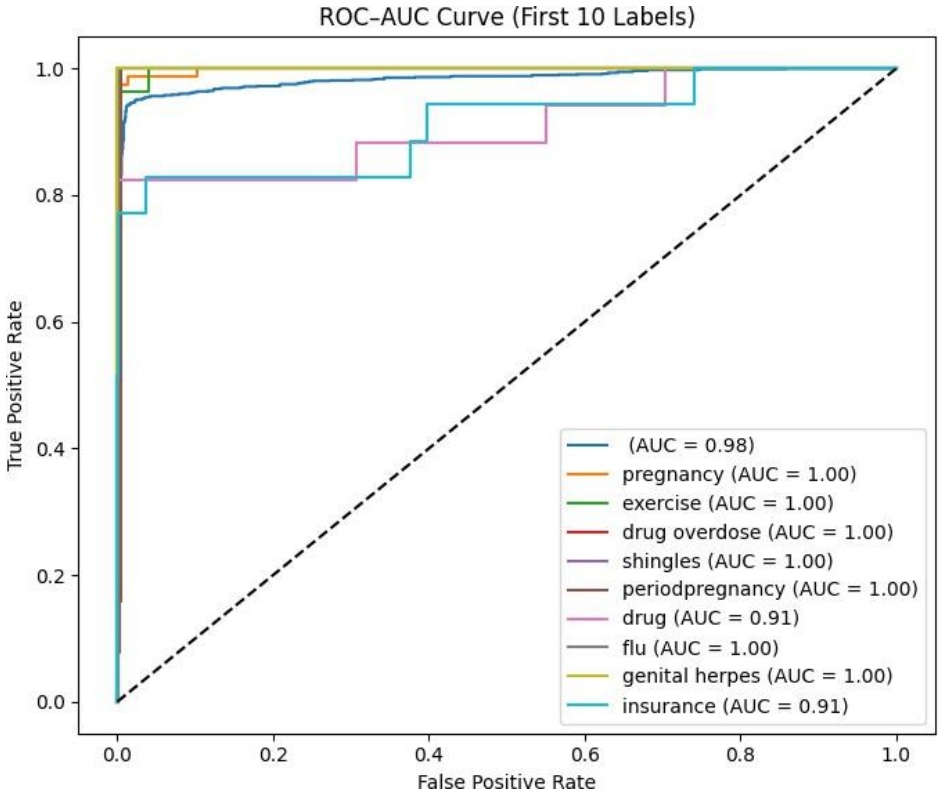
The aggregated confusion matrix gives a general overview of the disease labels classification accuracy of the Medical Chatbot System in a multi-label environment.

- True Negative (TN = 714,382): The extremely high TN value indicates that the model hardly labels disease needlessly. It has the capacity to predict when a label must not be predicted, which is a good attribute of discrimination.
- False Positive (FP = 227): The FP is quite small in relation to TN, which indicates high precision. The model is virtually impossible to predict a disease where it should not, and this is essential to safe medical outcomes.
- False Negative (FN = 181): A low number of FNs indicates that the model is consistent in identifying the useful medical situation, and it does not commonly overlook the actual illnesses.
- True Positive (TP = 3,410): The good TP value indicates that the model is capable of identifying the disease labels on a variety of symptom descriptions.

On the whole, the confusion matrix shows that the model is capable of classifying medical conditions accurately with the least amount of misclassification. Its reliability and real-time usefulness in medical assistance are indicated by the extremely low FP and FN rates (Naviwala, 2024).

Figure 2

ROC-AUC Curve for First 10 Disease Tags



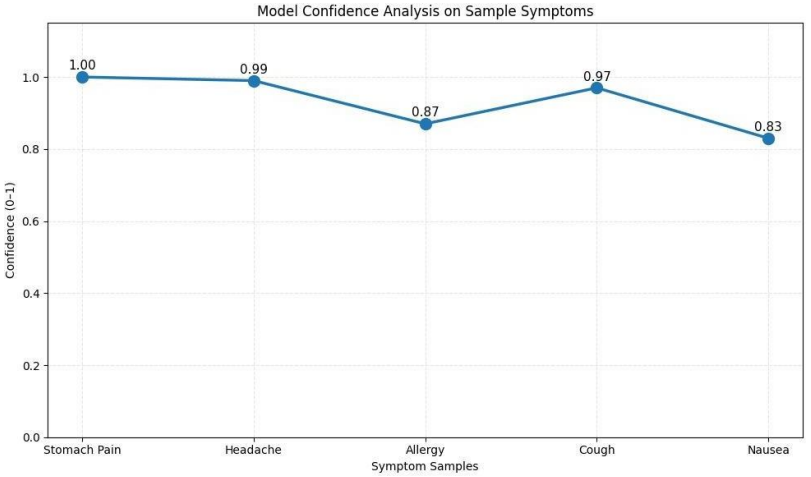
The ROC–AUC chart shows the separation of positive and negative classes of the model within the first ten medical conditions.

- High AUC Scores: The AUC values of most disease labels are close to 1.00, and this implies that they are nearly perfect in their classification behaviour. The model is useful in differentiating the symptoms that are relevant and those that are not relevant.
- Perfect Separation: Perfect separation is demonstrated with conditions like pregnancy, drug overdose, shingles, flu, and exercise; as well as genital herpes, which has an AUC of 1.00, indicating ideal detection.
- Slightly Lower Scores: There are labels with slightly lower AUC, such as drug and insurance, which have an AUC value of about 0.91. Although slightly lower, these scores still represent excellent predictive strength.

The ROC-AUC findings confirm that the system is highly sensitive and specific to various medical categories, and thus it can be depended on to detect diseases in real-life interactions.

Figure 3

Model Confidence Analysis on Sample Symptoms

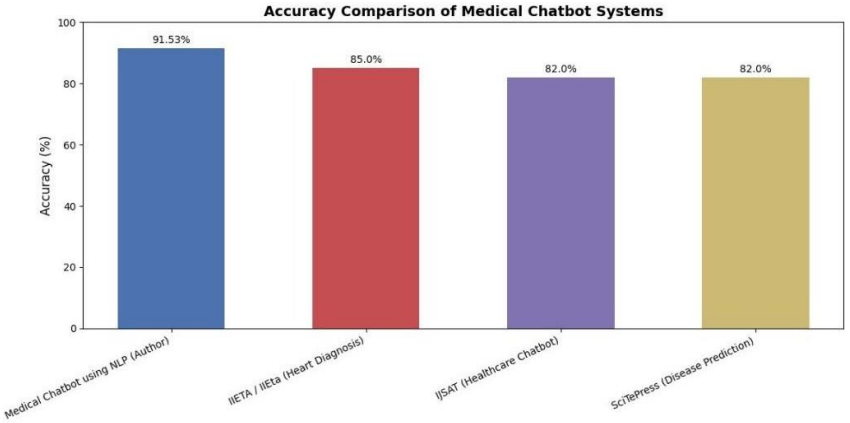


The line plot will show the confidence levels of five symptoms in the model, which are Stomach Pain, Headache, Allergy, Cough, and Nausea. In general, the BERT-based classifier provides almost stable confidence scores of between 0.83 and 1.0, which proves to be highly reliable in the interpretation of the symptoms.

The highest confidence scores (around 1.0) were given to Stomach Pain and Headache, which implies that the model is very effective in identifying the clear and common patterns of symptoms. In like manner, Cough and Allergy had confidence levels of 0.87 to 0.97, which indicates consistency even to symptoms that can be shared by several conditions. The lowest confidence was recorded as nausea (0.83), which means that the less specific or the broader the symptoms are, the slight decrease in the certainty to predict, and the value of it is still considered to be a strong degree of certainty. Overall, the system achieved an accuracy of 91.53%, with a precision of 95.54% and a recall of 94.96, demonstrating strong predictive reliability.

Figure 4

Performance Comparison with previous Research



The comparison chart on accuracy points out the performance of the Medical Chatbot System in comparison with the performance of the healthcare diagnostic chatbots that have been published in the past.

a) Author – 91.53%: The accuracy of prediction is significantly improved by the use of a fine-tuned BERT transformer model, a highly selected dataset, and a successful multi-label technique, which becomes superior to previous systems.
b) IIETA Heart Diagnosis Chatbot – 85%: This model uses a more basic method of diagnosing conditions, which reduces its capacity to learn about the relationship between symptoms, and thus it is less accurate (Majeed & Hardan, 23 November 2023).
c) IJSAT Healthcare Chatbot – 82% and SciTePress Disease Prediction – 82%: These systems rely on traditional machine-learning techniques, which restrict their capacity to capture deeper contextual meaning from user symptoms. (Shaikh et al., 2025) (Oltean Anisia Veronica & Coriou, 2025).

The comparison is an obvious demonstration that the suggested Medical Chatbot System is superior to the available models by a significant margin (6-10%). Its application of next-generation NLP-based transformer architecture allows grasping the pattern of symptoms with more accuracy, as well as making health condition predictions that are more assertive than in previous methods.

Discussion

The Medical Chatbot System successfully predicts potential medical conditions based on the user-provided symptoms and offers medical advice through an intelligent conversational interface. The fact that the system achieved 91.53% accuracy indicates that it is useful in systematizing a disease, but at the same time, being able to achieve high precision (95.54) and recall (94.96) rates indicates that the system can be utilized in real-life healthcare support.

The integration of BERT for classification and Mistral LLM for conversation generation creates a balanced system that combines analytical precision with natural interaction. m that is heavy on precision and analysis, yet light on conversation and natural interaction. The true negative (714,382) and false positive (227) are very high and low, respectively, which means the system rightly avoids over-diagnosis, which is paramount in clinical use to avoid unwarranted panic. Equally, the low false negative rate (181) indicates that the system does not miss actual disease conditions very often, which is critical to patient safety.

The emergency red-flag detection feature was used successfully to detect the most important symptoms and send out necessary alerts immediately, which proved that the system includes a safety mechanism. To ensure responsible design, in circumstances of low confidence, the system correctly reacted with fallback messages and general care tips instead of trying to make uncertain predictions.

However, it is important to note that the system stresses that it is not to substitute professional medical diagnosis and that it is always accompanied by relevant medical disclaimers and safety warnings to ensure that patients who seek healthcare information use

the system responsibly. The system serves as a prescreening device and an online access to professional healthcare, especially useful in a resource-restricted environment such as Nepal, where access to immediate medical care might be limited.

Despite strong performance, the system relies on secondary datasets and may not fully capture rare or region-specific medical conditions. Additionally, overlapping of the symptoms between diseases can undermine confidence of some predictions. Future work may focus on increasing the dataset and the inclusion of clinical validation.

Overall, the results show that combining BERT-based classification with the LLM-driven conversational intelligence can considerably enhance the disease prediction accuracy without compromising the safety and usability of healthcare applications in practice.

Conclusion and Future Work

The paper recommended the Medical Chatbot System, which uses Natural Language Processing (NLP), to provide users with preliminary medical information by analysing user-reported symptoms and generating context-based medical advice. The experiment's results demonstrate that the proposed system can predict potential medical conditions with exceptional accuracy, ensuring a balanced level of precision and recall. By combining a disease classification model based on BERT with a Large Language Model (LLM), it is possible to make predictions with necessary precision and engage with the system through natural conversation, making it suitable for real-world healthcare scenarios.

The system can be especially useful in Nepal, where access to healthcare services is limited in rural and remote areas. The chatbot will help reduce unnecessary hospital visits and support high-risk cases with timely medical attention through early symptom assessment and simple medical consultations. Red-flag detection and fallback responses serve as safety mechanisms, ensuring users seek medical care when needed. Consequently, the system helps improve health awareness and encourages preventive health practices.

Future work could focus on making systems more accessible and functional. The mobile-based application has the potential to expand users' reach, while multilingual support, such as Nepali and local dialects, would enhance usability for rural communities. Additional features might include a doctor recommendation service, telemedicine options, and real-time hospital databases that enable users to make appointments and access online consultations. Furthermore, wearable health-monitoring devices could be integrated into the system, which would likely allow for more personalized and accurate health insights.

Overall, the Medical Chatbot System is an important step towards digital health innovation as it provides a reliable first-line diagnosis, provides an equal opportunity to deliver healthcare services, and improves health literacy. Further, it can be even more detailed and easy to use, which will help to modernize the healthcare system in Nepal and advance the health conditions of its citizens.

References

Azam, A., Naz, Z., & Muhammad Usman, G. K. (2024). PharmaLLM: A Medicine Prescriber Chatbot Exploiting Open-Source Large Language Models. *Human-Centric Intelligent Systems, 4*, 527–544.

Babu, A., & Boddu, S. B. (2024). Bert-based medical chatbot: Enhancing healthcare communication through natural language understanding. *Exploratory research in clinical and social pharmacy, 13*, 100419.

Bhartendu, T. (2021, May 7). *Modified cross-entropy loss for multi-label classification and handling imbalanced data*. *Medium*. https://medium.com/@matrixB/modified-cross-entropy-loss-for-multi-label-classification-with-class-a8afede21eb9

geeksforgeeks. (2025, July 23). *Machine-learning/derivative-of-the-sigmoid-function*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/derivative-of-the-sigmoid-function/

Kavyasrirelangi. (2025, September 9). *Understanding-model-evaluation-metrics-for-machine-learning-160d385b72c5*. *Medium*. https://medium.com/@kavyasrirelangi100/understanding-model-evaluation-metrics-for-machine-learning-160d385b72c5

Majeed, B. A., & Hardan, A. Y. (2023, November 23). *A study on [title of the article]*. *International Information and Engineering Technology Association (IIETA)*. https://www.iieta.org/journals/ria/paper/10.18280/ria.380121

Mhatre, A., Warhade, S. R., Pawar, O., Kokate, S., Jain, S., & Emmanuel, M. (2024). Leveraging LLM: implementing an advanced AI chatbot for healthcare. *Int. J. Innov. Sci. Res. Technol*, *9*, 3144-3151.

Naviwala, H. (2024, September 23). *Confusion-matrix*. Data Science Dojo. https://datasciencedojo.com/blog/confusion-matrix/

Oltean Anisia Veronica, Pop, I. D., & Coroiu, A. M. (2025). *Medical chatbot for disease prediction using machine learning and symptom analysis*. In *Proceedings of the 20th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2025)* (pp. 600–607). SCITEPRESS – Science and Technology Publications. https://www.scitepress.org /Papers/2025/133579/133579.pdf

PHC-Nepal. (2024). *Public health challenges and opportunities in Nepal: In-depth critical analysis*. https://phcnepal.com/public-health-in-nepal-challenges-and-opportunities

Saleem, M. (2022). *Medical chatbot* [Kaggle notebook]. Kaggle. https://www.kaggle.com/code/mohsinsial/medical-chatbot

Shaikh, U., Mustafa, S. M., Mujawar, S., Shaikh, H., & Pathan, Z. (2025). *AI healthcare chat bot. International Journal on Science and Technology (IJSAT), 16*(1), 1–6. https://www.ijsat.org/papers/2025/1/2313.pdf

Ultralytics. (2025). *confidence*. Ultralytics. https://www.ultralytics.com/glossary/confidence