

AI-DxMH- Artificial Intelligence Diagnosis for Modern Health A PROJECT REPORT

Submitted by,

Mr. Mohammed Fahad Pasha F	20201CAI0093
Mr. Vaishak G Kumar	20201CAI0138
Mr. Abhilash Prusty	20201CAI0139
Mr. Dhrupath Rajeev	20201CAI0136

Under the guidance of,

Mr. Gnanakumar Ganesan

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

At



PRESIDENCY UNIVERSITY

BENGALURU

JANUARY 2024

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

CERTIFICATE

This is to certify that the Project report “**AI-DxMH- Artificial Intelligence Diagnosis for Modern Health**” being submitted by “MOHAMMED FAHAD PASHA F” , “VAISHAK G KUMAR”, “ABHILASH PRUSTY” and “DHRUPATH RAJEEV” bearing roll number(s) “20201CAI0093”, “20201CAI0138”, “20201CAI0139” and “20201CAI0136” in partial fulfillment of requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering is a bonafide work carried out under my supervision.

Mr. GNANAKUMAR GANESAN

Assistant Professor

School of CSE

Presidency University

Dr. ZAFAR ALI KHAN

Associate Professor & HoD

School of CSE

Presidency University

Dr. C. KALAIARASAN

Associate Dean

School of CSE&IS

Presidency University

Dr. L. SHAKKEERA

Associate Dean

School of CSE&IS

Presidency University

Dr. SAMEERUDDIN KHAN

Dean

School of CSE&IS

Presidency University

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

DECLARATION

We hereby declare that the work, which is being presented in the project report entitled “**AI-DxMH- Artificial Intelligence Diagnosis for Modern Health**” in partial fulfilment for the award of Degree of **Bachelor of Technology in Computer Science and Engineering**, is a record of our own investigations carried under the guidance of **Gnanakumar Ganesan, Assistant Professor, School of Computer Science Engineering , Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

Name	Roll No	Signature
Mohammed Fahad Pasha F	20201CAI0093	
Vaishak G Kumar	20201CAI0138	
Abhilash Prusty	20201CAI0139	
Dhrupath Rajeev	20201CAI0136	

ABSTRACT

Our idea, **AI-DxMH-Artificial Intelligence Diagnosis for Modern Health** addresses the issue of limited access to healthcare in small towns and villages in India by using AI-powered healthcare systems. The solution uses artificial intelligence (AI), machine learning (ML), and natural language processing (NLP) to create a digital advisor that can diagnose serious diseases based on user feedback. The system uses a large-scale language model (LLM), similar to OpenAI and GPT, that integrates natural language understanding (NLU) to efficiently interpret user queries. Machine learning models analyze user input, compare it to clinical data, and create diagnoses that need to be continually improved through feedback. The user-friendly interface facilitates effective communication, and integration with telemedicine platforms enables greater interaction with human experts. Strong data protection and continuous improvement processes ensure reliability and compatibility with ongoing medical information. Our solution anticipates a transformative impact on access to healthcare, especially in areas where the availability of medical professionals is limited. By providing timely, accurate and personalized health advice, the AI-powered doctor aims to significantly improve human health outcomes in diverse populations. In summary, AI-DxMH not only addresses the immediate need for accessible healthcare in remote areas but also focuses on long-term sustainability, cultural sensitivity, and community empowerment, making it a comprehensive and innovative solution for improving healthcare outcomes in underserved population.

ACKNOWLEDGEMENT

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Dean, School of Computer Science and Engineering , Presidency University for getting us permission to undergo the project.

We record our heartfelt gratitude to our beloved Associate Deans **Dr. Kalaiarasan C** and **Dr. Shakkeera L**, School of Computer Science and Engineering , Presidency University and **Dr. Zafar Ali Khan**, Head of the Department, School of Computer Science and Engineering, Presidency University for rendering timely help for the successful completion of this project.

We are greatly indebted to our guide **Mr Gnanakumar Ganesan**, Assistant Professor, School of Computer Science and Engineering, Presidency University for her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the University Project-II Coordinators **Dr. Sanjeev P Kaulgud**, **Dr. Mrutyunjaya MS** and the department Project Coordinator **Dr Murali Parameswaran**.

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

Mohammed Fahad Pasha F
Vaishak G Kumar
Abhilash Prusty
Dhrupath Rajeev

LIST OF FIGURES

Sl. No.	Figure Name	Caption	Page No.
1	Figure 6.1	Architecture	22
2	Figure 7.1	Gantt Chart	25
3	Figure 9.3.1	While Training	37
4	Figure 9.3.2	Hardware Status	37
5	Figure 9.4.1	Phi-2 Learning	38
6	Figure 9.4.2	Hardware Status	38
7	Figure 9.5.1	StableLM Prompt Eval 1	39
8	Figure 9.5.2	Microsoft phi-2 fine-tuned on MedQuad	39

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	CERTIFICATE	ii
	DECLARATION	iii
	ABSTRACT	iv
	ACKNOWLEDGMENT	v
	LIST OF FIGURES	vi
1.	INTRODUCTION	1
	1.1 Problem Statement	1
	1.2 Key Components	2
	1.2.1 Comprehension of CNL	2
	1.2.2 Utilization of ML in diagnosis	2
	1.2.3 User Friendly Interface	2
	1.2.4 Seamless Incorporation with Telemedicine Plaforms	3
	1.2.5 Safeguarding Data Confidentiality	3
	1.2.6 Continuous Improvement	3
2.	LITERATURE SURVEY	4
	2.1 Application of AI in Medical Technologies: a systematic review of Main Trends.	4
	2.2 Applications of Explainable Artificial Intelligence in Diagnosis and Surgery.	5
	2.3 Artificial Intelligence in disease diagnosis- a systematic literature review, synthesizing framework, and future research agenda.	6
	2.4 Embracing LLMs for Medical Applications: Opportunities and Challenges	7
	2.5 CPLLM: Clinical Prediction using Large Language Models	8
	2.6 Large Language Models in Healthcare:Development, Applications and Challenges	9
	2.7 On the Limitations of Large Language Models in Clinical diagnosis	10
	2.8 Path to Medical AGI: Unify Domain -specific Medical LLMs with the lowest cost	11

CHAPTER NO.	TITLE	PAGE NO.
	2.9 Evaluating the Utility of Large Language Model in Answering Gastrointestinal Health-Related Questions: Are We There Yet?	12
	2.10 Creation and Adoption of LLMs in Medicine	13
3.	RESEARCH GAP IN EXISTING METHODS	14
	3.1 Key Research Gaps	14
4.	PROPOSED METHODOLOGY	16
	4.1 Data Collection	16
	4.2 Data Preprocessing	16
	4.3 Model Building	16
	4.4 AI Model Integration	16
	4.5 Continuous Learning	17
	4.6 Scalability Planning	17
	4.7 Education and Awareness	17
	4.8 Collaboration	17
5.	OBJECTIVES	18
	5.1 Create an AI powered Diagnostic Model	18
	5.2 Improve Natural Language Processing(NLP) skills	18
	5.3 Compile an Extensive Medical Dataset	19
	5.4 Employ Supervised Learning Methods	19
	5.5 Investigate Transfer Learning Possibilities	20
	5.6 Create a User-Friendly Interface	20
	5.7 Ensure Diagnostic Accuracy	20
	5.8 Telemedicine Platform Integration	21
	5.9 Model Performance Evaluation	21
	5.10 Scalability	21
6.	SYSTEM DESIGN AND IMPLEMENTATION	22
	6.1 Architecture	22
	6.1.1 The Phi-2 Model	22
	6.1.2 Mistral 7B	23
	6.1.3 StableLM Zephyr 3B	23

CHAPTER NO.	TITLE	PAGE NO.
	6.1.4 Fine Tuning	24
7.	TIMELINE FOR EXECUTION OF PROJECT (GANTT CHART)	25
	7.1 Literature Survey	25
	7.1.1 Purpose	25
	7.1.2 Tasks	26
	7.2 Design Process	26
	7.2.1 Objectives	26
	7.2.2 Tasks	26
	7.3 Data Collection	27
	7.3.1 Purpose	27
	7.3.2 Tasks	27
	7.4 Front End Development	27
	7.4.1 Purpose	27
	7.4.2 Tasks	28
	7.5 Backend Development	28
	7.5.1 Purpose	28
	7.5.2 Tasks	28
	7.6 Model Integration	29
	7.6.1 Objectives	29
	7.6.2 Tasks	29
	7.7 Deployment	29
	7.7.1 Objectives	29
	7.7.2 Tasks	29
8.	OUTCOMES	30
	8.1 Available Health Services	30
	8.2 Timely Diagnosis	30
	8.3 Reduce Health Differences	30
	8.4 Scalable Telemedicine Solutions	31
	8.5 Ease of Use	31
	8.6 Public Health Awareness	32

CHAPTER NO.	TITLE	PAGE NO.
	8.7 Data Driven Insights	32
9.	RESULTS AND DISCUSSION	33
	9.1 Training	33
	9.1.1 What is LoRA?	33
	9.1.2 Benefits of LoRA	33
	9.1.3 Quantized Low-Rank Adaptation(QLoRA)	34
	9.1.4 Benefitsof QLoRA	34
	9.2 LoRA Configuration(Used for all Models)	35
	9.3 Training Report: StableLM Zephyr 3B	37
	9.4 Training Report: Phi-2 (2.7B)	38
	9.5 Prompt Evaluation	39
10.	CONCLUSION	41
	REFERENCES	43
	APPENDIX-A PSEUDOCODE	45
	APPENDIX-B REFERENCES	50
	APPENDIX-C ENCLOSURES	52

CHAPTER-1

INTRODUCTION

1.1 Problem Statement:

India faces a shortage of doctors, particularly in smaller towns and villages, which hinders access to healthcare for many individuals. Prior attempts at implementing telemedicine and other solutions have struggled to scale due to this problem. In the digital age of voice assistants like Google and Alexa, is it possible to develop an artificial intelligence (AI) doctor that can effectively diagnose common acute illnesses such as the common cold or flu based on simple questions?

Solution:

To address the scarcity of doctors in India's rural areas, our proposed solution involves creating an advanced AI-powered healthcare system. By harnessing Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP), we aim to build an intelligent digital assistant capable of diagnosing prevalent acute diseases like the common cold or flu through user-friendly interactions.

The core technology behind our solution will rely on a sophisticated Large Language Model (LLM) similar to OpenAI's GPT platform. This LLM would be designed specifically for understanding natural language queries from users regarding their symptoms. The AI-based doctor will utilize a heuristic approach by integrating medical knowledge, symptom analysis techniques, and historical data records effectively providing accurate personalized diagnoses tailored towards each individual case scenario.

1.2 Key Components

1.2.1 Comprehension of Natural Language (CNL):

Objective: The main objective of CNL is to enhance the system's capacity to grasp and interpret user inquiries with precision.

Methods: Employ sophisticated techniques in Natural Language Processing (NLP) to train the model on recognizing and comprehending diverse linguistic subtleties such as context, intonation, and informal expressions.

Significance: This enables seamless processing and comprehension of user-provided information by the AI system, thus facilitating more precise diagnosis based on natural language intricacies.

1.2.2 Utilization of Machine Learning in Diagnosis:

Goal: Constructing a resilient model based on Machine Learning (ML) to examine user inputs, juxtapose them with an extensive repository encompassing medical records and expertise, thereby generating plausible diagnoses.

Operation: The ML algorithm perpetually assimilates new information and hones its diagnostic precision by incorporating feedback from users along with updated medical insights.

Benefits: By harnessing ML capabilities, the system becomes adaptable towards emerging medical advancements. It can seamlessly integrate novel data while refining its diagnostic aptitude through observing user outcomes and soliciting their valuable input.

1.2.3 User-Friendly Interface:

Goal: Develop a user-friendly interface that allows for smooth interaction between users and the AI doctor, prioritizing ease of use and intuitiveness.

User-Centric Design: With a focus on the end-user, prioritize the development of an interface that enhances user experience, thereby facilitating the widespread acceptance of the AI-driven healthcare solution.

1.2.4 Seamless Incorporation with Telemedicine Platforms

Aim: Streamline the incorporation of current telemedicine systems to allow individuals to connect with human medical experts for additional consultation.

Role of AI Doctor: The AI doctor plays a crucial role in the healthcare industry by serving as an initial diagnostic tool. Its purpose is to streamline the diagnostic process for both patients and healthcare providers. It facilitates a seamless transition between AI-powered diagnostics and human expertise, ensuring efficient and effective healthcare delivery.

1.2.5 Safeguarding Data Confidentiality

Goal: Establish strong security protocols to safeguard user information and guarantee adherence to healthcare privacy mandates.

Confidentiality is a crucial aspect that needs to be prioritized in order to establish trust and encourage the widespread adoption of the AI-driven healthcare solution.

Compliance: It is crucial to ensure strict compliance with healthcare privacy regulations in order to protect sensitive medical data and uphold ethical standards. By doing so, we can safeguard the confidentiality of patient information and maintain integrity in the healthcare industry.

1.2.6 Continuous Improvement:

Aim: Create a system for constant enhancement by routinely integrating the most recent advancements in healthcare research, treatment approaches, and feedback from users into the AI model.

Adaptability: To maintain the system's effectiveness and relevance as medical knowledge evolves, it is crucial to prioritize adaptability and ensure continuous updates. This will enhance the system's diagnostic accuracy over time.

Feedback loop: Foster a feedback loop that promotes user input to drive updates, establishing a dynamic system that consistently incorporates the latest developments in medical science and user preferences.

CHAPTER-2

LITERATURE SURVEY

2.1 Application of artificial intelligence in medical technologies : a systematic review of main trends.

Introduction:

The use of intelligence (AI), in technology has grown significantly especially in specialized fields such as oncology, pulmonology, cardiovascular medicine orthopedics, hepatology and neurology. This comprehensive review aims to explore the emerging trends in incorporating AI into healthcare practices.

Main Trends:

The review emphasizes areas where AI is making contributions. These include collecting and analyzing data for disease diagnosis well as assisting in active treatment processes. Notably deep learning methods like networks (CNNs) have shown promise, particularly in radiology and oncology by enabling image recognition. The development of AI models that rival the expertise of professionals is particularly noteworthy, in cancer detection highlighting how AI can enhance healthcare capabilities.

Future Developments:

Looking ahead the study calls for advancements to broaden the scope of AI in medicine with a focus on its applications, in surgical procedures. However it acknowledges that there is ambiguity surrounding the concept of AI and recognizes the challenges associated with keeping up with evolving research trends.

Conclusion:

Finally, this systematic review provides a comprehensive overview of the use of artificial intelligence in medical technology and highlights its potential for transforming health care. The findings reinforce the need for continuous efforts to improve integration and use of medical AI, which serves as a valuable teaching tool for healthcare professionals, developers.

2.2 Applications of Explainable Artificial Intelligence in Diagnosis and Surgery

Introduction:

First of all, healthcare has undergone a transformation thanks to the growth of medical artificial intelligence (AI) applications, which are driven by deep learning and machine learning. This paper addresses concerns regarding the opaque nature of some AI models and focuses on Explainable AI (XAI) applications in medical diagnostics and surgery.

Using a search from 2019 to 2021, important XAI techniques for medical applications were found. There was discussion of rule-based methods, XGBoost models for choosing laser surgery, and SVM models for surgical training.

XAI Techniques:

The study investigates various XAI techniques. Rule-based methods, which specify the criteria for making decisions, offer transparency. Surgical decision transparency is ensured by the practical application of XAI, as demonstrated by the employment of XGBoost models in laser surgery. SVM models in a surgical education demonstrate XAI versatility in medical education.

Goal and Significance:

Highlighting XAI's importance in healthcare, this concise assessment aims to bridge the gap between medical practitioners and AI.. The review provides guidance for future research, improving the integration and efficacy of AI in the medical profession by providing insights into various XAI techniques and their roles in diagnosis and surgery.

Conclusion and Upcoming Prospects:

A summary addressing the present shortcomings in medical XAI applications is provided at the end of the article. Prospective viewpoints underscore the continuous significance of study in surmounting obstacles, furnishing invaluable discernments for forthcoming advancements in the multidisciplinary domain of medical artificial intelligence applications.

2.3 Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda

Introduction:

The impact of artificial intelligence (AI) on disease detection across medical domains is investigated in this comprehensive literature review. The paper summarizes a framework for AI in disease detection modeling, reviews a few papers, and emphasizes the importance of empirical research with experimental results.

Applications and Benefits:

The paper highlights the potential advantages of AI technology in improving diagnostic accuracy and efficiency by discussing its applications in diabetes diagnosis, cancer detection, and healthcare automation.

Challenges , Considerations and Conclusion:

The paper highlights the significance of precise and timely data for AI-driven disease diagnosis, addressing issues in diagnosis, severity evaluations, and disease manifestations. The study underscores the critical role that artificial intelligence (AI) plays in disease diagnosis. It also emphasizes the importance of accuracy and pushes for continued progress in AI-based procedures.

Future Research Agenda:

The scope includes investigating decentralized federated learning models, continuing improvements in AI-based methods, and cooperative efforts between AI and doctors. The overall efficacy of healthcare as well as early disease diagnosis are the goals of these initiatives.

2.4 Embracing LLMs for Medical Applications: Opportunities and Challenges

Introduction:

Large language models (LLMs) have the potential to transform medicine, and this viewpoint piece examines that possibility with a particular focus on improving clinical decision-making and diagnostic precision. It discusses the particular difficulties and factors that must be taken into account for LLMs to be successfully integrated into the medical field.

Key Topics Covered:

The paper explores a number of topics, including interdisciplinary collaboration, education, assessment metrics, clinical validation, ethical issues, data privacy, and regulatory frameworks. It also covers transfer learning, domain-specific fine-tuning, domain adaptation, reinforcement learning, and dynamic training.

Method for Accuracy:

The paper suggests methods including domain adaptation, transfer learning, and fine-tuning on pertinent medical data to guarantee accuracy across a range of medical specialties. For sophisticated knowledge and relevance in the medical profession, dynamic training, ongoing updates, and reinforcement learning with expert input are recommended.

Conclusion:

In conclusion, this thorough analysis of many points of view offers a thorough examination of the benefits and drawbacks of utilizing LLMs in medical applications. In addition to discussing technical issues, it highlights the significance of ethical considerations, education, and teamwork in order to fully realize the promise of LLMs in the healthcare industry.

2.5 CPLLM: Clinical Prediction using Large Language Models

Introduction:

Using data from Electronic Health Records (EHRs), this work presents a novel method for improving clinical predictions: Clinical Prediction with Large Language Models (CPLLM). When it comes to managing sequential data, CPLLM overcomes the shortcomings of conventional models, especially when it comes to clinical predictions concerning patient diagnoses.

Methodology:

CPLLM uses prompts to fine-tune Large Language Models (LLMs), such as Llama2 and BioMedLM. The main goal is to forecast, based on past medical information, whether patients would receive a follow-up diagnosis or a diagnosis of a particular condition during their following visits. Across a range of prediction tasks and datasets, CPLLM notably shows notable improvements over state-of-the-art models, such as Logistic Regression, RETAIN, and Med-BERT.

Performance Evaluation:

CPLLM performs better than other illness prediction models, especially for adult respiratory failure, acute and unclear renal failure, and chronic kidney disease. The study shows how effective CPLLM is in raising diagnostic accuracy by comparing its performance to baseline models.

Conclusion:

In conclusion, CPLLM is a major development in clinical prediction models, providing better performance, effectiveness, and flexibility in a range of clinical settings. The approach has great potential for transforming the way Large Language Models are incorporated into clinical practice, handling the challenges posed by sequential clinical data and greatly enhancing diagnostic precision.

2.6 Large language models in health care: Development, applications, and challenges

Introduction:

The present article examines the significant influence of Large Language Models (LLMs) in the healthcare industry, emphasizing notable instances such as ChatGPT. It offers a thorough analysis of the LLMs' developmental landscape by classifying them into niche areas like ClinicalBERT for clinical settings and BioBERT for biomedical tasks.

Developmental Landscape:

The article describes how conversational LLMs, such as Med-PaLM 2 and ChatDoctor, are becoming more and more popular and shows how they may be used in patient contacts and medical question answering. It shows how flexible LLMs are in meeting particular healthcare demands by classifying them according to their areas of expertise.

Future Directions:

The story goes on to discuss how LLMs in healthcare should go forward, arguing in favor of decentralizing research initiatives and optimizing current models for particular therapeutic tasks. This highlights how LLM applications are dynamic and can provide customized solutions for a range of healthcare settings.

Conclusion:

The article's conclusion highlights the need of LLMs working in tandem with physicians to achieve the best possible clinical outcomes. This represents a paradigm change in the healthcare industry, as LLMs are now seen as useful tools that work in concert with medical professionals to improve patient care, diagnosis, and overall clinical decision-making, rather than as replacements.

2.7 On the limitations of large language models in clinical diagnosis

Introduction:

This article critically examines the diagnostic properties of the GPT-4 model using the New England Journal of Medicine (NEJM) 2021 and 2022.

Methodology:

Two different approaches are used in this study: a feature-based approach that concentrates on significant clinical anomalies and a narrative approach that follows the methodology of a prior study. These methods seek to reveal subtleties in the diagnostic thinking of the GPT-4.

Findings:

GPT-4 exhibits excellent diagnostic accuracy in the narrative approach, which is in line with earlier conclusions. Nevertheless, the feature-based method has lower diagnostic accuracy since it does not include narrative material. This highlights the shortcomings of big language models in clinical diagnosis and highlights how important a complete linguistic context is to reliable differential diagnosis.

Conclusion:

In summary, the paper draws attention to the complex interplay between language context and the diagnostic performance of the GPT-4. It highlights the necessity of continued investigation and improvement in medical applications to overcome the noted drawbacks. The results emphasize how crucial it is to take into account both feature-based and narrative techniques in order to use big language models in clinical diagnosis in a more reliable and accurate manner.

2.8 Path to Medical AGI: Unify Domain-specific Medical LLMs with the Lowest Cost

Introduction:

This article discusses significant drawbacks that current models, such as ChatGPT, have in the quest for Artificial General Intelligence (AGI) customized for medical applications. One particular issue that is brought to light is the sole dependence on text input, which is insufficient for the thorough examination of vital medical picture information.

MedAGI paradigm:

The paper presents the MedAGI paradigm, a unique strategy intended to successfully integrate domain-specific medical Large Language Models (LLMs) in order to address this issue. MedAGI handles queries and photos provided by users by utilizing the Vision Transformer (ViT) and Q-Transformer models.

Processing Methodology:

The Q-Transformer creates embeddings via a transformer-based design, while the ViT model extracts important information from images. MedAGI distinguishes itself by automatically determining which expert layer is best, matching visual representation to user inquiries, and producing text-based diagnosis.

Conclusion:

The paper concludes by highlighting the affordability, scalability, and adaptability of MedAGI and presenting it as a promising step toward the development of medical artificial general intelligence. The creative method of combining domain-specific medical LLMs solves important drawbacks and demonstrates how MedAGI has the ability to transform medical AI applications in a variety of specializations.

2.9 Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There yet?

Introduction:

Given the frequent frequency of gastrointestinal (GI) symptoms in medical consultations, this study explores the possible application of big language models, notably OpenAI's ChatGPT, in addressing patients' GI health queries.

Objective:

The study intends to assess the effectiveness of AI chatbots in the specialized field of gastroenterology, while acknowledging the encouraging outcomes demonstrated by ChatGPT in healthcare.

Methodology:

The methodology include asking ChatGPT 110 different patient questions on symptoms, diagnostic procedures, and gastrointestinal therapies. Subsequently, skilled gastroenterologists evaluate these answers according to standards like precision, lucidity, current understanding, and general efficacy.

Conclusion:

The study's conclusion emphasizes how critical it is to assess the benefits and drawbacks of AI chatbots in healthcare, especially in gastroenterology, in order to make the most use of them. It highlights how huge language models must be continuously improved upon and carefully considered when incorporated into the intricate web of medical information distribution.

2.10 Creation and Adoption of Large Language Models in Medicine

Introduction:

This essay examines the growing acceptance of large language models (LLMs), as demonstrated by ChatGPT, with an emphasis on the sector of medicine. Important topics covered in the conversation include applications, training approaches, and the necessity of value proposition verification in the healthcare industry.

Ethical Use and Pertinence:

The importance of the medical community's active involvement in influencing the creation and uptake of LLMs is emphasized in the essay. To guarantee their applicability, moral application, and compliance with medical guidelines, this engagement is thought to be essential.

Conclusion:

The importance of a cautious and cooperative approach to integrating LLMs into medical practices is highlighted in the article's conclusion. To maximize the beneficial effects of large language models in the medical field, it exhorts the community to actively participate in conversations, influence the course of LLM development, and give ethical issues top priority.

CHAPTER-3

RESEARCH GAPS OF EXISTING METHODS

3.1 Key Research Gaps

Based on the literature review presented in the paper, there are some key research gaps and limitations in existing methods for implementing AI in healthcare:

Generalizability:

Many AI models show strong performance in narrow academic datasets or benchmarks, but generalization is difficult. the complexity and variability of real clinical conditions. More testing on heterogeneous real-world data is needed.

Handling incomplete and sparse data:

Many models are based on large curated datasets, while clinical data are often incomplete or sparse, limiting applicability (ref. 1-10). More research is needed on low sample learning, data augmentation/synthesis, and handling missing inputs.

Explainability:

The transparency of the results of most AI models limits practitioner confidence and adoption. Developing interpretable models and explanatory methods is critical.

Model Bias and Fairness:

Inadequate Safeguards to Identify and Mitigate Bias That Maintain or Exacerbate Health Care Disparities. The development of bias-reducing techniques and representative data can help overcome this problem.

Model Validation:

Many studies highlight the lack of standardized methods for clinical validation of AI systems prior to deployment, which increases the potential for undetected defects or failures . Establishing comprehensive validation protocols is imperative.

Customization:

Most models use a one-size-fits-all approach, while medical diagnosis and treatment planning require customization. Advancing personalized and precision medicine with artificial intelligence is an open challenge.

Integration into Workflows:

There is a lack of substantial research to quantify the effects and ease of integration into existing clinical workflows in a minimally disruptive manner. More applied research is needed. In summary, although progress has been rapid, significant gaps remain around critical factors such as validity, utility, acceptability and safety in applied clinical contexts. Overcoming these limitations requires extensive interdisciplinary collaboration and innovation.

GPU price and availability:

As the models grow larger and larger, proprietary high-performance GPUs are required to train and maintain them, which are very expensive and of limited availability (ref 1-10). This presents obstacles for many healthcare systems with limited resources. Building efficient models and optimizing hardware requirements is an open challenge.

Lack of training recipes:

There are no clear standard recipes for training robust and reliable models that encode best practices for regularization, scaling, etc. This leads to the proliferation of fragile designs that fail unexpectedly. Establishing strict training protocols and benchmarks is important.

Prompt Engineering:

Appropriate prompts and examples are critical to encoding intended behavior, but best practices remain unclear. Bad signals increase the likelihood of unintended action. The development of rapid design techniques is crucial to minimize these risks.

CHAPTER-4

PROPOSED METHODOLOGY

4.1 Data Collection:

Our AI-powered healthcare initiative relies on meticulously curated, diverse datasets that transcend mere symptom records. We prioritize a comprehensive approach, including detailed documentation of diagnoses and therapeutic interventions. This conscientious effort extends beyond medical cases, considering factors like demographics, geographic location, and disease severity. This ensures our dataset is not only broad but also representative, strengthening the efficiency and reliability of our AI model. The goal is to empower our model to be a valuable tool supporting users in diverse healthcare situations.

4.2 Data Preprocessing:

The foundation of our AI model is laid through thorough data preprocessing. This involves cleaning and rectifying errors in our vast medical records. Beyond technical aspects, privacy is paramount. We adhere strictly to data protection laws, ensuring the integrity and accuracy of the data that forms the backbone of our AI model.

4.3 Model Building:

We adopt a two-pronged strategy, utilizing natural language processing (NLP) and machine learning (ML) in building our advanced AI model. Mistral Language Model (LLM) forms the core, enhancing the model's ability to understand complex healthcare queries. The integration of NLP and ML ensures our model comprehends intricacies in user questions, improving diagnostic capabilities and enabling more profound engagement in health-related conversations.

4.4 AI Model Integration:

User experience is at the forefront of our AI model integration. We prioritize an intuitive and user-friendly interface, allowing seamless input of symptoms. The system quickly generates initial diagnoses, providing users with an overview of their health status. Beyond diagnostics, we strategically integrate with telemedicine vendors, allowing users

to bridge the gap between digital communication and personalized care.

4.5 Continuous Learning:

Our commitment to excellence extends beyond initial implementation. We establish a continuous training system, leveraging user interactions as feedback for ongoing model development. This iterative process ensures our AI model evolves with the dynamic healthcare landscape, consistently improving in accuracy and diagnostic capabilities over time.

4.6 Scalability Planning:

Scalability is a focal point in our strategic planning. We design our AI-based healthcare system to be not only robust but also flexible enough to adapt to a growing user base. This involves creating an infrastructure that dynamically scales to accommodate increased data volumes, processing power, and user interactions while maintaining optimal performance.

4.7 Education and Awareness:

We place significant emphasis on education and awareness, recognizing that an informed user is an empowered user. Our resources go beyond technicalities, providing a nuanced understanding of our AI model's capabilities and limitations. We communicate openly about the collaborative role of our AI doctor, encouraging users to view it as a complement to traditional healthcare.

4.8 Collaboration:

Collaboration is ingrained in our approach to AI healthcare. Actively involving healthcare professionals in our development process ensures our model aligns with medical standards and benefits from industry expertise. This collaborative approach aims to redefine healthcare by merging technology and medicine, creating an innovative system deeply rooted in validated healthcare knowledge.

CHAPTER-5

OBJECTIVES

5.1 Create an AI-powered Diagnostic Model:

Introduction:

The initial step in developing an AI-based diagnostic model is to design and implement a reliable system capable of detecting common acute disorders. This approach makes use of artificial intelligence to assess user input and deliver reliable diagnoses, first focused on ailments such as the common cold and flu.

Methodology:

To accomplish this, use machine learning techniques such as decision trees or neural networks that have been trained on a broad collection of medical data and symptoms. Using supervised learning approaches, the model should learn to spot patterns and connections between symptoms and diseases.

Implementation:

Create the model in a programming language suitable for machine learning, such as Python, and make use of popular libraries like TensorFlow or PyTorch. Ascertain that the model's design can accommodate the model's complexity, enabling scalability and simple connection with other components.

5.2 Improve Natural Language Processing (NLP) Skills:

Boosting:

Boost the model's Natural Language Processing (NLP) capabilities to boost user interaction. This entails the system's ability to comprehend and interpret user inquiries about symptoms, medical history, and other pertinent data.

NLP Techniques:

Use advanced NLP techniques such as sentiment analysis, entity recognition, and semantic comprehension. Use pre-trained language models such as BERT or GPT to improve the system's ability to understand and respond to a variety of user inputs.

Integration:

Integrate the NLP advancements into the diagnostic model to ensure that the user and the AI system communicate seamlessly. This will make the experience more user-friendly by allowing users to enter information in a conversational manner.

5.3 Compile an Extensive Medical Dataset:

Curation of Datasets:

A rich dataset is required for the development of a robust diagnostic model. Curate a broad set of medical data, symptoms, diagnostic patterns, and treatment outcomes for prevalent acute disorders. Ascertain that the dataset is representative of a wide range of demographics and medical circumstances.

Privacy and Security:

Prioritize data privacy and security while complying with ethical norms and rules. Anonymize and safeguard sensitive data inside the dataset by getting the required permissions and consents.

5.4 Employ Supervised Learning Methods:

Model Education:

To train the AI model, use supervised learning approaches. By including labeled instances in the training dataset, you may emphasize the relationship between presented symptoms and accurate diagnoses.

Measures for Evaluation:

Define and monitor performance measures such as accuracy, sensitivity, and specificity to assess the model's effectiveness in disease diagnosis. Create a baseline for future comparison and improvement.

5.5 Investigate Transfer Learning Possibilities:

Transfer Learning Ideas:

Examine transfer learning options for leveraging prior medical knowledge. Investigate how pre-trained models in similar domains might be fine-tuned to improve the core model's diagnostic skills.

Perform a feasibility study to investigate the applicability and benefits of transfer learning in the context of medical diagnostics. Examine how knowledge transfer from one domain to another can increase the model's accuracy and efficiency.

5.6 Create a User-Friendly Interface:

Create an intuitive and user-friendly interface for individuals to interact with the AI-based diagnostic system. Prioritize user experience (UX) design concepts to ensure accessibility and simplicity of use for people of diverse technology literacy levels.

Iterative Design:

Use an iterative design method to refine and improve the interface by incorporating user feedback. Conduct usability testing to discover potential pain areas and make required changes to provide a smooth user experience.

5.7 Ensure Diagnostic Accuracy:

Continuous Learning:

Implement continuous learning methods to improve diagnostic accuracy over time. Include feedback loops in which the system learns from user inputs and updates its predictions based on changing medical knowledge.

Dynamic Adaptation:

Establish processes for the model's dynamic updating to accommodate new medical research, emergent disorders, and changes in diagnostic criteria. Maintain the system's accuracy and dependability in providing diagnostics.

5.8 Telemedicine Platform Integration:

Effortless Integration:

Allow for seamless interaction with telemedicine services, allowing consumers to consult healthcare specialists as needed. Interoperability with existing telehealth infrastructure and standards must be ensured.

Implement mechanisms for secure and confidential data flow between the AI-based diagnostic system and telemedicine platforms. Consider regulatory requirements to preserve patient information and maintain privacy standards.

5.9 Model Performance Evaluation:

Testing Methodology:

Use a variety of datasets to thoroughly test and validate the AI model. Assess accuracy, sensitivity, specificity, and overall performance using various testing approaches such as cross-validation and real-world scenario simulations.

Benchmarking:

To validate the model's efficacy, compare it to existing diagnostic procedures and expert opinions. To ensure the model's generalizability, compare results across different demographic groupings.

5.10 Scalability:

Architecture that can be scaled:

Create the system with scalability in mind so that it can support a big number of users. Optimize algorithms and infrastructure to deliver consistent performance even when bandwidth is limited, or user load is high.

Cloud Integration:

Investigate cloud-based solutions to improve scalability, considering the potential to scale resources based on demand. To ensure consistent performance, use caching methods and load balancing.

CHAPTER-6

SYSTEM DESIGN & IMPLEMENTATION

6.1 Architecture

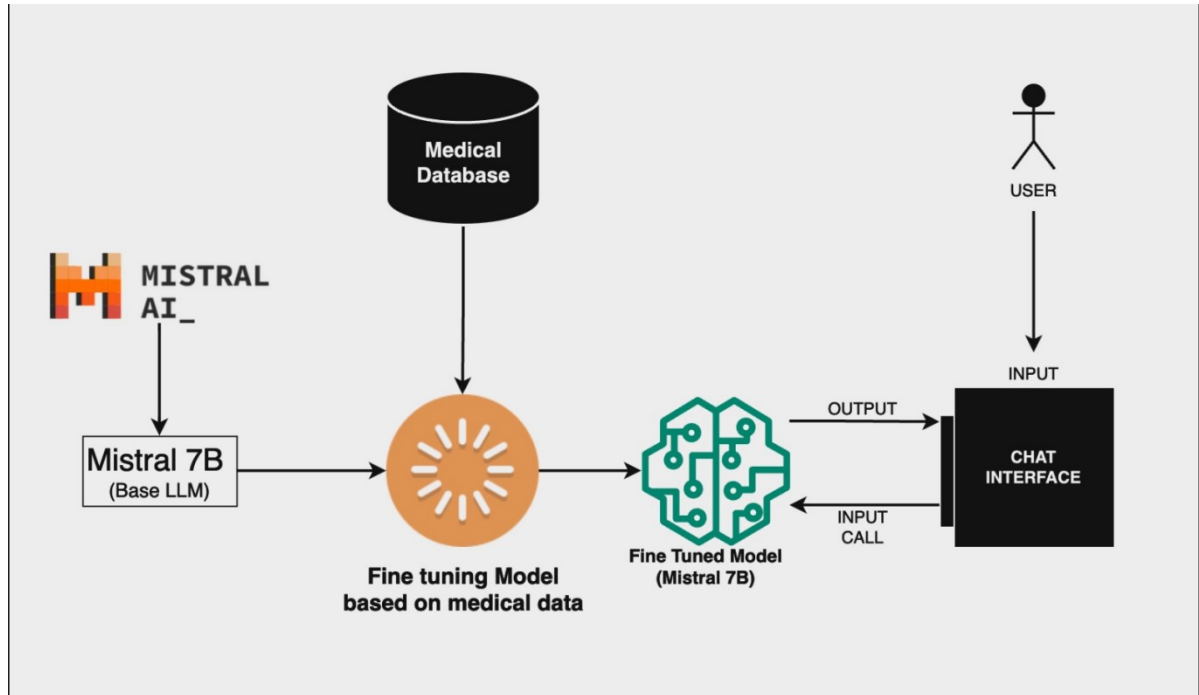


Figure 6.1

6.1.1 The Phi-2 Model:

A transformer with 2.7 billion parameters is the Phi-2. The same data sources as Phi-1.5 were used to teach it, along with an additional data source made up of different fake NLP texts and censored websites (for training and security). In terms of performance against benchmarks measuring logical thinking, language understanding, and common sense, Phi-2 demonstrated near-best results for models with less than 13 billion parameters. This model was not refined through feedback from humans. This open source model aims to give the scientific community a free and unrestricted tiny model to explore critical safety concerns like controlling toxicity, comprehending social biases, and more.

6.1.2 Mistral 7B:

Innovative attention methods including Sliding Window Attention (SWA) and Grouped Query Attention (GQA) are included in Mistral 7B. Larger batches and improved performance in real-time applications are made possible by GQA, which increases speed and lowers memory requirements during decoding. Large language models (LLM) have a general restriction that SWA aims to solve by handling longer sequences more effectively and at a lower computing cost. When these attention processes are combined, the Mistral 7B performance and efficiency are enhanced. Mistral 7B is available under the Apache 2.0 license and comes with a benchmark application that can be easily deployed using the vLLM inference server and SkyPilot to local computers or cloud platforms like AWS, GCP, or Azure. Hugging Face integration is more seamless, and the Mistral 7B is made to make fine-tuning for different jobs easier. The improved chat model Mistral 7B outperforms Llama 2 13B - Chat with its exceptional performance and versatility. Mistral 7B seeks to help the community produce more reasonably priced, effective, and superior models by striking a compromise between high performance and efficiency for big language models.

6.1.3 StableLM Zephyr 3B :

The 3 billion-parameter compact neural network model StableLM Zephyr 3B is designed to produce text on a variety of devices with efficiency and adaptability. With 60% fewer parameters than similar 7B models, it produces precise results without the need for sophisticated gear. The template, which is available under a non-commercial license, emphasizes asking questions and according to directions. The training process, which draws inspiration from the Zephyr 7B architecture, entails adjusting and personalizing training data resources based on users' preferences. The StableLM Zephyr 3B scores better in benchmarks than larger versions in terms of linguistic accuracy, consistency, and relevance. Its lightweight performance is available outside of sophisticated systems and is appropriate for a range of text creation activities, such as content creation and summarizing. All things considered, the StableLM Zephyr 3B provides a strong yet effective solution that prioritizes adaptability in practical applications across a range of computing settings. Its small form factor attempts to increase the number of people who can create effective text.

6.1.4 Fine Tuning:

In the context of machine learning, fine-tuning is the act of modifying a pre-trained model's parameters to make it more appropriate for a particular task or domain. A model might not function as well for a more specialized or nuanced task if it was first trained on a big dataset for a generic purpose. This pre-trained model is fine-tuned by retraining it on a smaller, task-specific dataset.

To better match the complexity of healthcare data and diagnostic requirements, fine-tuning the AI-based diagnostic system described earlier may entail modifying the model's parameters, which may include weights and biases. This procedure enhances the model's functionality, precision, and efficacy in managing particular medical diagnostics.

Fine-tuning is an essential component of our two-pronged strategy to building our sophisticated healthcare AI system, balancing Natural Language Processing (NLP) and Machine Learning (ML). This methodical approach is particularly important when working with Mistral, our Large Language Model (LLM), since it enhances the model's ability to go beyond traditional communication boundaries. In order to provide a thorough comprehension of user inquiries and to customize Mistral's contextualization to the complexities of healthcare interactions, fine-tuning is essential. The process of fine-tuning Mistral is crucial to improving its complex understanding of medical intricacies as we incorporate it into our chatbot, resulting in more sophisticated and productive user interactions. Fine-tuning goes beyond Mistral and involves our ML component as well, honing the model's recognition skills of complex patterns and relationships in medical data.

CHAPTER-7

TIMELINE FOR EXECUTION OF PROJECT (GANTT CHART)

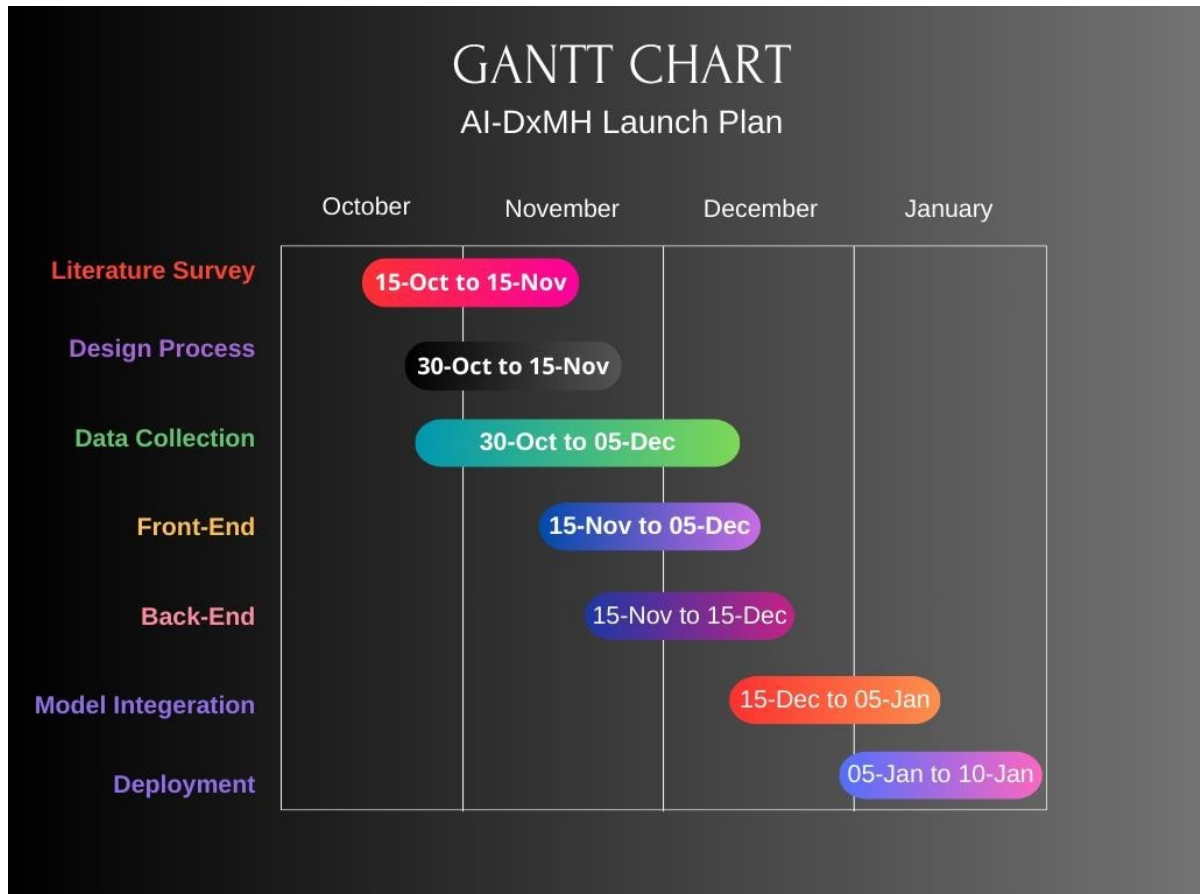


Figure7.1

7.1 Literature Survey:

7.1.1 Purpose:

The purpose of the literature survey stage is to provide a comprehensive understanding of existing research and technologies in the field of AI in healthcare, focusing in particular on diagnostic tools, natural language processing (NLP) and relevant methods.

7.1.2 Tasks:

Literature Review on Artificial Intelligence in Healthcare:

- Research and analyze academic papers, articles and case studies related to the application of artificial intelligence in healthcare.
- Summarize the main findings, emerging trends and challenges in this field.

Analyze available diagnostic tools and techniques:

- Examine currently available diagnostic tools, paying attention to their characteristics, strengths and limitations.
- Identify successful implementation cases and areas for improvement.

Identify key insights and possible methods:

- Synthesize data from the literature review to identify key insights.
- To present possible methods and approaches compatible with the objectives of the project.

7.2 Design Process:

7.2.1 Objective:

The design process phase focuses on outlining the structure of the AI-based diagnostic tool, which includes user interface design, system architecture, and overall system flow.

7.2.2 Tasks:

Define user interface requirements:

- Define functional and aesthetic user interface requirements.
- Consider user experience (UX) principles to improve usability.

Create wireframes and design patterns:

- Develop wireframes to illustrate the basic structure and layout of the user interface.
- Create mockups to visualize the final appearance of the tool.

System Architecture and Workflow Design:

- Define the overall architecture of the system, including the interaction of front-end and back-end components.
- Design a tool workflow considering user interaction and data flow.

7.3 Data collection:

7.3.1 Purpose:

The data collection phase focuses on obtaining the data sets needed to train the AI model, including medical data sets, user input samples, and relevant health data.

7.3.2 Tasks:

Identify and retrieve medical datasets:

- Research and identify appropriate medical datasets covering various diseases and symptoms.
- Obtain the necessary permissions and licenses to use the datasets.

Collect user input samples for training:

- Design a strategy to collect user input samples to train an AI model.
- Provide various input samples to improve reliability of model.

Data protection and compliance:

- Implement measures to ensure data protection, regulatory compliance and ethical considerations.
- Develop protocols to anonymize and securely store sensitive health data.

7.4 Front End Development:

7.4.1 Purpose:

The user interface development phase involves the transformation of design data into a functional and user-friendly user interface.

7.4.2 Tasks:

Development of user interface elements:

- Code and implement visual elements of the user interface according to design specifications.
- Ensure consistency of design elements and user interaction.

Apply design patterns to the user interface:

- Integrate previously created pattern patterns into user interface development.
- Make sure the applied model meets the original specifications.

Ensure Responsiveness and User-friendliness:

- Test UI responsiveness on different devices and screen sizes.
- Optimize user interaction for a smooth and friendly experience.

7.5 Back-end development:

7.5.1 Purpose:

The back-end development phase is dedicated to building the infrastructure supporting the artificial intelligence model, data management and data transfer through the user interface.

7.5.2 Tasks:

Configure the Server Infrastructure:

- Configure the server infrastructure required to host the backend components.
- Choose the right hosting solutions that meet your scalability requirements.

Develop the backend logic for data processing:

- Code the backend logic responsible for processing the data received from the user interface.
- Implement data analysis and AI model preparation algorithms.

Implementing Communication APIs:

- Creating application programming interfaces (APIs) enables communication between the user interface and the backend.

- Ensure secure data transmission and efficient data exchange.

7.6 Model Integration:

7.6.1 Objective:

The model integration phase involves embedding the trained AI model into the system, ensuring seamless communication with both front-end and back-end components.

7.6.2 Tasks:

Integrate the AI model into the backend:

- Embed the trained AI model into the backend infrastructure.
- Create communication channels between the model and other system components.

Perform model functionality testing:

- Perform extensive testing to ensure functionality and accuracy of the integrated AI model.
- Fix some model performance or interoperability issues.

7.7 Deployment:

7.7.1 Objective:

The deployment phase marks the final stage to make the AI-based diagnostic tool available to users in a real-time environment.

7.7.2 Tasks:

Hosting and Server Deployment:

- Deploy the system on live servers and configure hosting settings.
- Ensure the availability and reliability of the hosting environment.

Monitor and fix potential problems after deployment:

- Establish monitoring mechanisms to monitor system performance and user interactions.
- Be prepared to quickly fix any problems or defects that arise after deployment.

CHAPTER-8

OUTCOMES

8.1 Available health services:

Improved accessibility:

Adopting an AI-based diagnostic model will increase the availability of health services, especially in remote and underserved areas where doctor shortages are common. The system allows people in such areas to receive timely diagnostic knowledge without the need for immediate physical contact with health professionals.

Remote Monitoring:

The model enables remote monitoring of health facilities, allowing people to enter their symptoms and receive diagnostic recommendations. This not only reduces the burden on local health facilities, but also ensures that people in remote areas can benefit from medical knowledge without having to travel long distances.

8.2 Timely diagnosis:

The introduction of an AI-based diagnostic system enables quick and timely diagnosis of everyday acute diseases such as cold and flu. Users can enter their symptoms into the system, and the rapid analysis model provides rapid feedback, which allows treatment to begin faster.

8.3 Reduce health differences:

Geographic independence:

One of the important results is the reduction of health differences by providing consistent diagnostic services regardless of geographic location. This is particularly important in countries like India, where the disparity between urban and rural healthcare is clear.

Equitable access to information:

An AI-based diagnostic tool ensures that people have equal access to accurate and timely health information regardless of their location. This will help address disparities in access to and outcomes of health care between urban and rural populations.

8.4 Scalable Telemedicine Solutions:

Overcoming Previous Challenges:

Developing scalable telemedicine solutions augmented with artificial intelligence will address the challenges of previous telemedicine efforts. The diagnostic model based on artificial intelligence increases the efficiency and reliability of telemedicine services, making them even more scalable and sustainable.

Better coverage:

Scalable telemedicine solutions can reach larger populations even in remote areas. The integration of AI will not only improve diagnostic accuracy, but also enable more people to access telehealth services, leading to a wider impact.

Improved Healthcare Infrastructure:

Integrating AI into telemedicine will improve healthcare infrastructure, making it more adaptable to the changing needs of a growing and diverse population.

8.5 Ease of use:

User-friendly interfaces:

Development of user-friendly interfaces ensures that individuals, including those with limited technical knowledge, can easily interact with the AI-based diagnostic tool. It improves the accessibility of the health system by making it inclusive and user-centric.

Accessibility for all:

A focus on ease of use is critical to ensure that the benefits of an AI-based diagnostic system are available to a wide range of users. This inclusion includes people with varying levels of technical literacy, enabling them to benefit from health services.

8.6 Public health awareness:

Dissemination of information:

Implementation of an AI-based diagnostic system increases public health awareness by providing accurate information about common diseases and symptoms. By interacting with the tool, users can gain insight into various health conditions, promoting a proactive approach to health care.

Prevention and early intervention:

Greater awareness facilitates preventive measures and early interventions, thus reducing the overall burden on the health system. Users have the right to make informed decisions about their health, leading to healthier lifestyles and less preventable disease.

8.7 Data-Driven Insights:

Population Health Analysis:

AI-based diagnostic system produces valuable information about common diseases, symptoms, and diagnostic patterns. This knowledge advances the analysis of population health and enables health professionals and decision makers to make informed decisions about public health planning and resource allocation.

Evidence-based decision-making:

Evidence-based insights enable evidence-based decision-making in health care, which promotes a more proactive and effective approach to solving public health problems. This ensures that activities are adapted to the specific needs of the population.

CHAPTER-9

RESULTS AND DISCUSSIONS

9.1 Training:

As mentioned in CHAPTER-6, we considered three models namely;

1. **Phi – 2 (2.7B Parameter)**
2. **Mistral (7B Parameter)**
3. **StableLM Zephyr (3B Parameter, Instruct Model)**

To have a Standardized Training metric we used Q-LoRA Method from the LoRA

9.1.1 What is LoRA?

LoRA assumes that the weights of the pre-trained model are already rather near to the best answer for the jobs that come after. Therefore, LoRA concentrates on optimizing trainable low-rank matrices and freezes the pretrained model's weights.

- **Low-Rank Matrices:** LoRA adds matrices A and B, which are low-rank matrices, to each layer's self-attention module. By serving as adapters, these low-rank matrices minimize the amount of extra parameters required while enabling the model to adapt and specialize for certain tasks.
- **Rank-Deficiency:** The rank-deficiency of weight changes (ΔW) during adaptation is a key finding of LoRA. This implies that modifications made to the model can be adequately represented by weight matrices that are substantially lower in rank than the original ones. LoRA makes use of this finding to maximize parameter efficiency.

9.1.2 Benefits of LoRA

- **Reduced Parameter Overhead:** LoRA drastically lowers the amount of trainable parameters, which makes it far more memory-efficient and computationally less expensive by using low-rank matrices as an alternative to fine-tuning every parameter.
- **Effective Task-Switching:** LoRA minimizes the requirement to maintain distinct, fine-tuned instances for each task by enabling the pretrained model to be reused across many tasks. This lowers storage and switching costs by enabling smooth and rapid task transition during deployment.

9.1.3 Quantized Low-Rank Adaptation (QLoRA)

A additional quantization is introduced in QLoRA, an extension of LoRA, to improve parameter efficiency during fine-tuning. It introduces 4-bit NormalFloat (NF4) quantization and Double Quantization algorithms, building on the foundations of LoRA.

- **NF4 Quantization:** This method makes use of the pre-trained neural network weights' intrinsic distribution, which is typically a zero-centered normal distribution with a certain standard deviation. NF4 quantization efficiently quantifies the weights without the use of costly quantile estimation procedures by converting all weights to a fixed distribution that falls inside NF4's (-1 to 1) range.
- **Double Quantization:** This technique takes care of the quantization constants' memory cost. Double quantization, which quantizes the quantization constants themselves, drastically lowers the memory footprint without sacrificing performance. The second quantization phase of the process uses 8-bit Floats with a block size of 256, which saves a significant amount of memory.

9.1.4 Benefits of QLoRA:

- **Enhanced Memory Efficiency:** By including quantization, QLoRA expands on its memory efficiency and is especially helpful for deploying big models on devices with limited resources.
- **Maintaining Performance:** On a variety of downstream tasks, QLoRA outperforms completely fine-tuned models while maintaining good model quality due to its parameter-efficient design.
- **Consistency with Different LLMs:** With QLoRA, researchers can investigate parameter-efficient fine-tuning for a range of LLM designs. This technique is flexible and can be used to different language models, such as RoBERTa, DeBERTa, GPT-2, and GPT-3. Optimizing Big Language Models for Accuracy By Using PEFT

9.2 LoRA Configuration (used for all Models)

We have to apply some preprocessing to the model to prepare it for training. For that we use the `prepare_model_for_kbit_training` method from PEFT.

```
from peft import prepare_model_for_kbit_training
model.gradient_checkpointing_enable()
model = prepare_model_for_kbit_training(model)
def print_trainable_parameters(model):
    """
    Prints the number of trainable parameters in the model.
    """
    trainable_params = 0
    all_param = 0
    for _, param in model.named_parameters():
        all_param += param.numel()
        if param.requires_grad:
            trainable_params += param.numel()
    print(
        f'trainable params: {trainable_params} || all params: {all_param} || trainable%: '
        f'{100 * trainable_params / all_param}'
    )
```

Let's print the model to examine its layers, as we will apply **QLoRA** to all the linear layers of the model. Those layers are `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`, and `lm_head`.

```
print(model)
```

Using the `print model` we see the definable classes of the model.

Now we define the LoRA config:

r is the rank of the low-rank matrix used in the adapters, which thus controls the number of parameters trained. A higher rank will allow for more expressivity, but there is a compute tradeoff.

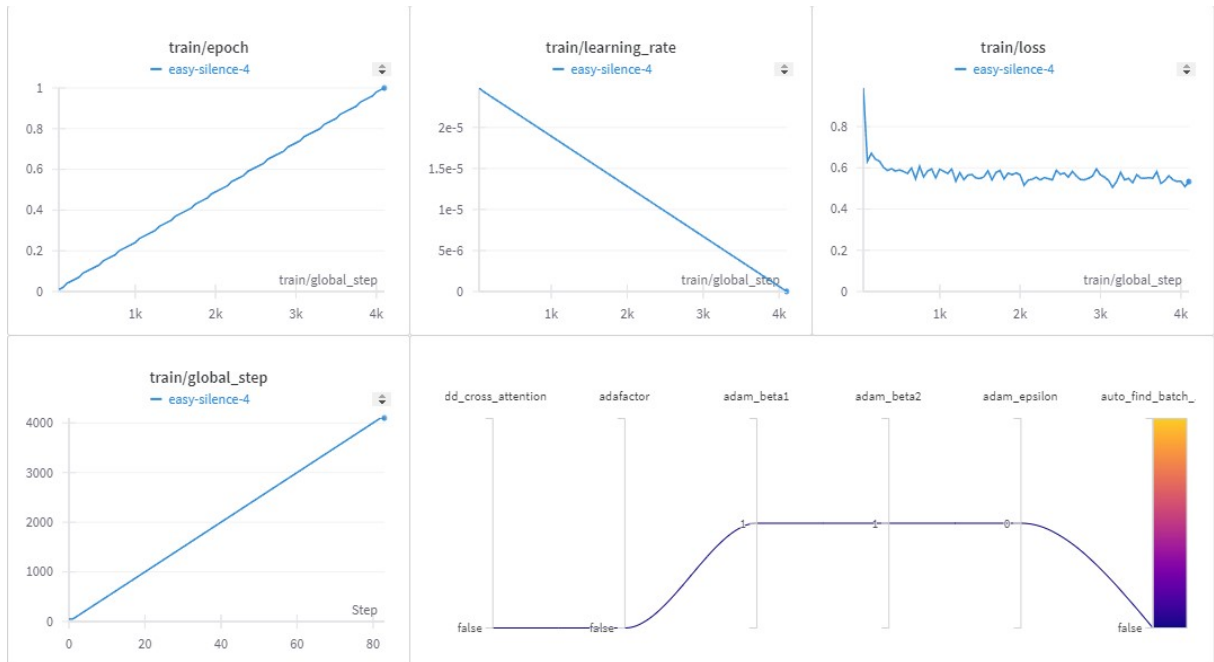
Alpha is the scaling factor for the learned weights. The weight matrix is scaled by α/r , and thus a higher value for alpha assigns more weight to the LoRA activations.

The values used in the **QLoRA** paper were $r=64$ and $\text{lora_alpha}=16$, and these are said to generalize well, but we will use $r=32$ and $\text{lora_alpha}=64$ so that we have more emphasis on the new fine-tuned data while also reducing computational complexity.

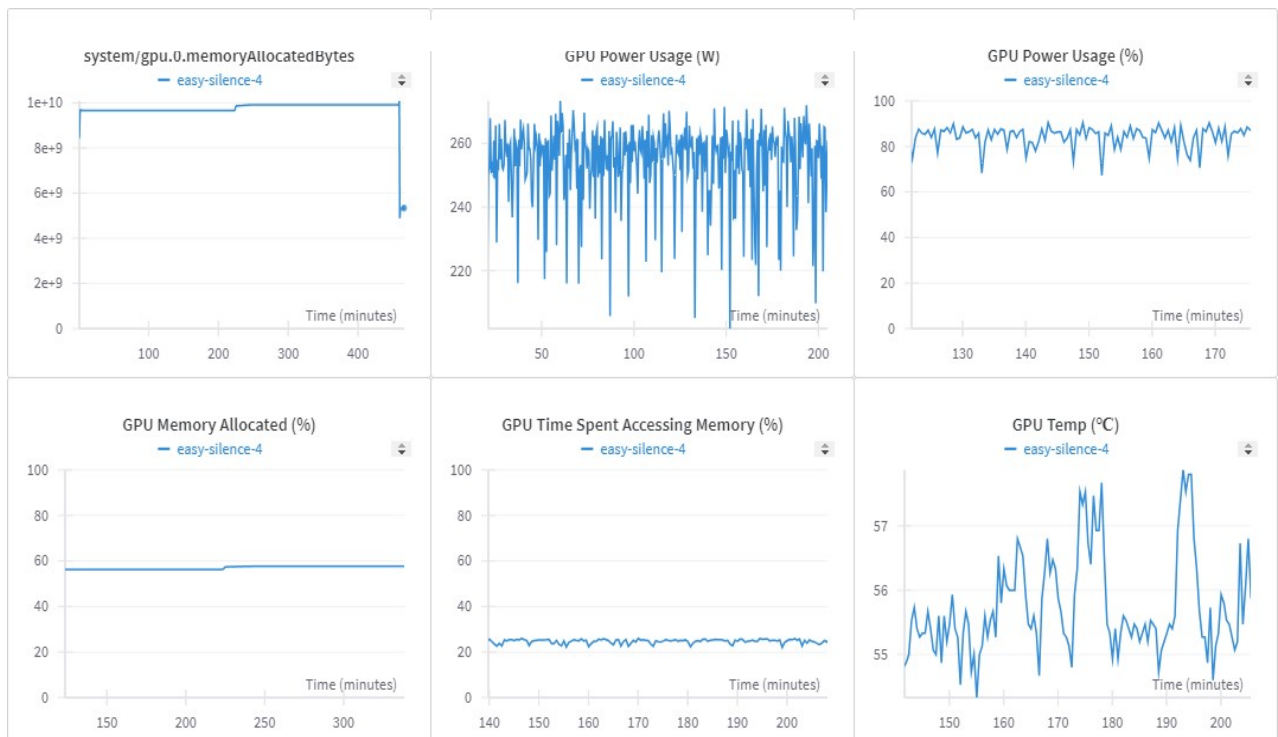
```
from peft import LoraConfig, get_peft_model
config = LoraConfig(
    r=32,
    lora_alpha=64,
    target_modules=[
        "q_proj",
        "k_proj",
        "v_proj",
        "o_proj",
        "gate_proj",
        "up_proj",
        "down_proj",
        "lm_head",
    ],
    bias="none",
    lora_dropout=0.05, # Conventional
)
model = get_peft_model(model, config)
print_trainable_parameters(model)
```

Note – “Play around with r and lora_alpha for finding the optimal setting of r & lora_alpha for each model”

9.3 Training Report: StableLM Zephyr 3B

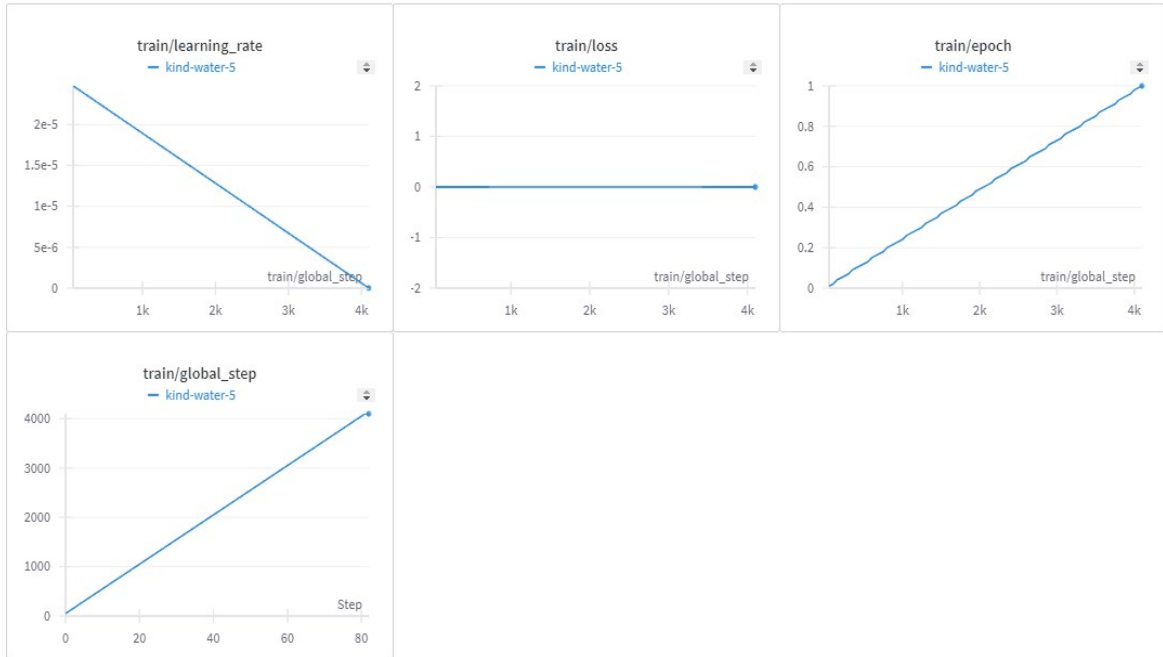


9.3.1 While Training

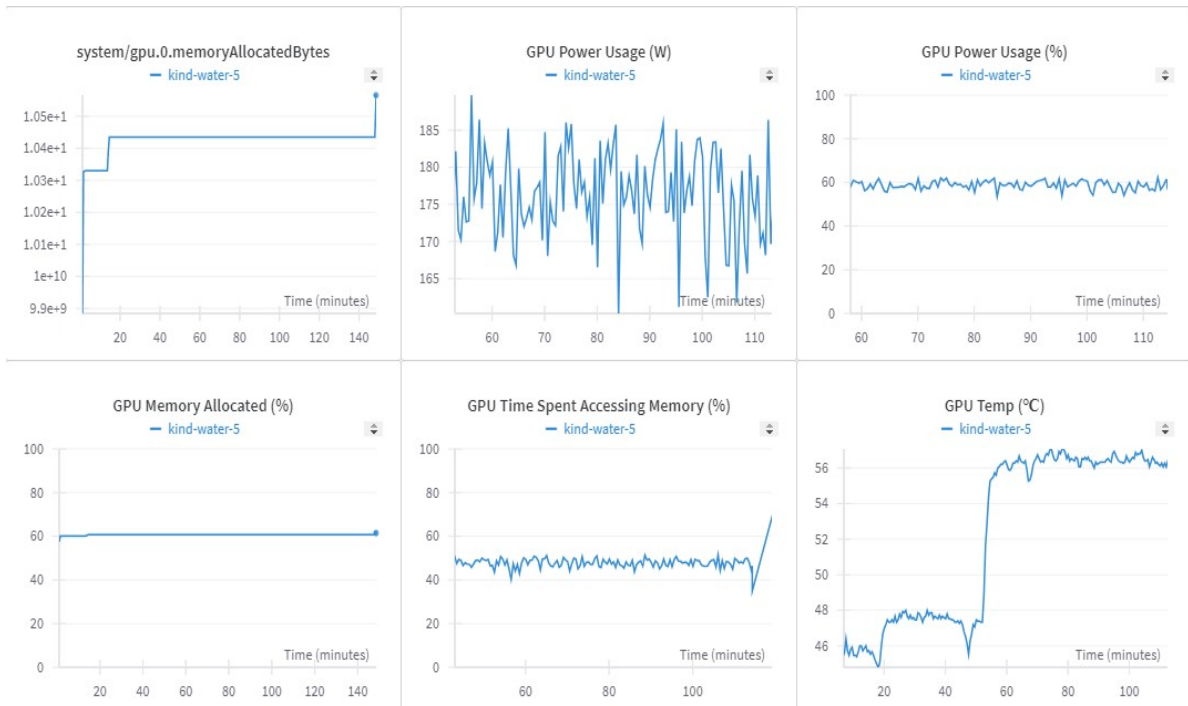


9.3.2 Hardware Status

9.4 Training Report: Phi – 2 (2.7 B)

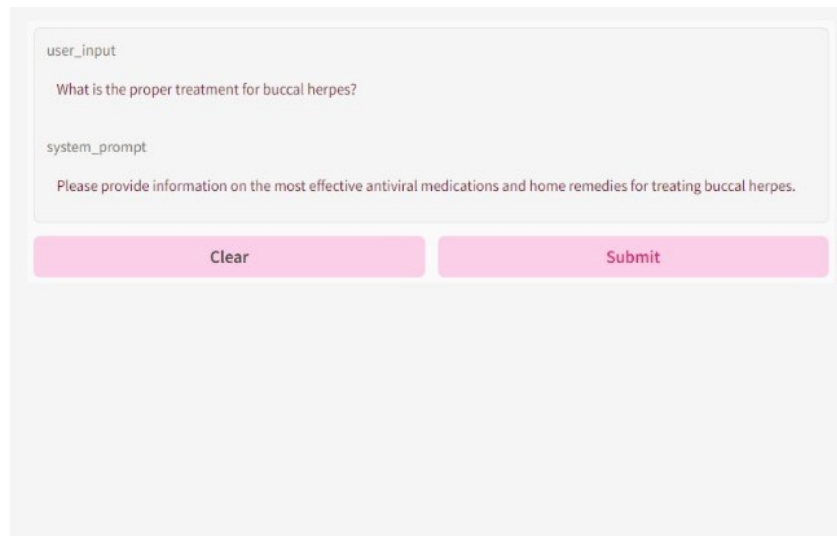


9.4.1 Phi – 2 Learning



9.4.2 Hardware Status

9.5 Prompt Evaluation:



user_input

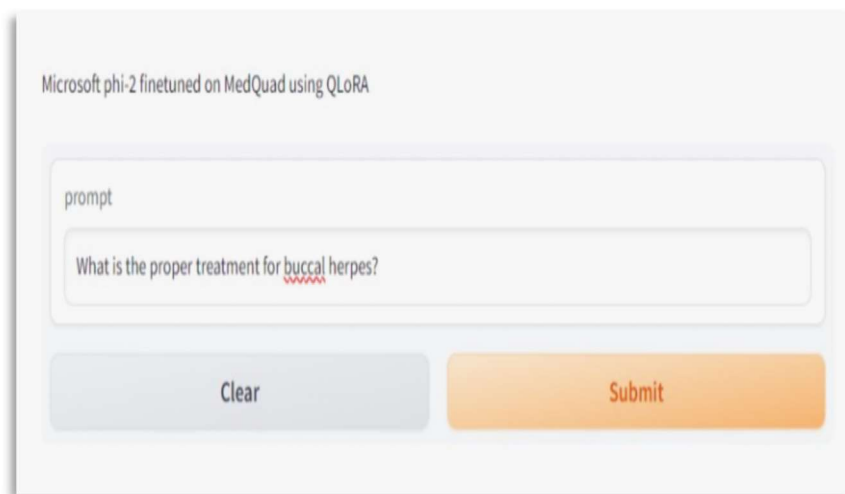
What is the proper treatment for buccal herpes?

system_prompt

Please provide information on the most effective antiviral medications and home remedies for treating buccal herpes.

Clear Submit

9.5.1 StableLM Prompt Eval 1



Microsoft phi-2 finetuned on MedQuad using QLoRA

prompt

What is the proper treatment for buccal herpes?

Clear Submit

9.5.2 Microsoft phi-2 finetuned on MedQuad

CHAPTER-10

CONCLUSION

An initiative to introduce an AI-based diagnostic tool in India has the potential to transform healthcare access, especially in underserved areas that face ongoing challenges with limited access to medical professionals. The intended outcomes include a broad set of improvements that together will shape the healthcare landscape to be more inclusive and efficient.

Tackling the shortage of doctors and improving accessibility:

The main objective of the initiative is to improve access to health services, especially in remote areas suffering from a lack of doctors. Offering a virtual "doctor"; The system aims to bridge the gap in medical services by providing timely and accurate diagnostic knowledge to residents of smaller towns and villages.

Quick and timely diagnosis for better health outcomes:

Quick and timely diagnosis of common illnesses such as colds and flu are a key component of the initiative. This ensures that people receive prompt medical care that improves health outcomes. Early intervention becomes a key preventive healthcare strategy that reduces disease severity and reduces the overall burden on the healthcare system.

Reducing health care disparities:

A notable result is a reduction in health disparities, which underscores the goal of achieving equitable health care for all regardless of geographic location. The initiative aims to break down barriers and provide consistent diagnostic services to individuals, whether they live in urban centers or remote villages.

Scalable telemedicine solutions for a connected healthcare system:

Developing scalable telemedicine solutions with artificial intelligence will not only solve previous scalability issues, but also lay the foundation for a more connected and digitally enabled healthcare system. This innovation promises to reach a larger population and improve health care.

User-friendly user interfaces for comprehensive health communication:

User-friendly user interfaces for an AI-based diagnostic tool are crucial when health communication is accessible to people at different levels. This inclusiveness ensures that a broad user base can take advantage of the tool, encouraging widespread adoption and use.

Health benefits and data insights:

In addition to individual health benefits, the initiative promotes public health awareness and practices. Increased awareness of common diseases and symptoms promotes a culture of preventive healthcare. In addition, building data-driven insights into common diseases supports public health planning by enabling targeted allocation of resources and intervention strategies.

Reliability and integration with existing systems:

The reliability of medical information provided by an AI-based tool is critical to building user trust. Integration with existing healthcare systems ensures a seamless healthcare experience by facilitating disease tracking and monitoring. This collaborative approach is consistent with the broader goals of strengthening health infrastructure.

Encouraging greater adoption of telemedicine:

The initiative plays a key role in promoting greater adoption of telemedicine in line with emerging trends in digital health. Such a change in health services not only increases the scope of health services, but also reflects a positive response to technological developments in the field. Integrating AI in healthcare will not only improve the diagnostic process but also lay the foundation for a more sustainable, accessible and patient-friendly healthcare ecosystem in the country. This initiative is a significant step towards democratizing access to quality medical services for all, as technology continues to play a key role in shaping the future of healthcare.

REFERENCES

- [1]. Zhou, J., Chen, X., & Gao, X. (2023). Path to Medical AGI: Unify Domain-specific Medical LLMs with the Lowest Cost. *arXiv preprint arXiv:2306.10765*.
- [2]. Shah, N. H., Entwistle, D., & Pfeffer, M.A. (2023). 'Creation and Adoption of Large Language Models in Medicine'. *JAMA*, 330, 866–869.
- [3]. Lahat, A., Shachar, E., Avidan, B., Glicksberg, B., & Klang, E. (2023). 'Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There Yet?' *Diagnostics*, 13(11), 1950.
- [4]. Shoham, O. B., & Rappoport, N. (2023). 'CPLLM: Clinical Prediction with Large Language Models'. Retrieved from <https://arxiv.org/abs/2309.11295v1>
- [5]. Reese, J. T., Danis, D., Caulfied, J. H., Casiraghi, E., Valentini, G., Mungall, C. J., & Robinson, P. N. (2023). 'On the limitations of large language models in clinical diagnosis'. *medRxiv*.
- [6]. Bitkina, O. V., Park, J., & Kim, H. K. (2023). 'Application of artificial intelligence in medical technologies: A systematic review of main trends.' *Digital Health*, Vol. 9. 20552076231189331
- [7]. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). 'Key challenges for delivering clinical impact with artificial intelligence'. *BMC Medicine*, 17, 1–9.
- [8]. Behara, K., Bhero, E., Agee, J. T., & Gonela, V. (2022). 'Artificial intelligence in medical diagnostics: A review from a South African context.' *Scientific African*, 17, e01360.
- [9]. Zhang, Y., Weng, Y., & Lund, J. (2022). 'Applications of Explainable Artificial Intelligence in Diagnosis and Surgery.' *Diagnostics 2022*, Vol. 12, Page237, 12, 237.
- [10]. Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2022). 'Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda'. *Journal of ambient intelligence and humanized computing* 1-28.

- [11]. Su, L. X., Weng, L., Li, W. X., & Long, Y. (2023). ‘Applications and challenges of large language models in critical care medicine’. *Zhonghua Yi Xue Za Zhi*, 103, 2361-2364.
- [12]. Yang, R., Ting, Tan, F., Lu, W., Arun, Thirunavukarasu, J., ... Ting, W. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2, 255–263.
- [13]. Karabacak, M., & Margetis, K. (2023). Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus*, 15(5).
- [14]. Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Natural Communications* 11(1), 3923.
- [15]. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature* 620(7972), 172-180

APPENDIX-A

PSUEDOCODE

A.1 Frontend pseudocode:

```
``jsx
// ChatSection component
function ChatSection() {

  // Get chat state and methods from useChat hook
  const {messages, input, handleInputChange, handleSubmit} = useChat();

  // Scroll to bottom on mount and message change
  useEffect(() => {
    scrollToBottom();
  }, [messages])

  // Render UI
  return (
    <Section>
      <Heading />

      <Card>
        <CardHeader />

        <CardBody>
          {messages.map(message => (
            <ChatMessage message={message} />
          ))}
        </CardBody>
      </Card>
    </Section>
  )
}
```

```
<CardFooter>
  <Form>
    <Input
      value={input}
      onChange={handleInputChange}
    />
    <Button
      onClick={handleSubmit}
    />
  </Form>
</CardFooter>

</Card>
</Section>
)
}
```

// ChatMessage component

```
function ChatMessage({message}) {
```

```
  if (message.from === 'bot') {
    return (
      <BotAvatar />
      <BotMessageText />
    )
  }
```

```
  return (
    <UserAvatar />
    <UserMessageText />
  )
}
```

```
// Avatar components
```

```
function BotAvatar() {  
  return <BotImg />  
}
```

```
function UserAvatar() {  
  return <UserIcon />  
}
```

```
``
```

A.2 Chat API Pseudocode:

```
```js
```

```
// Import HuggingFace and OpenAssistant modules
```

```
import { HfInference, HuggingFaceStream } from 'huggingface'
```

```
import { experimental_buildOpenAssistantPrompt } from 'openassistant'
```

```
// Create HuggingFace Inference instance
```

```
const hf = new HfInference(API_KEY)
```

```
// Set runtime to edge
```

```
hf.runtime = 'edge'
```

```
// Handler function
```

```
function chatbotHandler(request) {
```

```
 // Get messages from request body
```

```
 const {messages} = request.body
```

```
 // Build prompt from messages
```

```
 const prompt = experimental_buildOpenAssistantPrompt(messages)
```



```
// Call HF text generation stream
const response = hf.textGenerationStream({
 model: OPENASSISTANT_MODEL,
 inputs: prompt,
 parameters: {
 max_new_tokens: 200,
 typical_p: 0.2,
 // other params
 }
})

// Convert to text stream
const stream = HuggingFaceStream(response)

// Return streaming response
return new StreamingResponse(stream)

}
...
```

### **A.3 Backend Pseudocode:**

```
...

Load model and tokenizer
model = TransformerModel.from_pretrained(MODEL_NAME)
tokenizer = TransformerTokenizer.from_pretrained(MODEL_NAME)

Prepare model for efficient training
model = prepare_model_for_training(model)

Print number of trainable parameters
print_trainable_params(model)
```

**# Move model to GPU/TPU**

```
model = accelerator.prepare_model(model)
```

**# Enable model parallelism if multiple GPUs**

```
if num_gpus > 1:
```

```
 model.parallelize()
```

**# Load datasets**

```
train_dataset, val_dataset = load_datasets()
```

**# Tokenize datasets**

```
tokenized_train = tokenize_dataset(train_dataset, tokenizer)
```

```
tokenized_val = tokenize_dataset(val_dataset, tokenizer)
```

**# Configure training arguments**

```
args = TrainingArguments(
 output_dir = OUTPUT_DIR,
 num_train_epochs = NUM_EPOCHS,
 per_device_train_batch_size = BATCH_SIZE,
 learning_rate = LEARNING_RATE,
 # other arguments...)
```

**# Create Trainer**

```
trainer = Trainer(
 model = model,
 train_dataset = tokenized_train,
 val_dataset = tokenized_val,
 args = args,
 # other Trainer arguments)
```

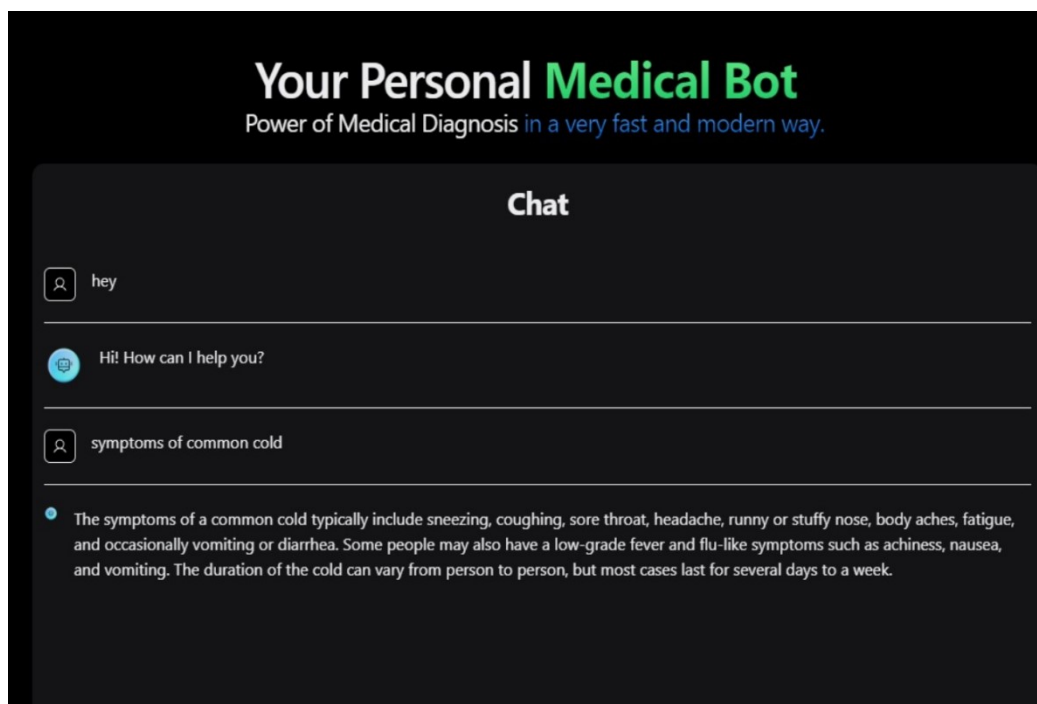
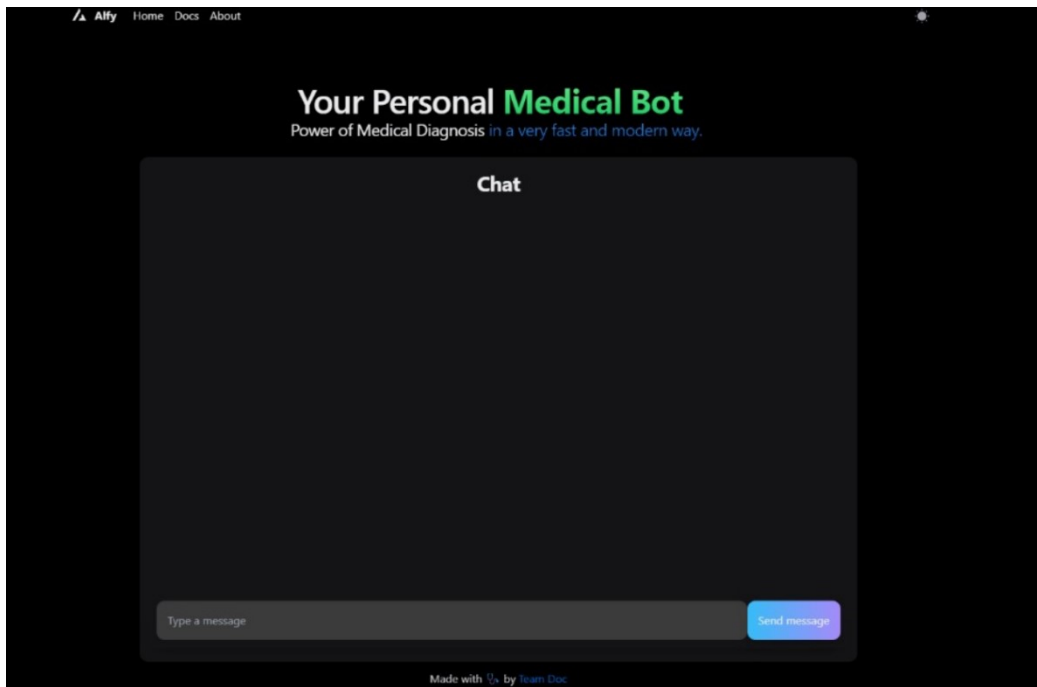
**# Train & Save model**

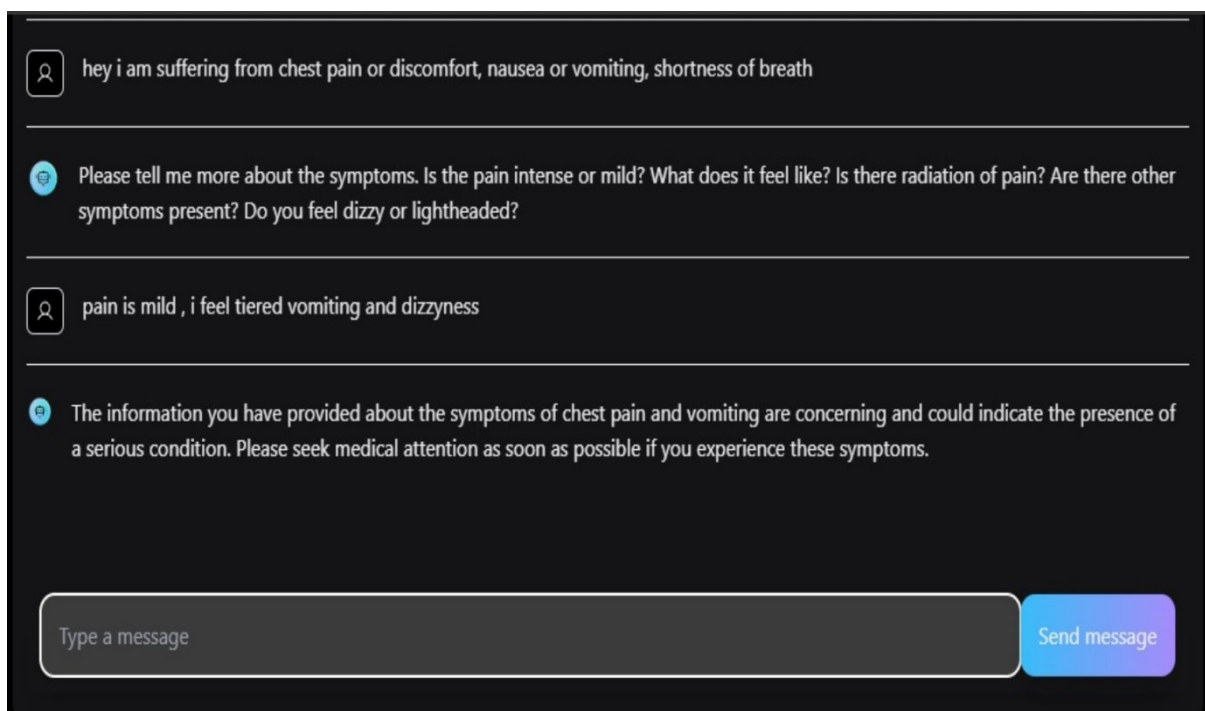
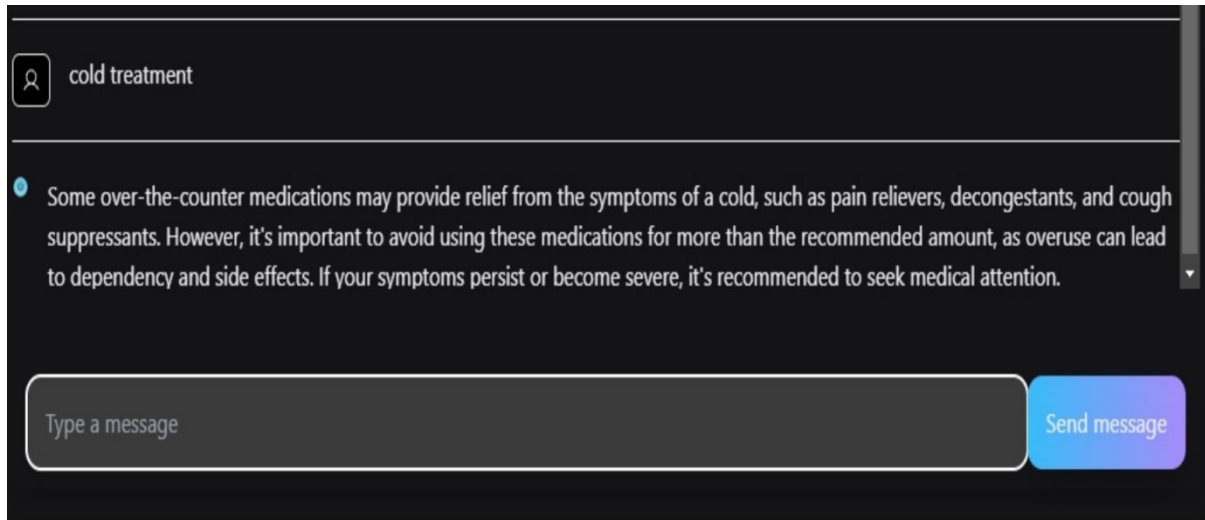
```
trainer.train()
```

```
trainer.save_model() ``
```

## APPENDIX-B

### SCREENSHOTS



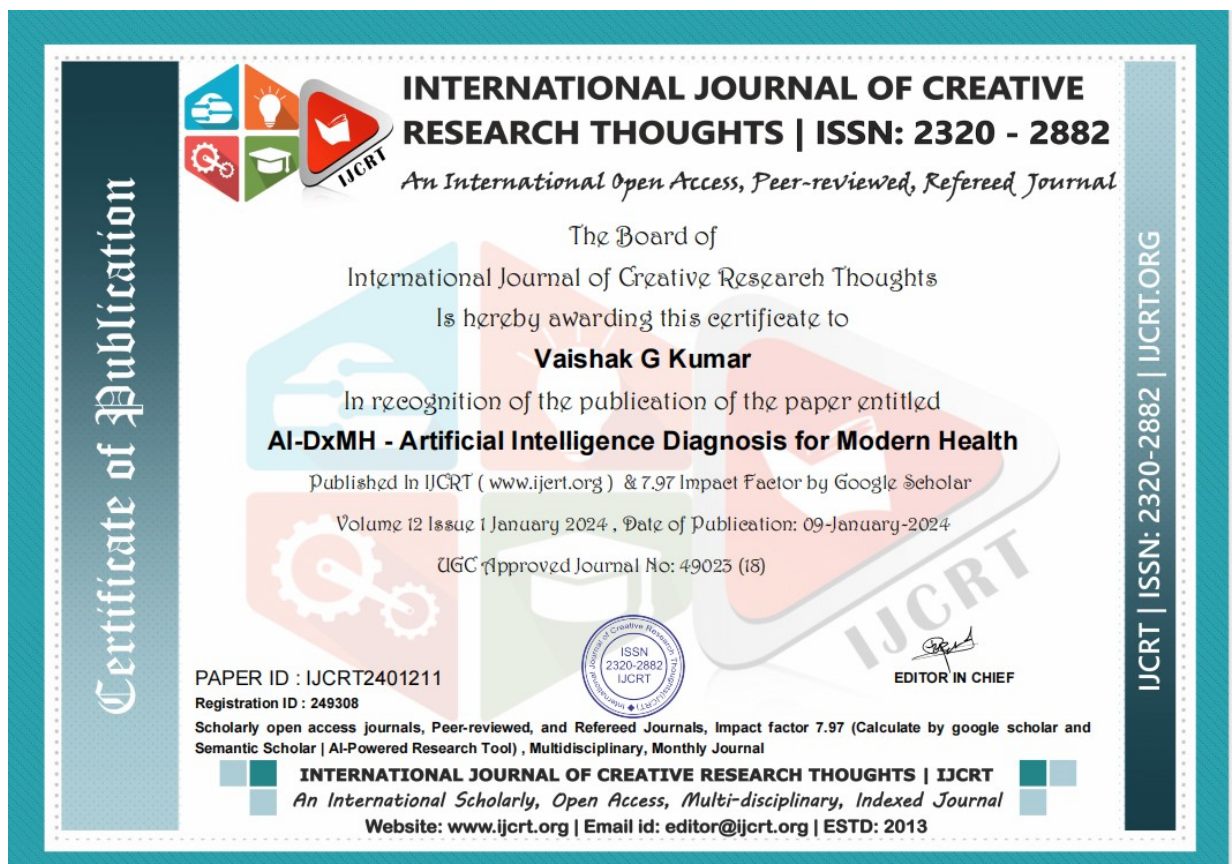


## APPENDIX-C

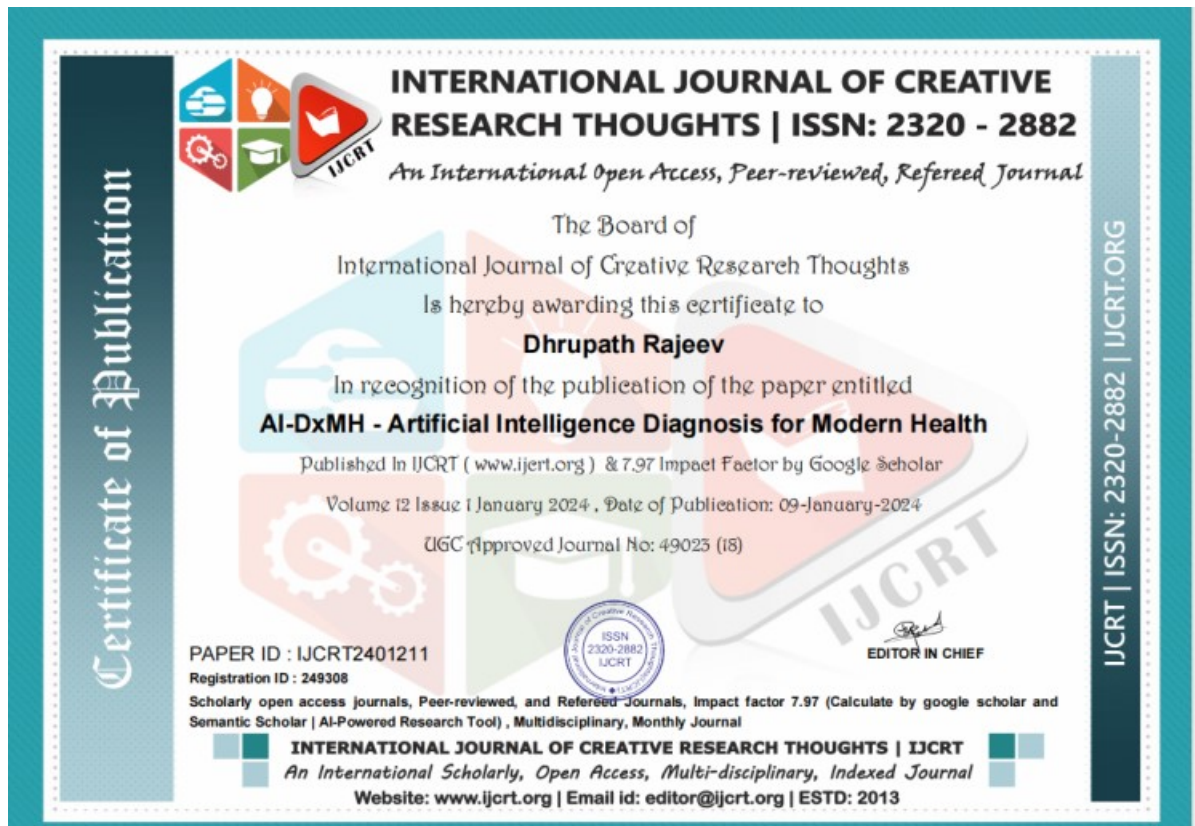
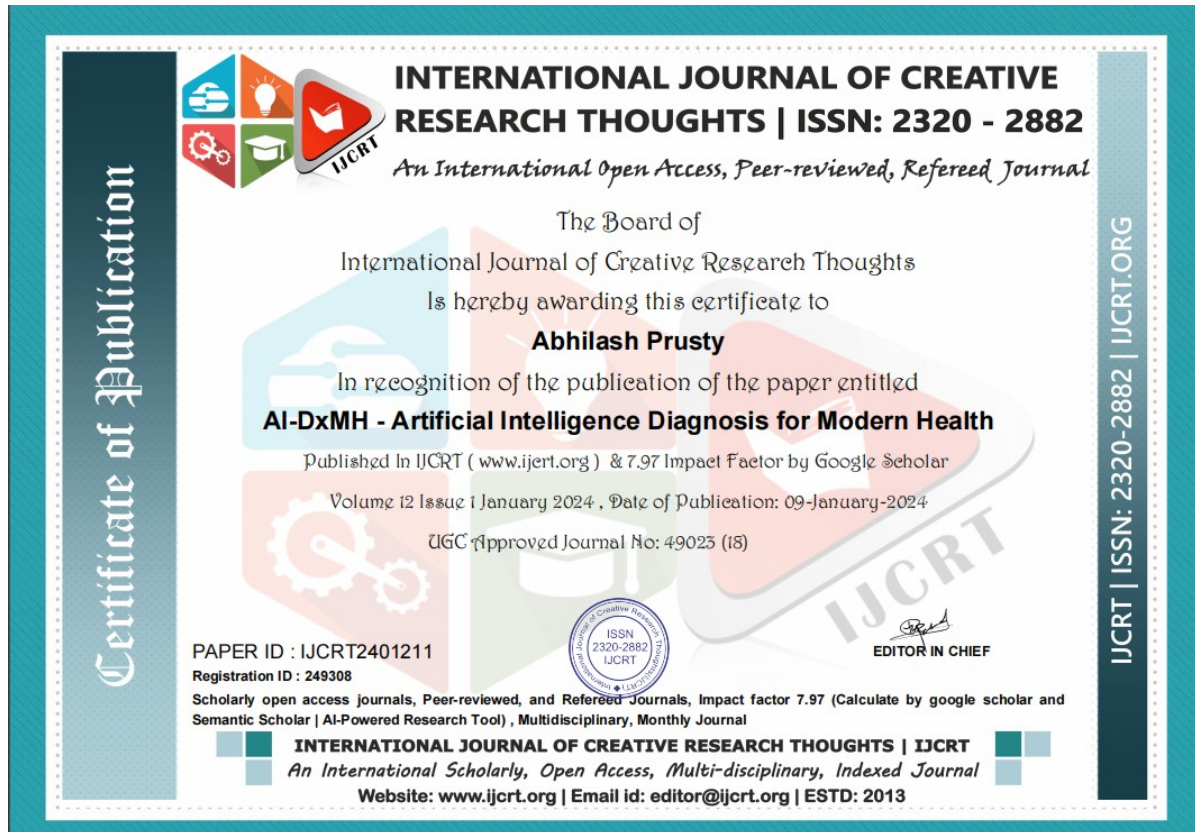
## ENCLOSURES











## final report

### ORIGINALITY REPORT

<b>16%</b>	<b>15%</b>	<b>8%</b>	<b>10%</b>
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

### PRIMARY SOURCES

<b>1</b>	<b>Submitted to Presidency University</b> Student Paper	<b>6%</b>
<b>2</b>	<b>brev.dev</b> Internet Source	<b>3%</b>
<b>3</b>	<b>www.coursehero.com</b> Internet Source	<b>1%</b>
<b>4</b>	<b>B Varshini, HR Yogesh, Syed Danish Pasha, Maaz Suhail, V Madhumitha, Archana Sasi. "IoT-Enabled Smart Doors for Monitoring Body Temperature and Face Mask Detection", Global Transitions Proceedings, 2021</b> Publication	<b>1%</b>
<b>5</b>	<b>Submitted to M S Ramaiah University of Applied Sciences</b> Student Paper	<b>1%</b>
<b>6</b>	<b>www.mdpi.com</b> Internet Source	<b>&lt;1%</b>
<b>7</b>	<b>dokumen.pub</b> Internet Source	<b>&lt;1%</b>





**The Project work carried out here is mapped to SDG-3 Good Health and Well-Being.**

The shortage of doctors and limited access to healthcare directly impacts the goal of ensuring healthy lives and promoting well-being. Developing an AI doctor could contribute to improving healthcare accessibility, providing timely and accurate diagnoses for common illnesses. The integration of technology in healthcare not only addresses immediate challenges but also supports the broader vision of ensuring healthy lives for all, emphasizing the importance of innovation and inclusivity in achieving sustainable health outcomes.

**The Project work carried out here is mapped to SDG-10 Good Reduced Inequalities.**

By enhancing the capabilities of our healthcare AI to understand and respond to user queries, including nuanced medical conversations, we aim to reduce inequalities in access to quality healthcare information and assistance. The introduction of an AI doctor aligns with SDG 10 by targeting the reduction of inequalities in healthcare access, ensuring that individuals, regardless of their geographic location or socio-economic status, have access to timely and preliminary healthcare services.