Vol. 1, No. 2, 2026
https://dkujournals.edu.et

Ras Dejen Journal of Sciences

Development of a Multi-Output Hybrid Artificial Intelligence Model for Medical Question Classification and Decision Support

Abraham Alamirew[1], Birhanu Alemneh[2],YeshiwasAdane[3]

[1] Department of Computer Science, Debark University, Debark, Ethiopia

Email:uogabrahamcsd@gmail.com

ORCID: https://orcid.org/0009-0009-4385-3568

[2] Department of Computer Science, Debark University, Debark, Ethiopia

Email:bright.alemneh@gmail.com

[3] Department of Computer Science, Debark University, Debark, Ethiopia

Email:yeshiwas.adane@dku.edu.et

corresponding author's email: uogabrahamcsd@gmail.com

Article History:

Received: September 23, 2026

Revised: October 26, 2026

Published: November 12, 2026

Abstract

Healthcare medical diagnosis is the process of identifying a disease, condition, or injury based on a patient's symptoms, physical examination, and diagnostic tests. The AI-based advisory model utilizes artificial intelligence techniques to provide recommendations, guidance, and decision-making support across various domains. This study explores the development and evaluation of a multi-output hybrid advisory model for healthcare medical diagnosis. Multi-Output Classifier models analyze datasets to aid healthcare decisions, helping diagnose conditions, suggest treatments, and provide medical information. We used a kaggle dataset (16,125) and an additional dataset (32,631) that was meticulously curated from 12 trusted websites of the National Institutes of Health (NIH), an official trusted website of the United States government, to build a medical diagnosis model via six experiments. The dataset was

©2026 Debark University
Corresponding author: uogabrahamcsd@gmail.com

partitioned into 80% for training and 20% for testing. To optimize performance, we have developed a hybrid model that combines the strengths of individual models to create a more robust model that overcomes their individual limitations. We evaluate the model classification performance using metrics such as accuracy, precision, recall, F1 score, and cross-validation. Additionally, we evaluated the model's performance with BERT and confidence scores for question answering. Six experiments were conducted using Extra Trees, random forests, LSTM, MLP, a hybrid model, and Extreme Gradient Boosting. The hybrid model achieved 97% accuracy in the defined classification task. As a result, the hybrid model was chosen as the best for medical diagnosis. The proposed multi-output hybrid model achieved an acceptance rate of 84% among domain experts.

Keywords- Healthcare, medical, medical diagnosis, multi output, hybrid model.

1.Introduction

Healthcare medical diagnosis is the process of identifying a disease, condition, or injury based on a patient's symptoms, physical examination, and diagnostic tests (Alowais et al., 2023). It is a critical step in healthcare that enables healthcare professionals to determine the appropriate treatment and management plan for a patient. The AI-based advisory model leverages artificial intelligence (AI) techniques to provide recommendations, guidance, or decision-making support in various domains (Aityan et al., 2024). These models are designed to analyze large amounts of data, identify patterns, and generate insights to assist users in making informed decisions. An AI-based advisory model for healthcare medical diagnosis utilizes artificial intelligence technologies to assist healthcare professionals in diagnosing medical conditions (Muscat et al., 2020). AI models analyze extensive patient data using sophisticated algorithms to identify patterns and suggest diagnoses and treatments. AI simulates human intelligence in machines, enabling them to learn reason, solve problems, understand language, perceive, and make decisions. While AI-based advisory models hold immense promise for revolutionizing patient care through improved diagnostic accuracy, the successful implementation of these systems hinges on their ability to interact with and process medical information in a practical clinical format. Current research indicates a clear need for continuous updates about new tests and treatments, and there is a scarcity of research focusing on the specific types of questions andanswers that are vital for quick and accurate data analysis in a clinical setting. Therefore, this study investigated a specific

task of developing a multi-output advisory model centered on the joint prediction of the question type (Qtype) and the structured Answer represents a critical, underexplored frontier in healthcare AI. By explicitly classifying the Qtype, the model ensures the output is targeted and evidence-based, directly addressing the clinician's immediate informational need and enforcing the principles of evidence-based practice. This approach is not merely an incremental enhancement of general medical language models, it is a fundamental shift toward delivering precision decision support. By recognizing frequent question types, this work aims to guide the creation of targeted training programs for clinicians and focus researchers on the most relevant areas. This dual-target approach allows the proposed model to not only categorize the patient's information needs but also to generate a precise, decision-supporting response, moving beyond generalized diagnostic support to address the concrete needs of practitioners and patients. This study employed Natural Language Processing (NLP) to analyze the text dataset.

NLP is a part of artificial intelligence that helps machines gets the hang of human language. It’s about making sense of what we say and write, and how to respond in a way that makes sense. NLP brings together rules about language along with machine learning and deep learning techniques to handle and analyze tons of language dat**a** (Ramgopal et al., 2023). NLP text processing uses the rules of language to help machines understand and work with human speech(Basha et al., 2023). This includes tasks like breaking words into tokens, analyzing sentence structure, figuring out feelings in text, and powering things like translation apps, chatbots, and summaries. By breaking language into smaller pieces and looking at how it all fits together, NLP helps machines get better at understanding us, which is a big step for AI and makes it easier for us to interact with technology. NLP plays a critical role in AI-based advisory models for medical diagnosis by enabling machines to understand and analyze human language in medical contexts(Rawas et al., 2024). It helps medical experts get better at diagnosing, makes their work easier, and supports them in making decisions that focus on patients. While NLP could really change how medical diagnosis works by automating data analysis and aiding decisions, it also has some tough challenges like misunderstandings, bias, privacy issues, and being complicated (Linardatos et al., 2021). Addressing these challenges requires robust data governance, ethical considerations, and continuous improvement of NLP models to ensure they are accurate, fair, and trustworthy. To bring AI advisory models into healthcare successfully, we

have to deal with some main challenges first. It's important to focus on tech advancements, ethics, and teamwork with healthcare workers to build better NLP advisory models. This way, we can help more people use AI in medical diagnosis and, in turn, improve patient care and how healthcare works overall.

Machine learning is a subset of artificial intelligence (AI) focused on developing algorithms for computers to learn and predict without explicit programming(Ibrahim et al., 2022). It uses statistical techniques to analyze data and identify patterns(Doreswamy & Nigus, 2020). Key points in machine learning include learning from data, training with big datasets, and creating models (Sarker, 2021). Machine learning techniques include supervised, unsupervised, Semi-supervised, and reinforcement learning (Pintelas & Livieris, 2020). Supervised learning maps input to output with correct values provided by a supervisor (Hasan et al., 2023). There are two main types of supervised learning those are classification and regression. Ensemble learning falls under this category. A number of studies have looked into how healthcare diagnoses are made using systematic reviews and LLM methods, focusing on the role of AI and machine learning. However, there hasn’t been much research on the types of questions and answers that are important for data analysis and helping doctors diagnose patients quickly and accurately. There’s a need to create training programs that can better prepare clinicians for their work. Due to this our main focus is on that identify the question types and answer, a multi output prediction. Predict question types with answer are essential for accurate and rapid medical diagnosis. Recognizing frequent question types allows medical educators to create targeted training programs, better preparing clinicians for practice. Furthermore, understanding common clinical questions guides researchers to focus on relevant areas, leading to impactful research that addresses clinical challenges. To this end, the current research tries to investigate and answer the following research questions.

1. Which algorithm provides the best performance for medical question classification?
2. To what extent does the proposed advisorymodel accurately identify the medical diagnosis?
3. How accurately does the proposed hybrid model classify medical questions and generate corresponding answers?

2.Related works

2.1.AI-Based Medical Diagnosis

Artificial intelligence (AI) has increasingly been applied to healthcare to support disease diagnosis, treatment planning, clinical risk prediction, and personalized medicine. (Alowais et al., 2023)reviewed the applications of AI in healthcare and highlighted its potential to improve diagnostic processes, support treatment decisions, facilitate laboratory workflows, and contribute to personalized patient care. Their discussion emphasized the importance of interdisciplinary collaboration, patient trust, appropriate implementation, and responsible use of AI technologies. However, the broad scope of their review does not specifically address the joint prediction of medical question types and corresponding answers in an advisory framework.

Similarly, (Aityan et al., 2024)investigated the application of AI-based systems in emergency medicine, focusing on the identification of urgent medical conditions and the potential use of large language models (LLMs) to support clinical decision-making. Their study examined emergency cases, including the identification of critical conditions such as sepsis. The work illustrates the potential of AI to provide rapid medical information and support decisions in time-sensitive clinical environments. Nevertheless, emergency diagnosis and clinical question answering represent different tasks. While emergency diagnostic systems focus on identifying medical conditions from clinical information, the present study investigates the classification of medical questions and the prediction of corresponding answers. This distinction highlights the need to evaluate medical advisory models according to their specific objectives and outputs.Although AI-based diagnostic approaches demonstrate considerable potential, their performance depends on the quality and representativeness of the data, the suitability of the algorithms, and the reliability of the evaluation procedures. Furthermore, high diagnostic accuracy alone does not necessarily establish the effectiveness of an advisory system designed to answer clinical questions. Therefore, additional research is required to investigate AI models that explicitly connect the identification of clinical information needs with the prediction of appropriate answers.

2.2.Medical Natural Language Processing

Natural language processing (NLP) enables computers to analyze, interpret, classify, and generate human language. In healthcare, NLP techniques are used to process electronic health records, extract information from clinical documents, classify medical text, identify symptoms, and support clinical information retrieval. These applications are particularly important because medical information is frequently expressed through unstructured text containing specialized terminology, abbreviations, and context-dependent expressions.(Ramgopal et al., 2023)reviewed the applications of AI and machine learning in clinical decision-making in pediatrics. Their review discussed how supervised and unsupervised learning approaches can identify clinically relevant patterns and support the integration of AI into clinical decision-support systems. The authors also highlighted challenges associated with limited datasets, low frequencies of clinically significant outcomes, data imbalance, and the generalizability of predictive models. These limitations demonstrate the importance of appropriate dataset construction, model validation, and careful assessment of performance across different clinical settings. Although the work by (Ramgopal et al., 2023)focuses on pediatric clinical decision support, its discussion of data limitations and model development is relevant to medical NLP research more broadly. However, clinical outcome prediction differs from the task of identifying question types and predicting corresponding answers. The present study addresses this distinction by investigating a multi-output advisory framework that processes medical text and produces two related outputs: the question type (Qtype) and its corresponding answer representation.Medical NLP systems must also address challenges involving ambiguous terminology, variations in question formulation, and the relationship between the linguistic structure of a query and the information required to answer it. Consequently, task-specific NLP models that explicitly represent clinical question categories may provide a useful approach to organizing medical information requests and evaluating their corresponding responses.

2.3.Medical Question Answering and Clinical Question Types

Medical question answering is an important application of NLP that aims to provide relevant information in response to questions about diseases, symptoms, diagnostic procedures, treatments, medications, and other healthcare topics. Such systems can assist healthcare professionals in retrieving medical information and may help patients access understandable

health-related information. However, generating a medically relevant answer requires more than identifying words or topics in a question. The system must recognize the information need expressed by the question and produce an answer that is relevant to that need.

(Muscat et al., 2020) investigated clinical questions submitted by Australian general practitioners to an information service supporting evidence-based practice. The study examined 126 clinical questions from 53 general practitioners and identified recurring categories, including questions concerning treatment, diagnosis, and epidemiology. The findings also highlighted the importance of keeping healthcare professionals informed about emerging diagnostic tests, treatments, and complementary therapies. The study provides useful evidence about the types of information general practitioners seek in clinical practice.

However, the study primarily focused on identifying and categorizing clinical questions and describing associated knowledge gaps. Its findings were based on a relatively small, non-random sample of practitioners and a particular clinical information service, which limits the generalizability of the results to other populations and healthcare settings. Furthermore, identifying the types of questions submitted by practitioners does not, by itself, provide an automated mechanism for predicting question categories and their corresponding answers.

In contrast, the present study investigates an NLP-based multi-output advisory model designed to predict both the clinical question type and its corresponding answer representation. This approach extends the research focus from descriptive analysis of clinical information needs to the development and evaluation of an automated prediction framework. By examining these outputs together, the study aims to investigate whether question classification and answer prediction can be integrated into a unified medical advisory process.

Explicit question-type identification may also help organize recurring clinical information needs into meaningful categories. Such information could potentially inform clinician training, medical information management, and future research priorities. Nevertheless, the effectiveness of these applications depends on the accuracy of the model outputs and requires appropriate evaluation.

2.4.Hybrid Models and Explainable Deep Learning in Healthcare

Hybrid machine learning models combine two or more algorithms or learning approaches to address limitations associated with individual models. In healthcare, hybrid approaches may combine feature extraction, classification, ensemble learning, neural networks, or language-processing techniques. Their purpose may include improving predictive performance, capturing complex patterns, or integrating complementary modeling capabilities. However, the benefits of a hybrid architecture depend on the specific task, dataset, model design, and evaluation methodology.

Ahmed and Islam (2025) investigated explainable deep learning approaches for medical diagnosis, including pneumonia detection and sepsis risk prediction. As reported in the study, the models achieved 94% accuracy in pneumonia detection and 87% sensitivity and 85% specificity in sepsis risk prediction. The authors also considered interpretability methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) to support the interpretation of model predictions. Their work highlights the importance of considering both predictive performance and interpretability in healthcare applications. The study also identified challenges involving computational costs, scalability, and the requirements of different application domains.

Although explainable deep learning and hybrid approaches may improve the analysis of medical data, their evaluation results must be interpreted in relation to their prediction targets. Pneumonia detection, sepsis risk prediction, clinical question classification, and medical answer prediction are distinct tasks. Therefore, the reported performance of diagnostic models cannot be directly compared with the accuracy of a multi-output question-answering model without considering the differences in datasets, output definitions, and evaluation measures.

The present study investigates a hybrid multi-output advisory model for medical question-type and answer prediction. Its primary focus is not simply to classify a disease or estimate clinical risk, but to analyze a medical query and predict its associated question category and answer representation. The model's performance should therefore be evaluated against suitable baseline approaches designed for the same or closely comparable prediction tasks. Where multiple

learning algorithms are combined, comparative experiments and ablation studies can help determine whether the hybrid architecture provides measurable benefits over individual models.

2.5.Clinical Decision Support and AI-Based Advisory Systems

Clinical decision-support systems provide healthcare professionals with information, alerts, recommendations, and analytical outputs to assist clinical decision-making. AI-based advisory systems extend these capabilities by applying machine learning, NLP, and other computational techniques to large amounts of medical data. Depending on their design, these systems may support diagnosis, treatment selection, clinical risk assessment, information retrieval, and the interpretation of medical records.

The work of Alowais et al. (2023) emphasizes the broad potential of AI to support diagnostic and treatment-related decisions, while Aityan et al. (2024) illustrate the application of AI-based approaches to time-sensitive emergency medical situations. Ramgopal et al. (2023) further discuss the integration of AI and machine learning into clinical decision support, particularly in pediatric care. Collectively, these studies demonstrate that AI-based systems can be investigated across different clinical contexts, although their effectiveness depends on the specific tasks they are designed to perform.

Despite these developments, a distinction should be made between systems that predict clinical outcomes and systems that respond to clinical information requests. A disease-prediction model estimates the likelihood of a particular condition based on patient information, whereas a medical advisory model processes a question and produces a relevant informational response. The latter requires attention to question interpretation, answer relevance, and the consistency between the question type and the predicted answer.

The proposed study addresses this task by developing a multi-output advisory model that jointly predicts the question type and the corresponding answer representation. The approach is intended to provide a structured mechanism for processing medical questions and producing task-specific outputs. However, the resulting predictions should be considered decision-support information rather than substitutes for professional medical judgment. Clinical deployment would require additional validation of factual accuracy, relevance, safety, and usefulness.

2.6.Research Gap and Contribution of the Proposed Study

The reviewed literature demonstrates the growing use of AI, machine learning, NLP, deep learning, and LLMs in medical diagnosis and clinical decision support. Alowais et al. (2023) discuss the broad applications of AI in healthcare, Aityan et al. (2024) examine AI-assisted emergency medical decision-making, Ramgopal et al. (2023) review AI-based clinical decision support in pediatrics, and Ahmed and Islam (2025) investigate explainable deep learning for medical diagnosis. Muscat et al. (2020), meanwhile, provide evidence about the types of clinical questions raised by general practitioners and the knowledge gaps associated with those questions.

Although these studies provide relevant foundations, they address different objectives, including disease diagnosis, emergency condition identification, clinical outcome prediction, explainable classification, and the descriptive categorization of clinical questions. Their findings cannot be directly compared without accounting for differences in prediction targets, datasets, and evaluation measures. In particular, the identification of clinical question types does not necessarily provide an automated mechanism for predicting corresponding answers, while conventional diagnostic models are not necessarily designed to classify the informational needs expressed in medical questions.

Accordingly, the research gap addressed in this study is the need to investigate a unified, NLP-based multi-output framework that jointly predicts clinical question types and their corresponding answer representations. The proposed model focuses on the relationship between question classification and answer prediction rather than exclusively on disease classification or the general generation of medical responses.

The main contribution of this study is the development and empirical evaluation of a multi-output medical advisory model that integrates question-type prediction and answer prediction within a single framework. The study also examines the model's predictive performance and the distribution of clinical question types in the dataset. These analyses may provide insights into recurring medical information needs and potential applications in clinical education and medical information support. The claimed advantages of the proposed model must ultimately be established through appropriate baseline comparisons, separate evaluation of each output, and,

where applicable, expert assessment of the medical relevance and factual accuracy of the predicted answers.

3.Methods

3.1.Data source and Preparation

This study used secondary medical question-and-answer data obtained from an existing Kaggle dataset (https://www.kaggle.com/datasets/thedevastator/comprehensive-medical-q-a-dataset) and supplementary medical information collected from reputable online health-information resources. The datasets were selected to support the development of a natural language processing (NLP)-based multi-output advisory model for predicting medical question types (Qtype) and their corresponding answers.The supplementary dataset was constructed from publicly accessible medical information resources, including the National Cancer Institute (NCI), the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), MedlinePlus, and the Genetic and Rare Diseases Information Center (GARD). These resources provide information on diseases, symptoms, diagnostic procedures, medications, treatments, prevention, and other healthcare topics.The relevant official websites include the following:

- ✓ Kaggle: https://www.kaggle.com/datasets
- ✓ National Cancer Institute: https://www.cancer.gov/
- ✓ National Institute of Diabetes and Digestive and Kidney Diseases: https://www.niddk.nih.gov/
- ✓ MedlinePlus: https://medlineplus.gov/healthtopics.html
- ✓ Genetic and Rare Diseases Information Center: https://rarediseases.info.nih.gov/

The original records were organized into three principal fields: Qtype, Question, and Answer. The Qtype field represents the category assigned to a medical question, the Question field contains the original medical query, and the Answer field contains the corresponding medical response. These fields were used to construct the prediction targets and input data required for the multi-output modeling framework.

Before combining the datasets, records were assessed for completeness, consistency, duplication, and suitability for the research objectives. Records that did not contain a usable question, a

corresponding answer, or sufficient information to assign a question type were considered for exclusion according to predefined data-quality criteria. Duplicate and near-duplicate question-answer records were identified and handled before model development to reduce redundancy and the risk of information leakage.

Question types were retained from the original dataset when reliable labels were available. For records requiring additional labeling, a documented classification procedure was applied using predefined category definitions. Any manual labeling or relabeling was recorded, and ambiguous cases were reviewed according to the study's labeling protocol. The final dataset was divided into training, validation, and testing subsets using a documented splitting strategy. Where appropriate, related or duplicate records were kept within the same partition to reduce the risk of overly optimistic performance estimates.

The final dataset composition, source-specific record counts, exclusion counts, duplicate-removal statistics, and the distribution of question-type categories should be reported after preprocessing. These steps provide a transparent description of dataset construction and support reproducibility of the proposed medical advisory model.

Table 1:feature description types used in the experiment

No	**Features**	**Description**	**Data types**	**Role in the model**
1	Qtype	The types of question	Categorical (string)	First prediction target
2	Question	The medical question posed by the patient	Text (String)	Input feature
3	Answer	The expert response to the medical question	Text (String)	Second prediction target

In machine learning, the model needs to know what it is trying to predict. This table lists your classes. By defining 12 distinct categories (Information, Symptoms, Treatment, etc.), we are showing that our experiment is a Multi-class Classification problem. Without the table 2 below, the reader wouldn't know if we were doing a simple search or a complex categorization.

Table 2: the question type description that was used as a class for the experiment

Question Type	Explanation
Information	General and foundational information about the condition, such as what it is, its definition, or how it is generally classified
Symptoms	Questions detailing the signs and symptoms a patient might experience, including their onset, severity, and frequency
Treatment	Inquiries about available treatments, therapies, management options, or potential cures for the condition
Inheritance	Questions regarding the genetic patterns of inheritance and how the condition is passed down through families
Frequency	Data and statistics on how frequently the condition occurs in the general population or specific demographic groups
genetic changes	Focuses on the specific genetic or molecular changes (mutations, chromosome abnormalities, etc.) that cause or are associated with the condition
Causes	Questions about the underlying etiology or risk factors that lead to the condition, which may be different from genetic changes (e.g., environmental, infectious, or lifestyle factors)
exams and tests	Details on the diagnostic exams and tests used to confirm the presence of the condition
Research	Information on current research efforts, clinical trials, or future directions in the understanding and management of the condition
Outlook	Questions about the prognosis or long-term course of the disease, including life expectancy and expected quality of life
susceptibility	Identifies who is at risk for developing the condition, often covering factors like age, gender, geographic location, or specific exposures
Prevention	Strategies and methods to prevent the onset or progression of the condition

3.2.Data Inclusion and Exclusion Criteria

The study included medical question-and-answer records obtained from documented and traceable data sources, including the Kaggle dataset and supplementary medical information resources. Records were considered eligible if they contained a meaningful medical question, a corresponding usable answer, and information relevant to healthcare or medical information needs. In addition, each record was required to have an available question-type (Qtype) label or sufficient information to be reliably assigned to one of the predefined question categories. If English was the language of the study, only English-language records were included to ensure consistency in text preprocessing, feature extraction, and model development.

Records were excluded if they contained missing or unusable questions, incomplete or unavailable answers, irrelevant information, malformed text, or insufficient information for reliable question-type classification. Records from sources whose origins could not be verified were also excluded. Where applicable, non-English records, advertisements, and responses containing only hyperlinks without meaningful medical information were removed. Exact and near-duplicate question-and-answer records were identified and handled according to the established deduplication procedure to minimize redundancy and reduce the risk of data leakage during model training and evaluation. These inclusion and exclusion criteria were applied during dataset preparation to improve data quality, consistency, and suitability for developing the proposed multi-output medical advisory model.

3.3.Data Pre-processing

The data pre-processing phase involves formatting and organizing data for analysis (Fan et al., 2021). Data preprocessing plays a crucial role in the performance of machine learning algorithms by eliminating noise and irrelevant features from the dataset. If irrelevant, redundant, noisy, or unreliable data is present in the dataset and used for model development, achieving the desired outcome during the training phase becomes more challenging(Ajuzieogu, 2024). In NLP, preprocessing includes text cleaning, tokenization, stops word removal, stemming, lemmatization, and feature extraction.The choice of data preprocessing steps is not arbitrary but is strategically aligned with the goal of creating a robust, multi-output advisory model from clinical and biomedical text data.

3.4.Text Cleaning

Cleaning text is an important step in working with language data. It helps make sure the text is in good shape for analysis and improves how models work(Ogunrinde, 2025). For example, changing everything to lowercase keeps things consistent, and getting rid of URLs, random symbols, and numbers cuts out the clutter. By applying these cleaning methods, we create a more structured and meaningful dataset, ultimately improving the accuracy and efficiency of NLP applications such as sentiment analysis and text classification.For our study, we used text cleaning to remove noise and standardize input.

3.5.Tokenization

Tokenization is the process of breaking down a text into smaller units called tokens(Bayram & Fincan, 2025). These tokens can be words, sub-words, or sentences, depending on the type of tokenization used. It is a fundamental step in NLP because it helps convert raw text into a structured format that machines can process.We used tokenization in our study due to its necessity for all NLP tasks to break the stream of text into meaningful units (tokens), such as words.

3.6.Stop word removal

Stop word removal is the process of filtering out commonly used words that do not add significant meaning to a text(Kaur, 2021). These words, known as stop words, include words like "the," "is," "and," "of,"and"to." Since they appear frequently in text but carry little meaningful information, removing them helps improve NLP model efficiency.For this study, we utilized the stopwords.words('english') function from the NLTK library to retrieve a predefined list of common English stop words. For our study, we used it to significantly reduce the dimensionality of the feature space without losing clinical meaning. In a high-stakes task like diagnosis, the focus must be entirely on meaningful terms (symptoms, conditions, treatments) to enhance model efficiency and prediction quality.

3.7.Stemming and Lemmatization

Stemmingandlemmatization are text normalization techniques in NLP used to reduce words to their root form(Sekhar, 2024). Tokens can be words, parts of words, or even whole sentences, depending on how you break down the text. This step is really important in natural language processing (NLP) because it turns messy text into a structure that computers can work with. It

makes analyzing text easier and simplifies how NLP models operate. Stemming and lemmatization are two key ways to prepare text for these models, as they both cut words down to their basic forms(Niveditha, 2021). Stemming chops words based on simple rules, which can sometimes make mistakes, while lemmatization relies on language knowledge to give more accurate results.For this study, we used Lemmatizationfortext normalization. In our study, we used it to preserve clinical meaning and context. In the medical domain, maintaining the correct root word is critical. Stemming is often too aggressive and can result in non-words or ambiguous stems that corrupt the clinical concept. Lemmatization ensures the semantic integrity of medical terms, which is essential for both the Qtype classifier and the final Answer generation.

3.8.Feature extraction

Feature extraction in NLP is a crucial, converting raw text into numerical representations that ML models can process(Ahidi et al., 2024). Traditional text representation techniques, such as Bag of Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF), represent text based on word occurrence and statistical importance, whereas embedding-based approaches, such as Word2Vec, GloVe, and BERT, capture semantic relationships and, in the case of contextual models, variations in meaning across different contexts. The selection of a text representation technique depends on the characteristics of the dataset, the prediction task, and the requirements of the learning algorithm. In this study, TF-IDF vectorization was used to transform the preprocessed medical text into numerical feature vectors for conventional machine learning algorithms. TF-IDF assigns weights to terms according to their frequency within a document and their distribution across the document collection. Terms that occur frequently in a particular question but less frequently throughout the dataset receive relatively higher weights, helping the model distinguish questions based on informative words and phrases. The TF-IDF representation was constructed using the training dataset, and the fitted vectorizer was subsequently applied to the validation and test datasets to prevent information leakage. The resulting sparse feature vectors were used as input representations for the conventional machine learning models.

For the Long Short-Term Memory (LSTM) model, the processed medical text was represented as a sequence of tokens rather than a conventional TF-IDF feature vector. First, a tokenizer was fitted to the training text to construct a vocabulary and map each word or token to a unique

integer index. Each question was then converted into a numerical sequence according to the vocabulary. To ensure consistent input dimensions, the sequences were padded or truncated to a predefined maximum length. The resulting token sequences were supplied to an embedding layer, which mapped the token indices to dense vector representations for sequential processing by the LSTM network. This representation enabled the model to learn patterns associated with token order and contextual dependencies within medical questions. The vocabulary and all preprocessing parameters were determined using the training data and then applied consistently to the validation and test sets.

3.9.Model Selection

After completing data preprocessing, ensemble machine learning models and deep learning models, namely random forest, extra tree, MLP, LSTM, extreme gradient boosting, and hybrid models, are selected to build a multi-output advisory model. Models are chosen for their effectiveness and adaptability in handling various types of data and generating precise predictions. These algorithms are popular due to their ability to produce accurate results and their versatility in handling complex relationships within the data.

Random Forest: As noted in (Zhao & Ye, 2024)**,** Random Forest trees are constructed independently using a random subset of features. They can handle categorical, NLP and continuous data, making them suitable for tabular datasets. Random Forest models are robust against over-fitting and demonstrate good generalization to unseen data, particularly in high-dimensional datasets.

Extreme Gradient Boosting (XGBoost): As pointed out by (Alshboul et al., 2022)**,**The XGBoost algorithm, which can work on both categorical and numerical data, is perfect for any NLP dataset. XGBoost uses regularization to make the model generalize better and to prevent overfitting. Gradient boosting can be used for a large number of features, as in this case, and it insulates you from the curse of dimensionality through a combination of "weak" learners to arrive at one "super-learner". XGBoost contains builtinregularisation techniques to avoid over-fitting, which makes sense if you have a high-dimensional dataset. XGBoost offers feature importance analysis to identify the most correlated features in datasets.

Extra Tree: As noted in (Ampomah et al., 2020)**,** The random feature selection helps to reduce over-fitting by introducing randomness and diversity among the decision trees. It can handle high-dimensional datasets efficiently by randomly selecting feature splits, reducing the impact of irrelevant or noisy features. Extra Tree is less prone to over-fitting compared to traditional decision trees, making it suitable for high-dimensional datasets. Extra Tree is suitable for NLP datasets as it can handle both categorical and textual data.

Long Short-Term Memory (LSTM): As noted in (Muhuri et al., 2020), LSTM has various benefits for classification tasks. It can learn intricate patterns and relationships in the data, leading to more precise predictions and classifications. Moreover, their capacity to manage sequences of varying lengths makes them versatile and suitable for classification issues. LSTM provides enhanced performance and flexibility over conventional machine learning models for classification tasks.

Multilayer Perceptron (MLP): pointed out by(Orrù et al., 2020) MLP is suitable for classification tasks. It is a type of feed forward artificial neural network that can be used to classify input data into different categories. The MLP consists of multiple layers of nodes, each connected to the next layer, and uses non-linear activation functions to process the input data. With proper training and optimization, the MLP model can effectively classify data into distinct classes, making it a powerful tool for various classification tasks.

Hybrid model:A hybrid model is more important to increase the performance of a model than an individual model (Meena et al., 2023). In our study, to optimize performance, we have developed a hybrid model that leverages the strengths of two top-performing models, MLP and Extra Tree, aiming to create a more robust and accurate predictive model. Through this hybrid approach, we can harness the strengths of each model to mitigate its weaknesses, resulting in a more reliable predictive model. By combining the MLP and extra tree models, we can leverage their complementary strengths to enhance the performance of the model.

How does the hybrid model work explained as follows

To construct the hybrid architecture, we combined an Extra Trees Sequential Ensemble Method with an MLP model. First, we define ExtraTreesClassifier and MLPClassifier models from scikit-learn with appropriate parameters. Next, we directly utilize the ExtraTreesClassifier for

feature importance or predictions as a meaningful transformation process, avoiding a FunctionTransformer. If feature augmentation is desired, we extract the Extra Trees model outputs to enhance the feature set before input to the MLP model. Subsequently, we construct a pipeline to sequentially combine the models. This pipeline includes the Extra Trees classifier processing input data, potentially yielding new feature representations for the MLP classifier. After defining the pipeline, we train the hybrid model using the fit method and training data. Once trained, we generate predictions on the test set using the pipeline's predict() method. Finally, we evaluate performance by generating and displaying a classification report using the classification_report function, which provides precision, recall, F1 score, and support for each class. This approach ensures a structured, reproducible, and clear implementation of the hybrid model.

3.10.Model development

We used three features for healthcare medical diagnosis. Experimental scenarios were conducted to assess prediction performance using these features. Six experiments namely Random Forest, Extra Tree, MLP, LSTM, hybrid model and Extreme Gradient Boosting were conducted. Training the model necessitated a dataset comprising 48,756 instances with 12 predefined categories of healthcare question type. To construct advisory model, two phases are required those are training and testing (Mavaie et al., 2023). During training, the model gets shown the same data over and over to learn how to respond correctly. Then, in the testing phase, it’s given new data it hasn’t seen before to check how well it performs (Mehrpour et al., 2023). Therefore, the advisory model should be developed by partitioning the complete dataset into 80% for training and 20% for testing subsets. We created a model that can handle multiple outputs. To check how well the model performed, we used various metrics for both classification and question answering.

3.11.Model Evaluation

Model evaluation is a crucial step in the machine learning process, as it allows to assess the performance and effectiveness of a model (Editor Wolfgang Walz, 2023). The predictive model is evaluated using objective-based evaluations. The most common objective-based evaluation metrics used include accuracy, precision, recall, cross-validation, confusion matrix, and fl-score.This advisory model is structured with a dual-target approach for each question it

processes. This means that for any given question, the model aims to predict two distinct target classes. The first of these target classes pertains to the question type itself, essentially categorizing the question based on its nature. The second target class, on the other hand, focuses on generating the answer to the question. Given this dual-target architecture, the evaluation strategy is also multifaceted, employing different sets of metrics tailored to each target class. Specifically, for the question type classification task, which involves assigning each question to its category, we leverage a suite of objective-based evaluation metrics that are widely recognized and used in the field of classification The most common objective-based evaluation metrics used include accuracy, precision, recall, cross-validation, confusion matrix, cross-validation, and fl-score.A confusion matrix is a vital tool for evaluating the classification performance of a model by summarizing predicted and actual values(Meena et al., 2023). Analyzing the matrix allows for the calculation of accuracy, precision, recall, and F1 score, which are crucial metrics for model evaluation (Düntsch & Gediga, 2019). Regular review and interpretation of the matrix helps identify model weaknesses for necessary improvements. The matrix displays actual and predicted classes in rows and columns, with diagonal elements showing correct predictions. Off-diagonal elements reveal frequently misclassified classes, highlighting areas for model enhancement. The confusion matrix allows for analysis of the model's performance for each class (Togunwa et al., 2023).

- ✓ Identify classes with high misclassification rates (high values off the diagonal)
- ✓ See how often the model confuses specific classes (e.g., Class 1 and Class 2)

By analyzing these errors, it is possible to improve model's performance by tuning the classification algorithm or collecting more data to better differentiate between classes.The following Table 3 is an example of a confusion matrix for a study that has three classes

Table 3: an example of a confusion matrix for a study that has three classes

		Predicted Class		
		Class 1	Class 2	Class 3
Actual Class	Class 1	C_{11}	C_{12}	C_{13}
	Class 2	C_{21}	C_{22}	C_{23}
	Class 3	C_{31}	C_{32}	C_{33}

- ✓ C11, C22, and C33 represent the number of correctly classified instances for each class.
- ✓ Others represent the number of instances that were misclassified between the respective classes.

As per the learning result correctly or incorrectly classified instances are further grouped as the following (Düntsch & Gediga, 2019).

- ✓ **True positives (TP)** - True positives are the cases when the actual class of the data point is positive and the predicted class is also positive it means this represents when the actual value and predicted value should be the same. E.g. from the above table C11, C22, and C33 are true positives. So when we are concerned Class 1 C11 is true positive.
- ✓ **False positives (FP)** - False positives occur when the actual class is negative but the predicted class is positive or the sum of values of the corresponding column except for the TP value. E.g. for Class 1false positives are C21 and C31.
- ✓ **False negatives (FN)**-False negatives happen when the actual class is positive but the predicted class is negative which means the sum of values of the associated rows except for the TP value. E.g. for Class 1 C12 and C13 are false negatives.
- ✓ **True negatives (TN)**- True negatives are cases when the actual class is negative and the predicted class is also negative it means The total of all columns and rows excluding the values of the class for TP, FP, and FN. E.g. for Class 1 C22,C23,C32, and C33 is TN.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP+TN+FP+FN)} \text{ ---------- (1)}$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \text{---------- (2)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \text{----------3)}$$

$$F1 - \text{score} = \frac{(2 * \text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \text{---------- (4)}$$

To evaluate the model's performance on the question answering task, we employ a different set of performance metrics that are more suitable for assessing the quality of generated text. These metrics are designed to capture the semantic similarity between the model's generated answer and the ground truth answer(Mastakouri et al., 2025). Among these metrics are BERTScore Precision, BERTScore Recall, and BERTScore (Avg F1), and Confidence Scores. These metrics

leverage the power of BERT, a pre-trained language model, to compute a similarity score between the generated and ground truth answers, taking into account both lexical and semantic aspects. BERTScore Precision measures the extent to which the words in the generated answer are also present in the ground truth answer, while BERTScore Recall measures the extent to which the words in the ground truth answer are also present in the generated answer. BERTScore (Avg F1) provides a balanced measure of the overall similarity between the two answers, combining both precision and recall.Confidence scores reflect a machine learning model's certainty in its predictions. In question answering models, these scores are essential for evaluating answer reliability.

Factors Influencing Confidence Scores-Several factors can affect the confidence score of a model:

- ✓ **Input Quality**: The presence of handwriting, the resolution of images, or the quality of scanned documents can significantly impact the score. Poor quality inputs may lead to lower confidence scores.
- ✓ **Model Performance**: The inherent capabilities of the model, including how well it has been trained on relevant data, also influence the confidence score.
- ✓ **Prompt Clarity**: The specificity and clarity of the input prompt can affect the model's ability to generate a confident response. Vague or ambiguous prompts may result in lower confidence scores.

3.12.Hyper parameter tuning

To ensure optimal performance, all individual machine learning and deep learning algorithms including Extra Trees, Random Forest, Extreme Gradient Boosting, Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM) were subjected to a comprehensive hyperparameter tuning process. Hyper-parameters are those settings that you choose before you start training a model. They aren’t learned from the data itself. The aim of tuning these parameters is to find the right mix that helps the model perform its best for a specific task (Hossain & Timmer, 2021).Adjusting the hyper parameters can improve the performance of the model and ultimately achieve better results (Shi et al., 2023). Hyperparameter tuning methods include grid search, random search, and Bayesian optimization. Grid search exhaustively checks all combinations random search samples them. Bayesian optimization uses models to guide the search. We used grid search to enhance our model, a traditional yet computationally intensive approach. The

optimization utilized Grid Search with 5-Fold Cross-Validation (CV=5) on the training dataset (80% of the total data).

Table 4: Hyper parameter for the selected model

No	Algorithm	Hyper Parameter name	Hyper parameter grid	Best parameter
1	Random forest	n_estimators	[100, 500, 1000]	500
		Criterion	['gini', 'entropy']	Entropy
		max_depth	[None, 10]	None
		min_samples_split	[5, 10]	5
		min_samples_leaf	[2, 5]	2
		max_features	['auto', 'sqrt']	Auto
2	Extreme gradient boosting	n_estimators	[100, 500, 1000]	100
		learning_rate	[0.1,0.3, 0.5,0.7, 1.0]	0.5
		max_depth	[None, 10]	None
		Subsample	[0.9, 0.8, 0.7]	0.9
		Gamma	[0, 0.1, 0.5]	0.1
		min_chlid_weight	[1, 3, 5]	3
3	Extra tree	n_estimators	[100, 500, 1000]	1000
		Criterion	['gini', 'entropy']	Gini
		min_samples_split	[5, 10]	5
		min_samples_leaf	[2, 5]	2
		max_depth	[None, 10]	None
		max_features	['auto', 'sqrt']	Auto
4	LSTM	Number of layers	[3, 4, 5]	5
		Activation Function	[Softmax ,relu]	Relu
		Learning rate	[0.001, 0.0001, 0.00001]	0.001
		Batch size	[8, 16, 32]	32
		lstm_units	[64, 128]	64
		return_sequences	[True, False]	True
		Dropout rate	[0.2, 0.3, 0.4]	0.2
5	MLP	Activation Function	[Softmax ,relu]	Softmax
		Learning Rate	[0.001, 0.0001, 0.00001]	0.001
		Number of layers	[3, 4, 5]	5
		hidden_layer_sizes	[50,100]	50
		Dropout rate	[0.2, 0.3, 0.4]	0.3
		l2_regs	[0.001, 0.0001]	0.001
		num_dense_layers	[1, 2]	2
		Batch size	[8, 16, 32]	8

4.Experimentation

In this study, we construct an advisory model to explorethe Design and Development of an intelligent multi-output Hybrid Advisory Model for Medical Diagnosis. Here, as a hardware tool, we use an Intel(R) Core(TM) i7-5300U CPU, clocked at 2.70 GHz, with 16 GB of memory and a 1 TB hard disk. Similarly, as software tools, we used Microsoft Office Word 2023 for writing documentation and Microsoft Office Excel 2023 for understanding the datasets manually. This research is implemented using Anaconda Navigator.

We used three features for healthcare medical diagnosis. Experimental scenarios were conducted to assess prediction performance using these features. Six experiments namely Random Forest, Extra Tree, MLP, LSTM, Extreme Gradient Boosting, and hybrid model were conducted. Training the model necessitated a dataset comprising 48,756 instances with 12 predefined categories of healthcare question type. This advisory model developed by partitioning the complete dataset into training and testing subsets. Out of the total datasets, we split them so that 80% was for training and 20% for testing.The figure 1 shows a bar chart that clearly presents the distribution of various question types. You can see that information questions are the most common, while symptoms and treatment questions also make up a good chunk. Other types, like stages, are not as frequent. The chart is easy to read, well-labeled, and the colors help to tell the different categories apart. The figure displayed is from descriptive analysis of the dataset used for model development.

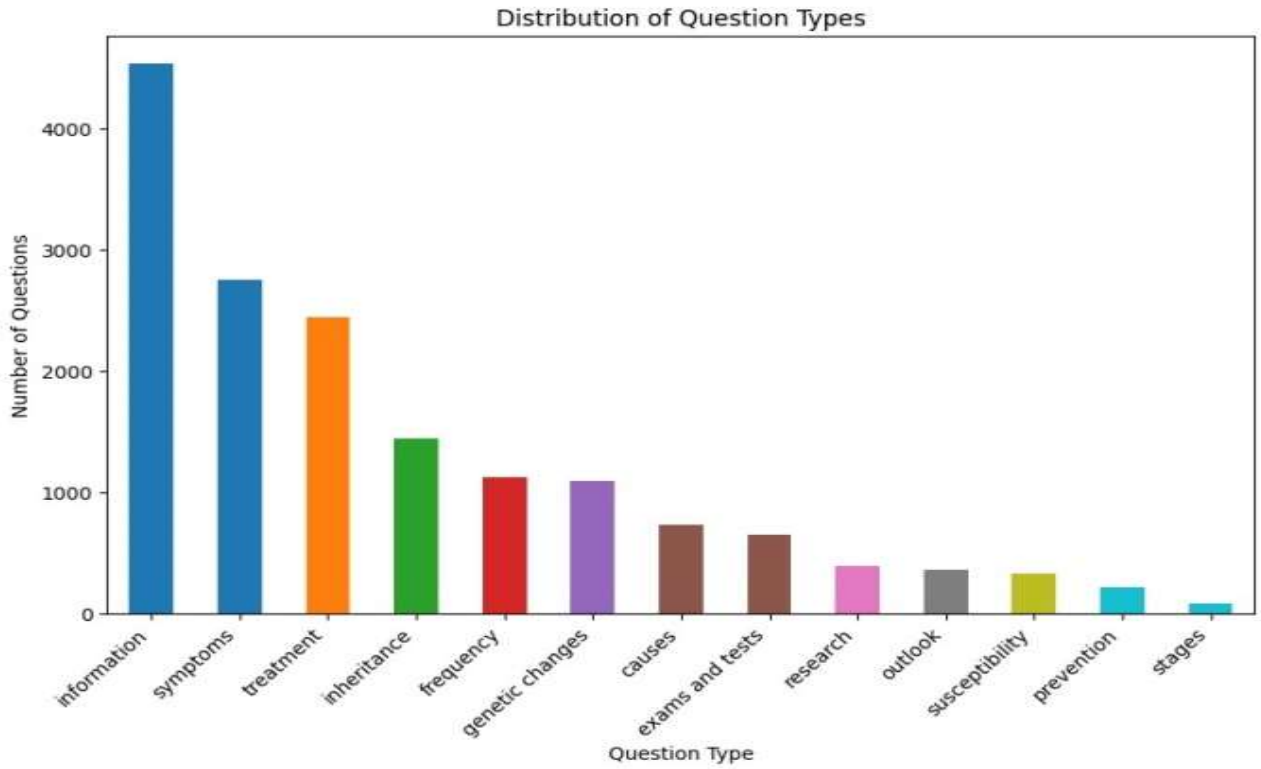


Fig.1. Distribution of 12 healthcare question categories with respective sample counts that used for model development it collected from different trusted database and website

In model evaluation, Fig. 2 serves as a Baseline Benchmark, illustrating the performance of a single-algorithm approach (Random Forest), allowing for a direct comparison of how "semantic understanding" changes with different architectures. This benchmark provides a clear, easily interpretable metric for assessing architectural improvements. Furthermore, the consistent application of Random Forest across all architectures mitigates potential biases introduced by algorithm-specific optimizations, ensuring a fair and reliable evaluation. The choice of Random Forest is motivated by its inherent robustness, ease of implementation, and relatively low computational cost, making it an ideal candidate for establishing a stable baseline. This allows us to isolate the impact of architectural variations on the model's ability to capture and utilize semantic information effectively.

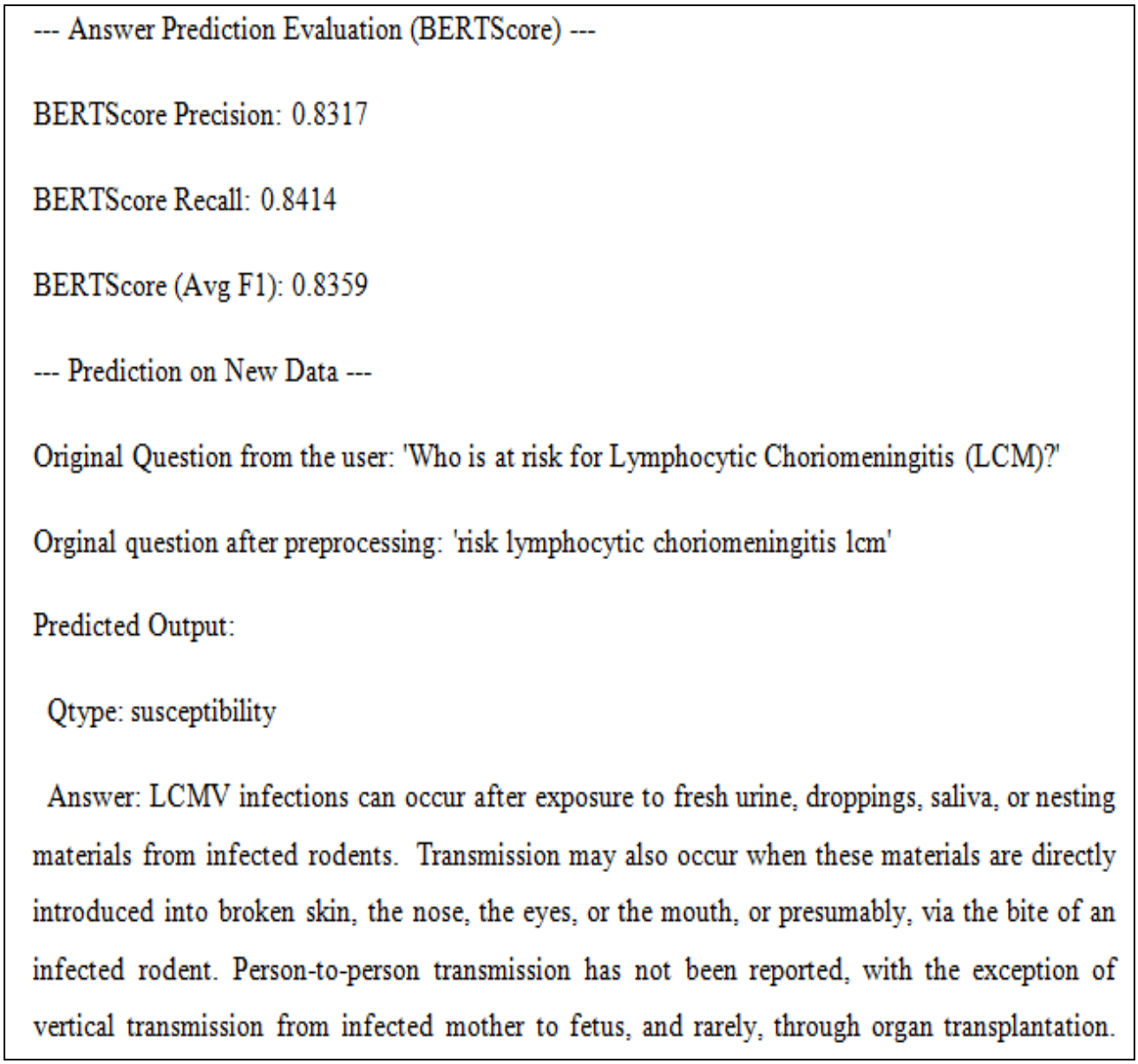


Fig.2.the question answering performance of random forest using BERTscore captures the semantic similarity between the model's generated answer and the ground truth answer

Fig. 3 is the Trust Meter. It validates that the Extra Trees component provides a reliable foundation before being fused into the final Hybrid model. This process helps the final model inherit trustworthiness. The Trust Meter constantly monitors the Extra Trees component's performance on a held-out validation set. If this performance drops below a defined threshold, the Trust Meter triggers an alert. This signals a potential reliability issue. Proactive intervention then becomes possible, preventing deployment of a compromised model. Metrics used in the

Trust Meter include precision, recall, and F1-score. They offer a comprehensive assessment of predictive capabilities.

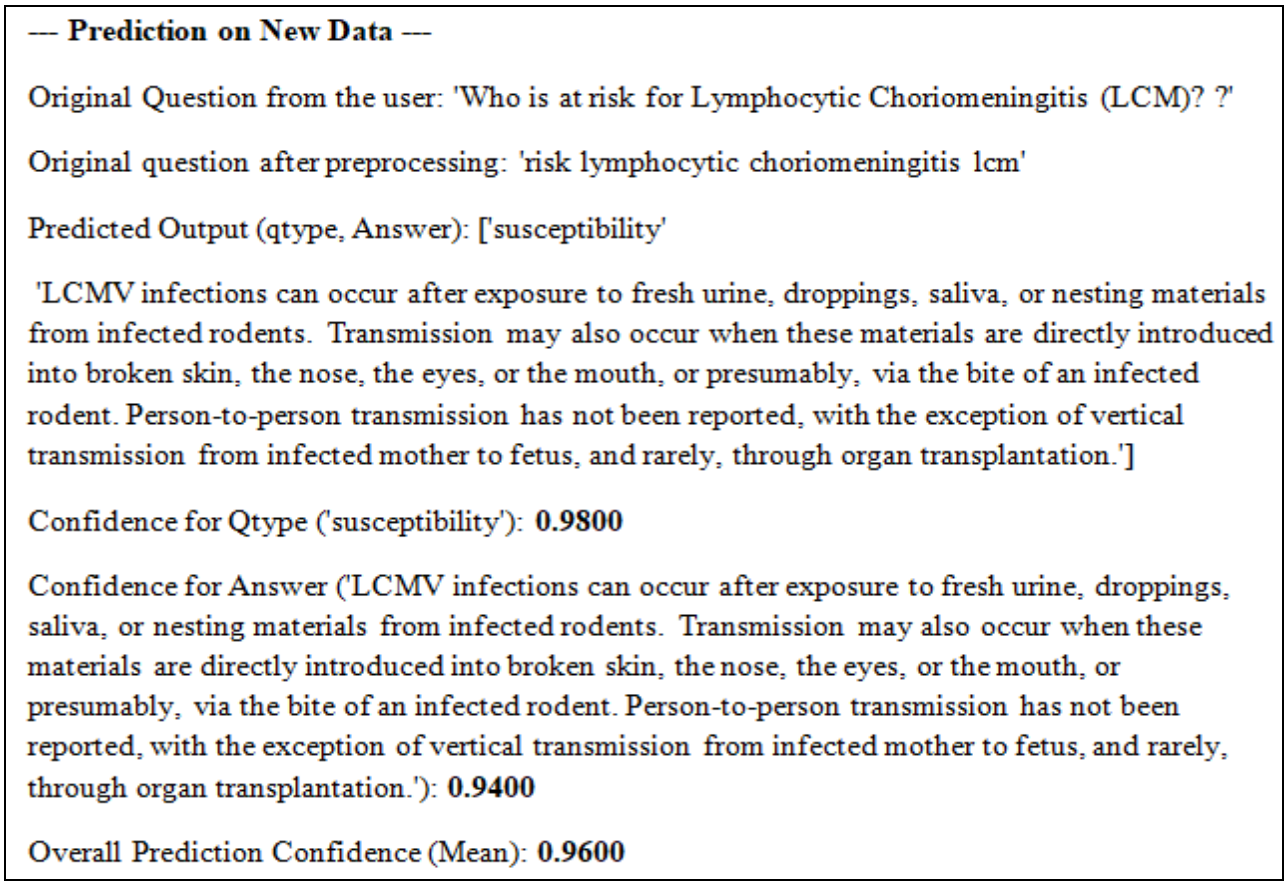


Fig.3.the question answering performance of an extra tree model based on confidence score, these scores are essential for evaluating answer reliability.

How does the hybrid model work explained as follows

First, define the extra trees and multi-layer perceptron (MLP) classifier modelusing the ExtraTreesClassifier and MLPClassifier class from scikit-learn. The models are configured with various parameters. Second, create Extra Trees Transformera FunctionTransformer is created without any specific transformation function. This transformer is used to incorporate the Extra tree model into the pipeline, allowing it to be seamlessly integrated with other models. Third, construct a Pipeline the Pipeline class is instantiated, which enables the combination of multiple models into a single pipeline. The pipeline is defined with a list of steps, where each step consists of a name and a model/transformer. Fourthly, the hybrid model(pipeline) is trained using the fit method, which takes the training data as input. This process trains the hybrid model

by fitting the extra trees transformer and the MLP model to the provided training data. After training the hybrid model, predictions are made on the test set using the predict () method of the pipeline. It is followed by model evaluation. Finally, generates a classification report using the classification_report function, which provides detailed metrics such as precision, recall, F1 score, and support for each class in the test set. The classification report is printed using the print statement.

Cross-validation is a performance evaluation metric that used to assess the performance and generalization ability of a model. It involves dividing the dataset into multiple subsets (k-fold) (Zhang et al., 2024). In k-fold cross-validation, the dataset is divided into k equal-sized folds. The model is then trained and evaluated k times, with each fold serving as the validation set once while the remaining k-1 folds are used for training. The performance metrics obtained from each iteration are averaged to get an overall performance of the model. Cross-validation is important because it gives a more reliable estimate of the model's performance, helps detect over-fitting, and allows for tuning the model's hyper-parameters by comparing the performance of different parameter settings. The k-fold cross-validation performance of the hybrid model, as depicted in the figure 4, is very good. It demonstrates high accuracy and generally good stability across different data splits, suggesting strong generalizability. The model appears robust, but a closer look at the data in the lower-performing folds could provide insights for further optimization. The model had optimal performance with k=5 for cross-validation.

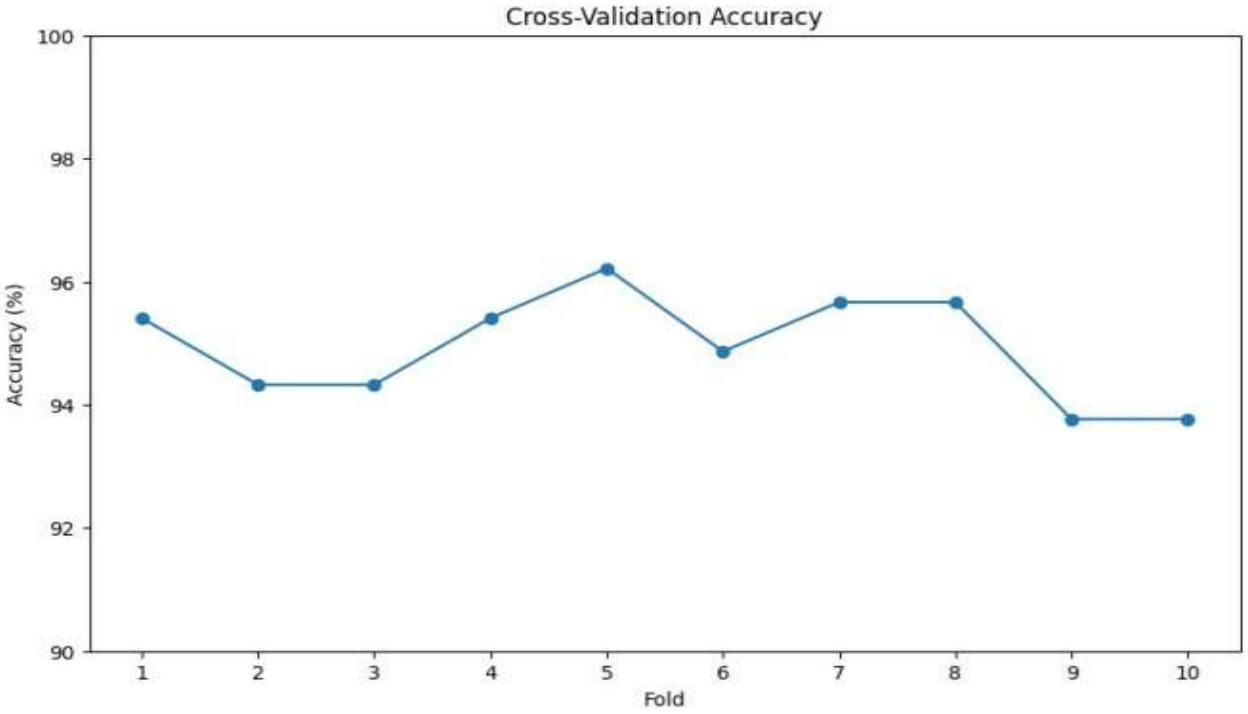


Fig.4. thehybrid model cross-validation performance, assess the performance and generalization ability of a hybrid model

As illustrated in the Figure 5 below,hybrid modelachieved 96% average accuracy.

Classification Report for hybrid model:

	precision	recall	f1-score	support
causes	0.98	0.96	0.97	333
exams and tests	0.99	0.99	0.99	438
frequency	0.98	0.99	0.99	554
genetic changes	0.99	0.98	0.97	211
information	0.97	0.96	0.98	4220
inheritance	0.96	0.97	0.95	278
outlook	0.94	0.95	0.95	60
prevention	0.97	0.93	0.96	141
research	0.99	0.98	0.96	250
susceptibility	0.95	0.94	0.97	186
symptoms	0.93	0.94	0.95	1639
treatment	0.94	0.95	0.93	1442
accuracy			0.97	9752
macro avg	0.97	0.96	0.96	9752
weighted avg	0.96	0.94	0.93	9752

Fig.5. Hybrid model classification report: evaluate per-class and overall model performance of a hybrid model.

The image shows a confusion matrix for a hybrid model with 11 classes. The matrix visualizes the performance of the model by showing the counts of true positive, true negative, false positive and false negative predictions.

4.1.Model Comparison

Comparing the performance of the models can determine which algorithm is best suited for a particular problem this involves training multiple models on the same dataset and evaluating their performance using metrics (Plaia et al., 2022). The researchers did six experiments to create an advisory model for healthcare medical diagnosis by compare its performance to select the optimal model.

Table 5Model Comparison

Evaluation metrics	**Extra tree**	**Random forest**	**XGBoosting**	**LSTM**	**MLP**	**Hybrid**
Accuracy	95%	86%	91%	92%	94%	97%
Precision	96%	91%	94%	92%	94%	98%
Recall	95%	90%	93%	92%	95%	97%
F1_score	96%	91%	91%	92%	96%	98%
BERTScore (Avg F1)	93%	84%	84%	91%	92%	94%
Confidence score	96%	88%	92%	93%	95%	97%

Then, the researchers used all performance metrics to evaluate and compare advisory models. As shown inTable 5, experiments were conducted on various models to diagnosis healthcare medical diagnosis. The hybrid model exhibited superior performance, leading to its selection as the optimal model.

4.2.Hybrid Model Superiority Performance Explanation

The hybrid model, which combines the strengths of the Extra Trees and Multi-Layer Perceptron (MLP) models, likely outperformed the other algorithms across all metrics due to the use of the hybrid learning approach. By mitigating the weaknesses of its constituent models, the hybrid

model achieves more robust, accurate predictions. This synergistic effect enables a more nuanced understanding of the data's underlying patterns, leading to superior generalization and predictive performance. In contrast, the individual Extra Trees and MLP models, while capable on their own, may be more susceptible to overfitting or failing to capture certain complex relationships. Other algorithms, lacking the combined advantages of the hybrid approach, likely exhibited limitations in handling the data's intricacies. Therefore, the hybrid model's success underscores the value of ensemble methods in leveraging diverse modeling techniques to achieve enhanced predictive accuracy and robustness. Table 6 is the evidence locker. It transforms the model from a theoretical concept into a validated solution by showing it is more accurate, more reliable, and safer than the standard models currently available.

Here is a brief explanation for the hybrid model's high performance in each metric listed in Table 4.

Table 6:hybrid model classification performance based on objective-based evaluation metrics, to evaluate its performance in question type prediction.

Metric	Hybrid Model Performance (vs. Others)	Design Explanation
Accuracy (97%)	Highest (vs. Extra Tree's 95%, MLP's 94%, etc.)	The combination of a strong tree-based model (Extra Trees) for feature-space partitioning and a powerful neural network (MLP) for non-linear pattern recognition reduces overall misclassification errors, leading to the highest rate of correct predictions.
Precision (98%)	Highest (vs. Extra Tree's 96%, MLP's 94%, etc.)	High precision suggests the model has a very low rate of False Positives (incorrectly predicting a question type or answer). The collective intelligence of the hybrid design ensures that when it makes a positive prediction, it is highly likely to be correct.
Recall (97%)	Highest (vs. Extra Tree's 95%, MLP's 95%, etc.)	High recall indicates the model is excellent at identifying most of the relevant cases, meaning a low rate of False Negatives (failing to identify a question type or answer when it should have). This is critical for medical diagnosis.
F1-Score (98%)	Highest (vs. Extra Tree's 96%, MLP's 96%, etc.)	The F1-score is the harmonic mean of Precision and Recall. The hybrid model's top scores in both components naturally result in the highest F1-score, confirming its overall balance and reliability for the classification task.

Table 7 is the Quality Control check. It moves the conversation from "Does the model work?" to "How good is the model's intelligence?" It confirms the hybrid model is both smart (BERTScore) and sure of itself (Confidence Score).

Table 7hybrid model question answering performance metrics captures the semantic similarity between the model's generated answer and the ground truth answer (BERTscore), and evaluates answer reliability (confidence score)

Metric	Hybrid Model Performance (vs. Others)	Design Explanation
BERTScore (Avg F1) (94%)	Highest (vs. Extra Tree's 93%, MLP's 92%, etc.)	BERTScore measures the semantic similarity between the predicted and ground truth answers using a pre-trained language model (BERT). The MLP component, being a neural network, is effective at learning complex, non-linear representations, which likely helps in generating an answer with a better semantic match to the expert response, thus achieving a higher BERTScore.
Confidence Score (97%)	Highest (vs. Extra Tree's 96%, MLP's 95%, etc.)	The confidence score reflects the model's certainty in its predictions. By combining the outputs of two different, high-performing models (Extra Trees and MLP) to make a single, fused prediction, the ensemble approach provides a more stable and verifiable prediction, which in turn leads to a consistently higher certainty level.

4.3.Domain Expert evaluation and user acceptance testing

Twenty domain experts, medical Doctors, were involved in assessing the model's effectiveness, efficiency, and usefulness in medical diagnosis to enhance accuracy and timeliness. The experts were chosen from Debark and Gondar hospital. Following a demonstration of the adversary model for medical diagnosis, the experts provided feedback based on set evaluation criteria aligned with the research objectives and questions. The evaluation concluded with domain

experts assigning a rating from 1 (poor) to 5 (excellent) based on predefined criteria. These ratings represent the number of evaluators who assessed the prototype as poor, fair, good, very good, or excellent for each criterion. In summary, the study achieved 84% of the domain expert's acceptance, signifying that the designed model is very effective at medical diagnosis. Ultimately, the user recommends that healthcare centers and users utilize and implement the findings for a multi-output hybrid advisory model in medical diagnosis.

5.Discussion

This study developed and evaluated an NLP-based multi-output healthcare medical advisory model using 48,756 instances and three principal textual variables: question type (Qtype), Question, and Answer. Six modeling approaches—Random Forest, Extreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP), Extra Trees, and a hybrid Extra Trees–MLP model—were investigated. The discussion of the findings is organized around the three research questions and considers the results in relation to previous studies, methodological differences, strengths, and limitations.

RQ1: Which algorithm provides the best performance for medical question classification?

The comparative experiments indicated that the proposed hybrid Extra Trees–MLP model achieved the strongest performance among the evaluated approaches. This finding suggests that combining complementary learning mechanisms may be beneficial for the present medical text classification task. Extra Trees can capture nonlinear relationships and interactions among extracted textual features, while the MLP can learn higher-level nonlinear representations from the features supplied to it. The observed improvement should therefore be interpreted as an empirical finding specific to the dataset, feature representation, and experimental configuration used in this study rather than as evidence that hybrid models are universally superior to individual algorithms.

This finding is broadly consistent with the literature showing interest in combining AI approaches for healthcare decision support. Aityan et al. (2024), for example, investigated the use of multiple large language models for emergency medical decision support and reported benefits from combining models rather than relying on a single model. However, the methodological settings are substantially different: their work involved LLM-based emergency decision support, whereas the present study uses conventional machine learning and neural-network components for medical question classification and answer prediction. Consequently,

the two studies provide complementary evidence for investigating model combinations, but their performance results should not be compared directly.

The finding also extends the work of Muscat et al. (2020), who analyzed 126 clinical questions submitted by 53 Australian general practitioners and found treatment, diagnosis, and epidemiology to be among the most common question categories. Their study used content analysis and established classification systems to characterize clinical information needs rather than developing an automated predictive model. The present study moves from descriptive classification of clinical questions toward automated prediction of Qtype and its associated answer representation. Nevertheless, the two studies address different objectives, and the present model should not be interpreted as replacing the qualitative or content-analysis approach used by Muscat et al.

The results should also be considered in light of broader evidence on AI in healthcare. Alowais et al. (2023) describe applications of AI across diagnosis, treatment, and clinical practice while emphasizing implementation, trust, and responsible use. Similarly, Ramgopal et al. (2023) emphasize that AI-based clinical decision-support systems need to be scientifically sound, clinically relevant, appropriately integrated into care pathways, and capable of addressing clinically meaningful questions. They also identify limitations related to data availability and generalizability. These considerations are relevant to the present model because superior performance on a retrospective dataset does not by itself establish clinical effectiveness.

RQ2: To what extent does the proposed model accurately perform the intended medical prediction task?

The proposed hybrid model achieved an average accuracy of 97% in the reported experiments. This result indicates strong predictive performance on the evaluation data under the experimental conditions of this study. However, the result should be interpreted according to the exact prediction target. If the model predicts medical question categories and answer categories, the result should be described as classification or advisory prediction accuracy, rather than as diagnostic accuracy. The term “diagnostic accuracy” should only be used if the model's target variable represents an actual clinical diagnosis.

The 97% accuracy should also not be directly described as superior to the 94% reported by Ahmed and Islam (2025) unless both studies use comparable datasets, target variables, evaluation procedures, and metrics. Their study concerns explainable deep-learning applications in medical diagnosis, whereas the present study focuses on a multi-output medical advisory task. Differences in prediction objectives and datasets mean that the numerical accuracy values are not directly equivalent. Therefore, the more appropriate interpretation is that the present study obtained 97% accuracy within its own experimental setting, while previous studies have reported different performance levels for different medical AI tasks.

The size of the present dataset is nevertheless an important methodological strength. The study uses 48,756 instances collected from multiple sources, potentially providing broader linguistic and medical coverage than a small single-source dataset. This is relevant to the concerns raised by Ramgopal et al. (2023), who note that sufficiently large datasets can be a challenge in developing and validating AI-based clinical decision-support models. However, a larger dataset does not automatically guarantee greater generalizability. The quality, independence, representativeness, labeling consistency, source diversity, and duplication structure of the records are also important. In particular, combining multiple online sources may introduce differences in writing style, terminology, answer structure, and source-specific biases.

Accordingly, the 97% result should be considered evidence of strong performance under the study conditions rather than evidence that the model is ready for clinical deployment. External validation using independent medical datasets and assessment by healthcare professionals would be necessary before determining whether the model can provide reliable clinical decision support. This is consistent with the broader clinical AI literature, which emphasizes that AI systems must be appropriately validated and integrated into clinical workflows and should support rather than replace professional judgment.

RQ3: How accurately does the proposed hybrid model classify medical questions and generate corresponding answers?

The third research question evaluates the multi-output nature of the proposed advisory model. The study considers both question-type classification and the prediction of corresponding

answers, requiring evaluation measures appropriate to each output. For question-type classification, accuracy, precision, recall, F1-score, and cross-validation provide complementary information about predictive performance. In particular, macro-averaged precision, recall, and F1-score are important when the 12 question-type classes are not equally represented because overall accuracy can conceal poor performance on less frequent categories.

For answer prediction, the evaluation approach should be interpreted according to whether the Answer output represents a predefined class or generated text. If the model selects predefined answer categories, classification metrics are appropriate. If it generates natural-language answers, metrics such as BERT-based similarity can provide information about semantic correspondence, but they do not establish medical factuality or clinical safety on their own. Expert evaluation is therefore important for determining whether generated responses are medically correct, relevant, complete, and appropriately expressed.

The multi-output formulation represents an important distinction from studies that only classify clinical questions. Muscat et al. (2020), for example, demonstrated that analyzing clinical questions can reveal important information needs among general practitioners, with treatment-related questions comprising the largest category in their dataset. The present study extends this general direction by attempting to automate both the identification of the question type and prediction of the corresponding answer. However, this should be described as an extension of the task rather than as proof that the model provides “precision decision support” in actual clinical practice. Clinical usefulness requires prospective and expert-based evaluation beyond computational performance.

The use of statistical comparisons, ablation analysis, and error analysis further strengthens the evaluation when these analyses have actually been performed. Statistical testing can determine whether observed differences between models are unlikely to have arisen from sampling variation, while ablation analysis can establish whether individual components of the hybrid architecture contribute to performance. Error analysis can additionally reveal difficult question types, ambiguous medical terminology, poorly represented classes, or systematic errors in answer prediction. These analyses are particularly important because a single aggregate accuracy value provides limited information about where and why a medical NLP model succeeds or fails.

6.Conclusion

This study developed and evaluated a multi-output hybrid NLP-based advisory model for medical question classification and corresponding answer prediction. Using a dataset of 48,756 instances, the study compared six approaches, including Random Forest, XGBoost, Extra Trees, LSTM, MLP, and a hybrid Extra Trees–MLP model. The results showed that the proposed hybrid model achieved the highest performance among the evaluated approaches, with an overall reported accuracy of 97% for the defined computational prediction task. These findings indicate that combining complementary machine learning approaches can be effective for the specific medical text classification and advisory task investigated in this study.The study contributes to medical NLP research by jointly considering medical question-type prediction and corresponding answer prediction rather than focusing exclusively on disease classification or question categorization. The use of different text representations and multiple machine learning architectures also provides a basis for comparing conventional machine learning, neural-network, and hybrid approaches. However, the reported performance should be interpreted within the context of the dataset, preprocessing procedures, feature representations, prediction targets, and evaluation methodology used in this study. Several limitations should be considered. The dataset consisted of secondary online medical question-and-answer information, which may differ from information collected during real clinical encounters. In addition, the use of data from multiple sources may introduce differences in terminology, answer structure, labeling practices, and data quality. Potential class imbalance and similarities or duplication among records may also affect model performance if not adequately controlled during preprocessing and dataset splitting. Furthermore, automated evaluation of answer prediction cannot by itself establish the factual correctness, clinical relevance, completeness, or safety of medical responses.Therefore, before the proposed model can be considered for practical clinical decision-support applications, additional validation is required. Future research should evaluate the model using independent and diverse clinical datasets from different institutions and healthcare settings to assess external validity and generalizability. Clinical experts should also assess the predicted answers for medical accuracy, relevance, completeness, and safety. Where natural-language answers are generated, evaluation should extend beyond automated similarity metrics to include expert-based assessment and clinically appropriate safety measures. Explainability techniques such as SHAP or LIME may additionally be investigated to provide greater insight into model predictions and

support responsible human oversight.Future studies should also investigate class-imbalance mitigation, robustness to changes in data distributions, error patterns across individual question types, and the effects of different dataset sizes and feature representations. Prospective or real-world usability studies could subsequently examine how such an advisory system interacts with clinical workflows. Ethical issues, including privacy, bias, transparency, accountability, and appropriate human oversight, should also be addressed before deployment.In summary, the proposed hybrid model demonstrated high computational performance on the evaluated medical question classification and multi-output advisory task, with a reported accuracy of 97%. The findings provide a basis for further research into hybrid NLP-based medical advisory systems, but external validation, expert assessment, and clinical safety evaluation are necessary before conclusions can be drawn about its effectiveness in real-world healthcare or clinical diagnosis.

Data availability

Some of the data (16,125 instances) used for this study is available on Kaggle and can be accessed at the following URL: https://www.kaggle.com/datasets/thedevastator/comprehensive-medical-q-a-dataset. The additional dataset (32,631 instances) was meticulously curated from 12 trusted websites of the National Institutes of Health (NIH), an official trusted website of the United States government. The dataset is available for medical research. These websites cover a wide range of health topics example, cancer.gov, niddk.nih.gov, MedlinePlus Health Topics to the Genetic and Rare Diseases Information Resource (GARD). The dataset collection covers different question types associated with diseases, drugs, and other medical entities such as tests. This variety reflects the diverse needs of individuals seeking medical information.

References

Ahidi, E., Lukwaro, E., Kalegele, K., & Nyambo, D. G. (2024). *A Review on NLP Techniques and Associated Challenges in Extracting Features from Education Data A Review on NLP Techniques and Associated Challenges in Extracting Features from Education Data*. *August*. https://doi.org/10.12785/ijcds/160170

Aityan, S. K., Mosaddegh, A., Herrero, R., Inchingolo, F., Nguyen, K. C. D., Balzanelli, M., Lazzaro, R., Iacovazzo, N., Cefalo, A., Carriero, L., Mersini, M., Legramante, J. M.,

Minieri, M., Santacroce, L., & Isacco, C. G. (2024). *Integrated AI Medical Emergency Diagnostics Advising System*.

Ajuzieogu, U. (2024). *AI Data Quality and Bias : Challenges , Implications , and Solutions in Modern Machine Learning*. *November*. https://doi.org/10.13140/RG.2.2.25830.02880

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., & Alshaya, A. I. (2023). *Revolutionizing healthcare : the role of artificial intelligence in clinical practice*. 1–15.

Alshboul, O., Shehadeh, A., Almasabha, G., & Almuflih, A. S. (2022). Extreme Gradient Boosting-Based Machine Learning Approach for Green Building Cost Prediction. *Sustainability (Switzerland)*, *14*(11). https://doi.org/10.3390/su14116651

Ampomah, E. K., Qin, Z., & Nyame, G. (2020). *Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement*.

Basha, M. J., Jayashankari, S. V. J., Hussein, A., & Pulatova, A. (2023). *Advancements in Natural Language Processing for Text Understanding*. *04031*.

Bayram, M. A., & Fincan, A. A. (2025). *Tokens with Meaning : A Hybrid Tokenization Approach for NLP Tokens with Meaning : A Hybrid Tokenization Approach for NLP*. 0–21.

Doreswamy, D., & Nigus, M. (2020). Feature Selection Methods for Household Food Insecurity Classification. *2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020*. https://doi.org/10.1109/ICCSEA49143.2020.9132945

Düntsch, I., & Gediga, G. (2019). Confusion Matrices and Rough Set Data Analysis. *Journal of Physics: Conference Series*, *1229*(1). https://doi.org/10.1088/1742-6596/1229/1/012055

Editor Wolfgang Walz, S. (2023). *Machine Learning for Brain Disorders* (Vol. 197). https://link.springer.com/10.1007/978-1-0716-3195-9

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research*, *9*(March), 1–17. https://doi.org/10.3389/fenrg.2021.652801

Hasan, M., Marjan, M. A., Uddin, M. P., Afjal, M. I., Kardy, S., Ma, S., & Nam, Y. (2023).

Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. *Frontiers in Plant Science*, *14*(August), 1–18. https://doi.org/10.3389/fpls.2023.1234555

Hossain, R., & Timmer, D. D. (2021). Machine learning model optimization with hyper parameter tuning approach. *Global Journal of Computer Science and Technology*, *21*(2), 7–13.

Ibrahim, A., Mirjalili, S., Zhang, Y., Elnazer, S., & Zaki, R. M. (2022). *Optimized Ensemble Algorithm for Predicting Metamaterial Antenna Parameters*. https://doi.org/10.32604/cmc.2022.023884

Kaur, J. (2021). *A Systematic Review on Stopword Removal Algorithms A Systematic Review on Stopword Removal Algorithms*. *April 2018*.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 1–45. https://doi.org/10.3390/e23010018

Mastakouri, A., Kirschbaum, E., & Kasiviswanathan, S. P. (2025). *OF L ANGUAGE M ODEL C ONFI* -. 1–24.

Mavaie, P., Holder, L., & Skinner, M. K. (2023). Hybrid deep learning approach to improve classification of low-volume high-dimensional data. *BMC Bioinformatics*, *24*(1), 1–20. https://doi.org/10.1186/s12859-023-05557-w

Meena, G., Kumar, K., Kumar, S., & Lokesh, K. (2023). A hybrid deep learning approach for detecting sentiment polarities and knowledge graph representation on monkeypox tweets. *Decision Analytics Journal*, *7*(April), 100243. https://doi.org/10.1016/j.dajour.2023.100243

Mehrpour, O., Saeedi, F., Nakhaee, S., Tavakkoli Khomeini, F., Hadianfar, A., Amirabadizadeh, A., & Hoyte, C. (2023). Comparison of decision tree with common machine learning models for prediction of biguanide and sulfonylurea poisoning in the United States: an analysis of the National Poison Data System. *BMC Medical Informatics and Decision Making*, *23*(1), 1–11. https://doi.org/10.1186/s12911-022-02095-y

Muhuri, P. S., Chatterjee, P., Yuan, X., Roy, K., & Esterline, A. (2020). Using a long short-term

memory recurrent neural network (LSTM-RNN) to classify network attacks. *Information (Switzerland)*, *11*(5), 1–21. https://doi.org/10.3390/INFO11050243

Muscat, D. M., Patel, P., Reid, S., Hoffmann, T., Albarqouni, L., & Trevena, L. (2020). *Content analysis of clinical questions from Australian general practice which are prioritised for answering : identifying common question types and perceived knowledge gaps*. *25*(1), 15–21.

Niveditha, N. M. (2021). *An Interpretation of Lemmatization and Stemming in Natural Language An Interpretation of Lemmatization and Stemming in Natural Language Processing*. *January*.

Ogunrinde, V. (2025). *Text Preprocessing Techniques in Python for Effective NLP Pipelines*. *May*.

Orrù, P. F., Zoccheddu, A., Sassu, L., Mattia, C., Cozza, R., & Arena, S. (2020). Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry. *Sustainability (Switzerland)*, *12*(11). https://doi.org/10.3390/su12114776

Pintelas, P., & Livieris, I. E. (2020). Special issue on ensemble learning and applications. *Algorithms*, *13*(6). https://doi.org/10.3390/A13060140

Plaia, A., Buscemi, S., Fürnkranz, J., & Mencía, E. L. (2022). Comparing Boosting and Bagging for Decision Trees of Rankings. *Journal of Classification*, *39*(1), 78–99. https://doi.org/10.1007/s00357-021-09397-2

Ramgopal, S., Sanchez-pinto, L. N., Horvat, C. M., Carroll, M. S., Luo, Y., & Florin, T. A. (2023). *Arti fi cial intelligence-based clinical decision support in pediatrics*. *June 2022*. https://doi.org/10.1038/s41390-022-02226-1

Rawas, S., Tafran, C., Alsaeed, D., & Al-ghreimil, N. (2024). *Transforming Healthcare : AI-NLP Fusion Framework for Precision Decision-Making and Personalized Care Optimization in the Era of IoMT*. https://doi.org/10.32604/cmc.2024.055307

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Sekhar, S. (2024). *Tokenization for Text Analysis*. *10*(1), 149–152.

Shi, C., Chiu, A. K., & Xu, H. (2023). Evaluating Designs for Hyperparameter Tuning in Deep Neural Networks. *The New England Journal of Statistics in Data Science*, *1*, 334–341. https://doi.org/10.51387/23-nejsds26

Togunwa, T. O., Babatunde, A. O., & Abdullah, K. U. R. (2023). Deep hybrid model for maternal health risk classification in pregnancy: synergy of ANN and random forest. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1213436

Zhang, Q., Ning, Z., Ding, X., Wu, J., Wang, Z., Tsangaratos, P., Ilia, I., Wang, Y., & Chen, W. (2024). *Landslide Susceptibility Mapping*.

Zhao, M., & Ye, N. (2024). High-Dimensional Ensemble Learning Classification: An Ensemble Learning Classification Algorithm Based on High-Dimensional Feature Space Reconstruction. *Applied Sciences*, *14*(5), 1956. https://doi.org/10.3390/app14051956

Author contributions

Abraham Alamirew: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Visualization, Writing – Original Draft, and Writing – Review & Editing.

YeshiwasAdane: Supervision, Validation, Methodology, Data Curation, Formal Analysis, and Writing – Review & Editing.

Birhanu Alemneh: Writing – Review & Editing, and manuscript preparation.

All authors contributed to the study, critically reviewed the manuscript, and read and approved the final version for publication.

Declarations

Competing interests

The authors declare no competing interests.

Ethics declaration

Ethical approval was not required for this study because it used publicly available secondary medical question-and-answer data that did not contain directly identifiable personal information. Permission was obtained from the relevant data providers/sources to use the datasets for research purposes. The study used the data solely for academic and research purposes and did not involve direct interaction with human participants or the collection of personally identifiable information.

Funding Declaration

The authors declare that no specific grant was received from any funding agency in the public, commercial, or not-for-profit sectors for the research reported in this article.