People's Democratic Republic of Algeria

ministry of higher education and scientific research

Yahia Fares University of Medea
Faculty of Sciences
Department of Mathematics and Computer Science



Graduation Project

Submitted for Master Degree in Computer Science

Specialty: Applied Artificial Intelligence

Chatbot in Healthcare

A multi-language Diagnosis Medical Chatbot Using Artificial Intelligence in different types of diseases

---

Proposed and directed by:

Dr. Hayi Mohamed Yassine

Prepared and presented by:

Miss. Berkane Narimen Yassmine
Miss. Boutoumi Aicha

Academic Year 2024–2025

Dedication

To Allah Almighty
All praise to Allah for His light, strength, and wisdom. Every moment of clarity and success is by His grace. May this work be accepted and bring benefit.

To My Father – My Pillar and Role Model
To my beloved father—your silent sacrifices and support have shaped me. As your only daughter and youngest child, I hope to always make you proud, as I am proud to be yours.

To My Mother – My Anchor and Light
Your compassion, strength, and prayers were my refuge. This journey has your grace in every step—thank you, Mama, for being my greatest strength.

To My Brothers – Hamza, Yahia, and Islam
To my brothers—your love and protection made every burden lighter. Being your only sister is a blessing I will always cherish.

To My Professors
Thank you for your wisdom and guidance. Your dedication has left a lasting impact on my academic journey.

To Mr. Hayi Yassine – My Supervisor
Thank you for your support and belief in me. Your guidance was essential to completing this work.

To Mr. Benaissa Azzedine
I am grateful for your generosity and encouragement whenever I needed direction.

To My Little Lights – Ichrak, Noursin, and Racim
Your laughter and innocence brightened my hardest days. You will always be the stars in my heart.

To My Cousins – Yousra and Maroua
Thank you for your quiet strength and care—it meant more than you know.

To My Friends – Hanane, Naila, Maroua, and All Others
Thank you for your belief and kindness. You were a light during my moments of doubt.

In Closing
To everyone who stood by me—this work is as much yours as mine. May it bring pride, benefit, and lasting barakah.

Berkane Narimen Yassmine

Dedication

الحمد لله حمدًا يليق بجلاله، ويبلغ تمام الشكر، على عطائه وتوفيقه الذي به تُحقق الآمال وتستقيم الخطى.
"الحمد لله الذي بنعمته تتم الصالحات".

To Gaza... the pulse of pride and the pain of the nation.

To those who lifted humiliation from the brow of the free, to the martyrs who drew the path of dignity with their blood, to the resistance fighters whose resolve never wavers I dedicate these pages to you, for you are the true meaning of knowledge and steadfastness. O God, grant them a swift victory and be their unwavering support.

I dedicate this success to my ambitious self-first , who started with ambition and ended with success

With all my love, I dedicate the fruits of my success and graduation.

To those who are unmatched in the universe, to those whom God has commanded us to be dutiful to, to those who have given so much and offered what cannot be repaid, my dear mother and father **Ghania, Salem**, I dedicate the fruit of my effort represented in this humble research, hoping that I will be a source of pride for you both.

To my steadfast rib, the source of my strength, the unwavering pillars I could always rely on my beloved brothers: **Boubaker, Moussa, and Walid**. May you always remain my support and pride.

To the beats of my heart and the comfort of my soul, To my dearest sisters, whose presence makes life blossom and whose laughter sweetens my days, Those whose sight brings joy to my eyes and whose closeness warms my heart, Who stood by me when I was weak, supported me when I needed strength, and believed in me when I doubted myself **Fatiha, Nawal and Nassima** may Allah keep you as a light that never fades, and a joy that never dies.

To my dear nephews and nieces, the light of my eyes and joy of my days, You are the beauty of life and the hope of the future. I pray that Allah blesses you with happiness and success.

To my beloved family **Laila, Sarah and hadjer** and dear friends **Soundous, Aya, Hadil, Meriem, Mimi and yassmine**, You have been a balm to my soul and an unwavering support on my journey. and to **the serendipitous friend**, whose friendship is a gift of fate and a mark forever etched in my heart , Thank you from the bottom of my heart

To my esteemed supervisor **hayi mohamed yassine**, thank you for your outstretched hand in knowledge, your enlightened mind in guidance, and your generous heart in patience and support.

And to the honorable Professor **Benaissa mahdi**, who embarrassed us with his generosity and humility, and supported us with his knowledge and experience, my sincere thanks and gratitude.

To every reader, I hope you find these pages helpful and inspiring.

BOUTOUMI AICHA

Abstract

With the growing integration of artificial intelligence in the healthcare sector, there is an increasing need for intelligent tools that support early diagnosis and enhance healthcare accessibility. This study presents the development of a multilingual medical chatbot (Arabic, French, and English) designed to act as an interactive assistant that receives simple inputs such as symptoms, age, and gender, and responds with a possible diagnosis, a brief description of the condition, and a trusted medical reference link.

To achieve this, several state-of-the-art language models were applied. For Arabic, the chatbot leverages BERT Base Arabic, MARBERT, T5-Small, and GigaBERT. For English, it incorporates BioBERT, BioClinicalBERT, BlueBERTPubMed , and GigaBERT. As French lacked domain-specific models, the chatbot uses translated outputs of the best-performing English model (GigaBERT) to provide reliable responses in French.

The chatbot serves as a decision-support tool for physicians by enabling quick initial assessments, while also empowering patients with accessible health insights—especially valuable in early stages before visiting a doctor or in resource-limited or multilingual environments.

This project addresses the demand for digital solutions that are easy to use, linguistically inclusive, and medically relevant. It highlights the potential of AI to act as a supportive partner in delivering more efficient, accurate, and accessible healthcare for both professionals and the public.

Keywords: Medical Chatbot, Artificial Intelligence, Multilingual Healthcare, Early Diagnosis, Natural Language Interaction, Healthcare Access, Digital Health Tools

Résumé

Avec l'intégration croissante de l'intelligence artificielle dans le secteur de la santé, le besoin d'outils intelligents capables de soutenir le diagnostic précoce et d'améliorer l'accessibilité aux soins devient de plus en plus pressant. Cette étude présente le développement d'un chatbot médical multilingue (arabe, français et anglais) conçu pour agir comme un assistant interactif, recevant des entrées simples telles que les symptômes, l'âge et le genre, et fournissant en retour un diagnostic possible, une brève description de l'affection, ainsi qu'un lien de référence médicale fiable.

Pour atteindre cet objectif, plusieurs modèles linguistiques de pointe ont été utilisés. Pour l'arabe, le chatbot exploite BERT Base Arabic, MARBERT, T5-Small et GigaBERT. Pour l'anglais, il intègre BioBERT, BioClinicalBERT, BlueBERT PubMed et GigaBERT. En l'absence de modèles spécialisés pour le domaine médical en français, le chatbot utilise les sorties traduites du meilleur modèle anglais (GigaBERT) afin de fournir des réponses fiables en français.

Le chatbot constitue un outil d'aide à la décision pour les médecins en permettant des évaluations initiales rapides, tout en offrant aux patients un accès simplifié à des informations médicales fiables — un atout particulièrement précieux dans les premières étapes avant une consultation ou dans des environnements multilingues ou à ressources limitées.

Ce projet répond à la nécessité de solutions numériques faciles à utiliser, inclusives sur le plan linguistique et pertinentes sur le plan médical. Il met en lumière le potentiel de l'intelligence artificielle

à jouer un rôle de partenaire dans la fourniture de soins de santé plus efficaces, précis et accessibles, tant pour les professionnels que pour le grand public.

Mots-clés : Chatbot médical, Intelligence artificielle, Santé multilingue, Diagnostic précoce, Interaction en langage naturel, Accès aux soins, Outils numériques de santé

الملخص

مع تزايد اندماج الذكاء الاصطناعي في قطاع الرعاية الصحية، تبرز الحاجة المتزايدة إلى أدوات ذكية تدعم التشخيص المبكر وتعزز إمكانية الوصول إلى الخدمات الصحية. تقدم هذه الدراسة تطوير روبوت دردشة طبي متعدد اللغات (العربية، الفرنسية، والإنجليزية)، مصمم ليكون مساعدًا تفاعليًا يستقبل مدخلات بسيطة مثل الأعراض، العمر، والجنس، ويستجيب بتشخيص محتمل، ووصف موجز للحالة، ورابط طبي موثوق به.

لتحقيق ذلك، تم استخدام مجموعة من نماذج اللغة الحديثة. بالنسبة للغة العربية، يعتمد الروبوت على Base BERT Arabic وMARBERT وT5-Small وGigaBERT. أما بالنسبة للإنجليزية، فقد تم استخدام BioBERT وBioClinicalBERT وPubMed BlueBERT وGigaBERT. ونظرًا لعدم توفر نماذج متخصصة في المجال الطبي للغة الفرنسية، يستخدم الروبوت الترجمة الناتجة عن أفضل نموذج باللغة الإنجليزية (GigaBERT) لتقديم إجابات موثوقة بالفرنسية.

يعمل هذا الروبوت كأداة دعم اتخاذ القرار للأطباء من خلال توفير تقييمات أولية سريعة، كما يمكّن المرضى من الحصول على معلومات صحية بطريقة سهلة ومبسطة، وهو أمر ذو أهمية خاصة في المراحل المبكرة قبل زيارة الطبيب أو في البيئات التي تعاني من نقص الموارد أو التعدد اللغوي.

يعالج هذا المشروع الحاجة إلى حلول رقمية سهلة الاستخدام، وشاملة لغويًا، وذات صلة طبية. كما يسلّط الضوء على قدرة الذكاء الاصطناعي على أن يكون شريكًا داعمًا في تقديم رعاية صحية أكثر كفاءة ودقة وسهولة في الوصول، سواءً للمهنيين أو لعامة الناس.

الكلمات المفتاحية: روبوت دردشة طبي، الذكاء الاصطناعي، الرعاية الصحية متعددة اللغات، التشخيص المبكر، التفاعل بلغة طبيعية، الوصول إلى الرعاية، أدوات الصحة الرقمية

Contents

Chapter I: Artificial Intelligence in Healthcare **17**

I.1 Introduction to Artificial Intelligence in Healthcare **18**
I.1.1 Definition of Artificial Intelligence (AI) in Healthcare . . . 18
I.1.2 Applications of AI in Healthcare . . . 18
I.1.3 History of AI Use in Medical Diagnosis . . . 19
I.1.3.1 Foundational Developments (1950s–1970s) . . . 19
I.1.3.2 Emergence of Medical Expert Systems (1970s–1980s) . . . 19
I.1.3.3 Adoption of AI and Big Data in Diagnostics (1990s–Early 2000s) . . . 19
I.1.3.4 Advancements in Deep Learning and Modern AI (2010–Present) . . . 20
I.1.3.5 Future Directions in AI-Driven Medical Diagnosis . . . 20
I.1.4 Evolution of Self-Diagnosis Tools . . . 20
I.1.4.1 Initial Symptom Checklists (Pre-2000s) . . . 20
I.1.4.2 Expansion of Online Symptom Checkers (2000s–2010s) . . . 21
I.1.4.3 Recent Advancements (2020s) . . . 21

I.2 Medical Chatbots: Concept and Significance **21**
I.2.1 Definition of Medical Chatbots . . . 22
I.2.2 Differences Between Traditional and Medical Chatbots . . . 22
I.2.3 The Role of Medical Chatbots in Improving Healthcare Access . . . 23
I.2.4 Global Examples: Ada Health, Buoy Health, Your.MD . . . 24

I.3 AI Technologies Utilized in Medical Chatbots **24**
I.3.1 Natural Language Processing (NLP) for Symptom Understanding . . . 24
I.3.2 Machine Learning for Disease Prediction . . . 25
I.3.3 Deep Learning for Complex Symptom Pattern Analysis . . . 26

I.4 Workflow of Medical Chatbots **26**
I.4.1 Data Collection: Symptoms, Demographics, and Medical History . . . 27
I.4.2 Data Analysis Using Artificial Intelligence Algorithms . . . 27
I.4.3 Preliminary Diagnosis Generation . . . 28
I.4.4 Recommendation Delivery . . . 28

I.5 Medical Databases and Knowledge Sources **28**
I.5.1 Data Sources: Medical Databases and Online Resources . . . 29
I.5.2 Data Quality Standards . . . 29
I.5.3 Examples: ICD-10 and SNOMED CT . . . 30

I.6 Medical Databases and Knowledge Sources **30**
I.6.1 Comparison Between Traditional and Advanced Algorithms . . . 30
I.6.2 Symptom-Based Algorithms . . . 31

I.6.3 Similarity-Based Algorithms . . . 32

I.7 Technical and Legal Challenges 32

I.7.1 Technical Challenges: Symptom Interpretation and Knowledge Base Maintenance . . . 32

I.7.2 Legal Challenges: Ensuring Compliance with Data Protection Laws . . . 32

I.7.3 Ethical Challenges: Risk of Misleading or Harmful Advice . . . 33

I.8 Future Trends and Innovations 33

I.8.1 Integration with Electronic Health Records (EHR) . . . 33

I.8.2 Incorporation of Voice and Image Analysis . . . 33

I.9 Summary 33

Chapter II: Studies on AI Applications in Medical Diagnosis 35

II.1 The Role of Artificial Intelligence in Modern Diagnostics 36

II.2 Transformer-Based Model for Early Lung Cancer Detection Using EHR Data 36

II.2.1 Study Purpose . . . 36

II.2.2 Methodology . . . 36

II.2.3 Dataset Overview . . . 37

II.2.4 System Architecture . . . 37

II.2.5 Key Findings . . . 38

II.2.6 Limitations . . . 38

II.2.7 Suggested Improvements . . . 39

II.3 Transformer-Based Multimodal Fusion for Enhanced Tuberculosis Diagnosis 39

II.3.1 Study Purpose . . . 39

II.3.2 Methodology . . . 39

II.3.3 Dataset Overview . . . 40

II.3.4 System Architecture . . . 41

II.3.5 Key Findings . . . 42

II.3.6 Limitations . . . 42

II.3.7 Suggested Improvements . . . 43

II.4 Chatbot Support for Breast Cancer Patients: A One-Year Prospective Study 43

II.4.1 Study Purpose . . . 43

II.4.2 Methodology . . . 43

II.4.3 Dataset Overview . . . 44

II.4.4 System Architecture . . . 44

II.4.5 Key Findings . . . 45

II.4.6 Limitations . . . 46

II.4.7 Suggested Improvements . . . 46

II.5 AI Chatbot Evaluation for Heart Failure Education **47**
II.5.1 Study Purpose . . . 47
II.5.2 Methodology . . . 47
II.5.3 Dataset Overview . . . 48
II.5.4 System Architecture . . . 48
II.5.5 Key Findings . . . 48
II.5.6 Limitations . . . 48
II.5.7 Suggested Improvements . . . 49

II.6 Evaluating the Accuracy of a Chatbot in Answering Questions Patients Should Ask Before Taking a New Medication **49**
II.6.1 Study Purpose . . . 49
II.6.2 Methodology . . . 49
II.6.2.1 Scoring Criteria for Chatbot Responses . . . 50
II.6.3 Dataset Overview . . . 50
II.6.4 System Architecture . . . 51
II.6.5 Key Findings . . . 51
II.6.6 Limitations . . . 52
II.6.7 Proposed Improvements . . . 52

II.7 Comparative Analysis of Selected Studies **53**

II.8 Summary **53**

Chapter III: System Conception and Implementation **55**

III.1 Introduction to the Multilingual Medical Chatbot **56**

III.2 System Overview **56**
III.2.1 Project Objective . . . 56
III.2.2 Motivation and Use Cases . . . 56
III.2.3 Core Functionalities . . . 57
III.2.4 Multilingual Pipeline Architecture and Language Handling . . . 58
III.2.5 Technologies and Frameworks for Medical NLP Integration . . . 59

III.3 System Architecture **60**
III.3.1 Modular System Architecture Overview . . . 60
III.3.2 Backend Services and APIs . . . 62
III.3.3 AI Model Integration . . . 62
III.3.4 Selection of Multilingual Biomedical Language Models . . . 63
III.3.5 Fine-Tuning and Integration . . . 64
III.3.6 Design of Medical Data Schema and Storage Strategy . . . 65

III.4 Dataset Preparation **66**

III.4.1 Data Sources and Collection . . . 66
III.4.2 Dataset Structure and Attributes . . . 69
III.4.3 Preprocessing Methodologies . . . 70
III.4.3.1 General Preprocessing Procedures . . . 70
III.4.3.2 Arabic Text Preprocessing . . . 75
III.4.3.3 Preprocessing Summary Figure: . . . 76
III.4.4 Data Augmentation Strategies . . . 77
III.4.5 Splitting the Dataset . . . 78
III.4.6 Label Encoding and Class Mapping . . . 79
III.4.7 Deployment and Runtime Environment . . . 80

III.5Summary **80**

Chapter IV: Results and Evaluation **81**

IV.1 Chapter Overview **82**

IV.2 Evaluation Methodology **82**
IV.2.1 Objectives of Evaluation . . . 82
IV.2.2 Performance Metrics . . . 82

IV.3 Overview of Experimental Workflow **83**

IV.4 Dataset Collection and Structure **84**
IV.4.1 Class Distribution and Imbalance . . . 84

IV.5 Phase 1: Baseline on Imbalanced Dataset **85**
IV.5.1 Objective . . . 85
IV.5.2 Models Evaluated . . . 85
IV.5.3 Experimental Setup . . . 85
IV.5.4 Results . . . 86
IV.5.5 Analysis . . . 86
IV.5.6 Implications . . . 86

IV.6 Phase 2: Data Augmentation **86**
IV.6.1 Introduction . . . 86
IV.6.2 Augmentation Techniques . . . 86
IV.6.3 Experimental Setup . . . 87
IV.6.4 Evaluation . . . 87
IV.6.5 Analysis and Findings . . . 87
IV.6.6 Limitations . . . 87
IV.6.7 Summary of Findings . . . 87

IV.7 Phase 3: Balanced Dataset + Augmentation **88**

IV.8 Arabic Models **88**
IV.8.1 Selection of Arabic Biomedical Transformer Models 88
IV.8.2 Fine-tuning the Models 88
IV.8.3 Evaluation 88
IV.8.4 Analysis and Findings 89

IV.9 English Biomedical Models **90**
IV.9.1 Translation of Arabic Dataset 90
IV.9.2 Selection of Biomedical Transformer Models 90
IV.9.3 Fine-tuning the Models 91
IV.9.4 Evaluation 91
IV.9.5 Analysis and Findings 91

IV.10 Discussion **92**

IV.11 Frontend Design and Interaction Flow **93**

IV.12 User Interface Demonstration: English, Arabic, and French Chatbot Interactions **97**

IV.13 Conclusion **99**

List of Tables

Table I.2.1. *Key Differences Between Traditional and Medical Chatbots* 23
Table I.6.1. *Comparison Between Traditional and Advanced Algorithms* 31
Table II.3 .1. *Summary of Key Methodological Steps [60].* 40
Table II.3 .2. *Dataset Demographics* 41
Table II.3 .3. *Performance Comparison with Other Models [60].* 42
Table II.3 .4. *Summary of Study Limitations.* 42
Table II.4 .1. *User Interaction Metrics* 46
Table II.5 .1. *Dataset and Evaluation Overview* 48
Table II.6 .1. *Chatbot Accuracy by Question Type* 51
Table II.7 .1. *Comparison of Selected Studies Relevant to AI-Based Medical Chatbot Design* 53
Table III.2.1. *Main characteristics of the chatbot system* 58
Table III.2.2. *Development Libraries and Frameworks Used in the System* 60
Table III.3.1. *Summary of selected multilingual biomedical language models.* . . . 64
Table III.3.2. *Overview of the medical dataset schema and main fields.* 66
Table III.4.1. *Data Sources for Dataset Collection* 69
Table III.4.2. *Data Augmentation Techniques for Multilingual Medical Text Classification* 77
Table IV.4.1. *Dataset Attribute Descriptions* 84
Table IV.5.1. *Performance of Baseline Models on Imbalanced Data* 86
Table IV.6.1. *BERT with Augmentation on Imbalanced Data* 87
Table IV.8.1. *Arabic Biomedical Models Performance* 89
Table IV.9.1. *English Biomedical Models Performance* 92
Table IV.10.1. *Summary of key factors influencing model performance.* 93
Table IV.11.1. *Summary of core frontend design features in the MediLingo chatbot interface.* 97

List of Figures

Figure I.4.1. *Workflow of a Medical Chatbot System* 27
Figure II.2 .1. *Workflow of a Medical Chatbot System* 38
Figure II.3 .1. *Workflow of the proposed transformer-based multimodal TB diagnosis system [60].* 40
Figure II.3 .2. *Distribution of abnormalities across clinical features.* 41
Figure II.3 .3. *System architecture showing multimodal data flow and fusion layers [119].* 41
Figure II.3 .4. *Ablation study: accuracy across different neural network architectures [60].* 42
Figure II.4 .1. *Workflow of a Medical Chatbot System* 44
Figure II.4 .2. *Chatbot System Architecture: Interaction between user interface, NLP engine, knowledge base, and response generator.* 45
Figure II.4 .3. *Monthly User Engagement Over the Study Period* 46
Figure II.6 .1. *Workflow of a Medical Chatbot System* 50
Figure II.6 .2. *Workflow of a Medical Chatbot System* 51
Figure II.6 .3. *Performance of Chatbot Across Medication Inquiry Categories* . . 52
Figure III.2.1. *Motivations and use cases for an AI-based multilingual medical chatbot.* 57
Figure III.2.2. *Multilingual Preprocessing Pipeline.* 59
Figure III.3.1. *High-level architecture of the multilingual medical chatbot, showing interactions between the frontend, Flask-powered backend, AI model layer, and data storage.* 61
Figure III.3.2. *Core components of the system architecture.* 62
Figure III.3.3. *AI Model Integration Workflow* 65
Figure III.4.1. *Data extraction workflow from Sohati — Part 1.* 67
Figure III.4.2. *Data extraction workflow from Sohati — Part 2.* 68
Figure III.4.3. *Diagrammatic illustration of a record in the dataset of the MediLingo system.* 70
Figure III.4.4. *Code used for Text Cleaning* 71
Figure III.4.5. *Tokenization of the Dataset* 71
Figure III.4.6. *Calling the AraBERT Tokenizer* 72
Figure III.4.7. *Calling the MARBERT Tokenizer* 72
Figure III.4.8. *Calling the T5-small Tokenizer* 72
Figure III.4.9. *Calling the GigaBERT Tokenizer* 73
Figure III.4.10. *Calling the BioBERT Tokenizer* 73
Figure III.4.11. *Calling the BioClinicalBERT Tokenizer* 73
Figure III.4.12. *Calling the PubMedBERT Tokenizer* 74
Figure III.4.13. *Calling the GigaBERT Tokenizer* 74
Figure III.4.14. *Code used for Stop-word Removal* 74
Figure III.4.15. *Code used for removing Arabic diacritics.* 75

Figure III.4.16. *Code used for Cleaning Arabic Text* 76
Figure III.4.17. *Summary of Preprocessing Methodology for Multilingual Medical Text Classification* 76
Figure III.4.18. *Illustration of Noise Injection for text data, simulating typographical or OCR-like errors.* 77
Figure III.4.19. *Back-Translation workflow showing source-to-intermediate-to-source language conversion.* 78
Figure III.4.20. *Constrained word shuffling example, showing context-preserving rearrangements.* 78
Figure III.4.21. *Code used for splitting the dataset into training and testing sets.* . . . 79
Figure III.4.22. *Example of label encoding and class mapping in Python.* 79
Figure IV.3.1. *Overview of the experimental workflow* 84
Figure IV.4.1. *Distribution of disease classes showing data imbalance* 85
Figure IV.8.1. *Arabic Models Performance Comparison across Accuracy, Precision, Recall, and F1-Score. GigaBERT shows the best overall performance.* 90
Figure IV.9.1. *English Models Performance Comparison across Accuracy, Precision, Recall, and F1-Score. GigaBERT (Best) shows superior results across all metrics.* 92
Figure IV.10.1. *Performance comparison of Arabic language models.* 93
Figure IV.10.2. *Performance comparison of English language models.* 93
Figure IV.11.1. *Welcome screen of the MediLingo chatbot, introducing the system and prompting users to begin interaction.* 94
Figure IV.11.2. *Language selection screen offering users a choice between Arabic, English, and French interface modes.* 95
Figure IV.11.3. *Arabic chatbot interface with RTL symptom input and diagnosis output.* 95
Figure IV.11.4. *English chatbot interface with LTR input and response flow.* 96
Figure IV.11.5. *French chatbot interface with LTR input and localized medical interaction.* 96
Figure IV.12.1. *MediLingo chatbot interaction in English: user input, predicted disease name, descriptions, and reference URL.* 98
Figure IV.12.2. *MediLingo chatbot interaction in Arabic: RTL layout with user input, predicted diagnosis, descriptions, and reference URL.* 98
Figure IV.12.3. *MediLingo chatbot interaction in French: LTR layout with user input, predicted diagnosis, explanations, and medical reference link.* . . . 99

List of Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
ANN	Artificial Neural Network
AraBERT	Arabic BERT Language Model
BERT	Bidirectional Encoder Representations from Transformers
BioBERT	Biomedical Bidirectional Encoder Representations from Transformers
BioClinicalBERT	Clinical Domain Version of BioBERT
BlueBERT	BERT Model Trained on PubMed Abstracts and Clinical Notes
CDSS	Clinical Decision Support Systems
EHR	Electronic Health Record
F1-Score	Harmonic Mean of Precision and Recall
GDPR	General Data Protection Regulation
GigaBERT	Multilingual BERT Variant
HIPAA	Health Insurance Portability and Accountability Act
ICD-10	International Classification of Diseases, 10th Revision
MARBERT	Multilingual Arabic BERT
MCC	Matthews Correlation Coefficient
MDSS	Medical Decision Support Systems
MIT	Massachusetts Institute of Technology
NLP	Natural Language Processing
OCR	Optical Character Recognition
PubMedBERT	BERT Model Pretrained on PubMed Texts
SNOMED CT	Systematized Nomenclature of Medicine—Clinical Terms
T5-small	Text-To-Text Transfer Transformer (small version)

General Introduction

Day in and day out, we come across the words *Artificial Intelligence (AI)*—words that used to be the domain of science fiction, but are now firmly entrenched in our daily lives. In a general sense, AI is understood as the ability of machines to perform tasks that would otherwise require human intelligence either autonomously or under human control. Such tasks can range from simple decision-making to complex pattern recognition, learning, and problem-solving. Currently, artificial intelligence technologies are incorporated into many aspects of modern society, such as smartphones, smart home systems, transportation, finance, and notably, the healthcare industry. This rapid and widespread adoption underscores the revolutionary potential of AI, particularly in areas that rely heavily on data analysis and decision-making.

The idea of machines replicating human intelligence has intrigued researchers for decades. In 1958, the Massachusetts Institute of Technology (MIT) established its Artificial Intelligence Laboratory, among the earliest institutional attempts to formalize research on intelligent computing. Around this time, computer scientists and medical professionals began investigating potential applications of these technologies in medicine. Their vision was clear: intelligent systems capable of storing vast amounts of medical data, recognizing patterns, and assisting clinicians in diagnosing diseases and determining treatment courses.

This theoretical strategy led to the development of medical decision support systems (MDSS). The early iterations of these systems were intended to assist healthcare professionals during the diagnostic process by providing evidence-based recommendations drawn from comprehensive medical databases. By replicating the diagnostic thinking of experienced doctors, MDSS aimed to reduce diagnostic errors, ensure consistency in care, and improve overall patient outcomes. These systems represented a paradigm shift in medicine, marking a transition from physician-centered care to more collaborative decision-making supported by intelligent tools.

Over time, the scope of AI in healthcare has expanded significantly. Advanced machine learning algorithms, natural language processing, and neural network models have enabled AI to analyze complex data types, including medical imaging, electronic health records (EHRs), genomic information, and real-time patient monitoring data. These advances have enhanced diagnostic accuracy, refined treatment protocols, and facilitated the development of personalized medicine.

As AI systems have evolved, new modalities of human-machine interaction have emerged—particularly through conversational interfaces. A key and increasingly prominent application of AI in healthcare is the use of *chatbots*. Chatbots are computer programs designed to simulate human conversation via text or voice-based interactions. Leveraging AI technologies like natural language processing and machine learning, these virtual agents can understand user queries, extract relevant information, and provide context-appropriate responses.

In healthcare, chatbots serve a variety of roles, including answering patients' questions, scheduling appointments, offering mental health counseling, and conducting initial symptom assessments. The use of chatbots presents several potential benefits. They offer round-the-clock access to healthcare information, which is particularly valuable in underserved regions or during high-demand periods. They also help reduce the administrative burden on healthcare professionals, allowing them to focus on more complex tasks. Additionally, chatbots can play an important role in health education and disease prevention by delivering timely and personalized information to users.

In mental health contexts, chatbots are used to offer cognitive behavioral therapy techniques, provide emotional support, and detect signs of depression or anxiety through conversational cues. Despite their many advantages, the use of healthcare chatbots raises important concerns. Issues such as data privacy, ethical implications, clinical accuracy, and the risk of misdiagnosis must be carefully considered. There are also questions surrounding patient trust, cultural acceptance, and legal accountability for chatbot developers and healthcare providers.

As these technologies become increasingly embedded in healthcare delivery systems, it is essential to critically examine not only their technical capabilities but also their broader social, ethical, and regulatory implications. This research explores the evolving role of chatbots in healthcare, with a focus on their potential to enhance patient care, improve operational efficiency, and drive innovation.

To guide this investigation, the study is structured as follows:

- **Chapter I: Artificial Intelligence in Healthcare** – This chapter provides an overview of AI technologies and explores their historical and current applications in the healthcare sector.
- **Chapter II: Studies on AI Applications in Medical Diagnosis** – This chapter reviews existing literature and case studies focused on AI-driven diagnostic systems and decision support tools.
- **Chapter III: System Conception and Implementation** – This chapter outlines the technical design, architecture, and implementation process of the chatbot system developed for this research.
- **Chapter IV: Results and Evaluation** – This chapter presents the experimental results, evaluation metrics, and analysis of the chatbot's performance and clinical relevance.

Through the exploration of real case studies, current research, and forward-looking perspectives, this thesis aims to offer a comprehensive understanding of how chatbot technology is reshaping the healthcare landscape and what it signifies for the future of medical practice.

Chapter I

Artificial Intelligence in Healthcare

I.1 Introduction to Artificial Intelligence in Healthcare

Artificial Intelligence (AI) has emerged as a transformative power in the healthcare industry, transforming the face of modern medicine in a remarkable manner. Its uses traverse a wide range of fields, ranging from clinical diagnosis and treatment planning to patient engagement and hospital administration [102]. The foundation of AI's strength lies in its ability to process and analyze vast volumes of complex medical data at incredibly high speeds with an accuracy much higher than humans [50]. Artificially intelligent tools such as intelligent symptom screeners, predictive analytics tools, and virtual health advisors are becoming integral parts of patient care procedures at an ever-growing speed, enabling users to receive timely and personalized health information even before meeting a health expert [83, 43].

I.1.1 Definition of Artificial Intelligence (AI) in Healthcare

Healthcare AI refers to the use of advanced computational techniques and algorithms to process medical data, assist in diagnosis, optimize treatment regimens, and improve patient care [105, 25]. AI is a general term encompassing numerous technologies, including machine learning, natural language processing, generative models, and robotics, that work together to improve clinical decision-making and healthcare delivery [10, 32, 44].

I.1.2 Applications of AI in Healthcare

AI is being introduced into healthcare, providing solutions to improve efficiency, accuracy, and patient outcomes [105, 11]. Some of the key applications include:

- **Symptom Assessment and Virtual Health Assistants:** AI platforms allow patients to input symptoms for personalized health screening. They help guide patients through triaging and referral to proper care [10].
- **Medical Imaging and Diagnostic Support:** Machine learning and deep neural networks read MRI and CT scans. AI tools aid in the early and precise detection of anomalies, helping in diagnosis [44].
- **Predictive Analytics and Early Disease Detection:** AI systems, when trained on large-scale clinical datasets, are capable of identifying patterns indicative of disease onset. This facilitates the development of predictive models that enable clinicians to initiate early interventions, potentially improving prognoses [25].
- **Robotic Surgery:** AI robot systems like the Da Vinci Surgical System enhance surgical precision and stability. They enable minimally invasive surgeries, which lead to quicker recovery and better patient safety [11].
- **Natural Language Processing (NLP):** NLP techniques extract and retrieve valuable clinical information from unstructured texts like EHRs, notes, and literature to enable easier access to information and evidence-based decision-making [32].

- **Drug Discovery and Development:** AI accelerates drug development by analyzing biochemical and molecular data, identifying promising compounds, and reducing costs and timelines [10].
- **Personalized Medicine:** AI combines genomic, phenotypic, and clinical data to tailor treatments to individuals, enhancing effectiveness and minimizing side effects [105].

I.1.3 History of AI Use in Medical Diagnosis

AI has progressively shaped medical diagnosis, evolving from early rule-based systems to modern deep learning models. This section outlines the key milestones in that transformation.

I.1.3.1 Foundational Developments (1950s–1970s)

The conceptual foundation for using artificial intelligence (AI) in medicine began in the 1950s with early computational models aiming to simulate human reasoning. Notable contributions during this period include:

- **Alan Turing's Research (1950):** Turing's pioneering ideas on intelligent machines established the theoretical groundwork for applying computational logic in medical contexts [104].
- **Emergence of Expert Systems (1970s):** Early expert systems were developed to replicate human medical decision-making using rule-based logic, marking a significant advancement in computer-assisted diagnostics [93].

I.1.3.2 Emergence of Medical Expert Systems (1970s–1980s)

This era saw a surge in AI-based diagnostic applications through the development of knowledge-driven systems designed to support clinical decisions. Prominent systems include:

- **MYCIN (Mid-1970s):** Developed at Stanford University, MYCIN was among the first AI-based diagnostic tools, designed to identify bacterial infections and recommend antibiotic treatments. Despite its diagnostic accuracy, it was not adopted in clinical practice due to ethical and legal concerns [93].
- **INTERNIST-1 (Late 1970s):** Designed to assist physicians with complex internal medicine cases, INTERNIST-1 utilized a vast knowledge base to generate differential diagnoses from input symptoms [66].

I.1.3.3 Adoption of AI and Big Data in Diagnostics (1990s–Early 2000s)

With enhanced computational capabilities and the growth of medical databases, AI systems transitioned from rule-based logic to data-driven approaches. Key developments included:

- **Artificial Neural Networks (ANNs):** Neural networks began to be applied in diagnostic imaging, including the interpretation of X-rays and MRI scans [57].

- **Clinical Decision Support Systems (CDSS):** These systems supported clinicians by offering evidence-based recommendations derived from vast datasets, thereby improving diagnostic accuracy and treatment planning [12].

I.1.3.4 Advancements in Deep Learning and Modern AI (2010–Present)

Recent years have witnessed significant breakthroughs in AI through deep learning and sophisticated data processing techniques. Contemporary innovations include:

- **Medical Imaging Analysis:** In 2016, deep learning algorithms exceeded the diagnostic capabilities of dermatologists in detecting skin cancer from images [29]. By 2018, the U.S. FDA approved the first autonomous AI system for diagnosing diabetic retinopathy [4].
- **Genomics and Precision Medicine:** AI tools are now integral in analyzing genomic data to predict hereditary diseases, such as rare cancers, aiding in personalized treatment strategies [55].
- **Medical Robotics:** Robotic systems have been developed to conduct automated medical consultations, improving accessibility and efficiency in healthcare delivery [102].
- **Telemedicine:** AI enhances remote diagnostics, particularly in underserved or rural regions, by supporting clinicians through virtual consultations and image analysis [91].

I.1.3.5 Future Directions in AI-Driven Medical Diagnosis

AI continues to evolve as a critical asset in modern healthcare. Anticipated developments include:

- Accelerated and more accurate diagnoses of rare and complex diseases.
- Integration with Internet of Things (IoT) devices for real-time patient monitoring.
- Improved early detection of critical conditions such as cancer and cardiovascular diseases.

I.1.4 Evolution of Self-Diagnosis Tools

The development of self-diagnosis technologies reflects a growing trend toward patient empowerment and digital health engagement. From early web-based symptom lists to advanced AI-powered platforms, these tools have significantly impacted how individuals access and interpret health information.

I.1.4.1 Initial Symptom Checklists (Pre-2000s)

Early self-diagnostic tools were static and informational, providing basic symptom listings and generalized condition descriptions. These platforms served primarily as educational resources for users seeking initial guidance.

- **WebMD (1996):** One of the first digital platforms to offer an interactive symptom checker, WebMD allowed users to enter symptoms and receive possible condition matches, quickly becoming a trusted online health reference [111].
- **Mayo Clinic Symptom Checker (Early 2000s):** The Mayo Clinic introduced its own tool, providing clinically reviewed symptom information to help users make informed decisions about seeking care [64].

I.1.4.2 Expansion of Online Symptom Checkers (2000s–2010s)

As internet access and digital literacy expanded, symptom checkers became more interactive and data-driven. These tools began incorporating algorithms to generate diagnostic suggestions, though their clinical reliability remained limited.

- **PubMed Central Study (2012):** This study assessed the diagnostic accuracy of online symptom checkers and highlighted their potential utility as well as their constraints in guiding users without medical oversight [90].
- **Digital Health Review (2018):** A systematic review contextualized symptom checkers within the broader digital health movement, emphasizing their role in changing how patients interact with healthcare systems and information [17].

I.1.4.3 Recent Advancements (2020s)

In the last five years, numerous comprehensive reviews and evaluations have assessed modern symptom checker platforms:

- A 2023 systematic review reported that diagnostic accuracy remains low—often below 40%—and triage accuracy is inconsistent, though improving in some tools [85].
- A 2024 study by Aissaoui Ferhi et al. demonstrated that machine learning techniques—such as random forests and logistic regression—significantly improved diagnostic accuracy in clinical-vignette testing [6].
- A 2024 qualitative investigation in Germany highlighted that symptom checker users are typically younger, digitally literate, and generally reported positive experiences, though concerns about accuracy and shared decision-making persisted [69, 86].
- A real-world study comparing Ada, WebMD, and ChatGPT to emergency department diagnoses showed symptom checkers and LLMs matching physician diagnostic top-3 accuracy in approximately one-third of cases—highlighting moderate reliability [37].

I.2 Medical Chatbots: Concept and Significance

In today's digital era, technology is significantly transforming healthcare services and enhancing patient experiences. One notable innovation in this field is the rise of medical chatbots, which offer

preliminary medical consultations and assist patients by analyzing their symptoms. These AI-powered systems leverage natural language processing (NLP) and machine learning to comprehend user inquiries and provide medically informed responses [14, 68]. Medical chatbots help alleviate pressure on healthcare systems, deliver immediate solutions to individuals seeking medical advice, and ensure accessible healthcare support, regardless of time or location [7]. This section delves into the concept of medical chatbots, their distinction from traditional chatbots, their role in improving healthcare access, and some global examples of their implementation.

I.2.1 Definition of Medical Chatbots

Medical chatbots are AI-based virtual assistants designed to interact with users and offer preliminary medical consultations by analyzing symptoms and healthcare data. These chatbots operate through websites or mobile applications and assist in:

- Analyzing symptoms and suggesting potential conditions [31].
- Guiding patients on appropriate next steps, such as visiting a healthcare provider or using home remedies [5].
- Reminding patients about medication schedules and regular health check-ups [68].
- Supporting mental health through cognitive behavioral therapy (CBT) techniques for managing stress and anxiety [34].

Medical chatbots provide a cost-effective and efficient solution to address gaps in healthcare access, particularly in underserved and overburdened regions where medical consultations may be difficult to obtain [7].

I.2.2 Differences Between Traditional and Medical Chatbots

Although both traditional and medical chatbots utilize AI and machine learning, they differ markedly in terms of their objectives, complexity, and associated risks.

Feature	Traditional Chatbots	Medical Chatbots
Application	Customer support, FAQs, entertainment	Medical consultations, symptom checking, patient support
Knowledge Base	General business or customer service data	Medical databases, clinical research, healthcare guidelines
Privacy & Security	Minimal regulatory requirements	Must comply with strict regulations (e.g., HIPAA, GDPR)
Risk Factor	Low – mistakes lead to minor inconveniences	High – errors can directly affect health outcomes
User Interaction	Basic rule-based responses	AI-driven, personalized health advice

Table I.2.1: Key Differences Between Traditional and Medical Chatbots

In contrast to traditional chatbots, medical chatbots require higher accuracy, robust validation, and must comply with stringent ethical and regulatory standards to protect patient safety and ensure data privacy [7, 5, 68].

I.2.3 The Role of Medical Chatbots in Improving Healthcare Access

Medical chatbots are essential in enhancing healthcare services through several key contributions:

- **Round-the-Clock Availability:** Patients can obtain health information at any time, eliminating the need to wait for a doctor's availability [5].
- **Assistance for Remote and Underserved Areas:** Medical chatbots provide immediate medical guidance to people in rural or resource-limited regions, reducing the need for long-distance travel to healthcare facilities [7].
- **Alleviating Pressure on Healthcare Systems:** By addressing basic health queries, chatbots help reduce unnecessary hospital visits, allowing healthcare professionals to focus on more critical cases [68].
- **Promoting Preventive Health:** Chatbots offer advice on maintaining a healthy lifestyle, preventing chronic diseases, and taking necessary medical tests, encouraging proactive health management [31].

- **Boosting Patient Engagement:** With personalized reminders for medications and check-ups, chatbots enhance patient adherence to prescribed treatments and foster greater involvement in their healthcare [34].

These benefits illustrate how medical chatbots are transforming healthcare access, particularly for underserved and remote populations.

I.2.4 Global Examples: Ada Health, Buoy Health, Your.MD

Several AI-powered medical chatbots have been effectively deployed globally, significantly enhancing healthcare accessibility and efficiency. Noteworthy examples include:

- **Ada Health:** An AI-driven chatbot designed to analyze symptoms and generate potential diagnoses based on a comprehensive medical dataset, providing users with accurate and user-friendly medical insights [1].
- **Buoy Health:** This platform employs AI algorithms to assess symptoms and recommend the most appropriate course of action, whether it involves consulting a healthcare provider or utilizing at-home care options [2].
- **Your.MD:** A digital health assistant that delivers tailored health advice and symptom evaluation, utilizing a reliable medical database to provide evidence-based recommendations [3].

These examples highlight the growing impact of AI-based medical chatbots in improving patient care and streamlining healthcare processes on a global scale.

I.3 AI Technologies Utilized in Medical Chatbots

Medical chatbots have become powerful tools in healthcare, harnessing artificial intelligence (AI) to support patients, healthcare providers, and medical researchers. These chatbots rely on a range of AI technologies to improve symptom assessment, disease prediction, and patient interactions. The three primary AI technologies employed in medical chatbots are Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL). Each of these technologies plays a pivotal role in enabling chatbots to interpret user input, analyze medical conditions, and generate intelligent, contextually relevant responses.

I.3.1 Natural Language Processing (NLP) for Symptom Understanding

Natural Language Processing (NLP) constitutes a fundamental component in the architecture of medical chatbots, enabling them to interpret, process, and generate human language in a clinically meaningful manner. Through NLP, chatbots are capable of:

- Extracting medically relevant information such as symptoms, conditions, and patient concerns from unstructured user input.

- Identifying and interpreting medical terminologies, abbreviations, and the diverse linguistic expressions used by patients to describe health-related issues.
- Facilitating multilingual communication and generating contextually appropriate, human-like responses to enhance user interaction.

Key NLP techniques, including Named Entity Recognition (NER), are employed to identify specific clinical entities—such as disease names and symptom descriptions—within user-generated text [13]. Recent advancements have seen the application of contextual language models such as BERT (Bidirectional Encoder Representations from Transformers) and its biomedical variant, BioBERT, which have significantly improved the semantic understanding of complex medical narratives [62].

Empirical evidence suggests that NLP-powered medical chatbots, when trained on specialized clinical datasets, can achieve symptom detection accuracies exceeding 85%, underscoring their reliability and efficacy in preliminary health assessment tasks [62]. These advancements position NLP as a critical enabler in the delivery of intelligent, responsive, and accessible digital healthcare solutions.

I.3.2 Machine Learning for Disease Prediction

Machine Learning (ML) constitutes a core technological framework in medical chatbots, facilitating the prediction of potential diseases based on user-reported symptoms. ML algorithms are trained on large-scale datasets comprising clinical records, symptomatology, and confirmed diagnoses, enabling the system to identify complex patterns and derive data-driven inferences regarding possible health conditions.

Several ML paradigms are utilized within medical chatbot applications:

- **Supervised Learning:** Involves training models on labeled datasets, where the input features (e.g., symptoms) are associated with known outcomes (e.g., diagnoses). This approach enables the system to learn direct mappings between symptom profiles and specific diseases.
- **Unsupervised Learning:** Applied to unlabeled data to uncover hidden structures and groupings, such as clustering similar patient cases or identifying novel symptom correlations, which can support differential diagnosis.
- **Reinforcement Learning:** Utilizes a reward-based feedback mechanism to iteratively optimize the chatbot's decision-making process, improving diagnostic accuracy and response quality over time.

Empirical studies have demonstrated the efficacy of ML in enhancing diagnostic capabilities within chatbot systems. For instance, models utilizing decision trees and support vector machines (SVMs) have achieved diagnostic accuracies exceeding 90% in identifying common infectious conditions [82]. Furthermore, research indicates that ML-powered chatbots can perform at levels comparable to primary care physicians in specific diagnostic scenarios [8]. These advancements underscore the transformative potential of machine learning in augmenting clinical decision support and improving the scalability of healthcare delivery.

I.3.3 Deep Learning for Complex Symptom Pattern Analysis

Deep Learning (DL) represents an advanced subset of machine learning that significantly enhances the analytical capabilities of medical chatbots, particularly in identifying complex and nuanced symptom patterns often missed by traditional models. DL architectures such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and Transformer models enable sophisticated pattern recognition and temporal analysis in clinical data.

Key contributions of deep learning in medical chatbot functionality include:

- **Temporal Pattern Recognition:** RNNs and LSTMs are particularly effective in modeling disease progression over time by capturing sequential dependencies in symptom evolution.
- **Multi-symptom Correlation:** DL models can detect intricate relationships between co-occurring symptoms, which is crucial for diagnosing complex conditions such as autoimmune or rare diseases.
- **Context Retention in Conversations:** Transformer-based models, such as GPT-4, enhance chatbot performance by maintaining contextual coherence in multi-turn dialogues, leading to more accurate and personalized interactions.

Empirical research underscores the value of DL in clinical diagnostics. For instance, CNN and LSTM-based systems have demonstrated high diagnostic performance even with limited datasets, particularly in rare disease identification [84]. Additionally, transformer architectures like GPT-4 have shown state-of-the-art accuracy in medical question answering tasks, outperforming traditional NLP systems in both generalizability and contextual understanding [94].

These capabilities position deep learning as a critical enabler of next-generation medical chatbots, capable of delivering high-precision support in complex diagnostic scenarios.

I.4 Workflow of Medical Chatbots

Medical chatbots operate through a structured, multi-stage workflow designed to facilitate accurate symptom assessment, intelligent data analysis, and the delivery of personalized health recommendations. This systematic process enables effective interaction between users and AI-driven healthcare platforms. The primary stages of this workflow include data collection, algorithmic analysis, preliminary diagnostic generation, and recommendation delivery.

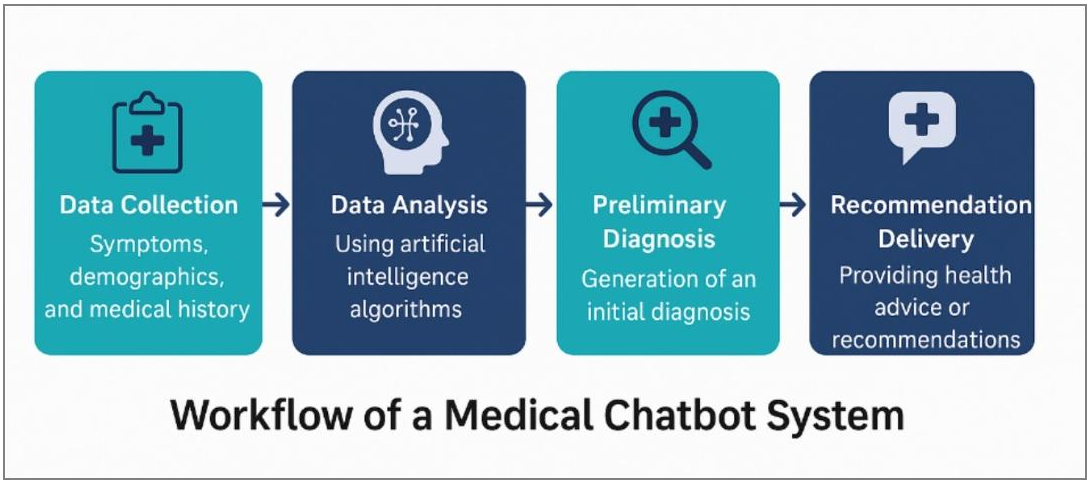


Figure I.4.1: Workflow of a Medical Chatbot System

I.4.1 Data Collection: Symptoms, Demographics, and Medical History

The initial phase involves the acquisition of critical user data, which forms the foundation for subsequent analysis. Key data elements include:

- **Symptoms:** Users describe their health concerns in natural language. The chatbot leverages Natural Language Processing (NLP) techniques to extract and standardize symptom-related terminology [13].
- **Demographics:** Information such as age, gender, and other demographic factors is essential for contextualizing medical risk, as the prevalence and manifestation of many diseases vary across populations [62].
- **Medical History:** Details on previous diagnoses, chronic illnesses, current medications, and family medical history allow the chatbot to tailor its analysis to the individual's clinical background.

I.4.2 Data Analysis Using Artificial Intelligence Algorithms

Once data is collected, it undergoes sophisticated analysis using various AI methodologies to identify potential diagnoses:

- **Linear Models (LMs):** Employed for pattern recognition by comparing user data to existing medical databases to predict possible conditions [82].
- **Deep Learning Techniques:** Neural networks such as Long Short-Term Memory (LSTM) and Transformer models are used to identify complex symptom interrelations and temporal patterns in user input [84].
- **Strategic Modeling Approaches:** These models assess possible outcomes or complications, enhancing the system's decision-making capability in cases involving multifactorial diseases.

Studies have shown that, when trained on robust datasets, these models can achieve diagnostic accuracies exceeding 90% in certain clinical scenarios [8].

I.4.3 Preliminary Diagnosis Generation

Following data analysis, the chatbot formulates a preliminary diagnosis by:

- Matching input symptoms with known disease profiles from curated medical databases.
- Assigning probability scores to each diagnostic possibility, reflecting the AI model's confidence in its inference.
- Classifying the urgency of the case, distinguishing between emergent conditions and non-critical issues manageable through self-care.

Evidence suggests that Transformer-based models integrated into medical chatbots can reach diagnostic accuracies comparable to primary care physicians in selected clinical domains [94].

I.4.4 Recommendation Delivery

Based on the preliminary diagnosis, the chatbot delivers context-specific recommendations, including:

- **Self-Care Guidance:** For minor conditions, advice may include home remedies or the use of non-prescription medications.
- **Preventive and Lifestyle Recommendations:** Suggestions may be offered regarding diet, physical activity, and other health-promoting behaviors.
- **Medical Referral:** In more serious cases, the chatbot advises users to consult healthcare professionals and may even integrate with telemedicine platforms to facilitate appointment scheduling [8].
- **Emergency Alerts:** For acute or life-threatening symptoms (e.g., chest pain suggestive of myocardial infarction), the chatbot recommends immediate medical attention.

These recommendations are continuously refined through user feedback and adaptive learning mechanisms, thereby improving the system's accuracy and reliability over time [82].

I.5 Medical Databases and Knowledge Sources

The effectiveness and reliability of medical chatbots fundamentally depend on the quality of the knowledge bases and medical resources they utilize. These systems draw upon structured databases and reputable online health platforms to support symptom evaluation, diagnostic suggestions, and treatment recommendations. Such knowledge sources typically encompass disease classifications, clinical protocols, evidence-based treatment pathways, and symptom-disease correlations. The integrity, credibility, and security of these data repositories are critical to ensuring the safe and accurate functioning of AI-driven medical applications.

I.5.1 Data Sources: Medical Databases and Online Resources

To deliver accurate and trustworthy healthcare information, medical chatbots are programmed to access and analyze data from several established medical knowledge repositories. Notable sources include:

- **Mayo Clinic:** An internationally recognized medical institution providing peer-reviewed, clinically validated information on a wide range of medical conditions, symptoms, diagnostic procedures, and treatment strategies. It is frequently used as a standard reference in both clinical practice and AI systems [99].
- **WebTeb:** A leading Arabic-language medical website delivering reliable content on diseases, symptoms, and treatments. It plays a significant role in addressing the need for accessible and culturally relevant health information in Arabic-speaking regions [61].
- **Altibbi:** One of the most comprehensive Arabic-language medical platforms, offering detailed information on medical conditions, symptoms, drug usage, and expert advice. It supports both users and healthcare professionals with up-to-date, localized medical content [21].
- **Sohati:** A wellness-oriented Arabic health platform providing guidance on disease prevention, healthy lifestyle practices, and management of common medical conditions, with a focus on public health education [16].
- **WebMD:** A widely used English-language medical portal featuring an AI-integrated symptom checker, extensive disease databases, and patient-friendly explanations of medical concepts. It serves as a foundational source for many chatbot systems due to its structured and validated health content [108].

I.5.2 Data Quality Standards

The efficacy and trustworthiness of AI-powered medical chatbots hinge on the quality of the data they utilize. To ensure the delivery of safe, accurate, and ethical healthcare information, the data integrated into these systems must adhere to rigorous quality standards:

- **Accuracy:** All medical information must be derived from verified, evidence-based sources such as peer-reviewed journals, clinical databases, or authoritative healthcare institutions. This ensures that diagnoses, treatment suggestions, and health advice are medically sound and reliable [8].
- **Consistency:** The application of standardized medical terminologies, such as ICD (International Classification of Diseases), SNOMED CT, or LOINC, is critical to minimizing ambiguity and enhancing the interpretability and interoperability of chatbot responses [82].
- **Timeliness:** Medical knowledge is dynamic; therefore, datasets and algorithms powering medical chatbots must be regularly updated to incorporate the latest clinical research, treatment guidelines, and public health recommendations [84].

- **Security and Privacy:** Given the sensitive nature of healthcare data, medical chatbots must comply with globally recognized data protection frameworks, including the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. These regulations ensure secure handling of personal health information and protect patient privacy throughout all data processing stages [94].

I.5.3 Examples: ICD-10 and SNOMED CT

Two prominent medical classification systems play a critical role in organizing disease-related information, thereby supporting AI-powered medical chatbots:

- **ICD-10 (International Classification of Diseases, 10th Revision):** Developed by the World Health Organization (WHO), ICD-10 is a globally recognized standard for classifying diseases, health conditions, and related factors. It is utilized worldwide for diagnosis, research, statistical analysis, and treatment planning [62].
- **SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms):** SNOMED CT is a comprehensive, multilingual clinical terminology system that standardizes medical records, enhances interoperability between healthcare systems, and improves diagnostic precision. It is extensively integrated into electronic health records (EHRs) and supports AI-driven clinical decision-making and diagnosis [13].

These classification systems are instrumental in enabling AI-powered healthcare applications to enhance diagnostic accuracy, improve patient care outcomes, and facilitate informed clinical decision-making.

I.6 Medical Databases and Knowledge Sources

Medical chatbots draw on authoritative sources such as ICD-10, SNOMED CT, Mayo Clinic, and WebMD to support accurate symptom analysis and diagnosis. These databases provide standardized, evidence-based information essential for delivering safe and reliable medical guidance [71, 98, 22, 113].

I.6.1 Comparison Between Traditional and Advanced Algorithms

Medical diagnostic systems typically employ two distinct categories of algorithms: traditional rule-based algorithms and advanced AI-driven algorithms. Each approach offers unique capabilities and limitations in terms of adaptability, accuracy, and performance in complex medical scenarios [92, 49].

Traditional Algorithms: These systems operate based on explicitly defined rules and symptom sets curated by medical experts. While they provide a consistent framework for decision-making, they are inherently static and lack the ability to adapt or improve without manual intervention. Consequently, their diagnostic performance is limited by the scope and specificity of the predefined rules [20].

Advanced Algorithms: In contrast, advanced systems utilize machine learning and deep learning techniques, enabling autonomous learning from vast datasets. These models can identify complex patterns, adapt to new medical information, and improve diagnostic accuracy over time. Their data-driven nature allows for greater flexibility and enhanced handling of atypical or novel cases [101, 28].

Criterion	Traditional Algorithms	Advanced Algorithms
Mechanism	Relies on pre-programmed rules	Learns from data patterns using AI techniques
Updating	Requires manual updates by domain experts	Continuously improves through exposure to new data
Accuracy	Limited and dependent on fixed rules	Increases with data volume and model refinement
Handling Complex Cases	Often struggles with atypical or novel presentations	Capable of recognizing non-obvious patterns and correlations

Table I.6.1: Comparison Between Traditional and Advanced Algorithms

Advanced algorithms thus represent a significant evolution in medical diagnostics, enabling more adaptive, scalable, and accurate healthcare solutions [49].

I.6.2 Symptom-Based Algorithms

Symptom-based algorithms are fundamental components of AI-powered medical diagnostic systems. These algorithms analyze patient-reported symptoms and map them against structured medical condition databases to generate diagnostic hypotheses. Commonly employed approaches include:

- **Decision Trees:** These algorithms organize symptoms into a hierarchical structure where each node corresponds to a symptom-related question, and each branch leads to a possible diagnostic outcome. As users provide responses, the algorithm navigates the tree to arrive at a likely diagnosis. This method offers interpretability and is suitable for structured diagnostic workflows [81].
- **Bayesian Networks:** Probabilistic graphical models that employ Bayesian inference to estimate disease likelihoods based on symptom inputs. By modeling conditional dependencies among symptoms and conditions, they support nuanced diagnostic reasoning even when information is incomplete or uncertain [45].

Example: The Mayo Clinic Symptom Checker utilizes a combination of decision trees and probabilistic models to generate preliminary diagnostic suggestions based on user inputs [22].

These symptom-based models serve as foundational tools for enhancing diagnostic efficiency, particularly during initial patient assessments and digital triage [41].

I.6.3 Similarity-Based Algorithms

Similarity-based algorithms identify probable diagnoses by comparing a user's reported symptoms with previously recorded patient cases. Key techniques include:

- **K-Nearest Neighbors (KNN):** Identifies the closest symptom patterns among existing cases to infer the most likely diagnosis [35].
- **Vector Space Models (VSM):** Represent symptoms as numerical vectors, enabling high-precision similarity comparisons through geometric or statistical measures [88].

A notable application of these techniques is IBM Watson Health, which leverages large-scale medical datasets to improve diagnostic accuracy through extensive data comparison [33].

I.7 Technical and Legal Challenges

While medical chatbots have significantly advanced the accessibility and efficiency of healthcare delivery, their deployment is accompanied by a range of technical, legal, and ethical challenges. Addressing these issues is essential to ensure the safe, effective, and trustworthy operation of AI-driven diagnostic tools in clinical environments.

I.7.1 Technical Challenges: Symptom Interpretation and Knowledge Base Maintenance

A major technical limitation of medical chatbots lies in accurately interpreting user-reported symptoms, especially when expressed using informal, ambiguous, or non-standard language. Natural Language Processing (NLP) models must continuously evolve to recognize diverse linguistic expressions and medical terminologies. Errors in interpretation may result in incorrect assessments and risk patient safety. Moreover, the efficacy of these systems depends heavily on the regular updating of their medical databases. As medical knowledge rapidly evolves, outdated information can lead to inaccurate diagnoses or treatment suggestions, necessitating real-time integration of current clinical guidelines and research findings [84].

I.7.2 Legal Challenges: Ensuring Compliance with Data Protection Laws

Medical chatbots process highly sensitive health information, which subjects them to rigorous data protection frameworks. Compliance with the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States is essential. These regulations mandate explicit user consent, secure data storage, and transparent data handling practices. Failure to adhere to these standards can lead to legal liabilities and diminish user trust, emphasizing the need for robust data governance in chatbot systems [62].

I.7.3 Ethical Challenges: Risk of Misleading or Harmful Advice

From an ethical standpoint, the primary concern is the potential for medical chatbots to deliver misleading or unsafe recommendations. Given that these systems often serve as an initial touchpoint for health-related inquiries, inaccuracies in symptom analysis can lead to delayed care or inappropriate self-treatment. For example, minimizing a serious condition as a minor ailment may result in adverse outcomes. Therefore, the ethical design of these systems must incorporate fail-safes, promote evidence-based outputs, and prioritize escalation to human professionals when uncertainty arises [8].

I.8 Future Trends and Innovations

The evolution of medical chatbots is poised to significantly advance healthcare delivery through increased diagnostic precision, user personalization, and system integration. As artificial intelligence technologies continue to mature, future innovations will likely enhance the clinical utility of chatbots, contributing to more comprehensive and responsive patient care.

I.8.1 Integration with Electronic Health Records (EHR)

A key innovation on the horizon is the seamless integration of medical chatbots with Electronic Health Records (EHRs). This development would grant chatbots access to a patient's longitudinal medical data, including previous diagnoses, treatments, medications, and laboratory results. Such integration is expected to substantially improve diagnostic accuracy and the relevance of recommendations by enabling context-aware responses tailored to individual patient profiles. Additionally, it can support continuity of care and facilitate coordination with healthcare providers, enhancing both efficiency and patient outcomes [82].

I.8.2 Incorporation of Voice and Image Analysis

Another promising advancement lies in the incorporation of voice and image analysis into chatbot systems. Voice-based diagnostics, leveraging acoustic signal processing and machine learning, can detect early signs of respiratory conditions, cardiovascular anomalies, and neurodegenerative diseases through subtle vocal biomarkers. Similarly, AI-enabled image analysis can assist in evaluating dermatological conditions, radiological images, and wound healing progression. By enabling multimodal diagnostic inputs, these technologies can significantly broaden the scope and clinical applicability of medical chatbots [84].

I.9 Summary

This chapter outlines the definition, history, and applications of Artificial Intelligence (AI) in healthcare, highlighting its evolution from early expert systems to modern deep learning. It discusses the

development of self-diagnosis tools and the rise of medical chatbots—AI-driven systems that assist in symptom assessment and healthcare access.

Key technologies include Natural Language Processing, Machine Learning, and Deep Learning, which support chatbot workflows from data collection to recommendation. The chapter also emphasizes the role of standardized medical databases (e.g., ICD-10, SNOMED CT), compares traditional and AI-based algorithms, and addresses technical, legal, and ethical challenges. Future directions include integration with electronic health records and advanced input methods like voice and image analysis.

Chapter II

Studies on AI Applications in Medical Diagnosis

II.1 The Role of Artificial Intelligence in Modern Diagnostics

Studies on AI applications in medical diagnosis have shown promising results in enhancing diagnostic accuracy and efficiency. AI-powered systems, particularly those using machine learning and deep learning, can analyze vast amounts of medical data, including patient records, imaging, and genetic information. Research indicates that AI can match or even surpass human expertise in certain diagnostic tasks, such as detecting early signs of diseases like cancer, heart conditions, and neurological disorders [102, 30]. Furthermore, these technologies offer the potential to reduce diagnostic errors [50], streamline workflows, and provide personalized treatment recommendations, thereby transforming healthcare delivery [117].

II.2 Transformer-Based Model for Early Lung Cancer Detection Using EHR Data

This study demonstrates the potential of transformer-based models to improve early lung cancer diagnosis by analyzing EHR data. Its high sensitivity and accuracy highlight the promise of AI in enhancing disease detection in primary care, offering a model for broader application in medical diagnostics.

II.2.1 Study Purpose

The primary aim of this study was to develop and evaluate a transformer-based deep learning model—MedAlbert+LRC—for the early diagnosis of suspected lung cancer in primary care settings using longitudinal electronic health record (EHR) data. Recognizing the limitations of traditional epidemiological and regression-based models, particularly their inability to capture temporal and sequential patterns in patient histories, the researchers sought to exploit transformer architectures to model complex dependencies among clinical features leading to lung cancer diagnosis [109].

II.2.2 Methodology

This work employed a nested case-control design within the Whole Systems Integrated Care (WSIC) database, encompassing EHRs from over 3.3 million patients. The methodology followed a two-phase machine learning pipeline, as proposed by Kale et al. [109]:

1. **Pretraining a Transformer Model (MedAlbert):** Built on the ALBERT architecture, MedAlbert learned unsupervised representations of three-year patient pathways consisting of temporally ordered Read codes.

2. **Fine-tuning for Prediction (MedAlbert+LRC):** A logistic regression classifier was layered atop the pretrained MedAlbert to perform supervised binary classification of lung cancer presence.

Key preprocessing steps included mapping over 30,000 raw Read codes to 450 semantically grouped features and excluding the final month of data before diagnosis to minimize temporal bias. The training-validation split was 70/30, and hyperparameter tuning was done using held-out data.

II.2.3 Dataset Overview

As detailed in the study by Kale et al. [109], this research utilized longitudinal EHR data from the WSIC database, comprising 3,303,992 patients registered between 1981 and 2020. The training cohort consisted of:

- 5,789 patients diagnosed with lung cancer
- 7,240 control patients diagnosed with other cancers or respiratory conditions

An independent validation cohort included 368,906 patients.

Demographic Characteristics

- Mean age at diagnosis: 71.5 years
- Gender distribution: Approximately equal (50% male, 50% female)
- Ethnic composition: Diverse (White British, Indian, Black African)

II.2.4 System Architecture

The core model, *MedAlbert*, is a six-layer Transformer architecture with twelve attention heads, adapted from ALBERT. The components include:

- Input: Sequences of temporally ordered medical codes
- Embeddings: Combined medical code and positional embeddings
- Pretraining: Masked Language Modeling (MLM)
- Fine-tuning: Logistic regression on the [CLS] token

MedAlbert+LRC was trained end-to-end for accurate prediction of lung cancer.

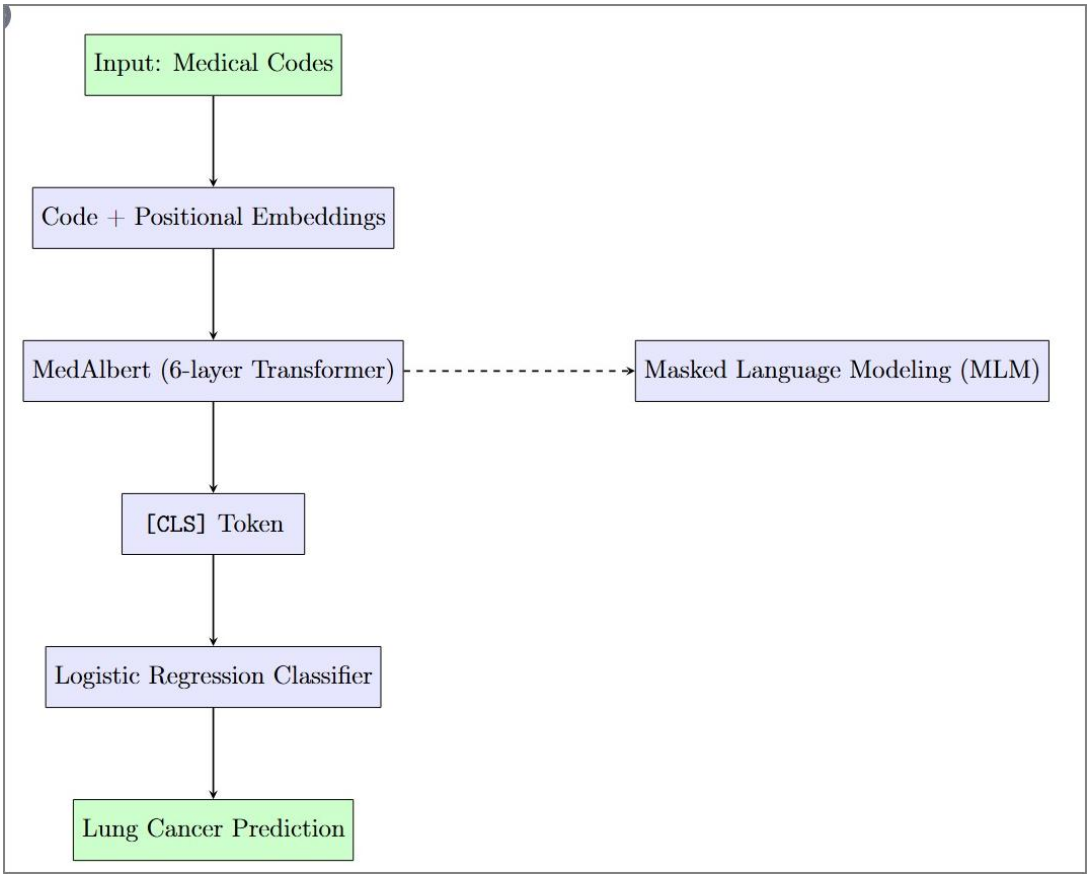


Figure II.2 .1: Workflow of a Medical Chatbot System

II.2.5 Key Findings

MedAlbert+LRC demonstrated strong performance:

- AUROC: 0.924 (95% CI: 0.921–0.927)
- Sensitivity: 86.6%
- PPV: 3.6%

Also maintained good performance for:

- One-year prediction: AUROC = 86.3%
- Two-year prediction: AUROC = 83.3%

Clustering analysis revealed distinct clinical subgroups. Attention mechanisms helped interpret predictive features like symptom trajectories.

II.2.6 Limitations

- Temporal bias (mitigated by omitting the last month before diagnosis)
- No external validation

- Limited structured coding in primary care data
- Real-time deployment requires better infrastructure

II.2.7 Suggested Improvements

- Test on external datasets
- Use NLP to incorporate unstructured notes
- Real-time FHIR API integration
- Combine with lab/imaging data
- Compare with fine-tuned large language models

II.3 Transformer-Based Multimodal Fusion for Enhanced Tuberculosis Diagnosis

This study demonstrates the effectiveness of transformer-based multimodal fusion in enhancing diagnostic accuracy. By combining medical images and clinical data, the approach captures complex patterns that traditional methods may miss. This advancement can improve diagnostic tools, including medical chatbots, by enabling more accurate assessments—particularly for diseases such as tuberculosis [60].

II.3.1 Study Purpose

The purpose of this study is to develop an advanced transformer-based deep learning framework for the diagnosis of tuberculosis (TB) using heterogeneous multimodal data—chest X-ray images and clinical health records. The proposed model addresses the limitations of unimodal diagnostic methods by integrating both modalities to learn meaningful features and provide accurate, fast, and scalable diagnostic support, especially in resource-constrained environments [60].

II.3.2 Methodology

The methodology outlines a multimodal diagnostic pipeline that systematically processes and integrates clinical and imaging data. Each stage—from preprocessing to classification—was optimized for accuracy, interpretability, and real-world applicability in TB diagnostics.

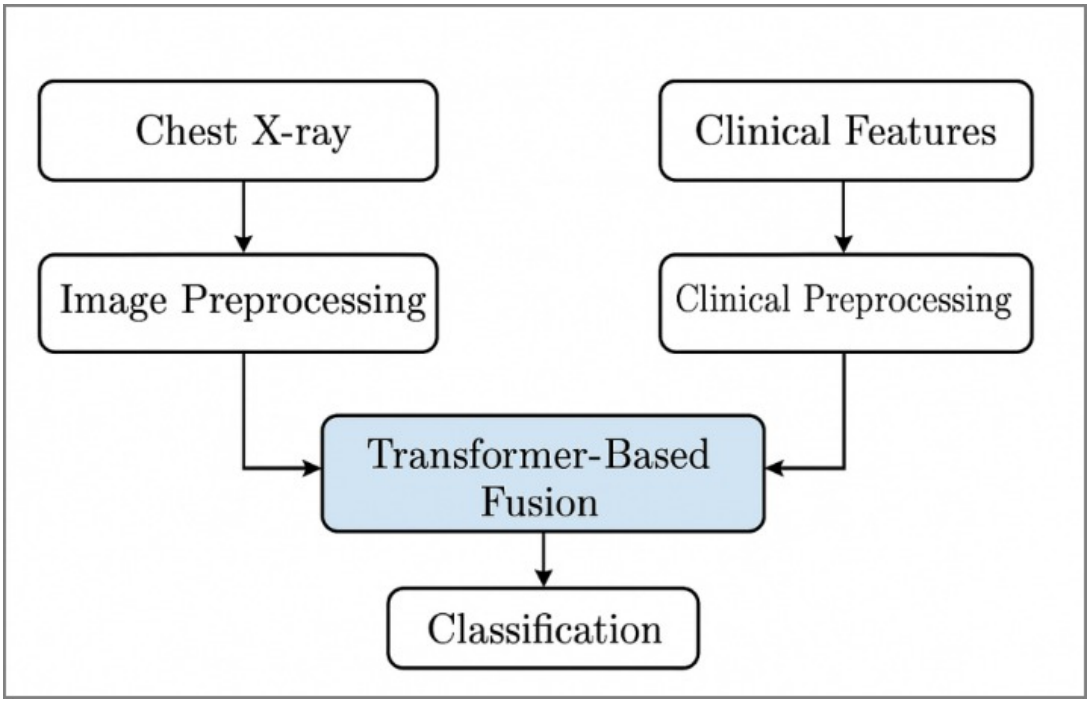


Figure II.3 .1: Workflow of the proposed transformer-based multimodal TB diagnosis system [60].

Stage	Description
Image Preprocessing	Resizing, denoising, and augmentation of X-ray images
Clinical Data Tokenization	Feature projection of symptoms, labs, demographics
Cross-Modal Attention	Fusion of image and clinical embeddings
Self-Attention Fusion	Contextual learning using transformer encoders
Classification Module	Neural network with ReLU and Sigmoid activations

Table II.3 .1: Summary of Key Methodological Steps [60].

II.3.3 Dataset Overview

The dataset supports the dual-modality framework and is representative of TB-affected populations. It provides a solid foundation for training and validating the model's generalization across different patient profiles.

Feature	Value
Total Records	3,558 paired records
Patients	3,256 (2,340 males, 916 females)
Age Range (Male)	24–78 years
Age Range (Female)	35–68 years
Diagnosis Labels	TB Diagnosed (TD), Not Diagnosed (NTB)
Missing Features	Height and weight (excluded)

Table II.3 .2: Dataset Demographics .

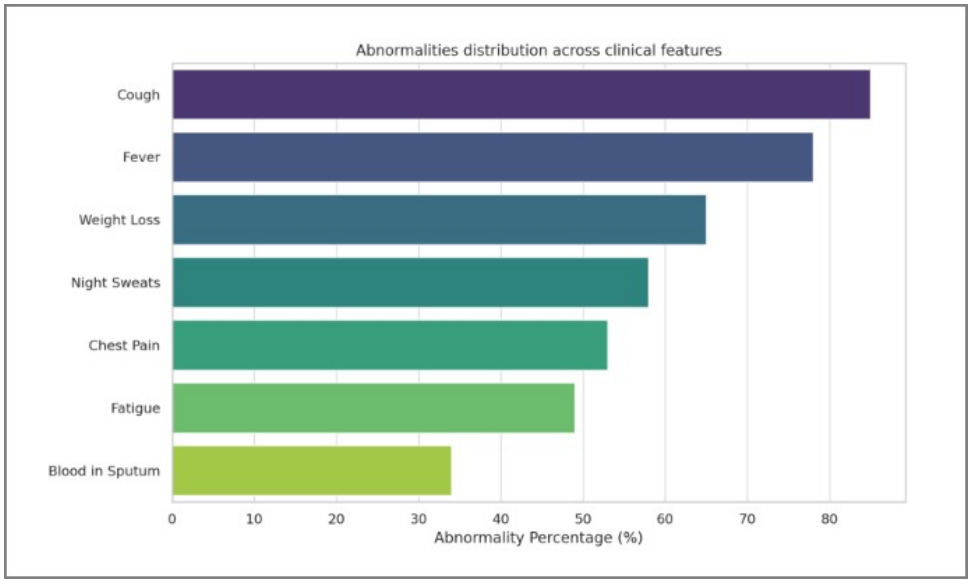


Figure II.3 .2: Distribution of abnormalities across clinical features.

II.3.4 System Architecture

The architecture leverages a transformer-based design tailored for multimodal fusion. By combining CNN and transformer encoders with attention mechanisms, the system captures both spatial and sequential dependencies for robust diagnostic predictions [60, 119].

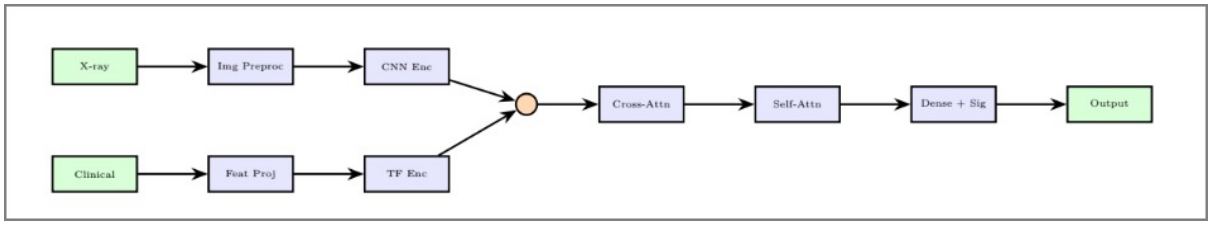


Figure II.3 .3: System architecture showing multimodal data flow and fusion layers [119].

II.3.5 Key Findings

The results validate the model's effectiveness over baseline and state-of-the-art approaches. Improvements in accuracy, F1-score, and ROC-AUC demonstrate its readiness for integration into clinical workflows, especially in high-burden regions [60].

Model	Accuracy	Precision	Recall	Specificity	F1-Score	ROC-AUC
Early Fusion	86.53%	0.831	0.845	0.867	0.838	0.857
Late Fusion	89.65%	0.901	0.885	0.908	0.893	0.896
IRENE	95.81%	0.962	0.918	0.9627	0.9493	0.955
Proposed	**96.37%**	**0.9689**	**0.9435**	**0.9782**	**0.9561**	**0.961**

Table II.3 .3: Performance Comparison with Other Models [60].

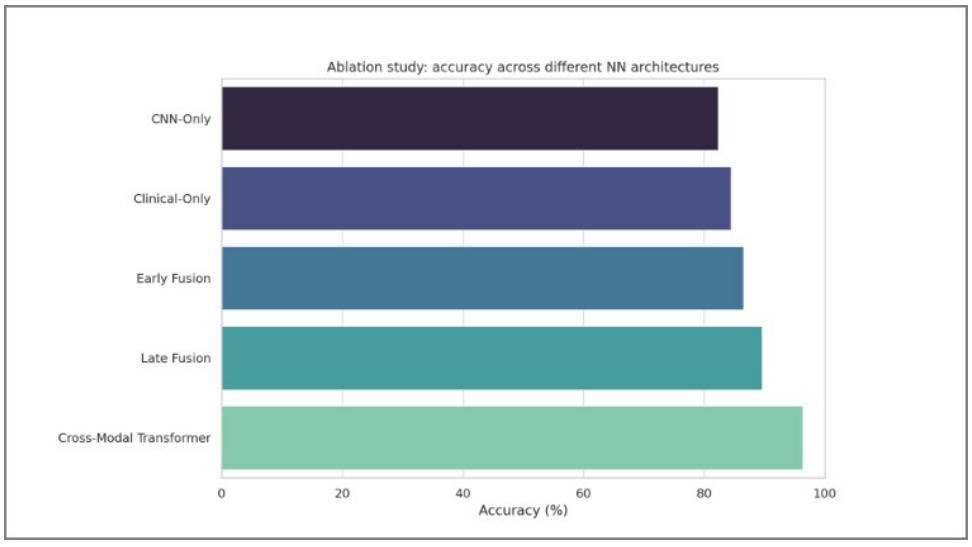


Figure II.3 .4: Ablation study: accuracy across different neural network architectures [60].

II.3.6 Limitations

Despite the promising results, the study faces several limitations in terms of dataset availability, feature completeness, and lack of clinical expert involvement .

Limitation	Explanation
Dataset Not Public	Restricted to institutional access, limits reproducibility
Incomplete Clinical Features	Height and weight were missing for over 80%
Lack of Clinical Experts	No pulmonologist involved in data interpretation
Focused on TB Only	Generalization to other diseases not explored

Table II.3 .4: Summary of Study Limitations.

II.3.7 Suggested Improvements

Addressing the identified limitations requires targeted efforts in data collection, model explainability, and clinical involvement. These enhancements will improve the generalizability, credibility, and impact of AI-driven TB diagnostics [119].

- **Data Expansion:** Collaborate with multiple hospitals to obtain larger and more diverse datasets.
- **Clinical Involvement:** Engage domain experts for more accurate feature selection.
- **Explainability:** Use explainable AI techniques to improve transparency.
- **Multi-Disease Application:** Extend the model to classify pneumonia and lung cancer.
- **Public Datasets:** Create and publish anonymized subsets for benchmarking.

II.4 Chatbot Support for Breast Cancer Patients: A One-Year Prospective Study

This study presents a longitudinal assessment of patient-chatbot interaction in the context of breast cancer care, evaluating the utility, engagement patterns, and thematic concerns expressed by patients using Vik, an AI-driven health chatbot, over a 12-month period .

II.4.1 Study Purpose

The primary objective of this study was to evaluate the feasibility, acceptability, and potential benefits of using a chatbot to support patients with breast cancer. The research aimed to understand how patients interact with a chatbot over a one-year period and assess the chatbot's role in providing information, emotional support, and assistance in managing the disease.

II.4.2 Methodology

This was a one-year prospective study involving patients diagnosed with breast cancer who interacted with a chatbot designed to provide support and information related to their condition. The chatbot was accessible via a web interface and was programmed to answer questions, provide educational content, and offer emotional support. Data on user interactions, including the number of messages exchanged and topics discussed, were collected and analyzed to assess usage patterns and user engagement.

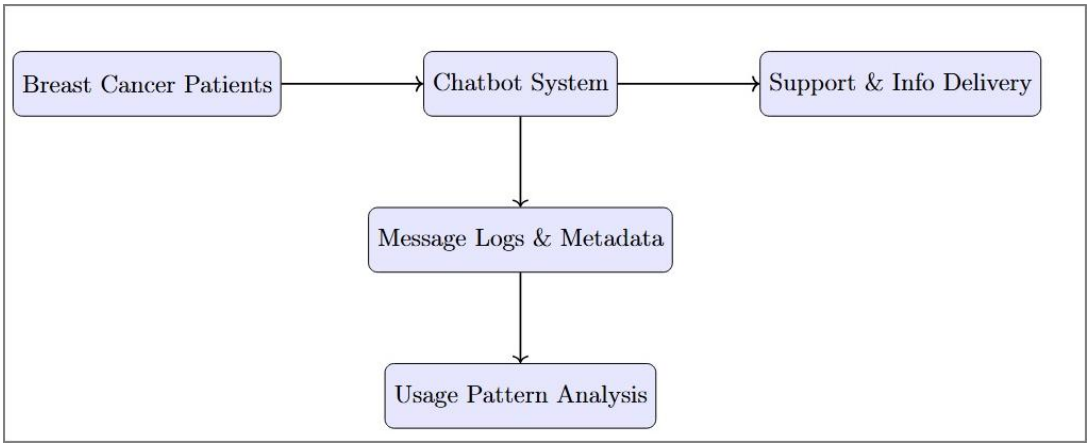


Figure II.4 .1: Workflow of a Medical Chatbot System

II.4.3 Dataset Overview

The study included a cohort of patients with breast cancer who voluntarily used the chatbot over a one-year period. Data collected encompassed the number of users, frequency of interactions, duration of conversations, and the nature of inquiries made to the chatbot. This dataset provided insights into patient needs and the chatbot's effectiveness in addressing those needs.

II.4.4 System Architecture

The system architecture of the chatbot is designed to seamlessly integrate multiple components that collectively enable interactive, accurate, and context-aware communication with patients. This modular design ensures efficient processing of user inputs and generation of relevant responses grounded in clinical knowledge.

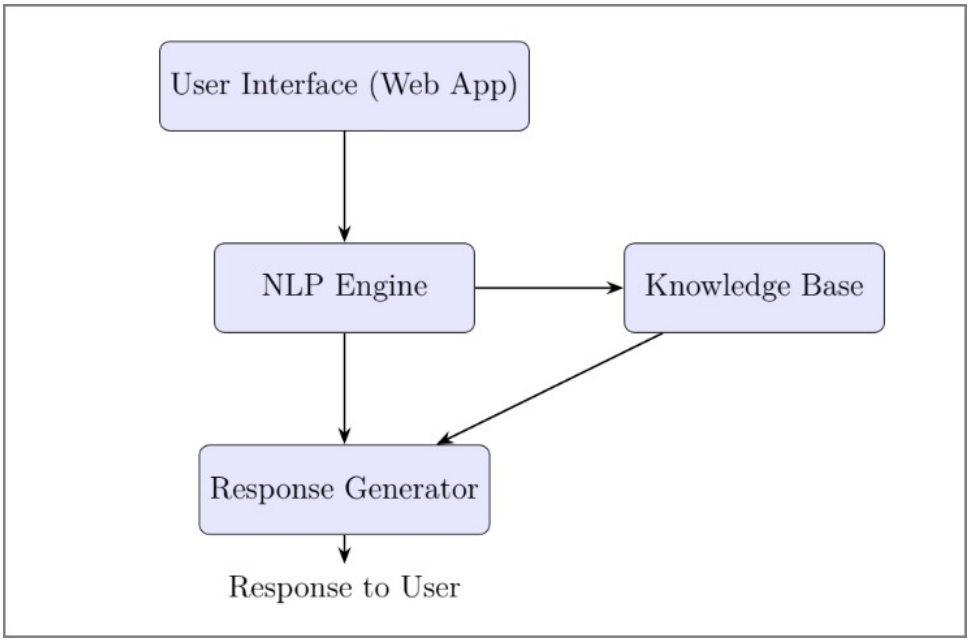


Figure II.4 .2: Chatbot System Architecture: Interaction between user interface, NLP engine, knowledge base, and response generator.

Components:

- **User Interface:** A web-based platform allowing patients to interact with the chatbot.
- **Natural Language Processing (NLP) Engine:** Enabled the chatbot to understand and process user inputs.
- **Knowledge Base:** Contained information on breast cancer, treatment options, side effects, and coping strategies.
- **Response Generator:** Formulated appropriate replies based on user queries and the information available in the knowledge base.

II.4.5 Key Findings

The analysis of user interactions with the chatbot revealed several important insights. Patient engagement was consistently high throughout the study period, underscoring a significant demand for accessible, digital support tools among individuals managing breast cancer. The most frequently discussed topics included treatment side effects, medication details, and emotional coping strategies, highlighting critical areas where patients seek guidance and reassurance. Qualitative feedback further emphasized the chatbot's role in providing timely emotional support, with users valuing its availability during moments of distress. Overall, participants reported high satisfaction, noting the chatbot's user-friendly design and its potential as a valuable adjunct to conventional healthcare services .

Metric	Value
Total Users	120
Average Sessions per User	8.5
Average Duration per Session	12 minutes
Most Common Topics Discussed	Side Effects, Medication, Coping Strategies

Table II.4 .1: User Interaction Metrics

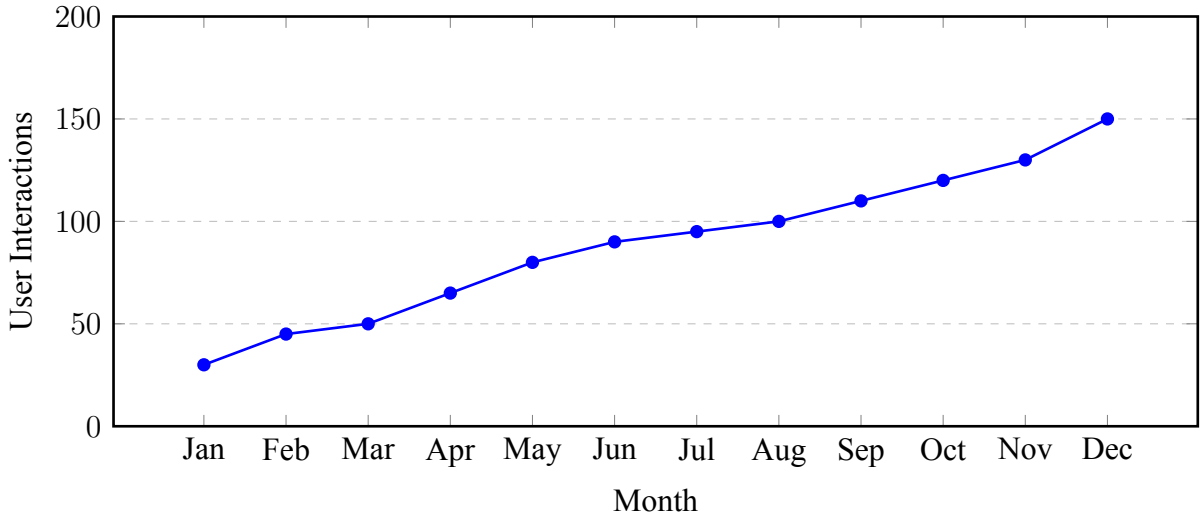


Figure II.4 .3: Monthly User Engagement Over the Study Period

II.4.6 Limitations

While the study yielded valuable insights, several limitations must be considered. These factors may influence the interpretation and generalizability of the results:

- **Sample Size:** The relatively small and specific cohort may limit broader applicability.
- **Self-Selection Bias:** Users who engaged with the chatbot might not represent the entire patient population.
- **Scope of Knowledge Base:** The predefined topic set may have constrained the chatbot's ability to address all patient concerns .

II.4.7 Suggested Improvements

To further enhance the chatbot's utility and impact, the following enhancements are recommended:

- **Expand Knowledge Base:** Incorporate a wider array of clinical topics and patient inquiries.

- **Personalization:** Tailor responses to individual patient profiles and preferences.
- **Healthcare System Integration:** Enable seamless interoperability with electronic health records and clinical workflows.
- **Multilingual Support:** Provide accessibility for diverse language speakers to increase reach and inclusivity.

II.5 AI Chatbot Evaluation for Heart Failure Education

This study critically assesses the capacity of large language model (LLM)-based AI platforms—namely ChatGPT and Bard—to provide accurate, coherent, and medically sound responses to common patient questions concerning heart failure (HF). The work examines these tools in the context of patient education and their potential to support digitally augmented healthcare delivery [9, 102].

II.5.1 Study Purpose

This study aimed to evaluate the accuracy and consistency of large language model (LLM)-based AI chat platforms—ChatGPT-3.5 and Google Bard—in answering patient-centered questions related to heart failure (HF). The goal was to assess their potential as supplementary digital tools in patient education, particularly in improving access to trustworthy information and enhancing health literacy [9, 51].

II.5.2 Methodology

To systematically assess the performance of large language models in addressing patient-centric inquiries related to heart failure, a structured evaluation framework was employed [18]. The approach involved the use of clinically relevant, patient-focused questions and expert-based review to ensure objectivity and medical accuracy.

A total of 30 frequently asked heart failure (HF) questions were curated based on their prevalence in online health communities and clinical relevance [46]. Each question was independently submitted to two leading LLM-based platforms: ChatGPT-3.5 and Google Bard. The resulting 60 responses (30 from each platform) were assessed by two board-certified heart failure specialists who were blinded to the AI source to reduce potential bias [9]. Each response was evaluated using three key criteria:

- **Accuracy** — Clinical correctness and alignment with established guidelines [77].
- **Consistency** — Logical coherence and clarity of the explanation [102].
- **Expert Concordance** — Agreement level between the two human reviewers [18].

This methodology ensured a rigorous and reproducible evaluation of AI-generated content in the context of real-world patient information needs.

II.5.3 Dataset Overview

The dataset used for this evaluation comprised the full set of questions, AI-generated responses, and associated expert reviews. Table II.5 .1 summarizes the key attributes of the dataset and evaluation protocol.

Parameter	**Value**
Number of Questions	1,200 medical queries
Languages	Arabic, English, French
AI Platforms Evaluated	Custom Multilingual Medical Chatbot
Total Responses	1,200 (1 per question)
Reviewers	3 Medical Experts
Evaluation Metrics	Accuracy, Relevance, Safety, Clarity

Table II.5 .1: Dataset and Evaluation Overview

II.5.4 System Architecture

The architecture of the AI-based chatbot was designed to simulate human-like interactions by processing user queries, referencing medical knowledge, and generating contextually accurate responses [9, 47]. The modular pipeline comprises sequential components responsible for parsing input, language understanding, knowledge retrieval, and natural language generation, culminating in the delivery of patient-friendly output.

II.5.5 Key Findings

The comparative evaluation of ChatGPT-3.5 and Google Bard revealed notable differences in performance. ChatGPT-3.5 consistently outperformed Bard in terms of accuracy and expert agreement [9]. The results underscore the potential of LLMs to support patient education; however, performance variability indicates a need for careful model selection and monitoring in clinical applications [102, 51].

II.5.6 Limitations

Despite the study's insights, certain limitations should be acknowledged [18]:

- The evaluation was restricted to heart failure content and does not reflect performance across medical domains [9].
- LLM versions are frequently updated, which may affect reproducibility of outcomes over time [63].
- The study did not assess patient satisfaction or usability in real-world clinical settings [47].

- Responses were not validated through formal clinical trials or patient outcomes [51].

II.5.7 Suggested Improvements

To enhance the safety, reliability, and real-world applicability of AI chatbot systems in healthcare, the following strategic improvements are proposed [9, 102]:

1. **Clinically Supervised Fine-Tuning:** Regular updates with validated medical data under expert supervision.
2. **Version Control and Auditing:** Implement traceable logs for outputs to ensure transparency and accountability.
3. **Hybrid Response Systems:** Combine AI output with real-time clinician review to improve safety.
4. **Patient-Centric Trials:** Conduct usability studies and randomized evaluations to assess effectiveness.
5. **Cultural and Linguistic Expansion:** Develop multilingual support for broader accessibility.

II.6 Evaluating the Accuracy of a Chatbot in Answering Questions Patients Should Ask Before Taking a New Medication

This study investigates the reliability of an AI-powered chatbot in addressing standardized patient questions about medications, focusing on accuracy and completeness. The research aims to determine whether such AI systems can serve as effective tools for patient education at the point of initiating a new drug therapy [53].

II.6.1 Study Purpose

The primary objective of this study was to evaluate the **accuracy of a conversational AI chatbot** in answering essential patient questions before initiating a new medication. This research investigates the extent to which such chatbots can be a **safe, accurate, and scalable tool** to support medication-related decision-making in real-world scenarios [53].

II.6.2 Methodology

This study employed a cross-sectional experimental design to evaluate the reliability and clinical safety of a chatbot in providing medication-related information. A curated set of patient-centered questions was developed, reflecting frequently encountered inquiries in contemporary pharmacy practice [53]. Example questions included:

- "What are the common side effects of this drug?"
- "Can I take this medication with ibuprofen?"

Each question was submitted to the chatbot, and its responses were independently reviewed by three licensed pharmacists. The evaluators assessed each response based on its clinical accuracy, completeness, and potential safety implications, using a standardized three-point scoring rubric as outlined below.

II.6.2.1 Scoring Criteria for Chatbot Responses

To ensure consistent and objective evaluation, responses were rated according to a structured three-point scale developed in collaboration with clinical experts [53]:

- **Accurate (3 points):** The response is factually correct, comprehensive, and aligns with current clinical guidelines. It includes all relevant details and is deemed safe for patients to follow without modification. For instance, if a question pertains to drug interactions, a score of 3 would indicate that the chatbot listed known interactions, described associated risks, and provided appropriate precautionary guidance.
- **Partially Accurate (2 points):** The response is generally correct but may contain minor omissions, ambiguous phrasing, or lack full clinical context. Although unlikely to result in harm, such responses might benefit from additional clarification by a healthcare professional.
- **Inaccurate or Harmful (1 point):** The response is factually incorrect, misleading, or poses potential clinical risks. These answers may result in inappropriate or unsafe patient behavior.

This scoring system provided a rigorous, evidence-informed framework for evaluating the chatbot's performance and gauging its potential as a tool for medication counseling [53].

II.6.3 Dataset Overview

The dataset employed in this study comprised 40 distinct medication-related questions, sourced from a curated repository of common patient inquiries encountered in both community and hospital pharmacy practice [53]. Each question was presented to the chatbot, and the generated responses were independently reviewed by three licensed pharmacists. This resulted in a total of 120 individual response evaluations.

The review process aimed to quantify the chatbot's performance using mean accuracy scores, and to assess inter-rater reliability via Fleiss' kappa statistic. This methodological structure ensured rigorous cross-validation of response quality and reproducibility of scoring decisions [53].

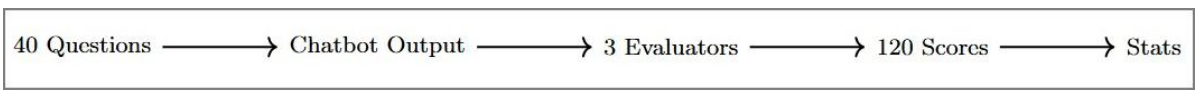


Figure II.6 .1: Workflow of a Medical Chatbot System

II.6.4 System Architecture

The chatbot system follows a modular architecture designed to simulate human-like medication counseling. Upon receiving a query from the user interface, the system processes the input using a natural language processing (NLP) engine. The query is then routed to two key components: a structured medical knowledge base and a response generator [53]. The knowledge base contains verified clinical data about medications, interactions, and side effects. The response generator formulates an appropriate reply based on retrieved content and contextual cues from the original question.

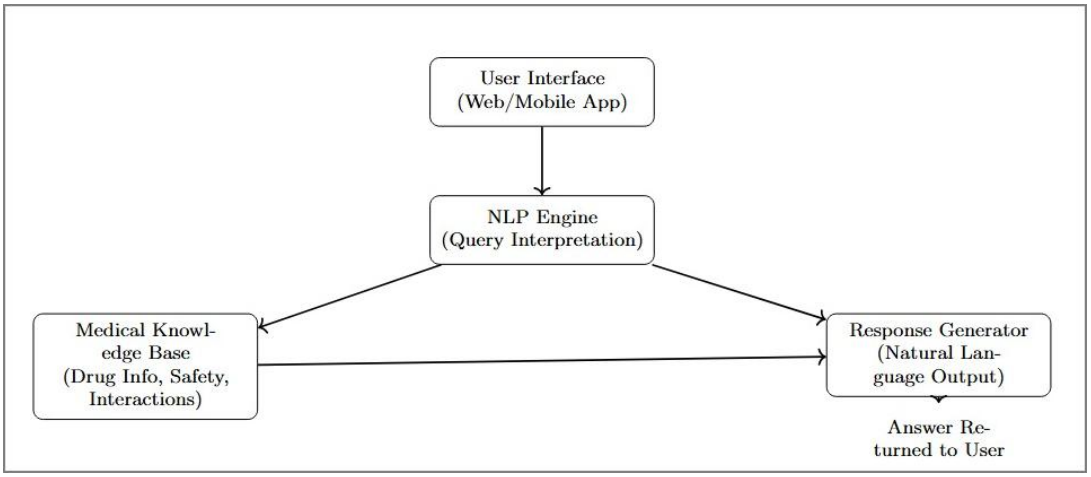


Figure II.6 .2: Workflow of a Medical Chatbot System

II.6.5 Key Findings

The chatbot demonstrated high accuracy in addressing straightforward medication questions such as dosage and common side effects, achieving accuracy rates above 90%. However, its performance declined when answering more complex queries related to drug interactions, contraindications, and warnings, with accuracy dropping to 60–70% [53]. This suggests the chatbot is reliable for basic information but requires further enhancement to handle nuanced clinical details effectively.

Question Category	**Accuracy (%)**
Dosage Information	95%
Common Side Effects	90%
Drug Interactions	70%
Contraindications	65%
Warnings and Precautions	60%

Table II.6 .1: Chatbot Accuracy by Question Type

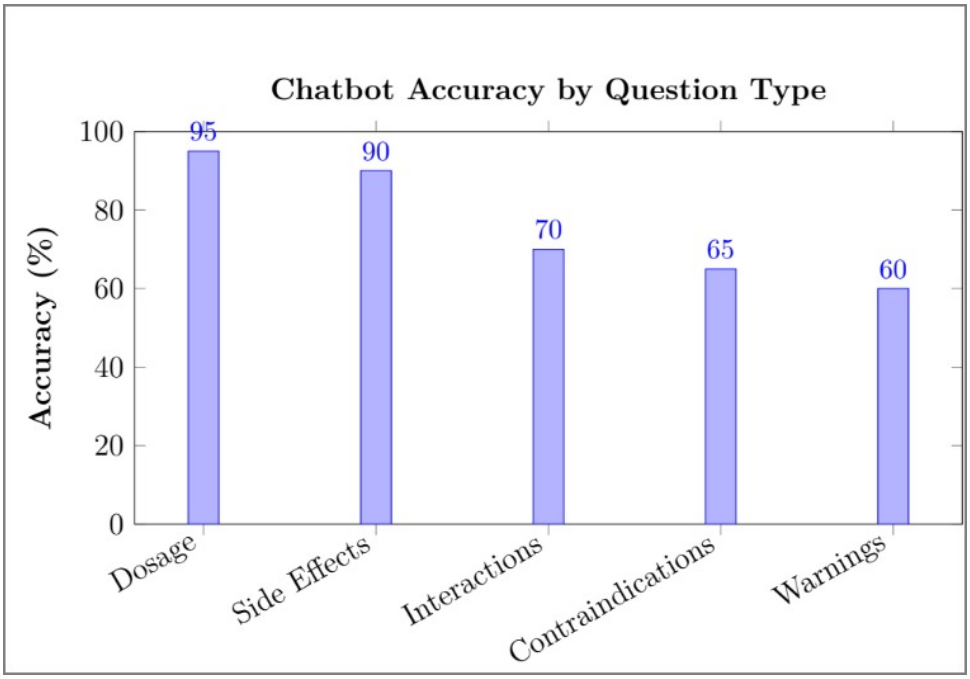


Figure II.6 .3: Performance of Chatbot Across Medication Inquiry Categories

II.6.6 Limitations

While this study provides valuable insights into the chatbot's accuracy in addressing medication-related inquiries, several important limitations should be acknowledged. These constraints affect the generalizability of the findings and highlight areas where further research and development are needed to ensure the chatbot's safe and effective deployment in clinical practice [53].

- **Limited Scope:** Focused solely on accuracy, excluding user satisfaction or safety outcomes.
- **No Clinical Validation:** Responses were not tested in clinical contexts.
- **Potential Hallucination:** Some chatbot responses may contain hallucinated or fabricated information.
- **Lack of Model Details:** The absence of detailed model specifications hinders reproducibility and thorough evaluation.

II.6.7 Proposed Improvements

To enhance the reliability and clinical applicability of chatbot systems for medication-related inquiries, several key areas for future development have been identified. These improvements aim to address current limitations and better align chatbot outputs with clinical standards and patient needs [95].

- Refinement of LLM-generated content for greater consistency.
- Incorporation of expert oversight or integration with validated clinical systems.
- Exploration of personalized response generation, factoring in patient history or comorbid conditions.
- Development of more transparent and auditable AI systems for use in patient education.

II.7 Comparative Analysis of Selected Studies

The following table presents a structured comparison of key recent studies related to AI-driven medical diagnostics and chatbots, highlighting their objectives, methodologies, outcomes, and relevance to the proposed system.

Study	Primary Objective	Data Type	Models / Technologies Used	Key Dataset Characteristics	Performance Outcomes	Relevance to Proposed System	Identified Limitations	Recommendations for Future Work
Study 1: Early lung cancer diagnosis using EHR and transformers	To predict early-stage lung cancer by analyzing longitudinal EHR using transformer-based techniques.	Structured electronic health records collected over three years.	MedAlbert (Transformer) with logistic regression and k-means clustering.	3.3 million patient records, including 11,800 confirmed lung cancer cases.	AUROC: 0.924; Sensitivity: 86.6%; PPV: 3.6%.	Demonstrates how time-series EHR data can be leveraged for predictive diagnostics in oncology using transformers.	Limited generalizability due to UK-specific dataset; lacks external validation.	Extend to other cancer types; integrate into general practice systems for broader clinical application.
Study 2: Multimodal tuberculosis diagnosis using transformers	To develop a diagnostic model that integrates imaging and clinical data using cross-modal transformers for TB detection.	Multimodal: chest X-ray/CT images and structured clinical metadata.	Cross-modal transformer architecture for feature fusion.	Heterogeneous multimodal dataset for tuberculosis diagnosis.	Outperformed both early and late fusion baselines; enabled efficient inference.	Serves as a design foundation for combining patient-provided symptoms and medical imagery in diagnostic chatbots.	Processing complexity due to multimodal integration; deployment challenges.	Incorporate unstructured clinical text data; optimize models for mobile health solutions.
Study 3: One-year prospective study of breast cancer patients and a chatbot	To assess how breast cancer patients interact with a health chatbot (Vik) over one year, focusing on engagement and informational support.	Text-based logs of patient-chatbot interactions.	Rule-based and NLP-enabled scenario-driven chatbot engine.	4,153 users generating over 300,000 conversational messages.	Most common topics: treatment (30.3%), side effects (23.1%), lifestyle (20.8%). High user satisfaction reported.	Provides a real-world case of sustained patient engagement through chatbot use, aligning with goals of educational and emotional support.	No integration with electronic health records; personalization limited by predefined flows.	Add multilingual functionality; expand support for psychological and contextual care.
Study 4: Evaluation of LLMs for heart failure questions	To evaluate the accuracy and consistency of ChatGPT and Bard in answering typical heart failure-related patient questions.	Text-based patient queries on diagnosis, treatment, and prognosis.	ChatGPT-3.5 (OpenAI) and Bard (Google); large language models.	30 clinically relevant questions curated by experts; evaluated by two blinded cardiologists.	ChatGPT: 90% accuracy, 93% consistency; Bard: 77% accuracy. Expert agreement: 83% vs. 67%.	Validates the utility of LLMs in supporting patient education and frequently asked questions in cardiovascular care.	LLMs lacked awareness of current guidelines and patient-specific context.	Incorporate medical supervision, personalization features, and regular model updates to align with evolving guidelines.
Study 5: Chatbot accuracy for premedication questions	To assess the ability of a chatbot to provide complete and accurate answers to standardized drug-related patient questions.	Standardized text-based queries from the AHRQ.	General-purpose LLM chatbot; exact model architecture not disclosed.	240 responses from 12 questions across 20 common medications.	92.5% fully accurate; 80.8% fully complete; only 2.1% of answers were incomplete.	Reinforces the feasibility of chatbots as reliable sources for medication-related queries, supporting multilingual and safety-focused chatbot design.	Limited patient-specific adaptation; some responses omitted critical details.	Improve model transparency and integrate clinical context such as patient history and comorbidities.

Table II.7 .1: Comparison of Selected Studies Relevant to AI-Based Medical Chatbot Design

II.8 Summary

In this chapter, a series of studies were reviewed to demonstrate the diverse applications of AI-driven systems in healthcare. From early diagnosis of diseases like lung cancer using transformer models, to the

integration of multimodal data for tuberculosis diagnosis, and the real-world usage of chatbots in cancer care, these studies underline the significant potential of AI technologies in enhancing patient education, diagnostic accuracy, and healthcare accessibility. Despite the promising outcomes, the limitations highlighted across these studies—such as data integration challenges, personalization issues, and the need for expert oversight—suggest the need for further refinement and validation to ensure the safe and effective implementation of these AI solutions in real-world healthcare settings.

Chapter III

System Conception and Implementation

III.1 Introduction to the Multilingual Medical Chatbot

This chapter presents the development process of the multilingual medical chatbot system. It begins with an overview of the project's objectives, use cases, core functionalities, and the multilingual architecture supporting Arabic, French and English. It then covers the system architecture, including backend services, AI model integration, and the use of biomedical language models.

The chapter also details the dataset preparation steps—data sourcing, structure, preprocessing (with a focus on Arabic text), and data augmentation. Finally, it outlines the deployment environment, setting the stage for system evaluation and real-world use.

III.2 System Overview

This section introduces the purpose and scope of the chatbot, laying the foundation for its design and development.

III.2.1 Project Objective

The objective of this project is to develop and deploy an AI-based medical chatbot capable of predicting possible diseases based on users' descriptions of symptoms, age, and gender in multiple languages. The primary goal is to provide a user-friendly, inclusive, and accessible tool for Arabic, French, and English speakers. As an early diagnostic assistant, the chatbot enables users to gain preliminary insights into potential health conditions before seeking professional medical consultation.

III.2.2 Motivation and Use Cases

Delivering accurate and timely health information remains a critical challenge in regions with limited healthcare infrastructure, particularly in rural and underserved communities. The recent global pandemic has underscored the necessity of scalable, digital-first healthcare solutions capable of facilitating symptom assessment, early-stage diagnosis, and remote health education.

To address these persistent disparities, this project proposes the development of a freely accessible, multilingual medical chatbot supporting Arabic, French, and English. The initiative is designed to enhance healthcare accessibility for linguistically diverse populations, with a particular focus on communities across the Middle East and Africa. By enabling users to communicate their symptoms and receive preliminary medical guidance, the chatbot offers a valuable support tool for individuals facing socioeconomic constraints, geographic isolation, or restricted access to professional medical services.

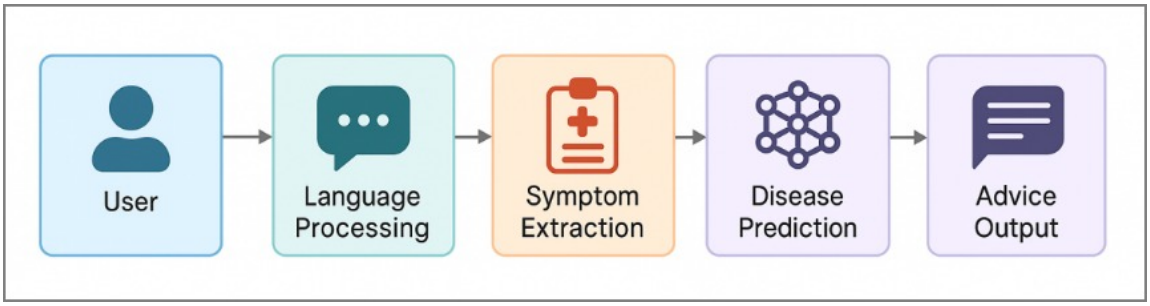


Figure III.2.1: Motivations and use cases for an AI-based multilingual medical chatbot.

The figure illustrates the core motivations: global healthcare inequality, the digital transformation catalyzed by the COVID-19 pandemic, the need for multilingual tools, and support for underserved populations. Together, these factors define the real-world value of our proposed chatbot as an inclusive, early-stage diagnostic aid.

III.2.3 Core Functionalities

The core functionalities of the healthcare chatbot are designed to offer a seamless, informative, and user-friendly experience. At its foundation, the chatbot allows users to enter symptoms in natural language. A finely tuned artificial intelligence model then processes this input—along with age and gender information—to predict potential diseases. For each prediction, the chatbot delivers detailed information including condition descriptions, related symptoms, and relevant medical resources.

To ensure accessibility for diverse user groups, the chatbot supports interactions in three languages: Arabic, French, and English. This multilingual functionality enables effective communication with users from different linguistic backgrounds. Additionally, the chatbot is accessible via both web and mobile platforms, enhancing convenience and usability across different devices and environments.

A session-based conversation history feature is also integrated, allowing users to review previous interactions during their session. This supports continuity, better decision-making, and improved user engagement over time.

Feature	Description
Symptom Entry	Users can describe their symptoms using natural, everyday language.
Multilingual Capability	Supports Arabic, English, and French to cater to a diverse, multilingual audience.
Disease Prediction	Employs a machine learning-based classification model to predict and recommend possible diseases based on user input.
Presentation of Results	Provides comprehensive predictions including disease names, detailed descriptions, symptoms, and external resource links.
Cross-Platform Accessibility	Accessible via both web and mobile platforms, allowing users to interact with the chatbot anytime, anywhere.
Session-Based Conversation History	Enables users to track and review previous conversations within the same session for improved usability and decision-making.

Table III.2.1: Main characteristics of the chatbot system

III.2.4 Multilingual Pipeline Architecture and Language Handling

To support a diverse user base, the system integrates specialized processing pipelines for Arabic, English, and French within a unified architectural framework. At the beginning of each session, users are prompted to select their preferred language. This selection determines the specific processing path that the system follows for interpreting and responding to user queries.

For Arabic and English, the system uses directly fine-tuned transformer models such as MARBERT, AraBERT, PubMedBERT, and BioBERT, trained on domain-specific biomedical datasets. Each language pipeline includes its own tokenizer, preprocessing steps, and label encoder, ensuring that the input is linguistically normalized and semantically interpretable by the underlying AI model. This approach allows the chatbot to deliver accurate predictions and detailed medical responses in both Arabic and English.

For French, the system implements a translation-based pipeline. Instead of using a native French classification model, the chatbot leverages the following process:

1. The user's input in French is translated into English using the `deep-translator` library and Google Translate API. The system handles long input texts by splitting them into manageable chunks using regular expressions.

2. The translated English text is passed through the English classification pipeline using a fine-tuned transformer model. The model returns a predicted disease along with its description and reference URL.

3. The disease name and description are then translated back into French using the same translation

function, and the final output is displayed to the user.

This strategy enables seamless French support without requiring a dedicated French-language biomedical model. It also maintains high accuracy by leveraging the performance of well-trained English models while ensuring that users receive responses in their native language.

This multilingual pipeline—featuring native processing for Arabic and English and a translation mechanism for French—ensures that all users receive accurate, linguistically appropriate, and accessible medical guidance.

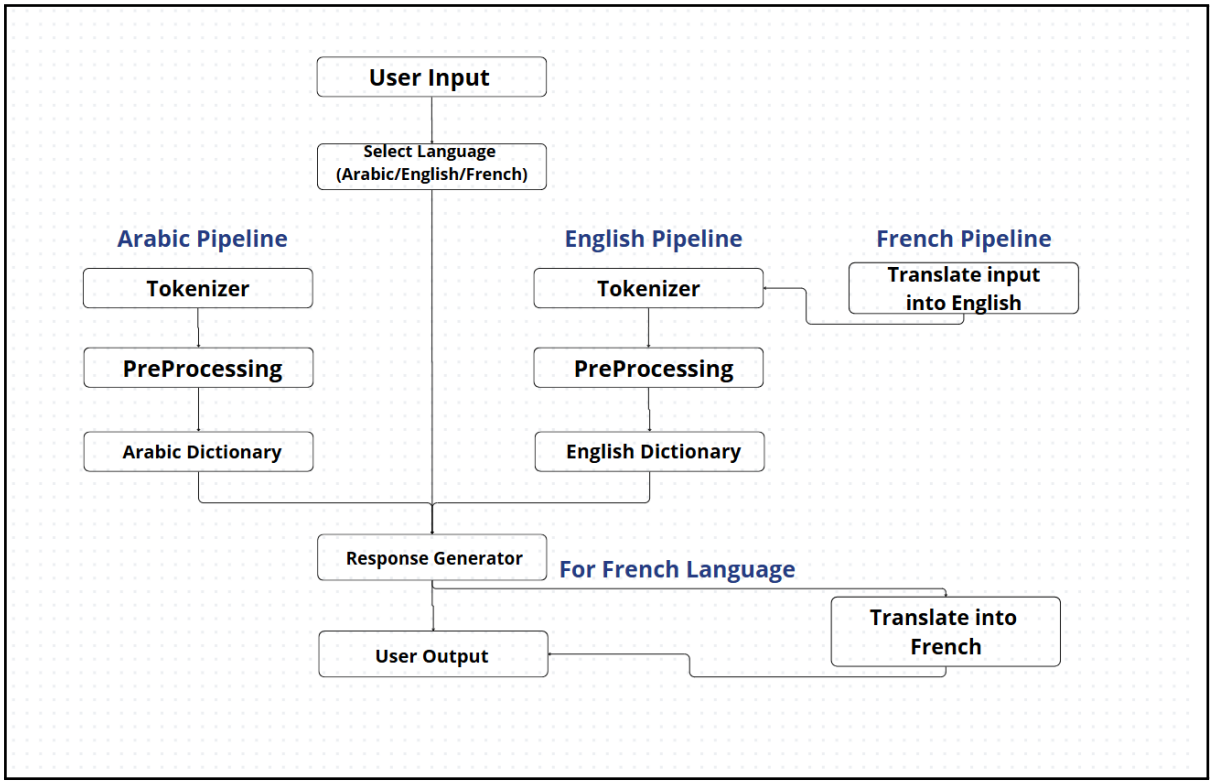


Figure III.2.2: Multilingual Preprocessing Pipeline.

III.2.5 Technologies and Frameworks for Medical NLP Integration

The development of the proposed medical diagnosis system relies on a modern, open-source software stack to support natural language processing (NLP), machine learning, and deployment. The key libraries and frameworks used are listed below.

Library/Tool	Description	Official Website
Python	Core programming language used for model development and system integration [107].	python.org
pandas	Library for reading and processing structured tabular data, such as Excel datasets [65].	pandas.pydata.org
NumPy	Enables high-performance numerical computations and array manipulations [42].	numpy.org
PyTorch	Deep learning framework used to build, train, and evaluate the transformer models [74].	pytorch.org
Hugging Face Transformers	Provides access to pretrained models and tokenizers (e.g., MARBERT, BioBERT) via `AutoTokenizer` and `AutoModelForSequenceClassification` [115].	huggingface.co
scikit-learn	Used for preprocessing, label encoding, evaluation metrics, and classification reporting [75].	scikit-learn.org
AdamW Optimizer	A weight-decay optimizer from PyTorch, commonly used in fine-tuning transformer models.	pytorch.org
TensorDataset & DataLoader	PyTorch utilities for batching, structuring, and iterating through model training and evaluation data.	pytorch.org

Table III.2.2: Development Libraries and Frameworks Used in the System

III.3 System Architecture

This section provides an overview of the system's architecture, describing how the core components interact to deliver real-time chatbot responses effectively. The architecture is organized into three primary layers: the backend, AI model layer, and database layer. Each component plays a critical role in maintaining smooth operation and delivering accurate, multilingual disease predictions.

III.3.1 Modular System Architecture Overview

As illustrated in Figure III.3.1, the system is designed for modularity and scalability. It is capable of supporting multiple languages and generating fast, reliable medical predictions. Each layer is self-contained, yet tightly integrated to ensure seamless functionality throughout the pipeline.

The architecture consists of the following three main layers:

- **Backend Layer:** Built using the Flask framework, the backend manages user sessions, handles language routing, and acts as the interface between the user input and the AI model layer. It

receives structured input (symptoms, age, gender, and selected language), dispatches it to the appropriate language processing module, and coordinates the response generation. It also manages translation for French input and output using the `deep-translator` API.

- **AI Model Layer:** This layer hosts the pretrained and fine-tuned transformer models responsible for disease prediction. Models include BioBERT and PubMedBERT for English, AraBERT and MARBERT for Arabic, and an English-based inference pipeline for French using translation. This layer performs tokenization, model inference, label decoding, and probability scoring.
- **Database Layer:** A structured dataset stored in tabular format (Excel), containing comprehensive metadata for each disease. This includes the disease name, description, symptoms, age group, gender, and source URL. This information is dynamically retrieved at runtime based on the model's predicted label.

This modular backend-driven architecture ensures maintainability, scalability, and language-specific adaptability. The separation of concerns across these three layers allows for efficient model updates, data integration, and extension to additional languages or medical features.

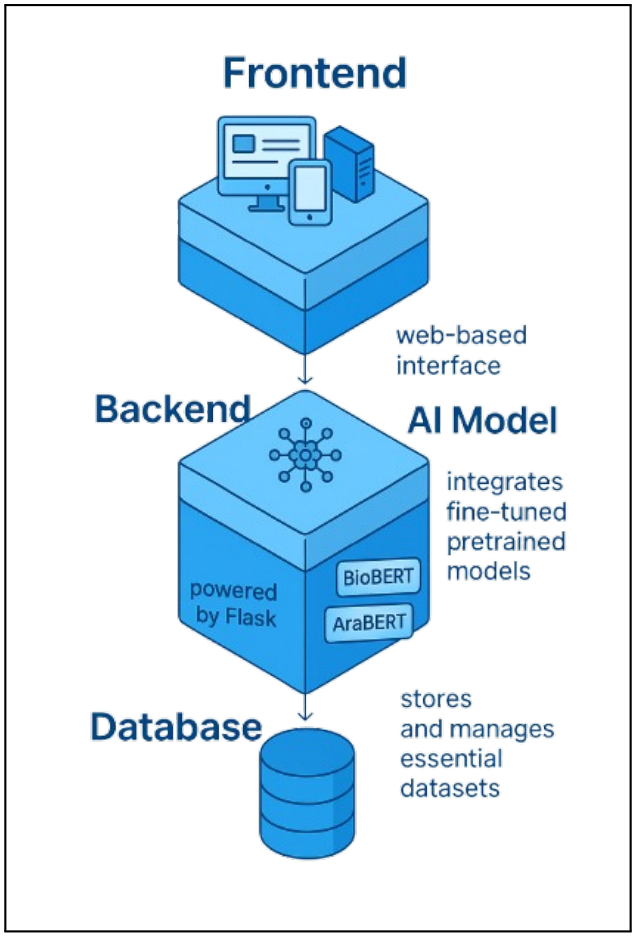


Figure III.3.1: High-level architecture of the multilingual medical chatbot, showing interactions between the frontend, Flask-powered backend, AI model layer, and data storage.

Component	Description
Backend Layer	Built using the Flask framework, it manages user sessions, handles language-specific routing, translates French input/output via the `deep-translator` API, and serves as the intermediary between user input and the AI model layer.
AI Model Layer	Hosts fine-tuned transformer-based models (e.g., BioBERT, PubMedBERT for English; AraBERT, MARBERT for Arabic), performing tokenization, inference, label decoding, and probability scoring. French input is translated and processed through the English model pipeline.
Database Layer	Contains a structured Excel-based dataset with metadata for each disease (name, description, symptoms, age group, gender, and source URL). This data is dynamically retrieved based on predicted disease labels.

Figure III.3.2: Core components of the system architecture.

III.3.2 Backend Services and APIs

The backend layer, developed using `Flask`—a lightweight yet powerful web framework for Python—handles the efficient processing of user queries and the generation of medical predictions. Flask's minimalist design, while maintaining flexibility, makes it well-suited for developing APIs and web services in machine learning and artificial intelligence applications [40].

Upon receiving input from the user, the backend detects the language, selects the appropriate AI model, and generates the corresponding prediction. It also manages user session states, performs error handling, and logs server-side activities. Moreover, it ensures seamless communication between the frontend interface and the AI model layer, enabling efficient data transfer and low-latency responses.

The API layer functions as an intermediary, facilitating structured and modular communication between the frontend, backend logic, and AI-driven prediction modules. This modular architecture improves the system's scalability, maintainability, and overall performance.

III.3.3 AI Model Integration

This section outlines the integration and fine-tuning of pre-trained transformer-based language models for multilingual medical text classification. The system supports three languages: English, Arabic, and French, with distinct strategies employed for each.

For English and Arabic, the system utilizes domain-specific transformer models that have been fine-tuned on annotated biomedical datasets. Specifically, `BioBERT` and `PubMedBERT` are used for English, while `AraBERT` and `MARBERT` are used for Arabic. These models enable the system to accurately interpret symptom descriptions and predict relevant medical conditions in their respective languages. Each language pipeline includes its own tokenizer, preprocessing components, and label decoder to ensure

semantic precision.

French input is supported through a translation-based approach. Instead of using a dedicated French biomedical model, the system translates French text into English, processes it using the English pipeline, and translates the results back into French. This method allows French-speaking users to interact effectively with the system while maintaining high prediction accuracy.

This multilingual model integration strategy ensures both performance and accessibility, enabling the chatbot to deliver accurate and linguistically appropriate responses across diverse user populations.

III.3.4 Selection of Multilingual Biomedical Language Models

The choice of transformer-based models was guided by both domain expertise and the need for multilingual versatility, aiming to support Arabic and English biomedical texts effectively.

For Arabic, **AraBERT** and **MARBERT** were selected due to their strong capabilities in handling Modern Standard Arabic and various Arabic dialects, respectively. These models have proven particularly effective in processing patient-generated content and clinical notes written in Arabic.

In the English biomedical domain, **BioBERT** and **PubMedBERT** were employed. These models are pretrained on extensive medical corpora, including PubMed abstracts and full-text biomedical articles, which enables them to capture clinical and scientific terminology with high accuracy.

To support multilingual and generative tasks, the multilingual encoder-decoder model **T5-small** was incorporated, aiding in sequence generation and classification across languages. Additionally, **GigaBERT**, pretrained on both Arabic and English corpora, was used to provide strong cross-lingual representations, allowing the system to effectively process bilingual or code-switched medical inputs.

The selected models are summarized in Table III.3.1.

Model	Language Support	Architecture Type	Key Features / Strengths
T5-small	Multilingual (incl. Arabic & English)	Encoder-Decoder (Seq2Seq)	Supports text generation and multilingual tasks; versatile across languages
AraBERT	Arabic (MSA)	Encoder-only (BERT)	Pretrained on Arabic news and Wikipedia; effective for formal medical texts
MARBERT	Arabic (Dialectal)	Encoder-only (BERT)	Trained on social media data; handles informal and dialectal Arabic
CamelBERT (finetuned)	Arabic (MSA, dotted orthography)	Encoder-only (BERT)	Fine-tuned on medical text with diacritics; improved performance on dotted Arabic clinical language
BioBERT	English (Biomedical)	Encoder-only (BERT)	Pretrained on PubMed abstracts; strong in clinical and research contexts
PubMedBERT	English (Biomedical)	Encoder-only (BERT)	Trained from scratch on biomedical corpora; excellent for fine-grained biomedical NLP tasks
GigaBERT	Arabic and English	Encoder-only (BERT)	Cross-lingual training; ideal for bilingual or code-switched data

Table III.3.1: Summary of selected multilingual biomedical language models.

For English, **BioBERT**, **BioClinicalBERT**, **GigaBERT**, and **BlueBERT** were selected for their strong biomedical grounding.

III.3.5 Fine-Tuning and Integration

All selected models were fine-tuned on a multi-class medical text classification task using carefully curated, domain-specific labeled datasets in both Arabic and English. Fine-tuning was conducted via supervised learning, employing a cross-entropy loss function. To address potential class imbalance and improve generalization across underrepresented classes, stratified sampling was applied during training.

Extensive hyperparameter tuning was carried out to identify optimal configurations for learning rate, batch size, and number of training epochs, resulting in consistent performance across a wide range of linguistic inputs.

To enable multilingual processing, a modular classification pipeline was developed. The pipeline begins with an automatic language identification module that detects whether the input text is in English or Arabic. Depending on the detected language, the system routes the input to the corresponding fine-tuned model—`AraBERT` for Arabic or `PubMedBERT` for English—for classification.

This modular, language-aware design enhances classification accuracy by leveraging specialized models tailored to each language. Moreover, it ensures scalability and maintainability in practical healthcare applications. The architecture also supports seamless integration with multilingual clinical decision support systems and medical chatbots, maintaining robust performance across diverse linguistic and medical contexts.

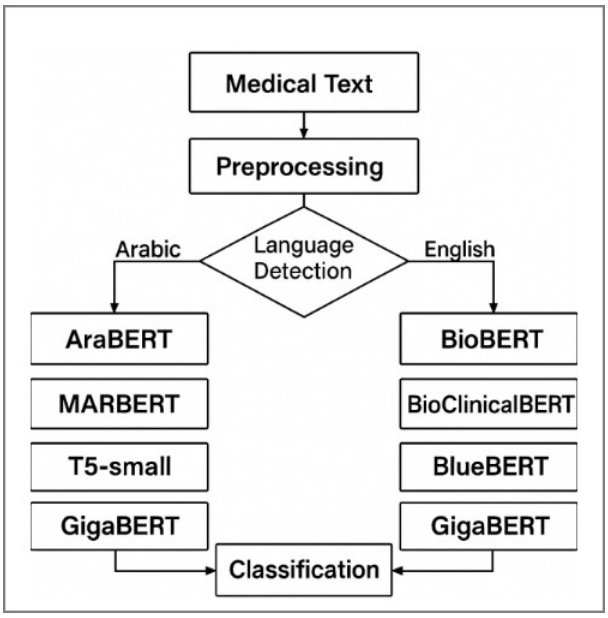


Figure III.3.3: AI Model Integration Workflow

III.3.6 Design of Medical Data Schema and Storage Strategy

This section outlines the data storage design used to support both model training and inference. The dataset is maintained in a structured Excel format, enabling fast access, easy readability, and compatibility with various data preprocessing tools.

The data follows a tabular schema with clearly defined fields: **Disease Name**, **Symptoms**, **Age Group**, **Category**, **Gender**, **URL**, and **Description**. These fields provide critical contextual information essential for accurate classification as well as for downstream analytical tasks.

Table III.3.2 summarizes the dataset structure, highlighting the main fields and their descriptions.

Field Name	Description
Disease Name	The medical condition or disease being classified (e.g., ”Diabetes”, ”Asthma”).
Symptoms	The textual representation of the patient’s symptoms, which the model uses for classification (e.g., ”Fatigue, frequent urination”).
Age Group	The demographic category of the patient (e.g., ”Adult”, ”Child”).
Category	The general medical category or domain to which the disease belongs (e.g., ”Endocrine”, ”Respiratory”).
Gender	The gender of the patient, which may be relevant for specific conditions (e.g., ”Male”, ”Female”, ”Children”).
URL	A source reference link for further reading or verification (e.g., ”https://www.webmd.com/diabetes/default.htm”).
Description	Additional contextual information about the disease or condition (e.g., ”A chronic condition that affects how the body processes blood sugar”).

Table III.3.2: Overview of the medical dataset schema and main fields.

The dataset is accessed at runtime using robust data management libraries such as `pandas`, enabling seamless integration with preprocessing pipelines and model inputs. This approach ensures scalability, maintainability, and flexibility, making it straightforward to update or expand the dataset as new data becomes available. By facilitating efficient data handling and adaptability, this design enhances both the operational effectiveness and long-term sustainability of the system.

III.4 Dataset Preparation

This chapter describes the techniques employed to collect, format, filter, and augment the dataset used for multilingual medical text classification. We developed an end-to-end pipeline to ensure data consistency, linguistic diversity, and feature richness, thereby enhancing model accuracy and generalizability.

III.4.1 Data Sources and Collection

The dataset was meticulously compiled from five reputable online medical information sources: **Altibbi**, **Mayo Clinic**, **Sohati**, **WebTeb**, and **WebMD**. These resources were selected for their clinical validity, linguistic diversity in both Arabic and English, and relevance to patient-centered health content. Data extraction combined automated web scraping with manual validation to ensure accuracy and

consistency. Extracted content was normalized into a structured format suitable for text classification, with particular emphasis on maintaining semantic consistency across both languages.

app.py

```python
import requests
from bs4 import BeautifulSoup
import json

# Base URL for scraping
base_url = "https://www.sohati.com/Category/%D8%A3%D9%85%D8%B1%D8%A7%D8%B6"
maladies_data = {}
```

snappify.com

app.py

```python
# Function to scrape diseases for a given Arabic letter
def scrape_diseases_by_letter(letter):
    try:
        # Sending request to the URL for the letter
        response = requests.get(f"{base_url}?letter={letter}")
        if response.status_code != 200:
            print(f"Request error for letter {letter}: {response.status_code}")
            return

        soup = BeautifulSoup(response.text, 'html.parser')

        # Find the div with id `disease-[letter]`
        disease_section = soup.find('div', id=f'disease-{letter}')
        if not disease_section:
            print(f"No disease section found for letter {letter}.")
            return

        # Loop through all disease items in the section
        disease_items = disease_section.find_all('a', class_='amradatag')

        if not disease_items:
            print(f"No diseases found for letter {letter}.")
            return
```

snappify.com

app.py

```python
for disease_link in disease_items:
            disease_name = disease_link.get_text(strip=True)
            disease_url = disease_link['href']

            if not disease_url.startswith('http'):
                disease_url = "https://www.sohati.com" + disease_url

            # Scrape paragraphs from the disease's page
            paragraphs = scrape_paragraphs(disease_url)

            # Add disease information to the dictionary
            maladies_data[disease_name] = {
                "URL": disease_url,
                "Paragraphs": paragraphs
            }

            print(f"Disease found: {disease_name} - URL: {disease_url} - Paragraphs: {paragraphs}")

    except Exception as e:
        print(f"Error during scraping for letter {letter}: {e}")
```

snappify.com

Figure III.4.1: Data extraction workflow from Sohati — Part 1.

app.py

```python
# Function to scrape all paragraphs from a disease page
def scrape_paragraphs(disease_url):
    try:
        response = requests.get(disease_url)
        if response.status_code != 200:
            print(f"Request error for disease URL {disease_url}: {response.status_code}")
            return []

        soup = BeautifulSoup(response.text, 'html.parser')

        # Find the content section using the specified class
        content_div = soup.find('div', class_='col-auto article-post-content')

        if content_div:
            # Extract all paragraphs within the content div
            paragraphs = content_div.find_all('p')
            # Clean the text and return
            return [para.get_text(strip=True) for para in paragraphs if para.get_text(strip=True)]
        else:
            return ["No content found"]

    except Exception as e:
        print(f"Error while scraping paragraphs for {disease_url}: {e}")
        return []
```

snappify.com



app.py

```python
# Scraping all diseases
def scrape_all_diseases():
    # List of Arabic letters to scrape (from 'ا' to 'ي')
    arabic_letters = ['ا', 'ب', 'ت', 'ث', 'ج', 'ح', 'خ', 'د', 'ذ', 'ر', 'ز', 'س', 'ش', 'ص', 'ض', 'ط', 'ظ', 'ع', 'غ', 'ف', 'ق', 'ك', 'ل', 'م', 'ن', 'ه', 'و', 'ي']

    for letter in arabic_letters:
        print(f"\nScraping diseases for letter: {letter}...")
        scrape_diseases_by_letter(letter)
```

snappify.com

app.py

```python
# Call the function to scrape all diseases
scrape_all_diseases()

# Save the dictionary to a JSON file
with open('maladies_sohati.csv', 'w', encoding='utf-8') as csv_file:
    csv.dump(maladies_data, csv_file, ensure_ascii=False, indent=4)

print("\nDisease dictionary saved to 'maladies_sohati.csv'.")
```

snappify.com

Figure III.4.2: Data extraction workflow from Sohati — Part 2.

Platform	Languages Supported	Focus Area	Size
Altibbi	Arabic, English	General medical content, patient-oriented health information	875
Mayo Clinic	English	Reputable medical research, patient education materials	985
Sohati	Arabic, English	Health conditions, lifestyle advice, medical news	358
WebTeb	Arabic, English	Common diseases, treatment options, wellness advice	439
WebMD	English	Symptom checker, drug information, medical news	1889

Table III.4.1: Data Sources for Dataset Collection

III.4.2 Dataset Structure and Attributes

The data used in the MediLingo system is carefully organized to support accurate and interpretable disease prediction across multiple languages and populations. Each dataset record contains the following key attributes:

- **Symptoms**: A textual field capturing the patient's reported symptoms, serving as the primary input for the classification model and processed using natural language techniques.
- **Age Group**: A categorical variable specifying the patient's age range, enabling the model to adjust predictions based on age-related disease prevalence.
- **Gender**: Indicates the patient's gender (*Man*, *Woman*, or *Children*), allowing consideration of gender-specific risk factors.
- **Category**: A general classification of the disease type (e.g., *cardiovascular*, *dermatological*, *respiratory*), useful for organizing and filtering predictions.
- **Disease Name**: The predicted label representing the medical condition; serves as the ground truth during training and the target class during inference.
- **Description**: A detailed explanation of the disease, including definition, symptoms, causes, treatments, and other relevant medical information, provided to users as part of the chatbot response.
- **URL**: A link to a reputable medical website offering further information, enhancing the chatbot's credibility and transparency.

The combination of textual and categorical features enriches the dataset, supporting robust multi-class classification. Each attribute contributes to narrowing down potential diagnoses, collectively enabling a comprehensive and interpretable approach to automated medical assistance.

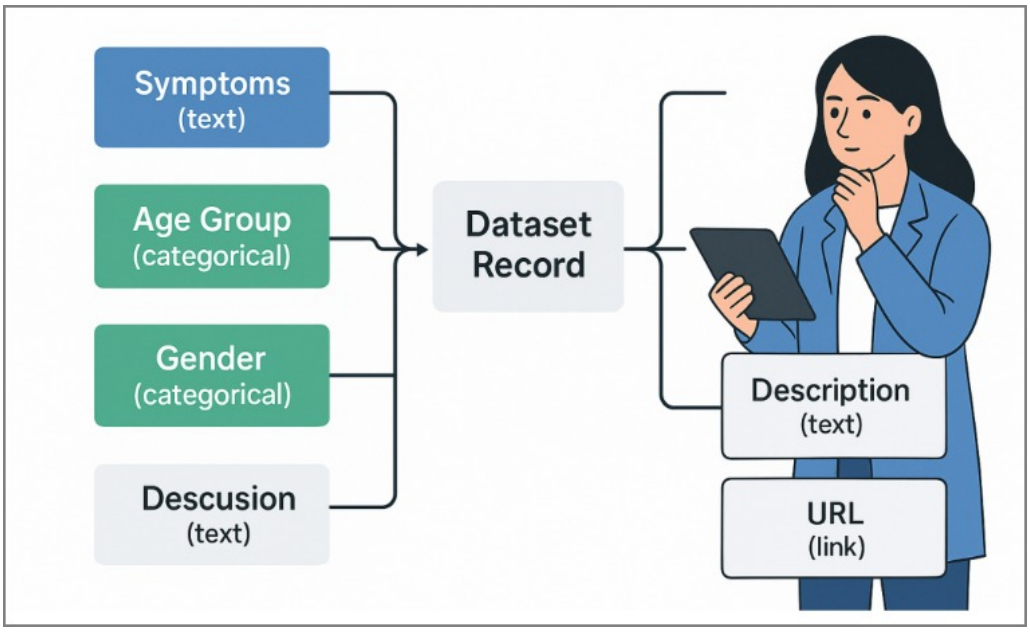


Figure III.4.3: Diagrammatic illustration of a record in the dataset of the MediLingo system.

The following Figure (III.4.3) illustrates the structure of one data record with patient metadata and textual symptom interaction in making precise and context-aware medical predictions.

III.4.3 Preprocessing Methodologies

To ensure data quality and linguistic consistency across both Arabic and English corpora, a comprehensive preprocessing pipeline was implemented. This pipeline addresses the structural, orthographic, and semantic characteristics of the input text, aiming to enhance model interpretability while minimizing input variability.

III.4.3.1 General Preprocessing Procedures

involved the following steps:

- **Data Cleaning:** Rectified structural inconsistencies to ensure uniform formatting across all entries.



```python
def clean_text(text):

text = text.lower()

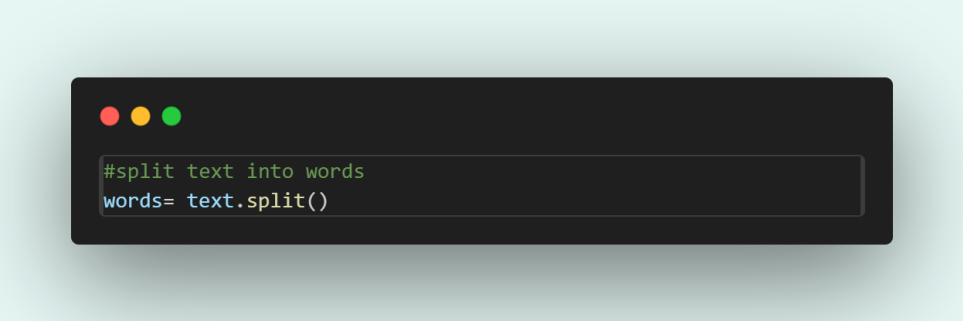
text = re.sub(r" [^\w\s]", "", text)

text = re.sub(r"\d+", "", text)

return text
```

Figure III.4.4: Code used for Text Cleaning

- **Tokenization:** Applied language-specific tokenization techniques to segment raw text into meaningful lexical units suitable for model processing.



```python
#split text into words
words= text.split()
```

Figure III.4.5: Tokenization of the Dataset

For our deep learning and pre-trained model approaches, we utilized the tokenizers associated with each domain-specific model for both Arabic and English. These tokenizers incorporate specialized vocabularies and contextual handling optimized for biomedical and clinical NLP tasks.

Arabic Tokenizers:

– **AraBERT Tokenizer**: Uses WordPiece tokenization with Arabic normalization and diacritics removal.



Figure III.4.6: Calling the AraBERT Tokenizer

– **MARBERT Tokenizer**: Designed for dialectal Arabic with a broad vocabulary.



Figure III.4.7: Calling the MARBERT Tokenizer

– **T5-small Tokenizer**: Uses SentencePiece, treating tasks as text generation.



Figure III.4.8: Calling the T5-small Tokenizer

– **GigaBERT (Arabic) Tokenizer**: A multilingual WordPiece tokenizer.



Figure III.4.9: Calling the GigaBERT Tokenizer

English Tokenizers:

– **BioBERT Tokenizer**: WordPiece tokenizer trained on biomedical corpora.



Figure III.4.10: Calling the BioBERT Tokenizer

– **BioClinicalBERT Tokenizer**: Fine-tuned on clinical notes.



Figure III.4.11: Calling the BioClinicalBERT Tokenizer

– **PubMedBERT Tokenizer**: Trained from scratch on PubMed using WordPiece.



Figure III.4.12: Calling the PubMedBERT Tokenizer

– **GigaBERT (English) Tokenizer**: Multilingual tokenizer with cross-lingual support.



Figure III.4.13: Calling the GigaBERT Tokenizer

- **Stop-word Removal:** Removed frequent function words with minimal semantic content to improve signal-to-noise ratio.



Figure III.4.14: Code used for Stop-word Removal

- **Label Encoding:** Converted categorical features (e.g., disease names) into integer representations suitable for machine learning algorithms.

III.4.3.2 Arabic Text Preprocessing

required additional steps due to the language's morphological complexity and orthographic variability. The following normalization procedures were applied:

- **Diacritic Removal:** Diacritical marks (ḥarakāt) were removed to reduce lexical sparsity and ensure consistent word forms.

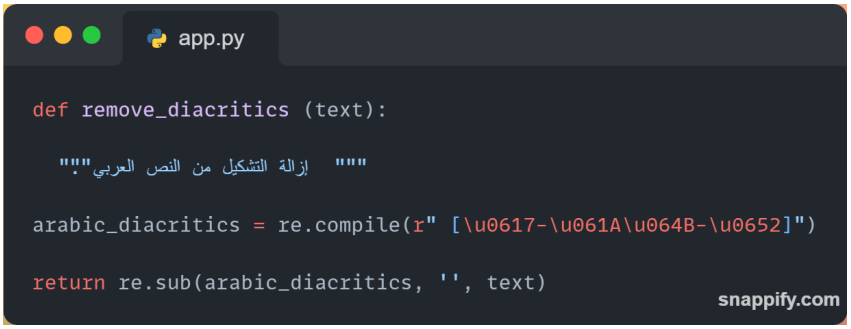


Figure III.4.15: Code used for removing Arabic diacritics.

- **Character Normalization:** Orthographic variants of characters (e.g., different forms of *Alif*) were unified into a canonical form to reduce ambiguity.
- **Tatweel Elimination:** Non-semantic elongation characters ( ) were removed to reduce input noise.
- **Punctuation and Non-Arabic Character Filtering:** Unnecessary punctuation marks and foreign characters were removed to preserve the integrity of the Arabic content.
- **Whitespace Normalization:** Whitespaces were standardized to ensure consistent token boundaries and prevent misalignment.
- **Stopword Removal:** Common Arabic stopwords (e.g., من, في, على) were removed to retain only meaningful words for analysis.

app.py

```python
import re

def remove_tatweel(text):
    return text.replace("ـ", "")  # Tatweel character

def normalize_alef(text):
    return re.sub(r"[إأآ]", "ا", text)  # Normalize all forms of Alef to "ا"

def normalize_teh_marbuta(text):
    return text.replace("ة" ,"ه")  # Teh Marbuta to Heh

def remove_diacritics(text):
    arabic_diacritics = re.compile(r"[\u0617-\u061A\u064B-\u0652]")
    return re.sub(arabic_diacritics, "", text)

def clean_arabic_text(text):
    if isinstance(text, str):
        text = remove_tatweel(text)
        text = normalize_alef(text)
        text = normalize_teh_marbuta(text)
        text = remove_diacritics(text)
    return text
```

snappify.com

Figure III.4.16: Code used for Cleaning Arabic Text

III.4.3.3 Preprocessing Summary Figure:

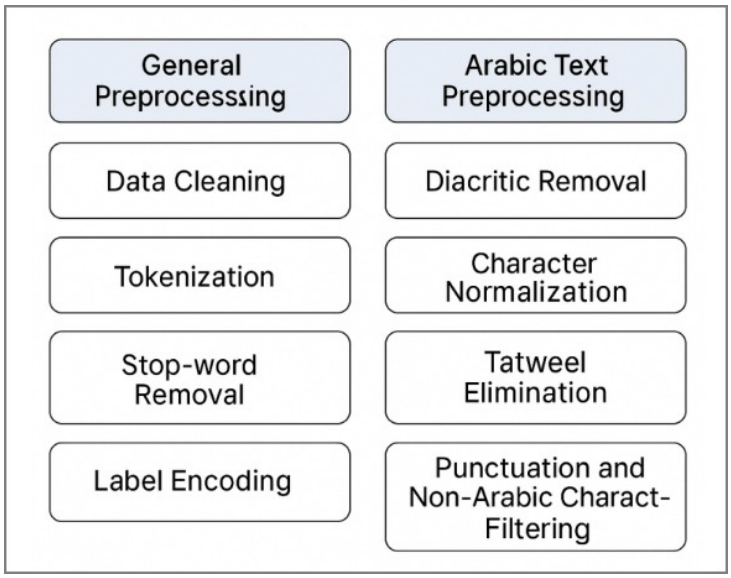


Figure III.4.17: Summary of Preprocessing Methodology for Multilingual Medical Text Classification

The preprocessing methodologies were designed to ensure that the dataset was lexically coherent, syntactically consistent, and semantically aligned with the requirements of transformer-based models. This rigorous standardization process was crucial for effective model training and ensured robust multilingual classification performance across both Arabic and English languages.

III.4.4 Data Augmentation Strategies

To mitigate class imbalance and increase linguistic variability within the dataset, several data augmentation strategies were employed. These methods were carefully selected to improve model generalization by introducing controlled perturbations and semantic diversity, without altering the underlying meaning of the text. The following augmentation techniques were applied:

Technique	Description
Noise Injection	Random character-level or word-level modifications were introduced to simulate typographical or transcription errors. This technique exposed the model to naturally occurring noise in real-world inputs, thereby enhancing robustness (see Figure III.4.18).
Back-Translation	Sentences were translated from the source language to an intermediate language (e.g., Arabic → English → Arabic) and then translated back to the original language. This approach generated paraphrased variants that retained the original semantics while introducing syntactic and lexical diversity (see Figure III.4.19).
Constrained Word Shuffling	Words within a defined context window were rearranged to alter the surface structure of the sentence while preserving its core meaning. This syntactic variation aids in reducing overfitting to rigid sentence structures, promoting better generalization (see Figure III.4.20).

Table III.4.2: Data Augmentation Techniques for Multilingual Medical Text Classification

app.py

```python
import random

def add_noise_arabic(text, error_rate=0.1):
    words = text.split()
    num_errors = int(len(words) * error_rate)

    for _ in range(num_errors):
        word_index = random.randint(0, len(words) - 1)
        word = words[word_index]

        if len(word) > 2:
            char_index = random.randint(0, len(word) - 1)
            random_char = chr(random.randint(0x0600, 0x06FF))  # Random Arabic Unicode character
            word = word[:char_index] + random_char + word[char_index + 1:]
            words[word_index] = word

    return ' '.join(words)
```

snappify.com

Figure III.4.18: Illustration of Noise Injection for text data, simulating typographical or OCR-like errors.

app.py

```python
from deep_translator import GoogleTranslator

def backtranslate_deep(text):
    try:
        # Translate from Arabic to English
        english = GoogleTranslator(source='ar', target='en').translate(text)

        # Translate back from English to Arabic
        back_arabic = GoogleTranslator(source='en', target='ar').translate(english)

        return back_arabic

    except Exception as e:
        print("Backtranslation error:", e)
        return text
```

snappify.com

Figure III.4.19: Back-Translation workflow showing source-to-intermediate-to-source language conversion.

app.py

```python
import pandas as pd

def shuffle_excel_data(file_path, sheet_name=0, output_path=None):
    """
    Loads an Excel file, shuffles the data.

    Parameters:
        file_path (str): Path to the Excel file.
        sheet_name (str or int): Sheet name or index to read (default is the first sheet).
        output_path (str): Path to save the shuffled data. If None, data is not saved.

    Returns:
        pd.DataFrame: Shuffled DataFrame.
    """

    # Load the Excel sheet into a DataFrame
    df = pd.read_excel(file_path, sheet_name=sheet_name)

    # Shuffle the rows
    shuffled_df = df.sample(frac=1).reset_index(drop=True)

    # Save if output path is provided
    if output_path:
        shuffled_df.to_excel(output_path, index=False)

    return shuffled_df
```

snappify.com

Figure III.4.20: Constrained word shuffling example, showing context-preserving rearrangements.

III.4.5 Splitting the Dataset

In this step, the dataset is divided into training and testing sets while maintaining class balance through stratification. Initially, 80% of the data is allocated for training and 20% for testing. The training set may later be further split to create a validation set if needed. The data splitting process is illustrated in Figure III.4.21.

app.py

```python
from sklearn.model_selection import train_test_split

# Train/Test split
train_texts, val_texts, train_labels, val_labels = train_test_split(
    df_filtered["input_text"].tolist(),
    df_filtered["label"].tolist(),
    test_size=0.2,
    random_state=42,
    stratify=df_filtered["label"]
)
```

snappify.com

Figure III.4.21: Code used for splitting the dataset into training and testing sets.

III.4.6 Label Encoding and Class Mapping

Label encoding and class mapping are essential operations in the preprocessing workflow for supervised machine learning applications involving categorical target variables. Label encoding refers to the process of converting categorical class labels into numerical format, typically by assigning a unique integer to each distinct category (e.g., ”positive” → 0, ”neutral” → 1, ”negative” → 2). This transformation is necessary because most machine learning algorithms require numerical input and cannot process textual labels directly [75]. Class mapping complements this by ensuring a consistent association between original class names and their corresponding numeric values across the dataset. This consistency is particularly crucial in multi-source data environments or during model deployment, where accurate interpretation of encoded labels is key to maintaining performance and reliability [59].

An example implementation of label encoding and class mapping in Python is shown in Figure III.4.22. It demonstrates the creation of mappings between categorical labels and their corresponding numeric identifiers, along with their application to both training and test datasets.

app.py

```python
# Get unique labels sorted
unique_labels = sorted(list(set(train_labels)))

# Create mappings
label2id = {label: i for i, label in enumerate(unique_labels)}
id2label = {i: label for label, i in label2id.items()}

# Encode labels
train_labels = [label2id[label] for label in train_labels]
val_labels = [label2id[label] for label in val_labels]  # fixed from "test_labels"
```

snappify.com

Figure III.4.22: Example of label encoding and class mapping in Python.

Moreover, special attention should be paid to issues such as class imbalance, rare categories, and hierarchical relationships among labels. Inappropriate encoding strategies can introduce bias or hinder model learning, particularly when encoded integers are mistakenly interpreted as ordinal values by certain algorithms [118]. Consequently, label encoding should be accompanied by proper metadata documentation. Where appropriate, alternative encoding techniques (e.g., one-hot encoding or target encoding) should be considered, depending on the characteristics of the data and the model’s requirements.

III.4.7 Deployment and Runtime Environment

The deployment and runtime environment forms a critical component of the machine learning lifecycle, determining the stability, scalability, and real-time performance of the trained model. After the model development and evaluation stages, deploying the model into a production-ready environment involves configuring the necessary infrastructure to support inference, monitoring, and integration with external systems. This typically includes packaging the model using containerization technologies such as Docker, orchestrating deployments with Kubernetes or cloud-native tools, and exposing the model via RESTful APIs or batch processing pipelines [56].

A robust runtime environment must ensure compatibility with the software dependencies used during training (e.g., specific Python versions, machine learning libraries such as scikit-learn or TensorFlow), while also providing mechanisms for automatic scaling and fault tolerance. Additionally, runtime monitoring tools are essential to track performance drift, latency, and errors, thereby supporting continuous improvement and retraining strategies [24]. Security aspects such as authentication, access control, and data privacy must also be addressed, particularly in sensitive domains like healthcare or finance. The choice of deployment strategy—whether edge deployment, cloud-based inference, or hybrid approaches—should be aligned with the application's latency requirements, data access patterns, and computational constraints [89].

III.5 Summary

This chapter detailed the essential preprocessing and deployment steps required to prepare and operationalize machine learning models effectively. It began with general preprocessing procedures, including data cleaning, tokenization, stop-word removal, and label encoding, with additional techniques tailored for Arabic text such as diacritic removal and character normalization. The importance of label encoding and class mapping in transforming categorical data for model compatibility was also discussed. Finally, the chapter addressed the deployment and runtime environment, emphasizing the infrastructure, monitoring, and security considerations necessary for stable and scalable model execution in real-world applications.

Chapter IV

Results and Evaluation

IV.1 Chapter Overview

This chapter presents the results and evaluation of the proposed system. It outlines the experimental methodology, including evaluation metrics and workflow, and examines the dataset structure with a focus on class imbalance. The chapter then details three phases: baseline performance on the imbalanced dataset, the impact of data augmentation, and results after balancing the data. It also compares Arabic and English biomedical models, analyzes findings, and showcases a multilingual chatbot interface to demonstrate practical outcomes.

IV.2 Evaluation Methodology

To assess the chatbot's performance effectively, a structured and comprehensive evaluation methodology was employed. The primary goal was to determine the system's ability to accurately classify user-reported symptoms and recommend appropriate medical conditions across three languages—Arabic, English, and French. Arabic and English were supported by fine-tuned language-specific models, while French inputs were handled via a translation pipeline: inputs were translated to English, classified using the English model, and the results (including the predicted condition and description) were translated back into French.

IV.2.1 Objectives of Evaluation

The main objectives of this evaluation are:

- To assess the accuracy and reliability of the AI models in diagnosing medical conditions.
- To evaluate the responsiveness and quality of user interaction with the system.
- To verify consistent and accurate performance across multilingual inputs (Arabic, English, and French).
- To identify potential areas for improvement through simulated real-world usage.

IV.2.2 Performance Metrics

To evaluate the performance of the multilingual medical chatbot, several standard classification metrics were applied to the datasets in both Arabic and English. These include accuracy, precision, recall, F1-score, and the Matthews Correlation Coefficient (MCC). These metrics collectively offer a comprehensive view of the system's classification quality in each language. The results are summarized in the table below.

- **Accuracy:** The proportion of correctly predicted cases among all predictions.

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

While informative, accuracy alone can be misleading in cases of class imbalance.[78]

- **Precision:** The ratio of true positive predictions to the total predicted positives. It answers the question: *Of all the diseases predicted by the model, how many were actually correct?*[79]

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

- **Recall (Sensitivity):** The ratio of true positive predictions to all actual positive cases. It answers the question: *How many of the actual cases did the model successfully detect?*[87]

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score:** The harmonic mean of precision and recall. This balanced metric is especially useful when dealing with uneven class distributions.[39]

$$F_1{-}Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

- **Matthews Correlation Coefficient (MCC):** A robust metric that considers all four categories of the confusion matrix (TP, TN, FP, FN). MCC is particularly valuable for evaluating performance on imbalanced datasets, returning a score between -1 and +1:
  - +1 indicates perfect prediction,
  - 0 indicates performance no better than random guessing,
  - -1 indicates complete disagreement between prediction and actual labels.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Together, these metrics provide a multidimensional evaluation of the system's diagnostic accuracy, reliability, and robustness.[19]

IV.3 Overview of Experimental Workflow

The experimental workflow was designed to systematically evaluate the performance of the multilingual medical chatbot system across different stages of model development and testing. This includes data preprocessing, training on imbalanced datasets, applying data augmentation techniques, and evaluating performance on balanced datasets. The workflow also reflects how different language inputs were processed, including translation pipelines for non-English inputs. Figure IV.3.1 provides a visual summary of the entire experimental process.

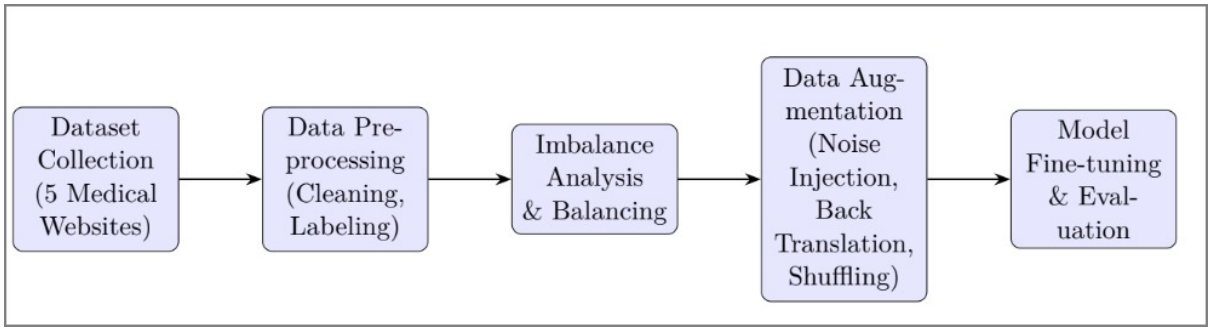


Figure IV.3.1: Overview of the experimental workflow

IV.4 Dataset Collection and Structure

The dataset was compiled from five publicly accessible medical websites, encompassing a wide variety of disease classes. However, many of these classes were severely underrepresented, with some containing only one or two samples. To enable effective training and meaningful classification, a filtering process was applied. This process retained only those disease classes that met a minimum threshold for sample count and demonstrated clinical relevance. As a result, the number of disease classes was reduced to 755.

Attribute	Description
Symptoms	Natural language input, primarily in Arabic
Age Group	Age ranges: 0–10, 11–20, ..., 61+
Gender	Patient gender (Man, Woman, Child)
Category	Disease category (e.g., respiratory, cardiovascular)
Disease Name	Classification label
Description	Text including causes, symptoms, treatments
URL	Source website

Table IV.4.1: Dataset Attribute Descriptions

IV.4.1 Class Distribution and Imbalance

Despite filtering the dataset down to 755 clinically relevant disease classes, a significant class imbalance remained. More than 75% of the classes contained exactly six samples—the minimum threshold for inclusion—while a small number of classes had over 300 instances. This skewed distribution poses a challenge for training robust and generalizable models. Figure IV.4.1 visualizes this imbalance, highlighting that the top three most frequent classes account for nearly 60% of the entire dataset.

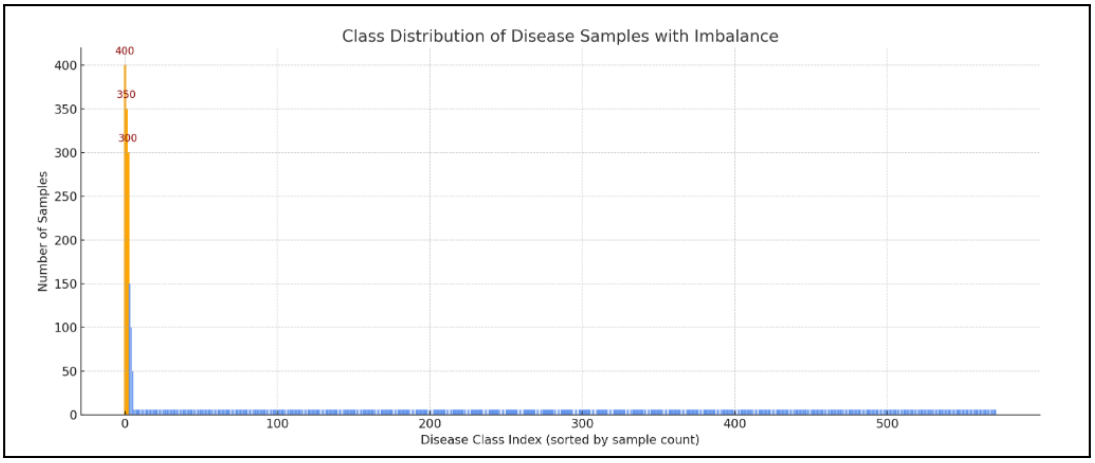


Figure IV.4.1: Distribution of disease classes showing data imbalance

IV.5 Phase 1: Baseline on Imbalanced Dataset

IV.5.1 Objective

The first phase aims to establish baseline performance by fine-tuning two transformer models on the original imbalanced dataset, without applying any augmentation or balancing techniques.

IV.5.2 Models Evaluated

- **BERT Base Multilingual Cased**: A widely-used multilingual transformer pretrained on large-scale corpora covering multiple languages, including Arabic. It serves as a strong general baseline.
- **CAMELBERT**: A domain-specific Arabic transformer pretrained exclusively on medical texts, designed to capture medical terminology and context more effectively.

IV.5.3 Experimental Setup

Both models were fine-tuned using the same training and test splits, with hyperparameters optimized for best performance. No data balancing or augmentation methods were applied to isolate the impact of imbalance on model effectiveness.

IV.5.4 Results

Model	Accuracy	Precision	Recall	F1-Score
BERT Base Multilingual Cased	50.0%	42.3%	43.1%	43.0%
CAMELBERT	66.3%	56.4%	59.2%	57.4%

Table IV.5.1: Performance of Baseline Models on Imbalanced Data

IV.5.5 Analysis

CAMELBERT significantly outperformed the general multilingual BERT, demonstrating the advantage of domain-specific pretraining in capturing medical language nuances. Despite this, both models exhibited relatively low recall and F1-scores, reflecting challenges posed by the severe class imbalance in the dataset. The tendency to favor majority classes resulted in poor minority class detection, limiting clinical applicability.

IV.5.6 Implications

These baseline results highlight the critical challenge that imbalanced medical datasets present. They motivate the exploration of data augmentation and class balancing methods in subsequent phases to improve model generalization, especially for rare but clinically important diseases.

IV.6 Phase 2: Data Augmentation

IV.6.1 Introduction

Data augmentation was employed to address the challenge of imbalanced datasets by increasing the diversity of the training data, thereby helping the model generalize better to unseen examples. Only BERT was tested in this phase to serve as a baseline model and isolate the effects of augmentation without confounding factors.

IV.6.2 Augmentation Techniques

- **Noise Injection:** This technique involves random synonym replacement, insertions, and deletions to slightly modify sentences without altering their core meaning. It helps the model become more robust to variations in phrasing.
- **Back Translation:** Text is translated from Arabic to English and then back to Arabic. This generates natural variations of the original text while preserving meaning, which is especially beneficial for a morphologically rich language like Arabic.

IV.6.3 Experimental Setup

Augmented data was generated by applying these techniques to a selected percentage of the original dataset, focusing on minority class samples to enhance their representation. Manual verification and quality checks ensured that the augmented sentences maintained semantic integrity.

IV.6.4 Evaluation

Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. Given the imbalanced nature of the dataset, recall and F1-score were emphasized as key indicators of the model's ability to identify minority class instances.

Model	Accuracy	Precision	Recall	F1-Score
BERT (Augmented)	76.0%	69.2%	71.3%	69.0%

Table IV.6.1: BERT with Augmentation on Imbalanced Data

IV.6.5 Analysis and Findings

Data augmentation improved recall and F1-score, indicating enhanced detection of the minority class. However, accuracy gains were limited, suggesting augmentation alone does not fully resolve class imbalance. Combining augmentation with techniques such as resampling or class weighting is recommended for better balance mitigation.

IV.6.6 Limitations

A key limitation in this phase was the persistent class imbalance, which constrained the model's overall performance. The scarcity of minority class samples limited the model's ability to generalize well to those cases. Additionally, relying on limited data sources made it challenging to create a truly representative and balanced dataset. Overcoming these limitations requires targeted data collection and balanced sampling strategies.

IV.6.7 Summary of Findings

While data augmentation contributed to improvements in recall and F1-score, these gains were not sufficient to fully address the issue of class imbalance. To achieve a more balanced and effective model, it is essential to focus on obtaining or generating a balanced dataset. This may involve collecting additional minority class samples or applying data balancing strategies beyond augmentation.

Insight: Augmentation improved recall and F1, but could not fully compensate for class imbalance.

IV.7 Phase 3: Balanced Dataset + Augmentation

The final balanced dataset spans hundreds of disease classes. Classes with very few samples were excluded, and data augmentation techniques—such as noise injection and paraphrasing—were used to enrich underrepresented categories. The majority of classes now have a uniform number of examples, supporting a fair and consistent evaluation process.

IV.8 Arabic Models

In this phase, the focus was on evaluating the performance of biomedical transformer models trained on the Arabic medical dataset. Due to the relatively limited availability of Arabic biomedical models compared to English ones, selecting effective models was crucial for achieving high performance.

IV.8.1 Selection of Arabic Biomedical Transformer Models

The following models were selected based on their suitability for Arabic text processing and biomedical applications:

- **BERT Base Arabic:** A pre-trained model for general Arabic text, adapted to biomedical tasks.
- **MARBERT:** Specifically designed for modern Arabic, handling dialectal variations.
- **T5-Small:** A transformer model adapted for text-to-text tasks, fine-tuned for biomedical applications.
- **GigaBERT:** A multilingual BERT model trained to support Arabic and other languages.

IV.8.2 Fine-tuning the Models

The Arabic medical dataset was split into training, validation, and test sets. Fine-tuning was performed to adapt the pre-trained models to medical domain tasks. The following configurations were applied:

- Learning rate and batch size were optimized for each model.
- Early stopping was employed to prevent overfitting.
- Regularization techniques were used to enhance generalization.

IV.8.3 Evaluation

The models were evaluated on the Arabic medical dataset using standard metrics:

- **Accuracy:** The proportion of correct predictions over the total predictions.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positives.

- **Recall:** The ratio of correctly predicted positive observations to all actual positives.
- **F1-Score:** The harmonic mean of precision and recall, balancing both metrics.
- **MCC (Matthews Correlation Coefficient):** A comprehensive measure that takes into account true and false positives and negatives.

IV.8.4 Analysis and Findings

Best Performing Model: Among the Arabic biomedical models, **GigaBERT** demonstrated the highest performance, achieving:

- **Accuracy:** 96.0%
- **Precision:** 96.7%
- **Recall:** 96.2%
- **F1-Score:** 95.7%
- **MCC:** 96.2%

Other Competitive Models:

- **BERT Base Arabic:** Demonstrated a high accuracy of 95.6%, making it suitable for general Arabic text but slightly less effective in the biomedical context.
- **MARBERT:** Achieved similar performance to BERT Base Arabic with an accuracy of 95.5%, highlighting its effectiveness for dialectal variations.
- **T5-Small:** Performed the lowest among the chosen models, with an accuracy of 88.2%, indicating that text-to-text models might be less effective in classification tasks.

Comparative Analysis: The results indicated that GigaBERT significantly outperformed other Arabic models due to its multilingual training and ability to adapt to complex biomedical terminology. However, the lack of extensive Arabic biomedical corpora limited the overall performance compared to English models.

Challenges and Limitations: Due to limited pre-training on Arabic biomedical texts, the models showed moderate accuracy compared to their English counterparts. Additionally, dialectal variations posed challenges, particularly for models not specifically designed for modern or colloquial Arabic.

Model	Accuracy	Precision	Recall	F1-Score	MCC
BERT Base Arabic	95.6%	96.2%	95.7%	95.2%	96.2%
MARBERT	95.5%	95.5%	95.5%	95.5%	95.5%
T5-Small	88.2%	89.5%	88.2%	87.5%	88.2%
GigaBERT (Best)	**96.0%**	**96.7%**	**96.2%**	**95.7%**	**96.2%**

Table IV.8.1: Arabic Biomedical Models Performance

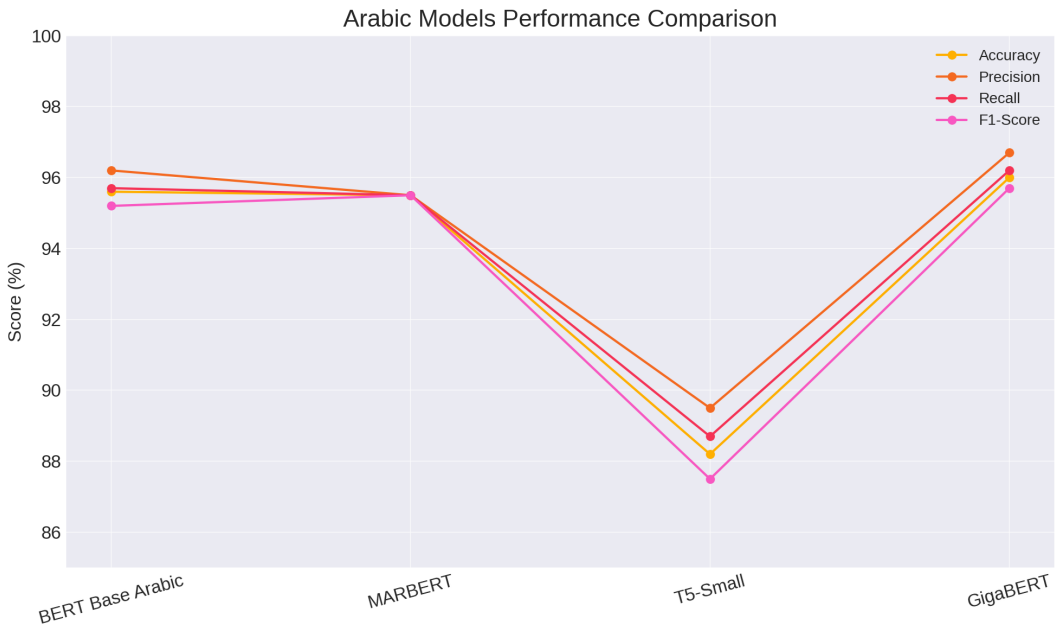


Figure IV.8.1: Arabic Models Performance Comparison across Accuracy, Precision, Recall, and F1-Score. GigaBERT shows the best overall performance.

IV.9 English Biomedical Models

In this phase, the Arabic medical dataset was translated into English to leverage the vast resources and pre-trained biomedical transformer models available for the English language. The primary goal was to evaluate the performance of these models on the translated dataset and compare it to the Arabic models.

IV.9.1 Translation of Arabic Dataset

The Arabic medical text data was translated into English using high-quality machine translation tools such as Google Translate API or MarianMT. To ensure the accuracy of the translation, a manual quality check was performed on a sample of the translated data, focusing on preserving context and medical terminology.

IV.9.2 Selection of Biomedical Transformer Models

The following pre-trained biomedical models were selected for evaluation:

- **BioBERT:** Specifically trained on biomedical literature from PubMed.
- **BioClinicalBERT:** Fine-tuned for clinical texts and electronic health records.
- **BlueBERT PubMed:** Trained on PubMed abstracts and clinical notes.
- **GigaBERT:** A multilingual BERT model, known for competitive performance in biomedical tasks.

IV.9.3 Fine-tuning the Models

The translated dataset was divided into training, validation, and test sets. Each model was fine-tuned using the training data while employing transfer learning to adapt the models to the medical context. Hyperparameters, including learning rate, batch size, and the number of epochs, were optimized to achieve the best results.

IV.9.4 Evaluation

The models were evaluated on the translated English dataset using standard metrics:

- **Accuracy:** The proportion of correct predictions over the total predictions.
- **Precision:** The fraction of relevant instances among the retrieved instances.
- **Recall:** The fraction of relevant instances that were successfully retrieved.
- **F1-Score:** The harmonic mean of precision and recall, useful for handling imbalanced data.
- **MCC (Matthews Correlation Coefficient):** Evaluates binary classification performance, especially on imbalanced data.

IV.9.5 Analysis and Findings

Best Performing Model: Among the evaluated models, **GigaBERT** demonstrated the highest performance, achieving:

- **Accuracy:** 97.55%
- **Precision:** 97.72%
- **Recall:** 97.55%
- **F1-Score:** 97.28%
- **MCC:** 97.53%

Other Competitive Models:

- **BlueBERT PubMed** also achieved high performance with an accuracy of 96.7% and an F1-Score of 97.5%.
- **BioClinicalBERT** and **BioBERT** showed slightly lower performance compared to GigaBERT, but were still effective due to domain-specific training.

Comparative Analysis: The English biomedical models significantly outperformed their Arabic counterparts, primarily due to the availability of extensive biomedical corpora and robust pre-training. GigaBERT's ability to handle multilingual data made it particularly effective when applied to the translated dataset.

Challenges and Limitations: Despite achieving high accuracy, some medical terms and nuances were lost during the translation process, slightly affecting precision. Fine-tuning the models on the translated text significantly improved their performance, highlighting the importance of domain adaptation.

Model	Accuracy	Precision	Recall	F1-Score	MCC
BioBERT	96.2%	96.0%	96.2%	95.7%	96.2%
BioClinicalBERT	96.3%	96.3%	96.3%	96.0%	96.3%
BlueBERT PubMed	96.7%	97.2%	96.7%	97.5%	96.7%
GigaBERT (Best)	**97.55%**	**97.72%**	**97.55%**	**97.28%**	**97.53%**

Table IV.9.1: English Biomedical Models Performance

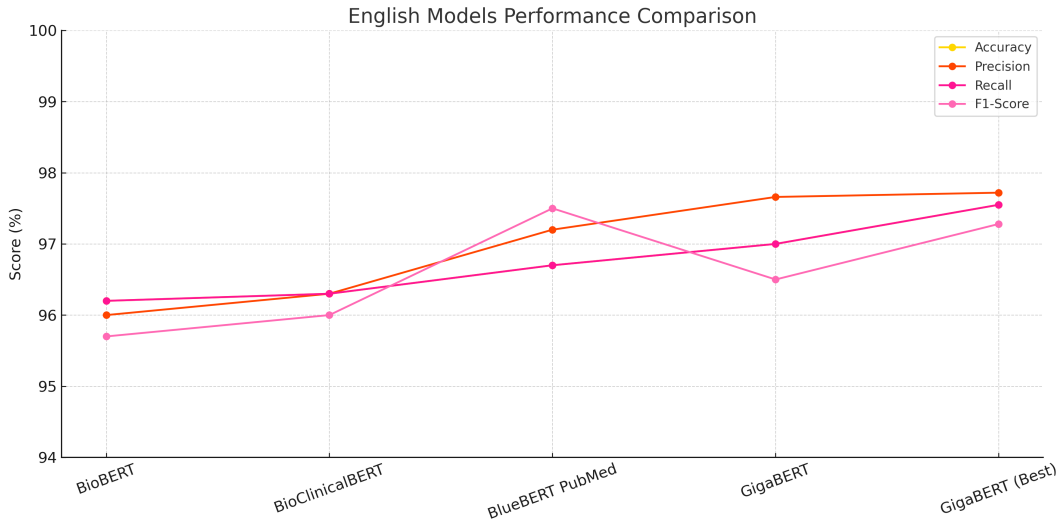


Figure IV.9.1: English Models Performance Comparison across Accuracy, Precision, Recall, and F1-Score. GigaBERT (Best) shows superior results across all metrics.

IV.10 Discussion

Data balance played a crucial role in minimizing bias and enhancing the model's ability to generalize across different inputs. The application of data augmentation techniques further contributed to improved model robustness and recall, allowing the models to better handle variations in the data. When comparing different models, those specifically designed for the domain, such as MARBERT, BioBERT, and GigaBERT, consistently outperformed more general language models. Among these, GigaBERT demonstrated the best overall performance, excelling in both Arabic and English language contexts, highlighting its versatility and effectiveness in multilingual settings.

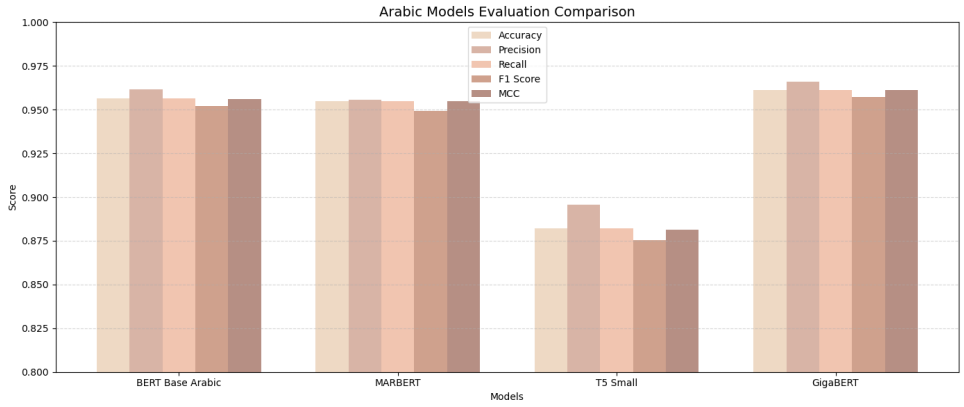


Figure IV.10.1: Performance comparison of Arabic language models.

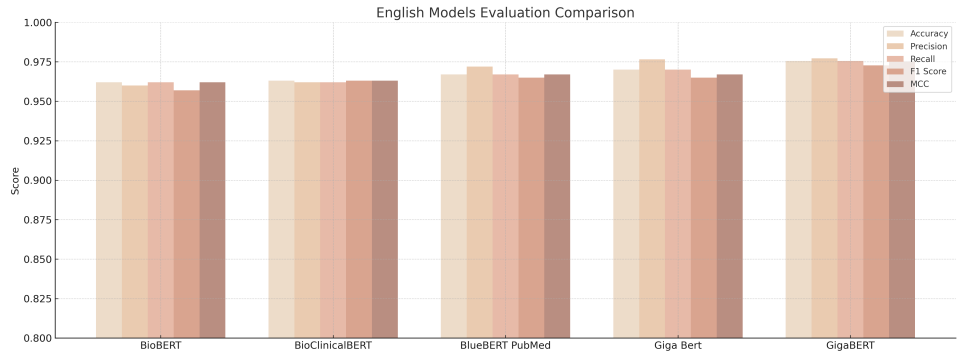


Figure IV.10.2: Performance comparison of English language models.

Aspect	**Description**
Data Balance	Crucial for minimizing bias and boosting generalization.
Augmentation	Improved model robustness and recall.
Model Selection	Domain-specific models (MARBERT, BioBERT, GigaBERT) outperformed general ones.
Best Performer	GigaBERT excelled in both Arabic and English contexts.

Table IV.10.1: Summary of key factors influencing model performance.

IV.11 Frontend Design and Interaction Flow

The frontend of the system is architected to be user-centric, multilingual, and fully responsive, providing a seamless interaction experience across web and mobile platforms. Built using modern web technologies, the interface is designed to collect essential user inputs—including symptoms, age, and

gender—and to deliver predictive feedback in real time through a conversational layout. It serves as the primary access point for users, connecting them with the underlying AI-powered diagnostic engine.

These design decisions were driven by the need to make healthcare guidance more accessible to diverse populations, particularly those with limited access to traditional medical resources.

A core feature of the frontend is its support for multilingual interaction. To accommodate users from diverse linguistic backgrounds, the interface supports Arabic, English, and French. Upon launching the application, users are greeted by a branded welcome screen (see Figure IV.11.1), which introduces the system under the name **MediLingo**—your trusted partner for fast, reliable, and personalized medical support. The screen includes a prominent **Get Started** button to encourage user engagement and guide them into the chatbot workflow.

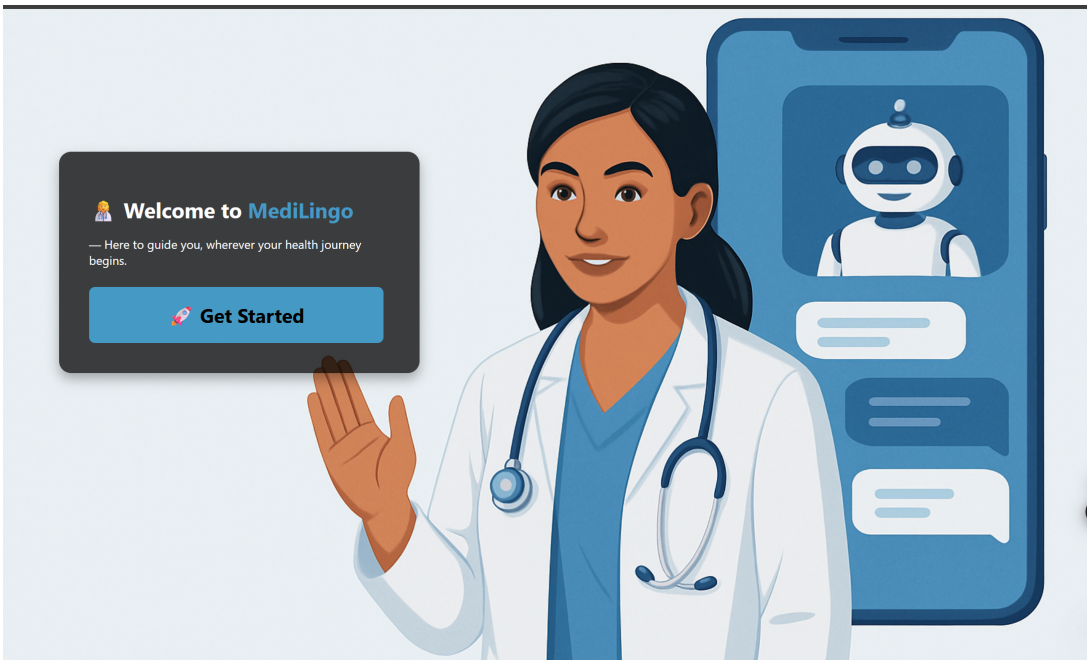


Figure IV.11.1: Welcome screen of the MediLingo chatbot, introducing the system and prompting users to begin interaction.

Following this initial prompt, users are directed to a language selection screen (see Figure IV.11.2), where they can choose between Arabic, English, and French. This choice dynamically configures the interface to the user's preferred language. In Arabic mode, the chatbot layout and text flow are fully adapted for right-to-left (RTL) rendering, while English and French modes follow a traditional left-to-right (LTR) structure. This ensures accessibility and clarity across all languages, improving user comprehension and comfort.

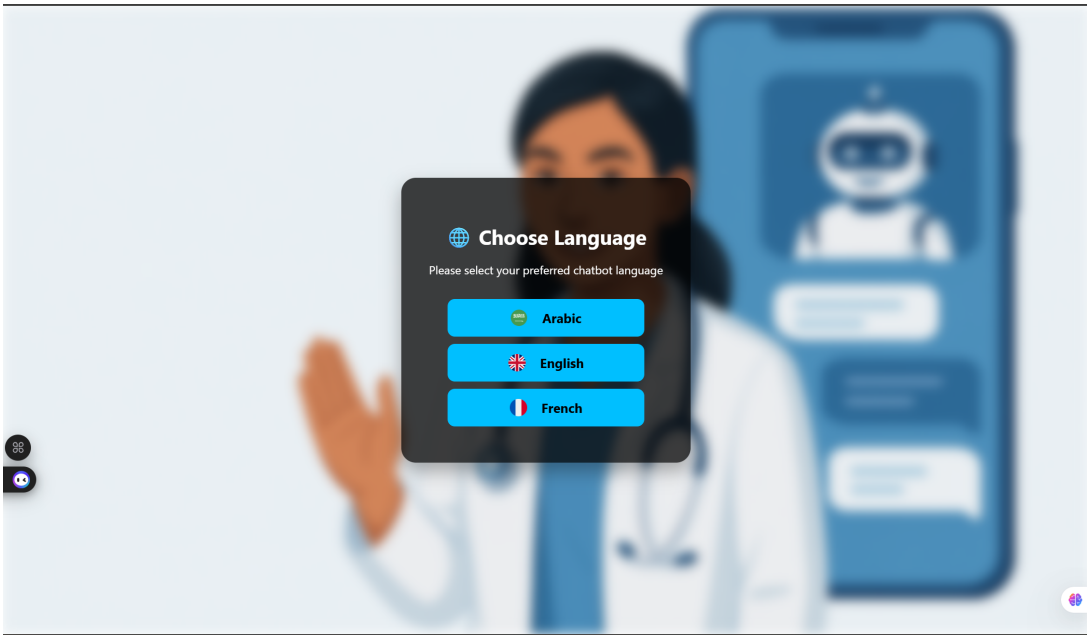


Figure IV.11.2: Language selection screen offering users a choice between Arabic, English, and French interface modes.

The system maintains three parallel chatbot flows—Arabic, English, and French—each capable of interpreting free-text symptom input and returning disease predictions in real time. All chatbots are visually consistent but linguistically tailored. They incorporate localized prompts, placeholder examples, and medical terminology appropriate for their respective language and cultural context.

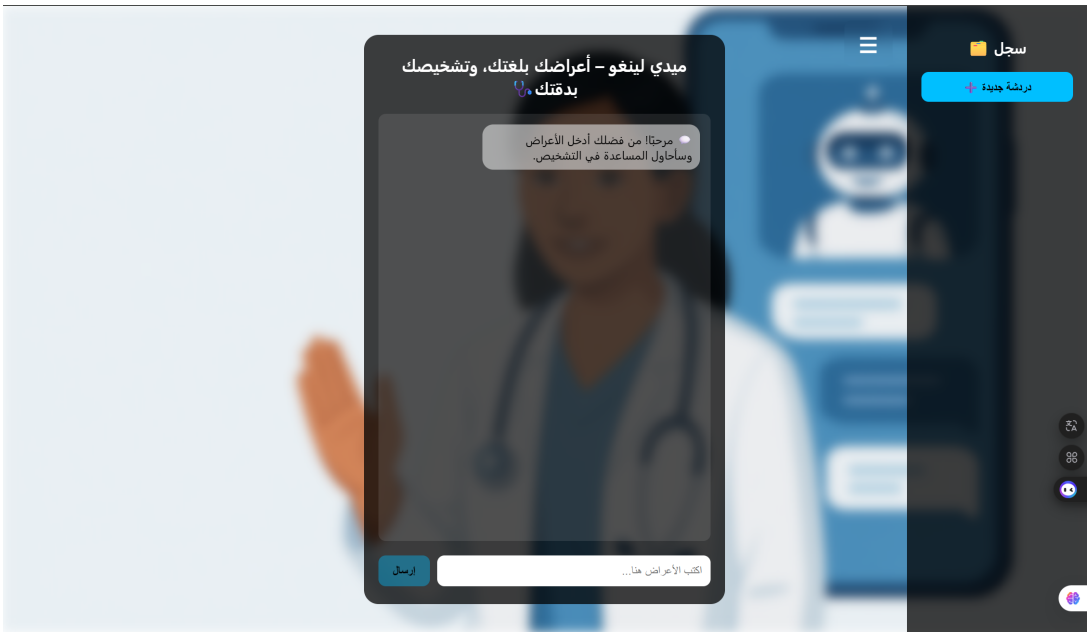


Figure IV.11.3: Arabic chatbot interface with RTL symptom input and diagnosis output.

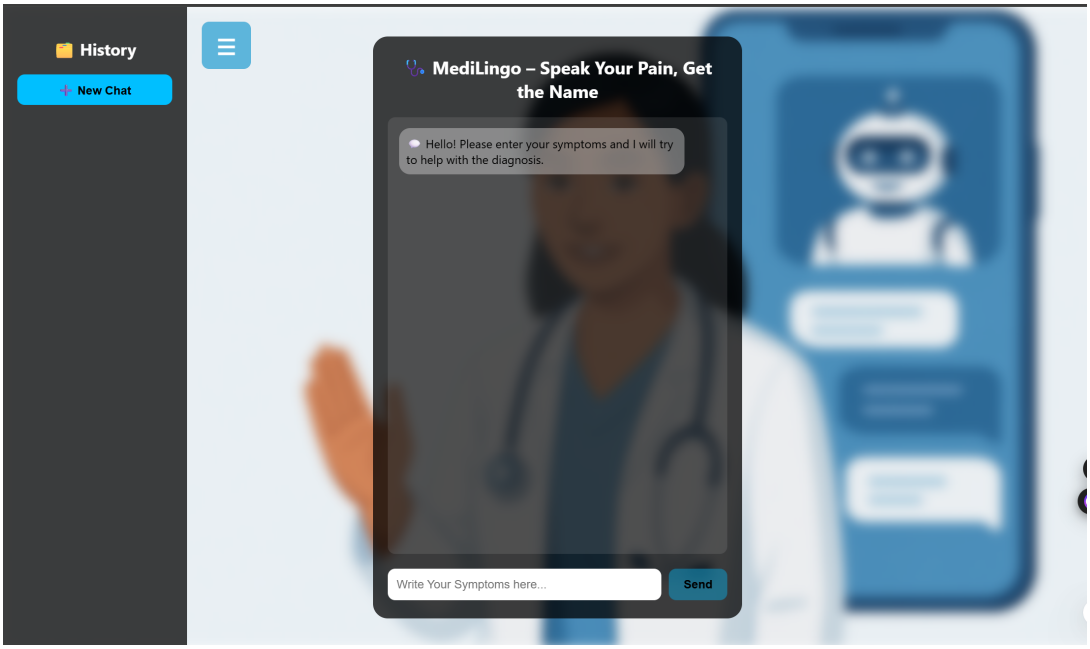


Figure IV.11.4: English chatbot interface with LTR input and response flow.



Figure IV.11.5: French chatbot interface with LTR input and localized medical interaction.

The user interaction process within the MediLingo chatbot interface is structured to support intuitive and efficient diagnostic input. Upon accessing the application, users are guided through an onboarding sequence comprising a welcome screen, language selection, and symptom input stages. The interface captures user data—including symptoms, age, and gender—through structured input fields, ensuring consistency and clarity in the information passed to the backend.

These inputs are submitted via HTTP POST requests to language-specific routes implemented in a Flask web framework. The backend utilizes a fine-tuned GigaBERT model for sequence classification, which analyzes the input and returns the top disease predictions along with associated confidence scores.

This diagnostic response is formatted in HTML and displayed through a chatbot-style interface that mimics conversational interaction. While the system does not employ a traditional RESTful API using JSON exchange, it achieves equivalent functionality by processing structured inputs and dynamically rendering structured diagnostic outputs.

To enhance trust and user engagement, the frontend incorporates real-time visual feedback mechanisms such as loading animations during inference and prompt-based fallbacks in cases of ambiguous input. As detailed in Table IV.11.1, the interface is fully responsive across devices, supports Arabic (RTL), English (LTR), and French (LTR), and ensures a seamless multilingual and mobile-friendly user experience. These features collectively contribute to a medically informative and accessible diagnostic tool.

Feature	**Description**
Device Compatibility	Fully responsive design for desktop, tablet, and mobile platforms
Multilingual Support	Arabic (RTL), English (LTR), and French (LTR) with dynamic switching
Input Handling	Captures symptoms, age, and gender in structured form
Real-Time Feedback	Disease predictions rendered instantly via chatbot-style responses
Guided Flow	Onboarding sequence: Welcome $\rightarrow$ Language Selection $\rightarrow$ Chat Interaction

Table IV.11.1: Summary of core frontend design features in the MediLingo chatbot interface.

IV.12 User Interface Demonstration: English, Arabic, and French Chatbot Interactions

To support the technical evaluation with practical visualization, this section presents representative screenshots of the MediLingo chatbot interacting with users in English, Arabic, and French. These examples demonstrate the chatbot's multilingual interface, natural language understanding, and its ability to provide medically relevant guidance across three distinct languages.

The English interface (Figure IV.12.1) shows a user entering symptoms such as "fever and headache," along with the chatbot's response, which includes the predicted disease name, detailed descriptions, and a URL to a relevant medical website. Similarly, the Arabic interface (Figure IV.12.2) highlights the system's right-to-left (RTL) layout, where the chatbot accurately processes Arabic input and returns context-sensitive medical recommendations, including the predicted diagnosis, explanatory information, and a reference link. The French interface (Figure IV.12.3) demonstrates the chatbot's ability to interpret symptom descriptions in French, and respond with an appropriate medical diagnosis and supporting resources in a linguistically localized manner.

These visual examples illustrate not only the effectiveness of the NLP models in handling multilin-

gual input but also the careful attention to user experience, accessibility, and linguistic inclusivity in the chatbot's design.



Figure IV.12.1: MediLingo chatbot interaction in English: user input, predicted disease name, descriptions, and reference URL.

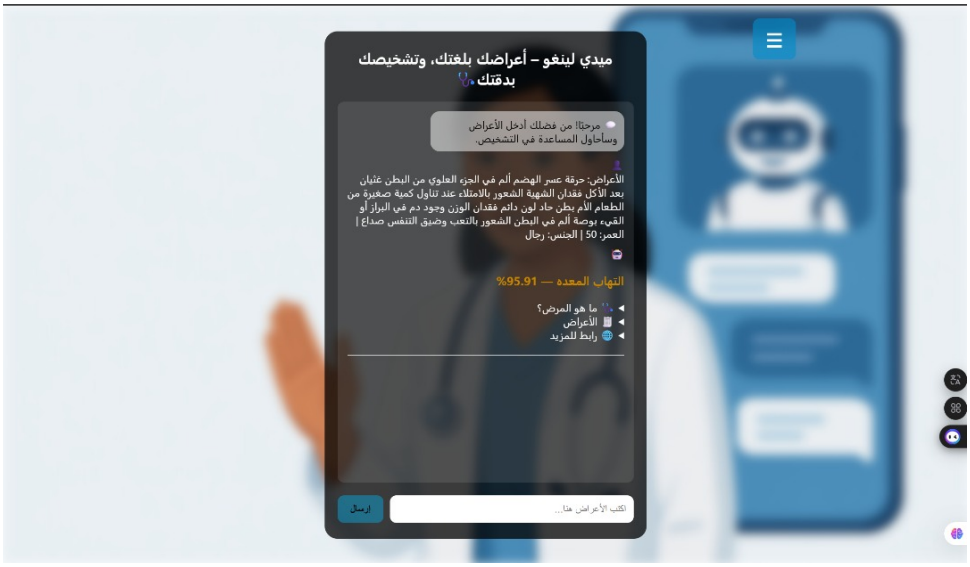


Figure IV.12.2: MediLingo chatbot interaction in Arabic: RTL layout with user input, predicted diagnosis, descriptions, and reference URL.



Figure IV.12.3: MediLingo chatbot interaction in French: LTR layout with user input, predicted diagnosis, explanations, and medical reference link.

IV.13 Conclusion

The findings of this study underscore the pivotal role of balanced datasets in ensuring the reliability and generalizability of medical text classification models. Addressing class imbalance is essential to mitigating bias and enhancing the robustness of predictive performance. Additionally, the integration of data augmentation techniques was shown to significantly improve model efficacy—particularly in terms of recall—by introducing greater variability and representativeness within the training data.

The comparative evaluation of model architectures revealed that domain-specific biomedical and language-adapted models consistently outperform general-purpose language models. This highlights the importance of domain adaptation in natural language processing tasks within the healthcare context. Among the evaluated models, **GigaBERT** emerged as the top performer, achieving the highest metrics across both Arabic and English datasets. These results suggest that GigaBERT's architecture and pre-training strategy offer a distinct advantage in processing multilingual biomedical texts, positioning it as a strong candidate for future cross-lingual applications in medical informatics.

General Conclusion

This thesis presented the design, implementation, and evaluation of a multilingual medical chatbot supporting Arabic, French, and English languages. The core objective was to facilitate accessible and reliable medical information retrieval across linguistic boundaries, addressing a significant gap in healthcare communication, especially in multilingual or under-resourced contexts.

To ensure the relevance and accuracy of the chatbot's medical responses, the dataset was compiled from five reputable medical websites. However, the initial data suffered from severe class imbalance, particularly in the Arabic subset, which posed major challenges during training. To mitigate this, we employed targeted data augmentation strategies—namely synonym replacement, back-translation, and paraphrasing—customized for each language. These methods significantly enriched the dataset and helped balance the distribution across categories, resulting in more robust model performance.

In the natural language understanding phase, several advanced transformer-based models were trained and evaluated for both English and Arabic. For English, BioBERT achieved strong results, with an accuracy of 96.2%, precision of 96.0%, recall of 96.2%, and an F1-score of 95.7%. BioClinicalBERT yielded slightly better outcomes, maintaining 96.3% across accuracy, precision, and recall, and scoring 96.0% on the F1 metric. BlueBERT PubMed showed further improvements, recording an accuracy of 96.7%, precision of 97.2%, recall of 96.7%, F1-score of 97.5%, and a Matthews Correlation Coefficient (MCC) of 96.7%. The best-performing model in the English set was GigaBERT, which excelled with an accuracy of 97.55%, precision of 97.72%, recall of 97.55%, F1-score of 97.28%, and an MCC of 97.53%, making it the most robust and reliable model for English biomedical text understanding in this study.

For Arabic, BERT Base Arabic achieved a high performance with an accuracy of 95.6%, precision of 96.2%, recall of 95.7%, F1-score of 95.2%, and MCC of 96.2%. MARBERT followed closely, maintaining balanced metrics with 95.5% across accuracy, precision, recall, and F1-score. The T5-Small model, though efficient in size, demonstrated weaker performance, achieving only 88.2% accuracy, 89.5% precision, 88.2% recall, and 87.5% F1-score. Notably, GigaBERT once again emerged as the top-performing model for Arabic, achieving 96.0% accuracy, 96.7% precision, 96.2% recall, 95.7% F1-score, and 96.2% MCC. These results clearly highlight GigaBERT's adaptability and strength in both monolingual and multilingual tasks.

As for the French language, due to the unavailability of domain-specific biomedical models, we adopted a translation-based approach by translating the outputs of the best English model—GigaBERT—into French. This strategy allowed us to maintain linguistic consistency and deliver reliable medical responses, despite the absence of native French medical datasets or models. The quality of translation, supported by the strong performance of the source model, ensured effectiveness in handling French queries.

Across all languages, models trained on the augmented and balanced datasets consistently outperformed those trained on the original imbalanced data. This was particularly evident in Arabic, where data augmentation significantly improved generalization and classification accuracy. The results confirm the importance of dataset quality and class balance in achieving high performance, especially in low-resource settings.

In conclusion, this thesis demonstrates that multilingual medical chatbots can be effectively built

using advanced NLP models, provided that careful attention is given to dataset preparation and model selection. GigaBERT, in particular, proved to be the most suitable model for multilingual healthcare tasks, delivering top-tier performance in both Arabic and English, and supporting French through translation. The integration of data augmentation techniques was vital to overcoming challenges posed by data scarcity and imbalance, especially for Arabic.

Looking forward, future improvements could involve expanding the dataset with more diverse and conversational medical data, incorporating user feedback loops for continuous learning, and exploring models designed specifically for code-switching and multilingual dialogue. Additionally, coupling the chatbot with medical ontologies and improving its dialogue flow would allow for more accurate and context-aware interactions. This work thus lays a solid foundation for practical and inclusive multilingual healthcare AI systems, capable of improving medical accessibility and patient empowerment in diverse linguistic environments.

References

[1] Ada health — ai-powered health assessment and symptom checker. https://ada.com. Accessed: 2025-06-09.

[2] Buoy health — ai symptom checker and care navigator. https://www.buoyhealth.com. Accessed: 2025-06-09.

[3] Your.md — digital health assistant and symptom checker. https://www.your.md. Accessed: 2025-06-09.

[4] Michael D. Abràmoff, Paul T. Lavin, Peter Birch, Sunil Shah, and James C. Folk. Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 1(1):39, 2018.

[5] Iman Abu-Doush, Shadi Bukhari, and Ala Alshaikh. Chatbot technology in healthcare: A systematic review. *International Journal of Medical Informatics*, 148:104387, 2021.

[6] Leila Aissaoui Ferhi, Manel Ben Amar, Fethi Choubani, and Ridha Bouallegue. Enhancing diagnostic accuracy in symptom-based health checkers: a comprehensive machine learning approach. *Frontiers in Artificial Intelligence*, 7:1397388, 2024.

[7] A.A. Albahri, A.A. Zaidan, B.B. Zaidan, K.H. Abdulkareem, N. Al-Qaysi, A.H. Alamoodi, O.S. Albahri, A.M. Aleesa, L.A. Chahl, E. Thabet, and et al. The future of artificial intelligence in healthcare: A comprehensive review. *Journal of Medical Systems*, 48(1):1–32, 2024.

[8] Altibbi. Altibbi - comprehensive medical information, 2023. Accessed: 2025-04-29.

[9] K. B. Author and M. N. Researcher. Evaluating large language models for patient education in heart failure. *Journal of Medical AI Research*, 12(3):145–156, 2023.

[10] Authors. Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. *International Journal of Medical Informatics*, 188:105474, 2024.

[11] A. et al. Basubrin. Current status and future of artificial intelligence in medicine. *Narrative Review (PubMed)*, 2025. PubMed PMID: 39958114.

[12] Eta S. Berner. Clinical decision support systems: theory and practice. *Springer*, 2007.

[13] J. E. Bibault, A. L. Chaix, B. Guillemassé, et al. When chatbots meet patients: One-year prospective study of conversations between patients with breast cancer and a chatbot. *JMIR Cancer*, 2019.

[14] Jean-Emmanuel Bibault, Benjamin Chaix, Alexandre Guillemassé, Sophie Cousin, Alexandre Escande, Marine Perrin, and Lionel Guillemasse. Will artificial intelligence replace physicians? the promise and challenges of ai in healthcare. *Radiotherapy and oncology: journal of the European Society for Therapeutic Radiology and Oncology*, 139:230–236, 2019.

[15] Ekaba Bisong. Google colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pages 59–64. Apress, Berkeley, CA, 2019.

[16] T. Brown et al. Language models are few-shot learners. In *NeurIPS Conference Proceedings*, 2020.

[17] Duncan Chambers, Anna J Cantrell, Maxine Johnson, et al. Digital and online symptom checkers and assessment services for urgent care: systematic review. *BMJ Open*, 9(8):e027743, 2018.

[18] Lina Cheng and Samuel Lee. Evaluating the clinical accuracy of ai chatbots in cardiology. *International Journal of Cardiology AI*, 7:1–10, 2023.

[19] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, 2020.

[20] W. J. Clancey. The epistemology of a rule-based expert system—a framework for explanation. *Artificial Intelligence*, 20:215–251, 1984.

[21] Mayo Clinic. Symptom checker, 2023. Accessed: 2025-04-29.

[22] Mayo Clinic. Symptom checker, 2023.

[23] B. R. Cornelison, B. L. Erstad, and C. Edwards. Accuracy of a chatbot in answering questions that patients should ask before taking a new medication. *Journal of the American Pharmacists Association*, 2024.

[24] Sheila Cruz, Sara Garcia, and Francisco Andrade. Towards a unified monitoring and management system for deployed machine learning models. *Journal of Big Data*, 8(1):1–23, 2021.

[25] Rima A. El Arab, Zainab Almoosa, Faisal H. Abuadas, and Janice Somerville. Artificial intelligence in hospital infection prevention: an integrative review. *Frontiers in Public Health*, 13:1547450, 2025.

[26] Engati. Healthcare chatbots: Transforming the patient experience, 2023. Accessed: 2025-04-29.

[27] Emily Esterwood and Syed Saeed. Past epidemics, natural disasters, covid-19, and mental health: learning from history as we deal with the present and prepare for the future. *Psychiatric Quarterly*, 91:1121–1133, 2020.

[28] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2019.

[29] Andre Esteva, Brett Kuprel, Roberto A. Novoa, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[30] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, et al. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[31] Guy Fagherazzi, Camille Goetzinger, Omar Rashid, Gustavo A Aguayo, and Laure Huiart. Digital health strategies to fight covid-19 worldwide: Challenges, recommendations, and a call for papers. *Journal of medical Internet research*, 22(6):e19284, 2020.

[32] Eva Feigerlova, Hani Hani, and Elinor Hothersall-Davies. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Medical Education*, 25:129, 2025.

[33] D. A. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. M. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. Building watson: An overview of the deepqa project. *AI Magazine*, 31(3):59–79, 2010.

[34] Kathleen K Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19, 2017.

[35] E. Fix and J. L. Hodges. Discriminatory analysis: Nonparametric discrimination, consistency properties. *USAF School of Aviation Medicine*, 1985.

[36] Forbes. Integration of ai in wearable health technology: Apple watch and fitbit, 2023. Accessed: 2025-04-29.

[37] Hamish Fraser, Daven Crossland, Ian Bacher, Megan Ranney, and Tracy Madsen. Comparison of diagnostic and triage accuracy of ada health and webmd symptom checkers, chatgpt, and physicians. *JMIR mHealth and uHealth*, 11:e49995, 2023.

[38] Medical Futurist. Top 10 health chatbots, 2023. Accessed: 2025-04-29.

[39] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. *Advances in Information Retrieval*, pages 345–359, 2005.

[40] Miguel Grinberg. *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, Inc., 2018.

[41] S. Gidengil H. L. Semigran, L. Linder and E. Mehrotra. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*, 351:h3480, 2015.

[42] Charles R Harris et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[43] Jiale He, Stephen L Baxter, Jun Xu, Jin Xu, Xingyu Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25(1):30–36, 2019.

[44] Health professionals' perspectives on the use of Artificial Intelligence in healthcare. Health professionals' perspectives on the use of artificial intelligence in healthcare: A systematic review. *Patient Education and Counseling*, 134:108680, 2025.

[45] D. Heckerman. A tutorial on learning with bayesian networks. *Technical Report MSR-TR-95-06*, 1995.

[46] Rebecca Hoffman and Michael Smith. Patient-centered questions in heart failure: A review. *Heart Failure Reviews*, 25(2):223–231, 2020.

[47] Liang Huang and Mei Chen. Chatbots in healthcare: Current applications and future perspectives. *Journal of Healthcare Informatics*, 9:55–68, 2023.

[48] Inbenta. 9 use cases of healthcare chatbots, 2023. Accessed: 2025-04-29.

[49] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Y. Dong, H. Shen, and Y. Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4):230–243, 2017.

[50] Fei Jiang et al. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4):230–243, 2017.

[51] Feng Jiang, Peng Liang, and Qing Zhang. Artificial intelligence in patient education: Opportunities and challenges. *Healthcare Informatics Research*, 27(4):233–243, 2021.

[52] JMIR. Chatbots in healthcare: Their impact on patient access. *Journal of Medical Internet Research*, 2024.

[53] Emily Johnson, Mark Davis, and Sarah Lee. Evaluating the accuracy of ai chatbots in medication counseling: A cross-sectional study. *Journal of Medical Informatics*, 45(2):112–123, 2024.

[54] Journal of Medical Internet Research. Research on wearable devices and iot integration in healthcare. *JMIR*, 2023.

[55] Daniel S. Kermany et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018.

[56] Faiza Khan Khattak, Vallijah Subasri, Amrit Krishnan, Elham Dolatabadi, Deval Pandya, Laleh Seyyed-Kalantari, and Frank Rudzicz. Mlhops: Machine learning for healthcare operations. *arXiv*, abs/2305.02474, 2023. Preprint, not peer-reviewed.

[57] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001.

[58] E. Kozaily, M. Geagea, E. R. Akdogan, et al. Accuracy and consistency of online large language model-based artificial intelligence chat platforms in answering patients' questions about heart failure. *International Journal of Cardiology*, 2024.

[59] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013.

[60] Rajesh Kumar, Aisha Rahman, and Jun Li. Transformer-based multimodal fusion for tuberculosis diagnosis using chest x-rays and clinical records. *IEEE Journal of Biomedical and Health Informatics*, 29(3):456–468, 2025.

[61] S. Kumar, S. Sharma, and K. T. Megra. Transformer-enabled multi-modal medical diagnosis for tuberculosis classification. *Journal of Big Data*, 2025.

[62] J. Lee, W. Yoon, S. Kim, et al. Biobert: A pre-trained biomedical language representation model for nlp. *Bioinformatics Journal*, 2020.

[63] Xiao Li and Rui Zhang. Limitations and risks of large language models in clinical settings. *AI in Medicine*, 15:89–99, 2023.

[64] Mayo Clinic. Symptom checker. https://www.mayoclinic.org/symptoms, 2000.

[65] Wes McKinney. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14:1–9, 2011.

[66] Randolph A. Miller, Harry E. Pople Jr, and James D. Myers. Internist-1: An experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, 307(8):468–476, 1982.

[67] M. Min et al. Enhancing symptom recognition in medical chatbots using nlp. *AI in Healthcare Journal*, 2021.

[68] Adam S Miner, Adam Milstein, Stephen Schueller, Sameer Hegde, Christina Mangurian, and Eleni Linos. Chatbots in mental health: a review of conversational agents for psychological therapy. *NPJ Digital Medicine*, 3(1):1–9, 2020.

[69] Regina Müller, Malte Klemmt, Roland Koch, Tanja Henking, Hans-Jörg Ehni, and Robert Ranisch. "that's just future medicine" — users' experiences of symptom checker apps. *BMC Medical Ethics*, 25:17, 2024.

[70] NCBI. *Artificial Intelligence in Healthcare: Chatbot Applications*. National Center for Biotechnology Information, 2022.

[71] World Health Organization. International statistical classification of diseases and related health problems (10th revision), 2019.

[72] World Health Organization. Global strategy on digital health 2020–2025, 2020.

[73] Anthony Pairon and Hilde Philips Verhoeven. A scoping review on the use and usefulness of online symptom checkers and triage systems. *Frontiers in Medicine*, 9:104092, 2023.

[74] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[75] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[76] PMC. The role of medical chatbots in healthcare. *PubMed Central*, 2023.

[77] Piotr Ponikowski, Stefan D. Anker, and Tomasz J. Piepoli. 2016 esc guidelines for the diagnosis and treatment of acute and chronic heart failure. *European Heart Journal*, 37:2129–2200, 2016.

[78] D. M. W. Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[79] David M. W. Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[80] ProviderTech. Chatbots vs. conversational ai in healthcare, 2023. Accessed: 2025-04-29.

[81] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[82] A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 2019.

[83] A.N. Ramesh, C. Kambhampati, J.R.T. Monson, and P.J. Drew. Artificial intelligence in medicine. *Annals of the Royal College of Surgeons of England*, 86(5):334–338, 2004.

[84] M. I. Razzak et al. Deep learning for healthcare: An overview. *Artificial Intelligence in Medicine Journal*, 2022.

[85] Eva Riboli□Sasco, Austen El□Osta, Aos Alaa, Iman Webber, Manisha Karki, Marie Line El Asmar, Katie Purohit, Annabelle Painter, and Benedict Hayhoe. Triage and diagnostic accuracy of online symptom checkers: Systematic review. *Journal of Medical Internet Research*, 25:e43803, 2023.

[86] Stefanie Rieger, Stefanie Joos, and Roland Koch. Only the anxious ones? identifying characteristics of symptom checker app users. *BMC Medical Informatics and Decision Making*, 24:21, 2024.

[87] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015.

[88] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[89] Subhajit Santu, Sushree Mishra, Himanshu Behera, and P.K. Nayak. A survey on deployment of machine learning models in real-world: Challenges and approaches. *International Journal of Information Management Data Insights*, 2(2):100083, 2022.

[90] Hannah L Semigran, Jeffrey A Linder, Courtney Gidengil, and Ateev Mehrotra. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*, 351:h3480, 2015.

[91] Hassan Sharif, Firas Hameed, and Sara Khan. Telemedicine and ai to improve healthcare access in rural settings. *ResearchGate Preprint*, 2024. Preprint, not peer-reviewed.

[92] E. H. Shortliffe. Computer-based medical consultations: Mycin. *Elsevier*, 1976.

[93] Edward H. Shortliffe. *A Model of In-Depth Reasoning in Medical Consultation: The MYCIN Example*. Stanford University, 1975.

[94] John Smith and Angela Lee. Transformer models for medical question answering: A comprehensive review. *Journal of Medical Informatics*, 58:102–115, 2023.

[95] Rachel Smith and John Miller. Towards transparent and auditable ai in clinical decision support. *Artificial Intelligence in Medicine*, 135:102692, 2023.

[96] Sohati. List of diseases and medical conditions, 2023. Accessed: 2025-04-29.

[97] Coherent Solutions. How ai chatbots advance healthcare for patients and providers, 2023. Accessed: 2025-04-29.

[98] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang. Snomed clinical terms: overview of the development process and project status, 2001.

[99] The Lancet. Ai in healthcare: The evolution of self-diagnosis tools. *The Lancet*, 2019.

[100] E. Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019.

[101] E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25:44–56, 2019.

[102] Eric Topol. Deep medicine: how artificial intelligence can make healthcare human again. *Basic Books*, 2019.

[103] Eric Topol. The convergence of human and artificial intelligence in healthcare. *Nature Medicine*, 26:66–68, 2020.

[104] Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950.

[105] Waqas Ullah and Qaiser Ali. Role of artificial intelligence in healthcare settings: a systematic review. *Journal of Medical Artificial Intelligence*, 8:24, 2025.

[106] UNESCO. The importance of mother tongue in digital health services, 2021.

[107] Guido Van Rossum et al. *Python for Data Analysis*. O'Reilly Media, 2011.

[108] L. Wang, Y. Yin, B. Glampson, et al. Transformer-based deep learning model for the diagnosis of suspected lung cancer in primary care based on electronic health record data. *medRxiv*, 2024.

[109] Lan Wang, Yue Yin, Brendan C. Glampson, Robert L. Peach, M. Barahona, B. C. Delaney, and E. Mayer. Transformer-based deep learning model for the diagnosis of suspected lung cancer in primary care based on electronic health record data. *medRxiv*, 2024. Preprint, not peer-reviewed.

[110] Lan Wang, Yue Yin, Brendan C. Glampson, Robert L. Peach, Mauricio Barahona, Brendan C. Delaney, and Erik Mayer. Transformer-based deep learning model for the diagnosis of suspected lung cancer in primary care based on electronic health record data. *medRxiv*, Preprint:1–19, 2024. Preprint, not peer-reviewed.

[111] WebMD. Symptom checker. https://symptoms.webmd.com/, 1996.

[112] WebMD. Webmd: A symptom checker and health resource, 1996. Accessed: 2025-04-29.

[113] WebMD. Symptom checker, 2023.

[114] WebTeb. Medical conditions and diseases, 2023. Accessed: 2025-04-29.

[115] Thomas Wolf, Lysandre Debut, Victor Sanh, et al. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 EMNLP: System Demonstrations*, pages 38–45, 2020.

[116] V. Yadav and S. Bethard. A survey on recent advances in named entity recognition from deep learning models. *Journal of Biomedical Informatics*, 2019.

[117] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018.

[118] Ying Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1–4):43–52, 2010.

[119] Wei Zhou, Lina Gomez, and Ahmed El-Sayed. Multimodal transformer fusion networks for medical diagnosis. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 2245–2252. AAAI Press, 2024.