

Faculty of Engineering & Technology

Electrical & Computer Engineering Department Circuits &

GRADUATION PROJECT ENCS5300-Section 38

Arabic Medical Semantic Qualifiers

Prepared by:

Katya Kobari 1201478

Rana Odeh 1201750

Hidaya Mustafa 1201910

Supervised By: Adnan Yahya

A graduation project submitted to the Department of Electrical and Computer Engineering in  
partial fulfillment of the requirements for the degree of B.Sc. in Computer Engineering

Birzeit

July, 2025

## Abstract

The integration of Artificial Intelligence (AI) and Natural Language Processing (NLP) in healthcare presents promising opportunities to bridge the communication gap between doctors and patients in Arabic-speaking regions. This project addresses both linguistic and diagnostic challenges by turning the absence of standardized semantic qualifiers in Arabic into an opportunity to build new structured medical resources. These qualifiers are essential for accurately interpreting symptoms and supporting effective medical decision-making.

The project focuses on developing an intelligent system capable of extracting patient symptoms from Arabic doctor–patient dialogues and mapping them to structured semantic qualifiers. To ensure both clinical relevance and linguistic diversity, the system is trained on a hybrid dataset that combines real medical consultations with AI-generated dialogues in Arabic.

The extracted symptoms and qualifiers are used as input for a secondary diagnostic model that predicts the most likely cardiac condition. The system also produces a structured, bilingual repository of symptoms and qualifiers, enabling precise, machine-readable outputs. By developing and evaluating this prototype in the context of heart disease, the project aims to assist healthcare professionals with AI-powered insights, improve diagnostic accuracy, and enhance the overall quality of care in Arabic clinical environments.

## المستخلص

يُعدّ دمج تقنيات الذكاء الاصطناعي ومعالجة اللغة الطبيعية في مجال الرعاية الصحية فرصةً واعدة لسدّ الفجوة في التواصل بين الأطباء والمرضى في البيئات الناطقة بالعربية. يتناول هذا المشروع التحديات اللغوية والتشخيصية من خلال تحويل غياب المؤهّلات الدلالية الموحدة في اللغة العربية إلى فرصة لبناء موارد طيبة منظمة وحديثة. وتعتبر هذه المؤهّلات ضرورية لفهم الأعراض بدقة ودعم عملية اتخاذ القرار الطبي بشكل فعال.

يركّز المشروع على تطوير نظام ذكي قادر على استخراج أعراض المرضى من حوارات بين الأطباء والمرضى باللغة العربية، وربطها بمؤهّلات دلالية منظمة. ولضمان الموثوقية السريرية والتنوع اللغوي، تم تدريب النظام على مجموعة بيانات هجينّة تجمع بين استشارات طيبة حقيقية وحوارات مُؤلدة باستخدام الذكاء الاصطناعي.

تُستخدم الأعراض والمؤهّلات المستخرجة كمدخلات لنموذج تشخيصي ثانوي يتوقع الحالة القلبية الأكثر احتمالاً. كما يُنتج النظام مستودعاً ثانياً للغة يحتوي على الأعراض والمؤهّلات بشكل منظم وقابل للقراءة الآلية بدقة. ومن خلال تطوير هذا النموذج واختباره في سياق أمراض القلب، يهدف المشروع إلى دعم المتخصصين في الرعاية الصحية بروء قائمة على الذكاء الاصطناعي، وتحسين دقة التشخيص، والارتقاء بجودة الرعاية الصحية في البيئات الطيبة الناطقة بالعربية.

## Table of Contents

<b>Abstract.....</b>	I
<b>المستخلص .....</b>	II
<b>Keywords .....</b>	VIII
<b>Chapter1   Introduction and Motivation .....</b>	10
<b>1.1 Motivation.....</b>	10
<b>1.2 Problem Statement.....</b>	10
<b>1.3 Methodology .....</b>	11
<b>1.4 Contribution .....</b>	12
<b>Chapter2   Background .....</b>	13
<b>2.1 Specialized Diseases .....</b>	13
<b>2.2 Artificial Intelligence .....</b>	13
<b>2.2.1 Machine Learning.....</b>	14
<b>2.2.2 Applications of Machine Learning in Medicine.....</b>	14
<b>2.2.3 Transformer-Based Models (AraBERT) .....</b>	15
<b>2.2.4 Sequence Modeling Using BiLSTM .....</b>	15
<b>2.2.5 Sequence Optimization Using CRF .....</b>	15
<b>2.2.6 Benefits of AI and ML in Medical Research .....</b>	16
<b>2.3 NLP in Medicine .....</b>	16
<b>2.3.1 Importance of NLP in Healthcare .....</b>	17
<b>2.3.2 Challenges in Arabic Medical Text Processing .....</b>	17
<b>2.4 Semantic Qualifiers in Medicine.....</b>	19
<b>2.4.1 Role of Semantic Qualifiers in Clinical Reasoning .....</b>	19
<b>2.4.2 Hypothetical-Deductive Reasoning and the Use of Semantic Qualifiers .....</b>	19

<b>2.4.2 Gaps in Arabic Semantic Qualifiers.....</b>	<b>20</b>
<b>2.5 Existing Diagnostic Systems.....</b>	<b>20</b>
<b>2.5.1 Current AI-Based Diagnostic Tools .....</b>	<b>20</b>
<b>2.5.2 Limitations of Current Systems.....</b>	<b>21</b>
<b>2.5.3 BIO Tagging in Clinical NLP.....</b>	<b>21</b>
<b>2.6 Related Works .....</b>	<b>22</b>
<b>2.6.1 Machine Learning Models for Efficient Automated Disease Diagnosis .....</b>	<b>22</b>
<b>Chapter3   System Implementation.....</b>	<b>23</b>
<b>3.1 Overview of the System Architecture .....</b>	<b>23</b>
<b>3.2 Data Collection .....</b>	<b>23</b>
<b>3.2.1 Real-World Dataset .....</b>	<b>23</b>
<b>3.2.2 AI-Generated Dataset.....</b>	<b>24</b>
<b>3.3 Data Preprocessing .....</b>	<b>25</b>
<b>3.3.1 Text Cleaning .....</b>	<b>26</b>
<b>3.4 Feature Extraction .....</b>	<b>27</b>
<b>3.4.1 Symptoms Extraction .....</b>	<b>27</b>
<b>3.4.2 Demographic Features.....</b>	<b>29</b>
<b>3.5 Bio Tagging.....</b>	<b>29</b>
<b>3.6 Symptoms and Qualifiers Extraction Model.....</b>	<b>30</b>
<b>3.6.1 INPUT DATA.....</b>	<b>30</b>
<b>3.6.2 MODEL ARCHATUCTURE .....</b>	<b>31</b>
<b>3.6.3 POST PROCESSING .....</b>	<b>33</b>
<b>3.7 Medical Concept Standardization .....</b>	<b>34</b>
<b>3.8 Diagnosis Classification Model .....</b>	<b>35</b>
<b>3.8.1 Diagnosis Prediction Pipeline.....</b>	<b>35</b>

<b>3.8.2 Final Output and Integration .....</b>	<b>36</b>
<b>3.9 Database Design .....</b>	<b>36</b>
<b>3.10 Web Development .....</b>	<b>37</b>
<b>3.10.1 Web Application Backend.....</b>	<b>37</b>
<b>3.10.2 Frontend User Interface .....</b>	<b>38</b>
<b>3.11 Testing Models and Evaluation .....</b>	<b>39</b>
<b>3.11.1 Symptom and qualifiers extraction .....</b>	<b>39</b>
<b>3.11.2 Diagnosis classification .....</b>	<b>52</b>
<b>3.12 System Flow.....</b>	<b>54</b>
<b>Chapter4   Future Work and Current Limitations and Conclusion .....</b>	<b>55</b>
<b>4.1 Possible Future Work .....</b>	<b>55</b>
<b>4.2 Limitations of the System:.....</b>	<b>56</b>
<b>4.3 Conclusion .....</b>	<b>56</b>
<b>References.....</b>	<b>57</b>

## Table of Figures

Figure 1:Model Archatucture Flow .....	33
Figure 2:SNOMED CT mapping output.....	35
Figure 3:Web application interface.....	39
Figure 4:Training Accuracies per Epoch for AI data in BiLSTM+CRF model .....	40
Figure 5:Training Loss per Epoch for AI data in BiLSTM+CRF model .....	40
Figure 6:Precision, Recall, F1-score for Validation for AI data in BiLSTM + CRF model .....	41
Figure 7:Precision, Recall, F1-score in Test for AI data in BiLSTM + CRF model .....	42
Figure 8:Training Accuracy per Epoch for Real data in BiLSTM+CRF model.....	43
Figure 9:Training Loss per Epoch for Real data in BiLSTM+ CRF model .....	43
Figure 10::Precision, Recall, F1-score for Validation for Real data in BiLSTM + CRF model ..	44
Figure 11:Precision, Recall, F1-score in Test for Real data in BiLSTM + CRF model.....	45
Figure 12:Training Accuracies per Epoch for AI data in AraBERT +BiLSTM+CRF model.....	45
Figure 13:Training Loss per Epoch for AI data in AraBERT +BiLSTM+CRF model .....	46
Figure 14:Precision, Recall, F1-score for Validation for AI data in AraBERT +BiLSTM + CRF model.....	46
Figure 15:Precision, Recall, F1-score for Test for AI data in AraBERT +BiLSTM + CRF model .....	47
Figure 16:Training Accuracy per Epoch for Real data in AraBERT +BiLSTM + CRF model ...	48
Figure 17:Training Loss per Epoch for Real data in AraBERT +BiLSTM + CRF model.....	48
Figure 18::Precision, Recall, F1-score for Validation for Real data in AraBERT +BiLSTM + CRF model .....	49
Figure 19::Precision, Recall, F1-score for Test for Real data in AraBERT + BiLSTM + CRF model.....	49
Figure 20:Diagnosis Model Accuracy Comparison.....	53
Figure 21:Sample Arabic result .....	53
Figure 22::Sample English result.....	54
Figure 23:System flow chart.....	54

## **Table of Tables**

Table 1:Comparison Between extraction models .....	50
Table 2:Comparison Between Real and AI data.....	51

## **Keywords**

1. Natural Language Processor (NLP)
2. Artificial Intelligence (AI)
3. Machine Learning (ML)
4. User Interface (UI)
5. Structured Query Language (SQL)
6. Conditional Random Field (CRF)
7. Bi-directional Long Short-Term Memory (BiLSTM)
8. Logistic Regression (LR)
9. Arabic Bidirectional Encoder Representations from Transformers (AraBERT)
10. Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)



# **Chapter1 | Introduction and Motivation**

## **1.1 Motivation**

Semantic qualifiers are essential linguistic tools that enable clinicians to communicate medical information with clarity and precision, particularly by describing symptoms in terms of intensity, duration, frequency, and progression. In Arabic-speaking medical contexts, the absence of a standardized framework for these qualifiers has traditionally been viewed as a limitation. However, this gap also presents a unique opportunity to innovate by introducing new forms of structured linguistic data that align with Arabic's medical and conversational landscape.

Rather than adapting existing models from other languages, this project takes advantage of the unmet need by building foundational resources specifically for Arabic. This approach not only fills a critical void in medical communication but also supports the development of more culturally and linguistically appropriate diagnostic tools.

The project begins by constructing a comprehensive framework for Arabic semantic qualifiers, systematically organizing expressions related to symptom intensity, duration, frequency, and progression. This framework bridges the gap between informal patient language and formal medical terminology, ensuring clearer understanding in clinical settings.

Building on this foundation, an AI-based diagnostic engine is developed to interpret patient-reported symptoms in real time and map them to the most likely cardiac condition using probabilistic modeling. By integrating semantic qualifiers with advanced natural language processing techniques, the system aims to improve diagnostic precision and enhance decision-making in Arabic healthcare environments.

## **1.2 Problem Statement**

In Arabic-speaking regions, healthcare professionals face significant challenges in accurately diagnosing diseases during doctor–patient interactions due to the absence of a system that standardizes and automates the diagnostic process in Arabic. Medical consultations often involve patients describing their symptoms in spoken Arabic or local dialects, which lack the precision and standardization found in English medical terminology. This creates a gap in effectively linking patient descriptions to possible diagnoses.

For Palestinian medical students and junior doctors, this gap is particularly challenging. While they are trained using English medical textbooks with well-defined semantic qualifiers and diagnostic frameworks, they often struggle to apply this knowledge in real-world scenarios where patients describe their conditions in Arabic. The disconnect between theoretical knowledge and practical application can lead to delays in decision-making, diagnostic uncertainty, and inefficiencies in patient care.

At the same time, this gap has presented a valuable opportunity. The lack of an Arabic-based diagnostic support system has motivated the development of a novel solution that addresses the region's linguistic and clinical needs. Rather than relying on translation or adaptation of existing tools, this project seeks to create an AI-powered platform built natively for Arabic, effectively bridging the current divide. The system is built on a dataset that combines authentic clinical dialogues with AI-generated conversations, ensuring both realism and linguistic diversity.

The proposed system leverages AI and NLP to analyze patient-reported symptoms in Arabic and generate the most likely diagnosis using probabilistic modeling. By introducing a tool that interprets spoken symptom descriptions and maps them to probable cardiac conditions, the solution aims to enhance diagnostic accuracy, support medical professionals, and streamline doctor–patient interactions—ultimately contributing to improved healthcare delivery in Arabic-speaking contexts.

### **1.3 Methodology**

This project focuses on developing a comprehensive system to enhance clinical reasoning in Arabic by analyzing both real doctor–patient conversations and AI-generated dialogues. Authentic interactions collected from medical professionals form the core of the dataset, offering valuable insights into the semantic qualifiers commonly used in actual clinical settings. In addition, ChatGPT was employed to generate realistic Arabic conversations, further expanding the database while ensuring alignment with real-world medical contexts.

Rather than viewing the absence of Arabic diagnostic resources as a limitation, the project embraces it as a key opportunity to build structured linguistic and clinical tools from the ground up. This proactive approach allowed us to create a domain-specific dataset that fills a critical gap in Arabic-language medical informatics.

A major focus of the project is data cleaning and preprocessing, which ensures that the collected information is both accurate and reliable. The system specifically targets heart-related diseases, using this focused medical domain to develop, train, and validate its functionality in a practical and high-impact context.

The curated and processed data are organized into a structured semantic qualifier database, which is then analyzed and modeled using NLP techniques and ML algorithms. The ultimate objective is to design a system capable of identifying, mapping, and translating semantic qualifiers in Arabic, thereby supporting effective diagnosis and improving clinical communication for Arabic-speaking healthcare providers.

This work is conducted under the supervision of a specialized medical expert, ensuring clinical accuracy and alignment with current standards of care. The expert provides critical guidance in selecting relevant semantic qualifiers and validating the system's diagnostic outputs. Their involvement guarantees that the system delivers meaningful, actionable insights that enhance diagnostic precision and support more effective patient care in real clinical environments.

## 1.4 Contribution

This project contributes to the advancement of Arabic-language healthcare by addressing the absence of standardized semantic representations for symptoms and diseases. Rather than treating this absence as a limitation, the project embraces it as an opportunity to create new linguistic and clinical resources that fill a critical gap in Arabic medical communication.

The system was developed using a rich dataset that combines real doctor–patient interactions with AI-generated dialogues. This hybrid approach not only expands the linguistic diversity of the data but also ensures clinical relevance and robustness. By leveraging artificial intelligence and natural language processing techniques, the system enables the structured organization of Arabic medical knowledge and supports more accurate and consistent diagnostic decision-making.

In addition, the project offers strong educational value, helping medical students and early-career practitioners bridge the gap between English-language academic training and Arabic real-world interactions. Ultimately, this work contributes to improving healthcare delivery and lays the foundation for intelligent, culturally adapted medical systems in Arabic.

## **Chapter2 | Background**

### **2.1 Specialized Diseases**

The initial focus of this project is on heart-related diseases due to their prevalence and clinical significance. Based on the available resources and time constraints, the scope may expand to include additional conditions such as gastrointestinal and orthopedic diseases.

1. Heart Diseases: Heart diseases represent a critical area of focus, encompassing a broad range of disorders affecting the heart's structure and function. These conditions, including coronary artery disease, heart failure, and arrhythmias, are a leading cause of mortality worldwide. Common risk factors include high cholesterol, hypertension, obesity, smoking, and a sedentary lifestyle. Symptoms often involve chest pain, shortness of breath, and fatigue. Addressing heart diseases through accurate diagnosis and timely interventions is vital for improving patient outcomes and reducing the burden of cardiovascular complications [1].
2. Potential Expansion to Other Conditions: In addition to heart diseases, the project may extend its scope to gastrointestinal and orthopedic conditions, depending on the progress and available resources. Gastrointestinal diseases often involve digestive tract disorders, while orthopedic diseases focus on musculoskeletal issues such as fractures, arthritis, and joint disorders. Including these categories would enhance the versatility and applicability of the diagnostic model, addressing a broader range of healthcare challenges in Arabic-speaking communities.

This structured approach ensures the system is both effective and adaptable, prioritizing critical areas of need while allowing for scalable development.

### **2.2 Artificial Intelligence**

Artificial intelligence (AI) has emerged as a transformative tool in various disciplines, particularly in the medical field. Scientists and researchers utilize AI to develop innovative frameworks, models, and systems that emulate and enhance the cognitive functions of the human brain. The application of AI in healthcare has been steadily growing, offering promising advancements in precision medicine and improving the overall efficiency of medical diagnostics and treatment planning.

AI techniques are highly adaptable and have been successfully applied to tackle numerous challenges in medical research. Through its ability to analyze extensive datasets and uncover complex patterns, AI provides a robust foundation for future developments aimed at improving healthcare outcomes and advancing the state of medical science [2].

### **2.2.1 Machine Learning**

Machine learning is a branch of artificial intelligence focused on developing algorithms that learn and adapt from data in their environment, mimicking aspects of human intelligence. In today's data-driven world, these algorithms form the backbone of numerous advancements across various fields. Machine learning methods excel in analyzing large, multidimensional datasets, often prioritizing prediction accuracy over hypothesis-driven approaches. This makes them particularly well-suited for complex problems where traditional methods may struggle. In the context of medical research, machine learning has proven invaluable, offering powerful tools for processing patient data, identifying patterns, and providing actionable insights.

The success of machine learning in these areas highlights its potential for improving diagnostic precision and supporting healthcare professionals in making more informed decisions. As a key component of this project, machine learning contributes to building a robust system that bridges linguistic and diagnostic gaps in Arabic-speaking medical practice [3].

### **2.2.2 Applications of Machine Learning in Medicine**

ML has significantly transformed the medical field by enhancing diagnostic accuracy and predictive capabilities. By analyzing vast datasets, ML algorithms can identify patterns and correlations that might not be immediately evident to human practitioners. This has proven especially valuable in early disease detection, personalized treatment plans, and predicting patient outcomes, ultimately improving the quality of healthcare.

One of the most critical applications of machine learning in medicine is its ability to support diagnosis and prediction. ML models are trained on extensive datasets comprising patient records, imaging results, and genetic information to predict the likelihood of diseases. For instance, machine learning algorithms have been successfully employed to predict the onset of diabetes, cancer, and neurological disorders by analyzing patient history and lifestyle factors. These

predictive models assist healthcare providers in making informed decisions, thus facilitating timely interventions and personalized treatment strategies [3].

Heart disease diagnosis has notably benefited from the advancements in machine learning. ML algorithms analyze electrocardiograms (ECGs), medical imaging, and patient data to detect heart conditions more accurately and swiftly. For example, machine learning models can identify subtle changes in ECG patterns that may indicate early stages of heart disease, allowing for prompt medical attention. Additionally, ML tools have been utilized to predict the risk of heart attacks by assessing factors such as blood pressure, cholesterol levels, and genetic predisposition [4].

The use of machine learning in medicine is an important advancement, helping to improve patient care and manage diseases more effectively.

### **2.2.3 Transformer-Based Models (AraBERT)**

Transformer-based models have revolutionized Natural Language Processing by enabling deep contextual understanding of language. Bidirectional Encoder Representations from Transformers is a widely adopted model that processes input text in both directions to capture full context. AraBERT is a variant specifically pretrained on a large Arabic corpus, making it suitable for understanding complex morphological and syntactic patterns in Arabic. In this project, AraBERT serves as the core contextual encoder that transforms medical dialogues into rich vector representations, enhancing the accuracy of symptom and qualifier extraction.

### **2.2.4 Sequence Modeling Using BiLSTM**

BiLSTM networks are a form of recurrent neural networks that learn temporal dependencies in both forward and backward directions. Unlike standard LSTM models that only consider past context, BiLSTMs incorporate future information, making them ideal for processing medical dialogues where the meaning of a word depends on both preceding and following words. This capability improves the system's understanding of sentence structure and aids in extracting symptoms and related qualifiers from free-form text.

### **2.2.5 Sequence Optimization Using CRF**

CRF are statistical modeling techniques used to optimize predicted label sequences in tasks such as NER. When used as a final layer in sequence models, CRF considers dependencies between neighboring tags to ensure globally coherent predictions. For example, it prevents invalid tag

transitions like "I-SYMPTOM" following an "O" without a "B-SYMPTOM" in between. In this project, CRF improves the consistency of symptom labeling in Arabic text, particularly when combined with AraBERT and BiLSTM layers.

### **2.2.6 Benefits of AI and ML in Medical Research**

AI and ML have revolutionized medical research by enabling the analysis of large-scale datasets to uncover patterns and insights that were previously inaccessible. By processing vast amounts of data from diverse sources such as electronic health records, medical imaging, and genomic sequences, AI and ML algorithms can identify correlations and trends that help in understanding disease mechanisms and predicting patient outcomes. This capability to analyze big data not only accelerates research but also opens new avenues for personalized medicine and targeted therapies [5].

In addition to uncovering patterns, AI and ML have significantly improved the efficiency of healthcare delivery and diagnostics. These technologies automate routine tasks and enhance decision-making processes, allowing healthcare professionals to focus on more complex and critical aspects of patient care. For instance, AI-powered diagnostic tools can rapidly and accurately interpret medical images, such as MRIs and CT scans, reducing the time needed for diagnosis and increasing the accuracy of detecting conditions like cancer or cardiovascular diseases. Furthermore, AI-driven predictive analytics can anticipate patient needs and optimize resource allocation, leading to more efficient and effective healthcare services [6].

Integrating AI and ML into medical research and healthcare is set to transform the field, making it more data-focused, efficient, and centered on patient needs.

### **2.3 NLP in Medicine**

NLP is a transformative technology that enables computers to understand, analyze, and generate human language. Using machine learning algorithms, NLP processes large volumes of text and speech data to uncover meaningful insights. Its applications span various fields, from analyzing sentiment in social media posts to powering automated customer support systems. By interpreting

complex language patterns and context, NLP enhances human-computer interaction, fostering innovation and efficiency across industries.

### **2.3.1 Importance of NLP in Healthcare**

NLP is a vital tool in modern healthcare, significantly enhancing the extraction and interpretation of meaningful information from unstructured clinical text. Its importance is highlighted through several key applications:

1. Medical Entity Recognition: NLP helps identify essential medical terms such as symptoms, diagnoses, and treatments from patient records. This capability makes critical information readily accessible, supporting accurate and timely clinical decisions [7].
2. Clinical Text Understanding: NLP analyzes medical reports and other textual data, assisting in decision-making and improving workflow efficiency. By streamlining the interpretation of clinical documents, NLP enhances the overall efficiency of healthcare delivery [8].
3. Semantic Qualifiers in Clinical Reasoning: Terms like "acute/chronic" or "mild/severe" are crucial for understanding patient conditions and making accurate diagnoses. NLP techniques can standardize and analyze these qualifiers, effectively bridging the gap between textual descriptions and structured clinical data. This standardization is particularly important for consistent and accurate medical assessments, leading to better patient outcomes [7].

Overall, the use of NLP in healthcare leads to improved data management and more precise medical insights, ultimately enhancing patient care and the efficiency of healthcare services.

### **2.3.2 Challenges in Arabic Medical Text Processing**

Arabic medical text processing presents unique challenges due to the language's complex linguistic features and the scarcity of resources.

#### **1. Unique Linguistic Challenges of Arabic**

Arabic presents several unique linguistic challenges when working with NLP tasks.

**Script and Tokenization:** Arabic is written without spaces between words, making tokenization a significant challenge. For example, the phrase "البيت الكبير" (which means "the big house") combines the article "ال" (the) with the noun "بيت" (house) and the adjective "كبير" (big), requiring sophisticated algorithms to separate the components accurately.

**Orthographic Conventions:** There are multiple ways to represent vowels in Arabic, either explicitly using diacritical marks (e.g., "كتاب" for "kitab") or omitting them entirely ("كتاب"). This can lead to ambiguity in understanding the word's pronunciation and meaning. For instance, "كتب" can mean "books" (plural noun) or "he wrote" (past tense verb) depending on context.

**Directionality:** Arabic's right-to-left script requires special handling in text processing, especially when mixed with left-to-right languages like English. For example, rendering a bilingual sentence like "I visited the house" (البيت) involves complex alignment to preserve both directions.

**Morphology:** Arabic is highly inflected and agglutinative. Words often include prefixes, suffixes, and infixes, complicating the identification of their base forms. For instance, the word "وسأكتبها" (meaning "and I will write it") contains the conjunction "و" (and), the future tense marker "س", the root verb "كتب" (write), and the pronoun "ها" (it), all within one token.

**Syntactic Ambiguity:** Rich verbal and nominal suffixes creates syntactic ambiguity. For example, the phrase "كتب الكتاب" could mean "he wrote the book" or "the book wrote" depending on the context, requiring sophisticated disambiguation methods.

**Resource Scarcity:** Unlike English, Arabic suffers from a lack of annotated corpora and NLP tools. For example, while English has extensive sentiment analysis datasets like IMDb reviews, equivalent datasets in Arabic are scarce and often dialect-specific, limiting their usefulness [9].

**2. Lack of Annotated Datasets for Arabic Healthcare:** Another significant challenge is the lack of annotated datasets for Arabic healthcare. Annotated datasets are crucial for training machine learning models, especially in NLP tasks. While there are some efforts to create such datasets, like the Arabic Healthcare Dataset, the availability of high-quality, annotated data remains limited. This scarcity hinders the development and optimization of NLP models for Arabic medical text processing [10].

## **2.4 Semantic Qualifiers in Medicine**

### **2.4.1 Role of Semantic Qualifiers in Clinical Reasoning**

In clinical reasoning, semantic qualifiers play a vital role in gathering and processing information. The process begins by obtaining and filtering information, which can be derived from various sensory inputs such as reading, visual imagery, and listening, alongside other sensory experiences like tactile or olfactory cues. These pieces of information are essential in building an initial set of hypotheses grounded in both the identified problems of the case and prior knowledge from similar instances.

This process often involves the use of illness scripts, which are mental frameworks or organized sets of knowledge that clinicians develop over time. Illness scripts incorporate typical presentations, causes, and consequences of diseases, as well as the contextual factors that distinguish one condition from another. These scripts help in recognizing patterns and associating clinical findings with possible diagnoses.

Experts can rapidly develop a set of hypotheses, often within a few minutes of the encounter, using minimal clinical data. This contrasts with novice or intermediate learners, who typically require more time and additional information. The initial hypotheses guide further information collection and are refined through repeated iterations of observation and testing. This iterative process continues as more data is gathered, particularly as the diagnostic process progresses from the early stages of hypothesis generation to the more refined analysis and validation of these hypotheses [11].

### **2.4.2 Hypothetical-Deductive Reasoning and the Use of Semantic Qualifiers**

A key strategy in this reasoning process is hypothetical-deductive reasoning, where hypotheses are developed based on the collected data and then tested. This method follows the "if - then - but - therefore" approach, using available information to test and confirm or exclude certain hypotheses. This process, while similar to the scientific method, recognizes that the complexity of the human body often makes it impossible to definitively prove or disprove most hypotheses. Instead, the goal is to identify the most likely diagnosis and be ready to adjust if further information changes the case's trajectory.

Semantic qualifiers, such as acute/chronic, sharp/dull, or tender/non-tender, are essential in this reasoning process. These qualifiers help frame the clinical presentation in a way that is aligned with illness scripts, allowing for a structured comparison and contrasting of diagnostic possibilities. Using these qualifiers, medical professionals can effectively "search" for patterns within clinical data, much like how search terms are used to find information in a database.

Once hypotheses are developed, they are subjected to a more detailed analysis through probabilistic and cause-effect reasoning, applying biomedical and pathophysiological knowledge to ensure the hypothesis is scientifically sound. This process also includes testing the hypothesis against both positive and negative findings to ensure the final diagnosis is the most reasonable explanation for the clinical presentation.

As the reasoning process evolves, it is important to continue considering other possible diagnoses to avoid premature closure. The final diagnosis is not confirmed until it is rigorously evaluated against all available clinical data, ensuring that it fits within established biomedical knowledge and pathophysiological principles [11].

#### **2.4.2 Gaps in Arabic Semantic Qualifiers**

Arabic medical texts face unique challenges due to the lack of consistency and structure in the language's medical terminology. Unlike languages like English, which have well-established systems for using semantic qualifiers, Arabic often lacks uniformity, leading to confusion and errors in diagnosis and treatment. For example, terms like "acute" or "chronic" may be used differently, creating ambiguity. This inconsistency also affects the development of NLP tools, which rely on structured data. To overcome these challenges, there is a need for standardized guidelines and better training for healthcare professionals. This will help ensure clearer communication and improve patient care by using precise and consistent medical descriptions.

### **2.5 Existing Diagnostic Systems**

#### **2.5.1 Current AI-Based Diagnostic Tools**

Current AI-based diagnostic tools have revolutionized the healthcare sector by leveraging advanced technologies to analyze both structured data and unstructured text. Systems like Infermedica's AI-driven patient navigation platform and MedicalBot's deep learning model

demonstrate the capabilities of AI in processing complex medical data and providing accurate health assessments.

**Infermedica** is known for its robust symptom checker, achieving an average accuracy of 93% in identifying possible conditions based on user inputs. It also boasts a precision rate of 85% in triage recommendations, ensuring patients are directed to the appropriate level of care.

**MedicalBot**, on the other hand, utilizes deep learning techniques for disease prediction, with reported diagnostic accuracy rates of 91% for general conditions and over 87% for rare diseases. The model also demonstrates high reliability in natural language processing tasks, analyzing unstructured patient narratives with 89% precision in extracting key symptoms and medical events.

These tools focus on enhancing the diagnostic process, ensuring that both patients and healthcare professionals have access to precise and actionable health information. Their performance metrics underscore the significant role AI can play in improving diagnostic accuracy, reducing time-to-diagnosis, and supporting clinical decision-making [12].

### **2.5.2 Limitations of Current Systems**

Despite the advancements, existing AI diagnostic tools encounter significant challenges when dealing with Arabic medical data. The morphological complexity creates barriers that many current systems are not equipped to handle effectively. This underscores the necessity for a specialized system designed to address these specific linguistic and diagnostic gaps. Such a system would improve the accuracy and reliability of medical diagnoses for Arabic-speaking populations, bridging a critical gap in global healthcare accessibility [4].

### **2.5.3 BIO Tagging in Clinical NLP**

BIO tagging (Begin, Inside, Outside) is a standard labeling scheme in NLP used for annotating sequences such as symptoms, body locations, and time expressions. Each token is tagged based on whether it marks the beginning (B-) of an entity, is inside (I-) an entity, or outside (O) any relevant span. In Arabic medical dialogue processing, BIO tagging provides a structured format for training machine learning models to recognize medical entities in free text. This structured annotation

allows the system to extract detailed information from patient responses and convert unstructured narratives into machine-readable formats that support clinical diagnosis. [13]

## 2.6 Related Works

### 2.6.1 Machine Learning Models for Efficient Automated Disease Diagnosis

This study, titled "Efficient Automated Disease Diagnosis Using Machine Learning Models," authored by Naresh Kumar, Nripendra Narayan Das, Deepali Gupta, Kamali Gupta, and Jatin Bindra, focuses on creating an effective solution for the automated diagnosis of three significant health concerns: heart disease, coronavirus (COVID-19), and diabetes. The solution leverages an Android app where users input their data, and the app displays diagnostic results. The data is analyzed by a machine learning model hosted on Firebase, a cloud-based platform developed by Google that provides real-time database services. Firebase is used to store and manage the machine learning parameters, enabling efficient and immediate updates to the prediction models.

Early detection of heart disease, diabetes, and COVID-19 is critical for reducing the risk of severe complications and managing the overall impact of these diseases. This approach supports doctors in providing timely treatment. After cleaning and analyzing the datasets, logistic regression models were applied to each disease dataset. The intercepts and coefficients of the models are stored in Firebase, enabling the Android app to make real-time predictions using the most up-to-date model parameters. This architecture ensures that any updates to the datasets and models are immediately reflected in the app.

As datasets grow and models improve, the stored parameters can be seamlessly updated in Firebase, allowing continuous enhancement of the system's predictive accuracy. The Android app not only provides users with real-time diagnostic predictions after answering a few questions but also delivers up-to-date information and trends related to these diseases.

Experimental results demonstrate the effectiveness of the proposed model. For COVID-19, the model achieves a 1.4765% improvement in accuracy and a 1.2782% better F-measure compared to competing models. Similarly, for diabetes, it shows a 1.8274% increase in accuracy and a 1.7264% higher F-measure. For heart disease, the model outperforms competitors by 1.3821% in F-measure and 1.7362% in accuracy [14].

## **Chapter3 | System Implementation**

### **3.1 Overview of the System Architecture**

The system architecture integrates AI and NLP technologies to enhance the diagnostic process in Arabic-speaking healthcare settings. It consists of several key components that work together to transform user input into structured, clinically meaningful output.

The user interface is a web-based platform that allows users—including doctors, medical students, and patients—to input clinical data in Arabic and view diagnostic suggestions along with detailed symptom analysis. The design is fully responsive and accessible across different devices, providing a smooth user experience in both Arabic and English.

The data layer is composed of two sources: real-world Arabic medical dialogues collected from actual clinical interactions, and AI-generated dialogues that simulate typical scenarios in the cardiology domain. These datasets provide the foundation for training and evaluating the system's components.

The AI and NLP processing layer analyzes the input using machine learning models to extract symptoms and their qualifiers, such as severity, duration, and anatomical location. These elements are then mapped to standardized clinical terms using SNOMED CT to ensure semantic accuracy and consistency.

The semantic qualifier mapping layer organizes the extracted symptom information and links it to relevant disease categories, improving the precision of the final diagnosis. All structured outputs—including parsed dialogues, mapped qualifiers, and predicted diagnoses—are stored in the database layer, ensuring consistency, accessibility, and ease of integration with other components.

Overall, the system architecture is modular, scalable, and designed to support future extensions to additional diseases and languages. It is accessible via a web browser on any device with internet connectivity, making it suitable for practical deployment in diverse medical environments.

### **3.2 Data Collection**

The project utilizes two distinct datasets to support its objectives:

#### **3.2.1 Real-World Dataset**

This dataset consists of medical dialogues collected directly from hospitals and medical clinics. These dialogues capture real interactions between doctors and patients, providing authentic examples of how symptoms, diagnoses, and treatment options are discussed in Arabic. The dataset includes various scenarios covering a range of medical conditions, allowing the system to learn and identify common patterns in patient complaints, symptoms, and medical histories. By

incorporating real-world medical dialogues, the system can better understand the nuances of patient-doctor communication in Arabic and ensure that the AI-generated semantic qualifiers are relevant to actual clinical situations. This dataset ensures the system is grounded in real-world clinical practices and enhances its applicability to actual healthcare settings.

In addition to symptom-based interactions, the dataset also encompasses discussions around patient histories, age, gender, lifestyle factors, and environmental triggers that contribute to certain conditions, such as stress, diet, or medication use. These aspects play a significant role in diagnosis and treatment plans, further enriching the system's ability to provide accurate suggestions and diagnosis in real-world healthcare environments.

### **3.2.2 AI-Generated Dataset**

An AI-generated dataset is created using advanced artificial intelligence tools to simulate medical dialogues between doctors and patients. This dataset plays a pivotal role in enhancing the system's performance by addressing gaps in real-world data. AI can generate a wide range of medical dialogues, including scenarios with rare symptoms, less common diseases, and atypical patient responses. These additions ensure that the system is robust and adaptable, capable of handling a diverse array of medical cases.

Furthermore, AI enables the generation of various linguistic expressions for medical terms and semantic qualifiers. This is particularly valuable in understanding how patients from different regions or backgrounds might describe symptoms in diverse ways. Such variations are essential in building a system that is flexible and comprehensive enough to handle the complexities of patient communication, especially in Arabic-speaking contexts.

There are also instances where real-world data might be insufficient, particularly for newly discovered diseases or rare medical conditions. AI-generated dialogues can bridge this gap, ensuring that the system is trained on a more diverse set of cases, improving its accuracy and reliability. By integrating these AI-generated dialogues with real-world data, the system becomes more comprehensive, capable of offering accurate diagnostic support in Arabic healthcare settings and addressing the unique linguistic and cultural challenges of the region.

By concentrating on heart-related diseases, these datasets enable the development of an AI system tailored to accurately identify and map semantic qualifiers, facilitating effective diagnostic support and communication in Arabic.

The process of interacting with the ChatGPT model to generate medical dialogues was a crucial part of developing the dataset. By using clear and precise prompts, we were able to guide the model in creating medical dialogues that simulate real-life conversations between doctors and patients. Specifically, we directed the model to generate dialogues that included details such as symptoms and medical complaints related to heart diseases. Additionally, we focused on linguistic diversity in expressing symptoms to account for differences in various Arabic dialects.

For example, one of the prompts we used in Arabic was:

توليد حوار بين الطبيب والمريض حيث يصف المريض ألمًا في الصدر مع ضيق في التنفس، ويكون الألم أكثر حدة أثناء ممارسة الرياضة المريض يبلغ من العمر 45 عامًا ولديه تاريخ عائلي من أمراض القلب

Based on this prompt, the AI generated the following dialogue

المريض: دكتور، أنا شعرت بألم حاد في صدري، وأحياناً يمتد الألم إلى ذراعي الأيسر. الألم يزيد لما أعمل أي نشاط بدني، مثل صعود السلم.

الطبيب: أفهم. من متى بدأ هذا الألم؟ وهل هناك تاريخ عائلي لأمراض القلب؟

المريض: بدأ منذ أيام قليلة. والذي كان يعاني من مشاكل في القلب، وعمي تعرض لنوبة قلبية العام الماضي.

الطبيب: من المهم أن نقوم ببعض الفحوصات للتأكد من صحة قلبك. سأقوم بترتيب بعض التحاليل.

Through this process, we were able to produce realistic medical dialogues that mimic how patients communicate with doctors, helping to build a comprehensive and accurate dataset for analyzing medical symptoms in Arabic contexts.

### 3.3 Data Preprocessing

The preprocessing phase is a crucial step in preparing medical dialogues for analysis. Given that these dialogues contain valuable clinical information in natural language, it is essential to clean,

structure, and process the data to ensure accurate symptom extraction and disease prediction. The following preprocessing steps will be applied to the medical dialogues:

### 3.3.1 Text Cleaning

The first step involves cleaning raw dialogue text to prepare it for analysis. This process includes several sub-steps:

#### 3.3.1.1 Tokenization

The dialogue text was tokenized using a pretrained Arabic tokenizer from the **aubmindlab/bert-base-arabertv02 model**.

This tokenizer is specifically optimized for Arabic morphology and segmentation, making it well-suited for handling Arabic clinical narratives.

Tokenization allows the system to split sentences into individual units (tokens), enabling more precise tagging and recognition of symptoms and medical qualifiers.

Original: "أشعر بخفقان حاد و غثيان شديد بعد صعود الدرج"

Tokenized: [".", "أشعر", "بخ", "#فقا", "#ن", "حاد", "و", "غ", "شديـ", "بعد", "صعود", "الدرج"]

#### 3.3.1.2 Removing Stop Words

We removed non-essential words in Arabic, such as "و", and "على", which do not add semantic value. This step reduced noise and kept the focus on terms directly related to symptoms and diseases.

Original: "المريض يشعر بألم في صدره"

After removing stop words: "المريض يشعر بألم صدره"

#### 3.3.1.3 Normalization

To address variations in spelling, diacritics, and formatting, the text was normalized for consistency and to eliminate redundancy.

Original: "ألم في الصدر" and "ألم في الصدر"

Normalized: "ألم في صدر"

### **3.3.1.4 Standardization of Terms**

Medical terms in Arabic often appeared in multiple forms or abbreviations. We standardized these terms to their full, consistent forms to ensure uniformity across the dataset.

Original: "ألم صدر ي"

Standardized: "ألم صدر"

### **Outcome of Preprocessing**

We initially experimented with stemming and lemmatization as additional preprocessing steps. However, we observed that these techniques negatively affected model performance. In particular, they tended to distort key medical terms or reduce important contextual distinctions, leading to reduced precision and recall during entity recognition.

As a result, these methods were excluded from the final preprocessing pipeline in favor of preserving the original, context-rich expressions found in patient dialogues.

By applying these preprocessing steps, the dialogue data will be transformed into a structured and normalized format, enabling more accurate symptom extraction, disease prediction, and diagnostic reasoning. This step is foundational to the development of an effective and reliable system for assisting healthcare professionals in Arabic-speaking environments.

## **3.4 Feature Extraction**

Feature extraction is a critical step in transforming raw medical dialogues into structured and analyzable data. It focuses on identifying and extracting essential medical information, such as symptoms, risk factors, and patient demographics, while preserving the semantic nuances of Arabic language interactions. By applying these methods, the pipeline ensures accurate diagnostic reasoning, and personalized healthcare solutions in Arabic-speaking healthcare systems.

### **3.4.1 Symptoms Extraction**

We have implemented a detailed extraction pipeline to identify and analyze symptoms from Arabic doctor-patient dialogues. This phase focuses not only on recognizing symptom expressions but also on breaking them down into structured components such as symptom name, location, time,

severity, and spread. The extracted symptoms are stored in a structured JSON format to enable accurate BIO tagging and support diagnostic modeling.

#### 3.4.1.1 Named Symptom Recognition

Instead of conventional NER, we applied Named Symptom Recognition to detect the core symptom name within each patient statement. This includes identifying key terms like "ألم", "خفقان", or "ضيق تنفس", regardless of how they are embedded in longer phrases. The focus is on capturing the symptom itself, not just the full expression.

Original Phrase: "أشعر بخفقان في قلبي منذ فترة"

Name: "خفقان"

#### 3.4.1.2 Dependency Parsing

We also applied dependency parsing to identify grammatical relationships between words, enabling the extraction of both symptoms and their associated contexts. This step allows the system to link symptoms to triggers, severity expressions, or locations mentioned in the dialogue.

Sentence: "أشعر بألم في الصدر عند ممارسة الرياضة"

Parsed Output: "عند ممارسة الرياضة"

#### 3.4.1.3 Semantic Qualifier Extraction

Semantic qualifiers were extracted using rule-based methods tailored to Arabic patient expressions. These qualifiers added descriptive clinical context to each confirmed symptom. The extracted fields included:

"استمر الألم طوال اليوم", "الألم يأتي بشكل متقطع", "أشعر بألم حاد في صدرى"

"severity": Represents the intensity of the symptom, "حاد".

"duration": Describes how long the symptom lasted, "طوال اليوم".

"chronicity": Included implicitly in duration-related expressions, "متقطع".

#### **3.4.1.4 Temporal and Spatial Information Extraction**

Temporal and spatial information was also extracted using handcrafted rules to determine when and where symptoms occurred or spread. The following attributes were captured:

"أشعر بألم في صدرني منذ فترة"

"كأنه ضغط على صدرني يمتد إلى الذراع"

"أشعر به أحياناً في الكتف أيضاً"

"time": Indicates onset or recurrence timing, ["منذ فتره"], ["في بعض الأحيان"]

"location": Describes the affected body region, "الصدر".

"spread\_to": Describes symptom propagation to other regions, "الذراع".

#### **3.4.2 Demographic Features**

In this phase, we extracted key demographic features — namely age, gender, smoking status, personal medical history, and family medical history — to enhance diagnostic precision by accounting for condition-specific risks and comorbidities. For example, age and gender can influence symptom presentation and disease prevalence, while a positive family history of cardiovascular disease or diabetes elevates baseline risk. Similarly, knowing whether a patient is a current or former smoker allows the model to account for tobacco-related pathologies. By combining these variables, the system delivers a more personalized analysis, tailoring risk assessments and diagnostic suggestions to the individual's complete demographic and clinical profile.

### **3.5 Bio Tagging**

This stage transforms the richly structured annotations produced during feature extraction—namely, the identified symptoms, their associated temporal expressions and anatomical locations—into a token-level BIO format that is immediately suitable for training a sequence labeling model. Each dialogue is first normalized and split into sub-word units by the AraBERT tokenizer, then every token is assigned a “B-” tag at the start of an annotated span or an “I-” tag for its continuation, while all other tokens receive the “O” label. In doing so, we generate a complete, fine-grained corpus in which every word is explicitly linked to its role, empowering

the NER model to learn the contextual cues that distinguish medical entities from surrounding text without reliance on handcrafted rules. The outcome is a massive BIO-tagged dataset—comprising thousands of labeled tokens across hundreds of conversations—that forms a solid foundation for training a deep learning classifier adept at pinpointing symptoms, timing information and body sites within Arabic clinical dialogue.

Beyond pinpointing when and where each symptom appears, our JSON schema enriches each annotation with three critical dimensions: its duration, severity and spread to properties. By integrating these temporal, intensity and anatomical spread characteristics into our output, we produce a far more informative dataset that supports downstream diagnostic modeling and ultimately improves the precision of AI-driven medical assessments.

**Sample :** لدى ألم في صدرى عند الإستقاء يمتد إلى كتفى الأيسر مصحوب بضيق تنفس وأعاني من تعرق شديد :

**Output Sample :**

لدى	O
الم	B-SYMPOTM
في	O
صدرى	B-LOCATION
عند	B-TIME
استقاء	I-TIME
يمتد	B- SPREAD
إلى	O
كتفى	I- SPREAD
الأيسر	I- SPREAD
محبوب	O
تضيق	B-SYMPOTM
تنفس	I-SYMPOTM
أعاني	O
من	O
تعرق	B-SYMPOTM
شديد	B-SEVERITY

### 3.6 Symptoms and Qualifiers Extraction Model

#### 3.6.1 INPUT DATA

The dataset used to train the symptom and qualifier extraction model was prepared in a structured token-level format based on the BIO (Begin, Inside, Outside) tagging scheme. Each entry in the dataset corresponds to a word from patient responses in Arabic medical dialogues, annotated with a tag that indicates whether the word belongs to a medical entity such as a

symptom, time expression, severity, location, duration, or spread, or if it lies outside any relevant entity.

To ensure compatibility with transformer-based models, the textual data was preprocessed using a pretrained Arabic language tokenizer. This tokenizer segments the text into subword units, which is particularly important for capturing the morphological richness of Arabic. During this process, special attention was given to maintaining alignment between the original words and their associated tags. When a word was split into multiple subword tokens, all subwords of a symptom retained the original tag.

Each sequence was then encoded into numerical representations suitable for model input, including token identifiers, attention masks, and label indices. These sequences were padded to a fixed maximum length to enable efficient batch processing during training.

The final dataset was randomly divided into a training set (90%) and a validation set (10%) to allow the model to learn effectively while being evaluated on unseen data. Additionally, label-to-index mappings were generated and stored to ensure consistent encoding and decoding of labels throughout the training and inference pipeline.

This processed and annotated input served as the foundation for training the model described in the subsequent section.

### **3.6.2 MODEL ARCHATUCTURE**

The model architecture adopted in this work builds upon an existing Arabic Named Entity Recognition framework previously published in open-source form. [15]

While the foundational structure of the original model served as a starting point, it was significantly extended and customized to address the specific requirements of medical symptom and qualifier extraction in Arabic-language clinical dialogues.

The proposed architecture is designed to capture both the contextual understanding of Arabic medical text and the sequential nature of labeled entities. It integrates a transformer-based

encoder with a sequential tagging layer, enhanced by a CRF, to produce coherent and accurate predictions.

At the core of the model lies AraBERT v0.2, a pretrained Arabic language model based on the BERT architecture. This component generates contextualized word representations that account for the linguistic complexity of Arabic, including morphology, syntax, and semantic nuance. These embeddings are passed to a BiLSTM layer, which captures long-range dependencies across the sentence in both forward and backward directions, refining the understanding of entity boundaries and relationships.

To support classification over a customized set of entity labels tailored to the medical domain (e.g., symptom, time, duration, severity, location, spread), a fully connected (dense) layer is applied after the BiLSTM to project the hidden representations into the label space. To ensure the validity of the resulting label sequences, a CRF layer is applied on top of the model. This layer globally decodes the most likely sequence of labels, enforcing structural constraints such as disallowing illegal tag transitions (e.g., preventing “I-DURATION” from directly following “O”).

The model was trained for six epochs using a CRF-based loss function that emphasizes both token-level accuracy and sequence-level consistency. Training was conducted using mini-batches, and performance was monitored on a separate validation set to ensure generalization and stability.

This enhanced hybrid architecture—combining transformer-based embeddings, bidirectional sequence modeling, and structured decoding—proved highly effective in extracting clinically

relevant symptoms and their associated qualifiers from Arabic doctor-patient dialogues with a high degree of precision and coherence.

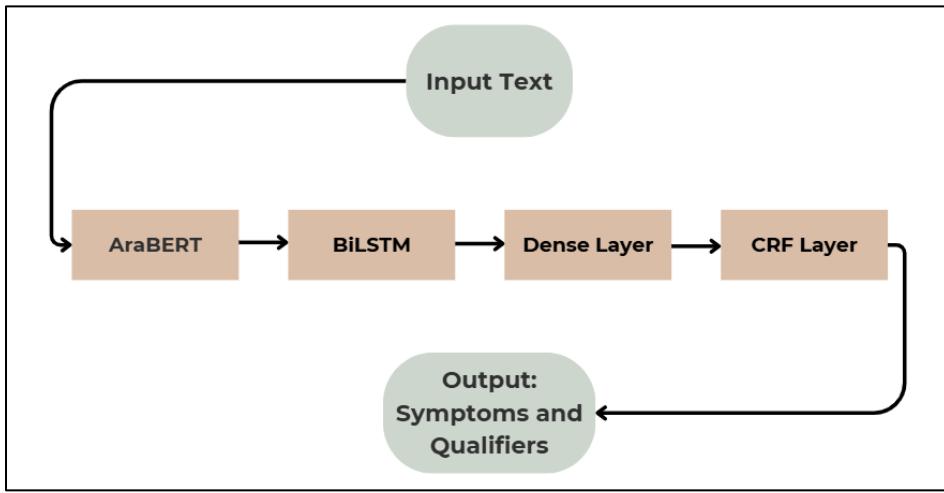


Figure 1: Model Architecture Flow

### 3.6.3 POST PROCESSING

Following the model's prediction phase, a comprehensive post-processing pipeline was employed to convert the raw BIO-tagged outputs into structured and clinically interpretable entities. This step is critical to ensure that the extracted information aligns with real-world diagnostic expectations and is presented in a consistent and machine-readable format.

Initially, the raw output is segmented into blocks corresponding to individual patient dialogues. Each block contains token-label pairs, which are first reassembled to correct for any subword tokenization artifacts. Specifically, fragmented tokens (e.g., prefixes such as "##") are concatenated to restore the original word structure and preserve semantic integrity. Tokens labeled with "O" (indicating non-entity) are excluded from further processing to focus solely on medically relevant components.

Each dialogue is then processed to extract contiguous sequences of labeled tokens representing complete medical entities. This includes not only symptoms, but also associated qualifiers such as location, time, duration, severity, and spread target.

In the presence of multiple mentions of the same symptom within a single dialogue, a merging strategy is applied to eliminate redundancy and consolidate information. The merging algorithm ensures that duplicate entries are unified, with priority given to preserving the most complete

attribute values available. For example, if one mention contains a severity qualifier while another includes a location, the merged output will retain both. Time expressions, when repeated, are collected as a list to capture all temporal references.

The final result is a structured and enriched JSON object, wherein each symptom is represented with the following attributes:

name: the canonical form of the extracted symptom

original-phrase: the literal expression found in the patient's dialogue

location, severity, duration, spread-to, and time: extracted qualifiers, if present

This post-processing framework plays a pivotal role in refining the model's raw output into a format suitable for clinical analysis, diagnostic support, and eventual linkage with standardized medical ontologies such as SNOMED CT.

### 3.7 Medical Concept Standardization

As part of the post-processing phase of our system, we carried out a medical concept standardization step to ensure that all extracted clinical information adhered to international medical standards. This step was essential for improving the clarity and clinical reliability of the data and for preparing it for downstream medical interpretation.

In this phase, we mapped each symptom and its related attributes, such as anatomical location, severity, and duration, to standardized identifiers from SNOMED CT, a globally recognized clinical terminology system. This allowed us to convert our structured outputs into medically precise and interoperable concepts.

We implemented this step to achieve two main objectives. First, to ensure consistency across different patient cases by assigning the same medical code to similar clinical findings. Second, to enable integration with medical platforms or decision-support tools that rely on structured, code-based inputs.

By completing this standardization process, we ensured that all symptom-related data in our system was not only clear and organized but also aligned with international clinical definitions, making the results more useful in real-world medical contexts.

```

"Dialogue_id": 1,
"Age": 52,
"Gender": "أنثى",
"Medical_history": "مريضة سكري منذ خمس سنوات",
"Family_history": "نعم",
"Smoking": "غير محدد",
"Symptoms": [
  {
    "name": "ضغط",
    "location": "صدر",
    "time": [
      "بعد بذل مجهود"
    ],
    "duration": "",
    "severity": "حقيق",
    "spread_to": "",
    "original_phrase": "بشق",
    "SNOMED_CT": "274640007",
    "location_SNOMED": "302551006",
    "severity_SNOMED": "255604002",
    "duration_SNOMED": ""
  }
]

```

Figure 2: SNOMED CT mapping output

### 3.8 Diagnosis Classification Model

In this phase, we began with an existing pre-trained model and a general-purpose dataset that included symptom-based classifications for a wide range of diseases. However, to tailor the model specifically to our project's objective—heart disease diagnosis—we created a separate, dedicated dataset containing only cardiac-related conditions and their associated symptoms. [16]

Rather than merging this new cardiac dataset with the general dataset, we treated it as an independent source and used it to fine-tune the existing model. This approach allowed us to adapt the model more accurately to the unique symptom patterns and diagnostic characteristics of heart diseases, without interference from unrelated medical data.

#### 3.8.1 Diagnosis Prediction Pipeline

Following the model selection, the trained classifier was integrated into the final stage of the system, which involves predicting the most likely diagnosis based on a patient's reported symptoms.

After the symptoms are extracted and normalized, they are mapped from English to standardized Arabic terms using a predefined mapping dictionary. This step ensures compatibility with the model, which was trained using English-labeled symptom features.

The mapped symptoms are then converted into a binary input vector, where each element represents whether a specific symptom is present (1) or absent (0). This input vector is passed into the trained model, which produces a prediction along with a confidence score representing the model's certainty.

### **3.8.2 Final Output and Integration**

The output of the diagnosis phase includes the following elements:

- A detailed list showing how each Arabic symptom was mapped to English.
- A list of symptoms that were successfully used as model input.
- The predicted diagnosis in English, along with its Arabic translation.
- The confidence percentage of the prediction.
- This structured output is saved in a JSON format and serves as the final result of the system, ready to be displayed in the web interface. It allows users—patients or healthcare providers—to receive an automated, interpretable diagnosis based on free-text symptom input.

## **3.9 Database Design**

The database design plays a pivotal role in the project, ensuring efficient data storage, retrieval, and management to support the system's operational needs. To handle unstructured medical data, the database must be flexible, scalable, and adaptable.

NoSQL databases, provide a schema-less architecture, enabling flexibility in handling diverse data formats. These databases are particularly suitable for healthcare applications where data evolves over time, such as semi-structured or unstructured data like patient notes, dynamic symptom entries, and real-time data processing requirements.

The primary advantages of NoSQL databases include their ability to handle flexible data structures, faster development cycles, and high scalability for managing large datasets and distributed systems. They are also highly efficient for real-time read and write operations, making them ideal for rapid data processing in healthcare scenarios.

After evaluating various database options, JSON, a NoSQL database, was selected as the most suitable solution for the project. This decision was made based on its flexibility in handling evolving data structures and its scalability for managing unstructured medical dialogues. JSON's schema-less architecture aligns perfectly with the dynamic nature of semantic qualifiers and Arabic medical data.

The schema consists of two primary collections:

**Dialogues:** Stores patient-doctor dialogues, including raw text and extracted medical information.

**Semantic Qualifiers File:** Stores symptoms name, location, time, spreade to, severity, duration.

## 3.10 Web Development

As part of the project deliverables, we developed a full web application that enables users to interact with the AI-based medical diagnosis system through a user-friendly interface. The platform supports Arabic input and output and offers both diagnostic results and detailed analysis in either Arabic or English, making it accessible and informative for a wide range of users.

### 3.10.1 Web Application Backend

The backend of the system was implemented using the Flask framework. It serves as the central controller that manages the flow of data between the user interface and the AI components. Once the user submits the required information, such as age, gender, medical history, smoking status, and symptom description, the backend initiates a series of processing steps. These include dialogue parsing, BIO-tag prediction, structured symptom extraction, SNOMED CT standardization, and final diagnosis classification. This multi-stage pipeline ensures that the input is transformed into structured, clinically meaningful output.

The backend handles multiple routes, including endpoints for submitting the input data and retrieving the analysis results. It integrates all components of the project—from rule-based parsing to machine learning predictions—and returns a structured diagnosis that reflects both the user's description and the model's clinical interpretation. Additionally, it supports multilingual output, allowing users to receive results in either Arabic or English.

### **3.10.2 Frontend User Interface**

The frontend was built using HTML, CSS, and JavaScript with RTL (right-to-left) styling tailored for Arabic-speaking users. It features a clean, modern interface where users can enter personal and medical information, describe their symptoms in free text, and choose the desired output language.

The interface includes multiple form fields for background data collection (age, gender, chronic conditions, smoking, etc.) and provides two core functionalities: submitting the data for diagnosis and requesting a detailed symptom analysis. The frontend communicates asynchronously with the backend, and the results are dynamically displayed within the same page to provide an interactive user experience.

Visual styling and layout were carefully designed to enhance readability and accessibility for native Arabic users, using appropriate fonts, directionality, and responsive elements. The overall design is fully responsive, ensuring compatibility with different screen sizes and devices.

**المؤهلات الدلالية للتشخيص الطبي بالعربية**

مرحباً بك في نظام يهدف لتحسين تجربة التشخيص الطبي باستخدام الملايين الأكبر دقة وموثوقية.

قال الله عالى: "وَإِذَا مَرَضْتُ فَهُوَ شَفِيْنِ" (السورة: ٨٠)

لمن هى تكون عمراً ملائى في رحلتك نحو الصحة والهداية، وتوفر أدوات شاملة على فهم حاليك بشكل أفضل.

العمر:

الجنس:

ذكر

ما هو التاريخ المرضي؟

أمثلة: سكري، ضغط منزلي، مشكل قلب...

هل يوجد تاريخ عائلي؟

نعم

هل تدخن؟

نعم

صف حالتك

الأعراض، التكرار، الشدة...

اللغة:

اللغة الإنجليزية  اللغة العربية

**التحليل** **التخمين**

© 2025 المؤهلات الدلالية للتشخيص الطبي بالعربية - جميع الحقوق محفوظة

Figure 3: Web application interface

## 3.11 Testing Models and Evaluation

### 3.11.1 Symptom and qualifiers extraction

The models was trained for 10 epochs.

#### BiLSTM + CRF Model Performance

##### For AI Data

###### ◊ Training Results

During training, the model showed consistent and steady improvement. The training accuracy increased from 75% in the first epoch to 97% by the tenth epoch. Simultaneously, the loss value decreased significantly from 120 to 17, suggesting that the model was learning effectively and minimizing its classification errors over time.

This behavior is typical of well-optimized deep learning models. The sharp drop in loss during

early epochs, followed by a gradual convergence, indicates that the model captured the underlying patterns in the training data. However, reaching very high training accuracy could also raise concerns about potential overfitting, especially if not matched by similar performance on validation data.

**Total training time:** 354.76 seconds, which is efficient given the model size and data complexity.

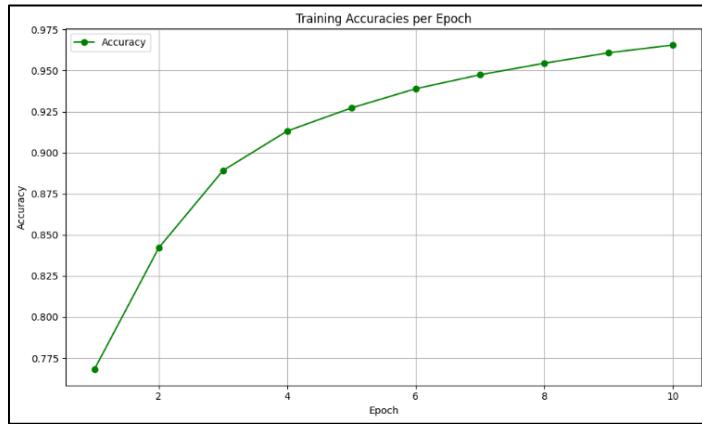


Figure 4: Training Accuracies per Epoch for AI data in BiLSTM+CRF model

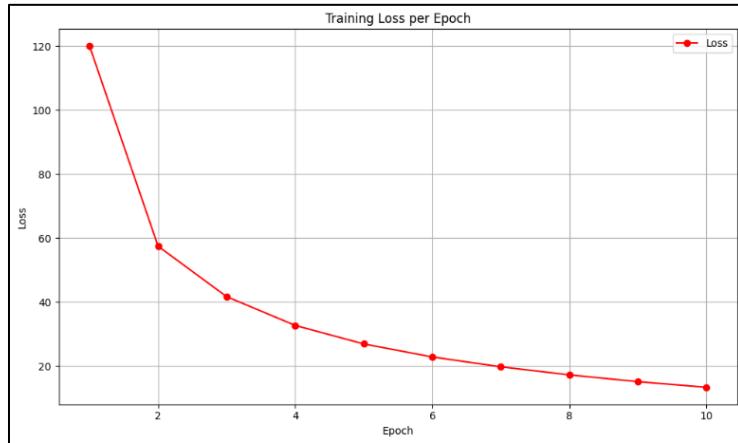


Figure 5: Training Loss per Epoch for AI data in BiLSTM+CRF model

## ❖ Validation Results

On the validation set, the model achieved a high accuracy of 95.09%, which is very close to its training accuracy. This indicates that the model generalized well to unseen data during training and did not memorize specific samples.

The similarity between training and validation performance suggests a well-generalized model. It implies that the validation data had a similar distribution to the training data, and the model was not overfitting. It also highlights that the model's architecture and hyperparameters were well-tuned for the task.

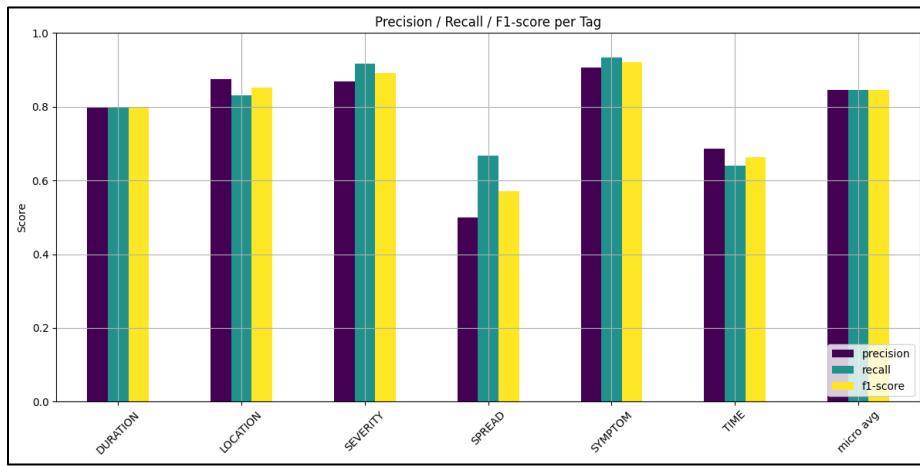


Figure 6: Precision, Recall, F1-score for Validation for AI data in BiLSTM + CRF model

## ◆ Test Results

The model achieved 85.81% accuracy on the test set at the token level, which is a strong result for sequence labeling tasks in medical NLP.

Entity-Level Metrics (seqeval):

- Precision: 65.79%
- Recall: 57.96%
- F1-Score: 61.63%

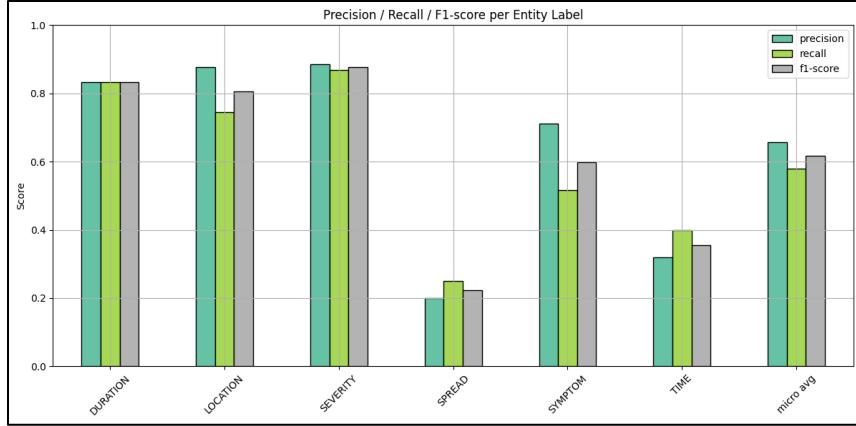


Figure 7:Precision, Recall, F1-score in Test for AI data in BiLSTM + CRF model

Compared to training and validation, the test performance—particularly in terms of precision and recall—was noticeably lower.

To address these limitations, may need to expanding the training dataset to include more diverse linguistic structures and better coverage of all entity types. Additionally, integrating a **pretrained language model with deeper medical domain adaptation** could further enhance generalization and tagging accuracy, especially for complex or rare symptom mentions.

## For Real Data

### ◊ Training Results

In Epoch 1, the model started with a training accuracy of 81.86% and a loss of 95.55, by Epoch 10, the accuracy improved to 96.58% and the loss decreased to 13.51.

The total training time was approximately 402.18 seconds.

This progressive drop in loss and increase in accuracy indicate that the model effectively learned from real-world medical dialogues over time.

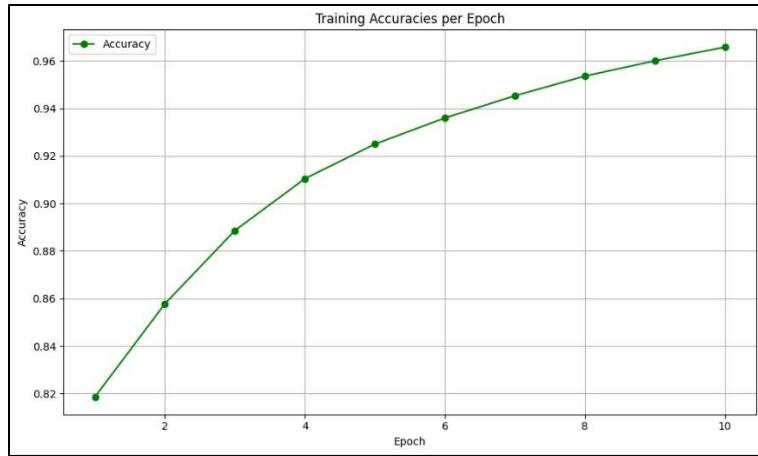


Figure 8: Training Accuracy per Epoch for Real data in BiLSTM+CRF model

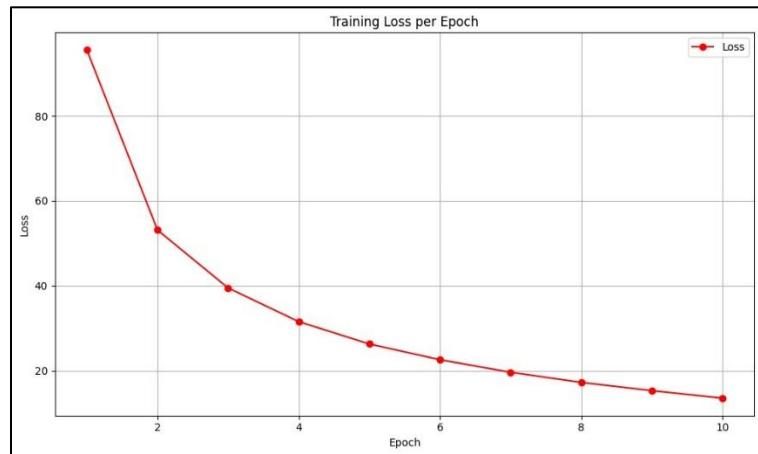


Figure 9: Training Loss per Epoch for Real data in BiLSTM+ CRF model

## ❖ Validation Results

The model performed well on the validation set, achieving:

Overall validation accuracy: 92.58%

Micro average F1-score: 0.77

The model performed excellently on well-represented and structurally clear entities like LOCATION and SEVERITY.

DURATION and TIME showed discrepancies between precision and recall, suggesting partial detection or mislabeling in some cases.

SPREAD remained challenging due to its lower frequency and contextual ambiguity.

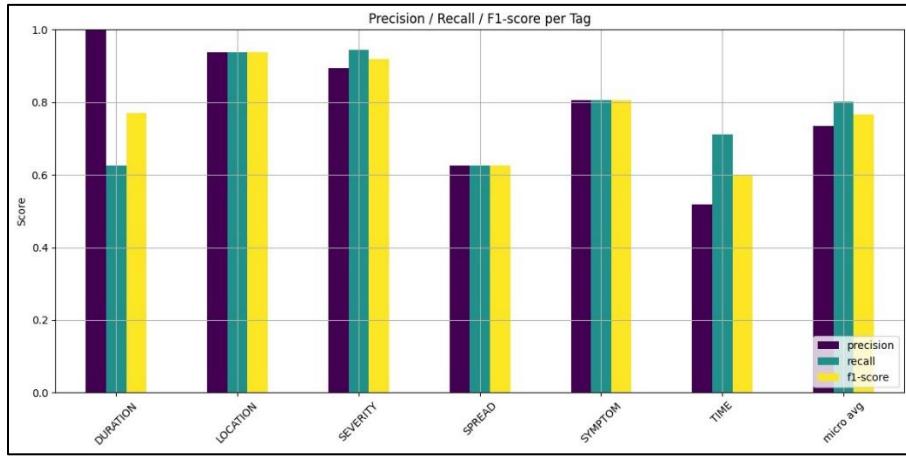


Figure 10::Precision, Recall, F1-score for Validation for Real data in BiLSTM + CRF model

## ◊ Test Results

On the held-out test set:

- Token-level Accuracy (sklearn): 85.67%
- Overall F1-score (seqeval): 56.12%

High performance was maintained for LOCATION, SEVERITY, and DURATION.

Lower scores for SPREAD, SYMPTOM, and TIME suggest difficulties in detecting these entities, likely due to:

- Data sparsity or imbalance.
- Complex or indirect phrasing by patients.
- Overlap or confusion between qualifiers.

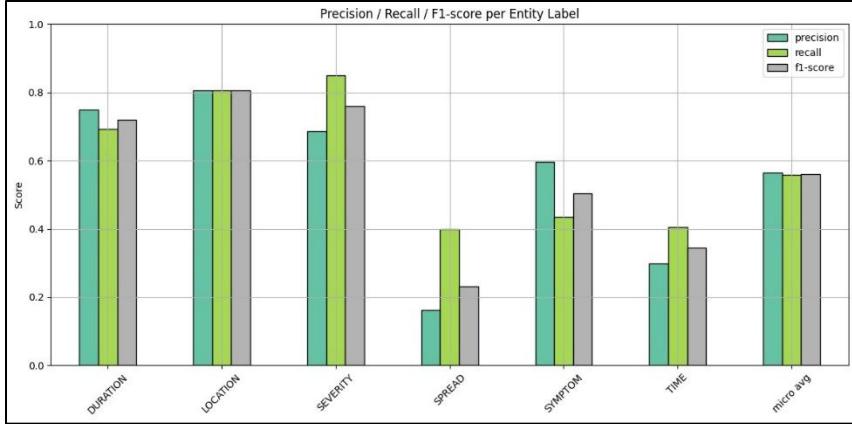


Figure 11:Precision, Recall, F1-score in Test for Real data in BiLSTM + CRF model

## AraBERT + BiLSTM + CRF Model

### For AI Data

#### ◊ Training Results

The AraBERT-based architecture enhanced with BiLSTM and a CRF layer showed a strong and stable training curve. Training accuracy began at 84% and rose steadily across epochs, reaching over 98% by the final iteration. Concurrently, the loss dropped sharply from around 70 to just under 5, indicating rapid early learning followed by gradual convergence.

This training pattern reflects successful feature extraction from AraBERT embeddings, refined further by BiLSTM's sequence modeling and CRF's structured decoding. The training process was completed in 542.14 seconds, which is reasonable given the model's complexity and deep architecture.

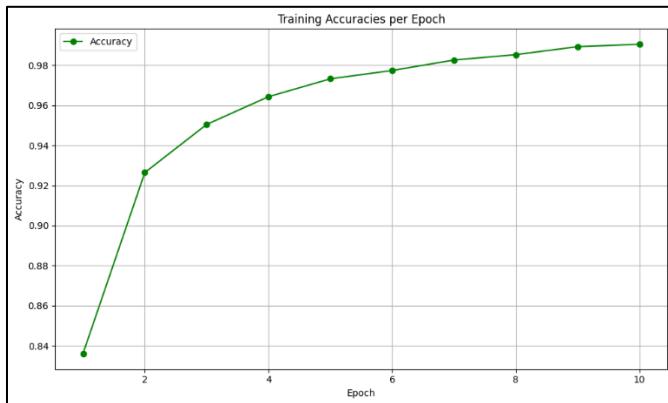


Figure 12:Training Accuracies per Epoch for AI data in AraBERT + BiLSTM + CRF model

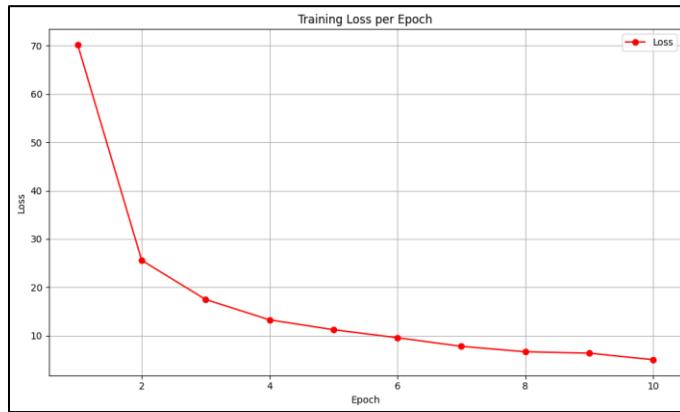


Figure 13: Training Loss per Epoch for AI data in AraBERT +BiLSTM+CRF model

## ❖ Validation Results

On the validation set, the model achieved an accuracy of 95.54%, closely matching the final training accuracy. This near-equal performance confirms that the model generalizes well to unseen samples and is not overfitting.

The consistent alignment between training and validation curves suggests that the model was well-tuned and benefited from the contextual richness of AraBERT, particularly in handling the morphological complexity of Arabic symptom expressions.

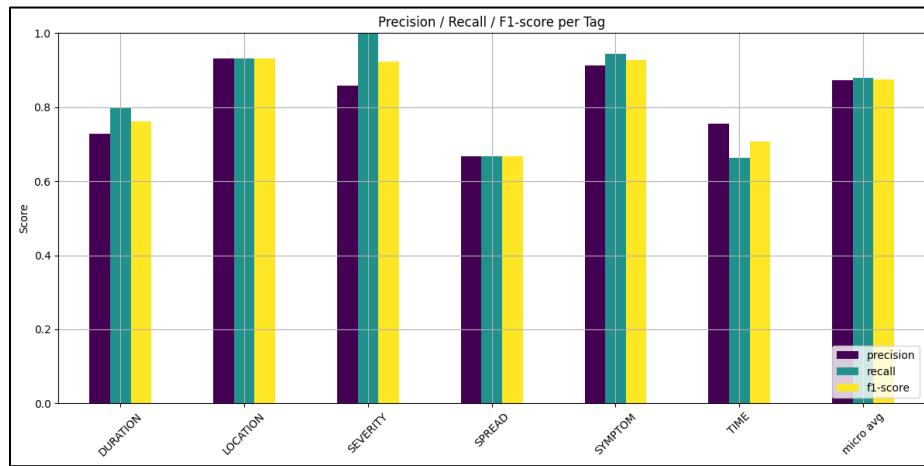


Figure 14: Precision, Recall, F1-score for Validation for AI data in AraBERT +BiLSTM + CRF model

## ◊ Test Results

The model reached a token-level accuracy of 89.33%, showing strong performance on real-world medical text outside of the training domain.

Entity-Level Performance (seqeval):

- Precision: 71.66%
- Recall: 72.25%
- F1-Score: 71.95%

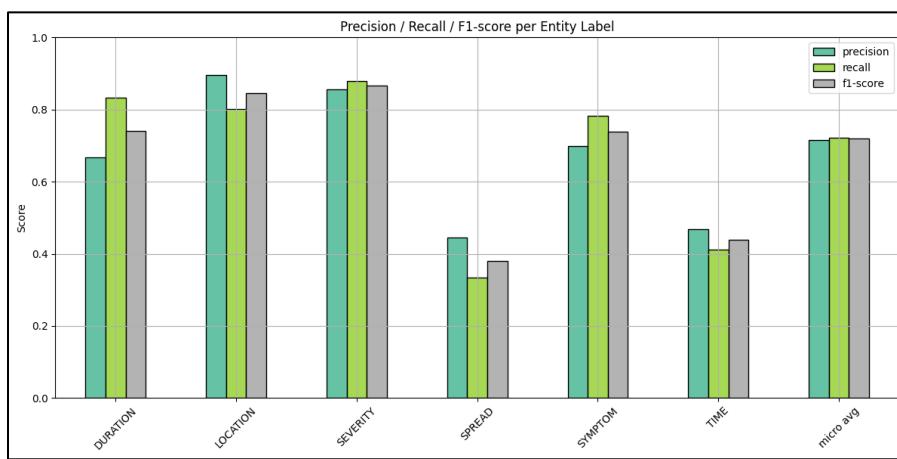


Figure 15:Precision, Recall, F1-score for Test for AI data in AraBERT +BiLSTM + CRF model

---

## For Real Data

### ◊ Training Results

Throughout training, the model showed steady improvements. The training accuracy increased from 83.5% in the first epoch to over 98.7% by epoch 10. Simultaneously, the loss decreased sharply from above 80 to just under 6, demonstrating effective learning and convergence.

Training time: 564.35seconds

This behavior suggests that the model successfully captured the structure and semantics of Arabic medical dialogues. The drop in loss was steep at the start, then gradually stabilized — a typical pattern for deep learning convergence.

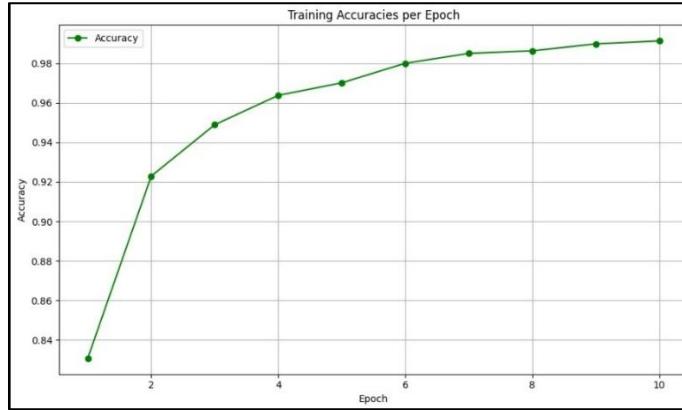


Figure 16: Training Accuracy per Epoch for Real data in AraBERT + BiLSTM + CRF model

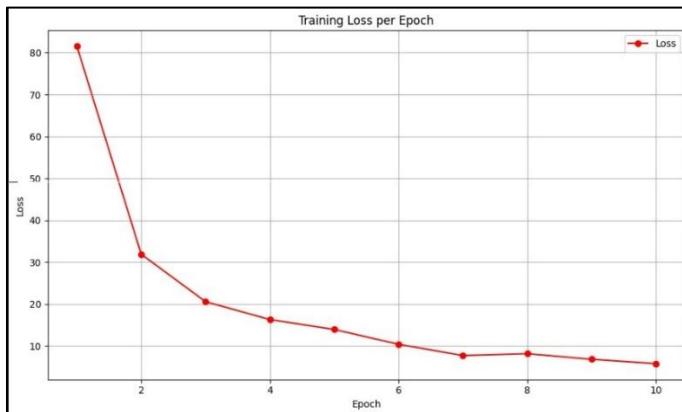


Figure 17: Training Loss per Epoch for Real data in AraBERT + BiLSTM + CRF model

## ❖ Validation Results

Validation performance remained strong and closely followed training trends. the model achieved an accuracy of 92%, Micro average metrics from the validation data show high scores across key entities

- The model performed best on well-represented entities like DURATION, LOCATION, and SEVERITY.
- Lower scores for SPREAD and TIME suggest a need for more data or improved handling of less frequent, more ambiguous qualifiers.

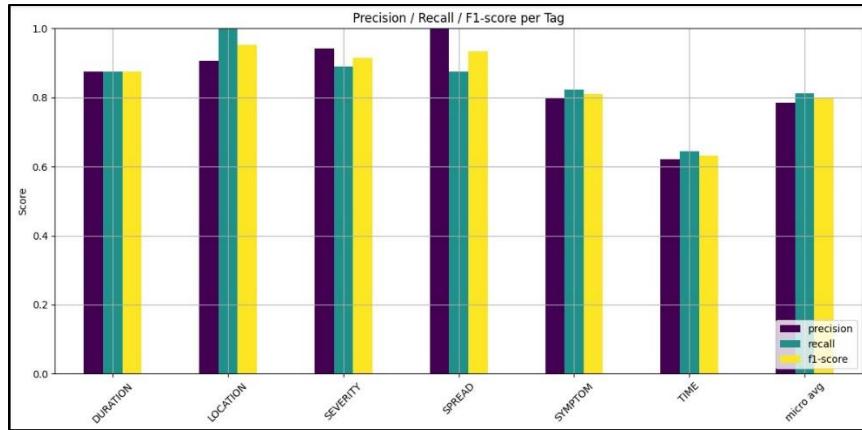


Figure 18::Precision, Recall, F1-score for Validation for Real data in AraBERT +BiLSTM + CRF model

## ❖ Test Results

From the seqeval classification report on the test data:

- Overall Precision: 0.6154
- Overall Recall: 0.6251
- Overall F1-score: 0.6202

In terms of token-level accuracy (using sklearn), the model achieved a solid 87.89%, indicating it was able to assign correct BIO labels to the majority of tokens in unseen dialogues.

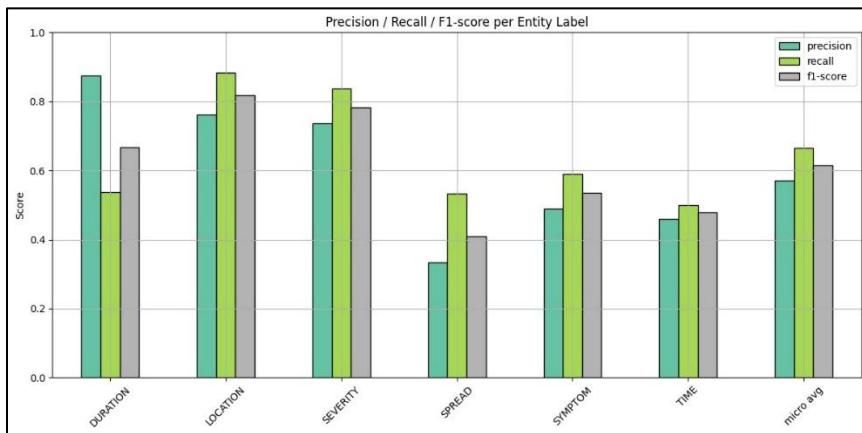


Figure 19::Precision, Recall, F1-score for Test for Real data in AraBERT + BiLSTM + CRF model

## Analysis and Interpretation Between AraBERT + BiLSTM + CRF and BiLSTM + CRF models

Metric	BiLSTM + CRF	AraBERT + BiLSTM + CRF
<b>Training Accuracy (Final)</b>	95.17%	98.6%
<b>Training Loss (Final)</b>	17.81	4.7
<b>Training Time (sec)</b>	354.76	542.14
<b>Validation Accuracy</b>	95.09%	95.54%
<b>Test Accuracy (Token-level)</b>	85.81%	89.33%
<b>Test Precision (seqeval)</b>	65.79%	71.66%
<b>Test Recall (seqeval)</b>	57.96%	72.25%
<b>Test F1-Score (seqeval)</b>	61.63%	71.95%

Table 1: Comparison Between extraction models

The AraBERT + BiLSTM + CRF model outperformed the standard BiLSTM + CRF model across nearly all key evaluation metrics. It achieved significantly higher test F1-score (71.95% vs 61.63%) and token-level accuracy (89.33% vs 85.81%), demonstrating better generalization and robustness, especially in recognizing diverse symptom expressions in Arabic medical dialogues.

While the training time for the AraBERT -based model was longer (~542s vs ~354s), this is expected due to the complexity of contextual embeddings. The gain in accuracy and recall justifies the additional computation time.

The most noticeable improvements are in recall, suggesting that AraBERT helped the model detect more correct entities, even if expressed in varied or subtle ways. This aligns with BERT's strength in capturing contextual and morphological nuances — a crucial factor in Arabic clinical text.

Integrating AraBERT significantly boosted the model's ability to generalize, making it better suited for real-world applications, especially when handling diverse patient narratives.

## Analysis and Interpretation Between Real and AI data

Metric	AI Data	Real Data
<b>Training Accuracy (Final)</b>	98%+	98.7%
<b>Training Loss (Final)</b>	~5	~6
<b>Training Time (sec)</b>	542.14	564.35
<b>Validation Accuracy</b>	95.54%	92%
<b>Test Accuracy (Token-level)</b>	89.33%	87.89%
<b>Test Precision (seqeval)</b>	71.66%	61.54%
<b>Test Recall (seqeval)</b>	72.25%	62.51%
<b>Test F1-Score (seqeval)</b>	71.95%	62.02%

Table 2: Comparison Between Real and AI data

The AraBERT + BiLSTM + CRF model showed consistently strong training performance across both datasets, with training accuracy exceeding 98% and a similarly low loss. However, test performance metrics (precision, recall, and F1-score) were significantly higher on the AI data compared to the real data.

This gap likely reflects the increased variability and ambiguity present in real-world medical dialogues. While AI-generated data can be cleaner and more uniform, real data introduces challenges such as inconsistent expressions, misspellings, and noise.

Despite these challenges, the model still achieved reasonable performance on real data, especially in terms of token-level accuracy, which stayed above 87%. These findings emphasize the importance of robust preprocessing, entity balancing, and potentially fine-tuning on larger, domain-specific real-world corpora.

## Final Interpretation

The analysis of both models across AI and real clinical data highlights clear differences in performance. The AraBERT + BiLSTM + CRF model consistently achieved higher results due

to AraBERT’s contextual understanding, especially on structured AI data. While performance declined slightly on real dialogues—due to informal language and variability—it remained strong for common entities like LOCATION and SEVERITY.

In contrast, the simpler BiLSTM + CRF model showed good training performance but struggled more on real data, particularly with low-frequency tags like SPREAD and TIME, due to limited contextual capacity.

Transformer-based models are better equipped for Arabic medical text, but challenges remain with rare entities. Future work should focus on expanding data diversity and exploring advanced architectures to further improve generalization.

### 3.11.2 Diagnosis classification

Using the specialized cardiac dataset, we retrained and evaluated a range of machine learning classifiers to determine which performs best in diagnosing heart conditions. The classifiers included:

Naive Bayes, Random Forest, K-Nearest Neighbors, LR, Support Vector Machine, Decision Tree, Multi-Layer Perceptron.

Each model was assessed using two performance metrics:

- **Test Accuracy:** Measures how well the model performs on unseen data from the cardiac dataset.
- **Cross-Validation Accuracy:** The average accuracy across five different training/testing splits (5-fold cross-validation), which evaluates the model’s generalizability.

The evaluation results are presented in the bar chart below. LR achieved the highest test accuracy at 87.94%, followed closely by SVM (87.7%) and MLP (87.06%). Based on these findings, LR was selected as the best-performing model and saved for deployment within the final diagnostic system.

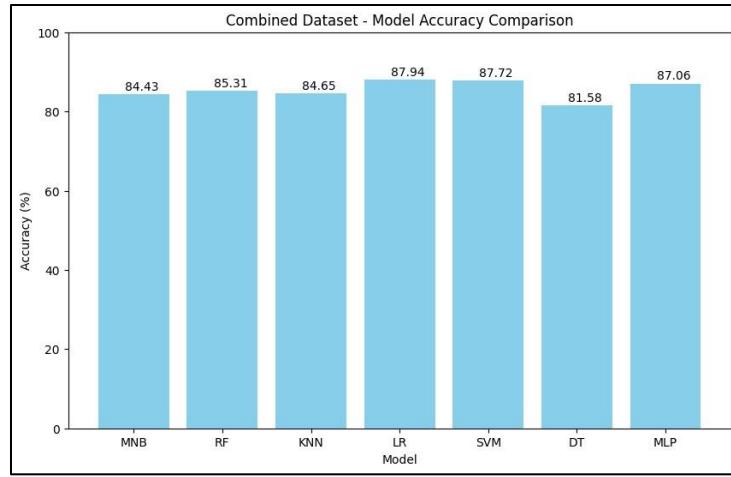


Figure 20: Diagnosis Model Accuracy Comparison

## Sample Output:

### In Arabic

**قال الله تعالى: "وَإِذَا مَرَأَتْهُ فَهُوَ شَقِيقٌ"** (المراء: 80)

حينها تكون عرضاً لك في حياتك غير الصحة والغاية، وتوفر أدوات تساعدك على فهم حياتك بشكل أفضل.

العمر:

الجنس:

ما هو تاريخ المرضي؟

إذا كان صحيحاً نعم  لا

هل يوجد تاريخ عائلي؟

هل تدخن؟

صحتك حالياً

أدعني مدة ثلاثة من دقائق لوى في القلب مخصصة عند سماع النسخة حتى يتم اغتصار المدة

اللغة:  اللغة العربية  اللغة الإنجليزية

**التحاليل** **التشخيص**

**التشخيص المقترن: قصور القلب الاحتياطي**

**تفاصيل التحليل الكامل**

- العمر: 45
- الجنس: ذكر
- التاريخ المرضي: إلقاء حجر خلط الدم.
- التاريخ العائلي: مم
- التدخين: نعم
- الرسم: متغير، يختلف قوي القلب خاصة عند سماع درج (أي اضطراب أهياً للوقت أخذ نفس عميق يعقب ببعض أوقات ألم ملطف، مظنة سريري لكنه ملطف) عموماً عدم كثرة أمثلة جرعة زوجة جرعة ساقين حالان فرقة سبب لك ذلك بصفتها بروبروم أمراض ذات روت على الشفاه يعني شيئاً غير طبيعي يحدث جسمك

**الأمراض**

- المرض: خللان
- المرض المرتبط: خللان
- SNOMED: 802130202
- المكان: قلب
- SNOMED: 80891009
- الرفق: آمنا
- اللذلة: حاد
- SNOMED: 24484000
- المرض: نام
- المرض المرتبط: نام
- SNOMED: 22253000
- المكان: صدر
- SNOMED: 302851008
- الرفق: عدم قدرة التنفس
- الأداء: ملطف
- SNOMED: 105480008
- المرض: تورم
- المرض المرتبط: تورم
- SNOMED: 230787002
- المكان: ساق
- SNOMED: 91609006
- الرفق: حلال فراء صباح

Figure 21: Sample Arabic result

## In English

قال الله تعالى: "وَمَا يَعْلَمُ فُورَتُهُ تَقْرِيبُهُ" (النَّارُ، ٨٠)  
لمن هنا تكون عبادتنا في سماتك غير الصالحة والكافرة، ونور آدوات شامتك في فهم حاليك أفضل.

العمر	45
الجنس	ذكر
ما هو تاريخ المرضي؟	ارتفاع ضغط الدم
هل يوجد تاريخ عائلي؟	نعم
هل تدخين؟	نعم
صف حالتك	اعاني من تضخم في القلب خاصة عند صعود السلالم، حتى اضطر لหย冤ا
اللغة	اللهجة العامية <input checked="" type="radio"/> اللهجة الكندية <input type="radio"/>

التحليل نشفاف

Expected diagnosis: Congestive heart disease

**Full Analysis Details**

- Age: 45
- Gender: male
- Medical History: High blood pressure.
- Family History: Yes
- Smoking: Yes

منذ فترة ملحوظ ارتفاع ضغط الدم، اضطر أحيناً لتناول أدوية لتنزيله. بعض الأوقات أنا متطلع للذهاب إلى المكتب للعمل، حيثما عند قيام وبعد أيام أحدها، جربت تزويج ساقين، حالاً فوجئت بارتفاع ضغط الدم، مما دفعني إلى زيارة طبيعي، حيثما في طبيعي يعطي جسم

**Symptoms**

- Symptom (Arabic): خلقان
- Normalized Symptom: Palpitations
- SNOMED: 85213002
- Location: Heart
- SNOMED: 81491109
- Time: sometimes
- Severity: sharp
- SNOMED: 24444000

- Symptom (Arabic): ألم
- Normalized Symptom: Pain
- SNOMED: 22233000
- Location: Chest
- SNOMED: 202331004
- Time: When there is an effort
- Duration: intermittent
- SNOMED: 15544000

- Symptom (Arabic): فرط
- Normalized Symptom: Swelling / Edema
- SNOMED: 220787002
- Location: Cal/Skin
- SNOMED: 81602004
- Time: During the morning period

Figure 22: Sample English result

## 3.12 System Flow

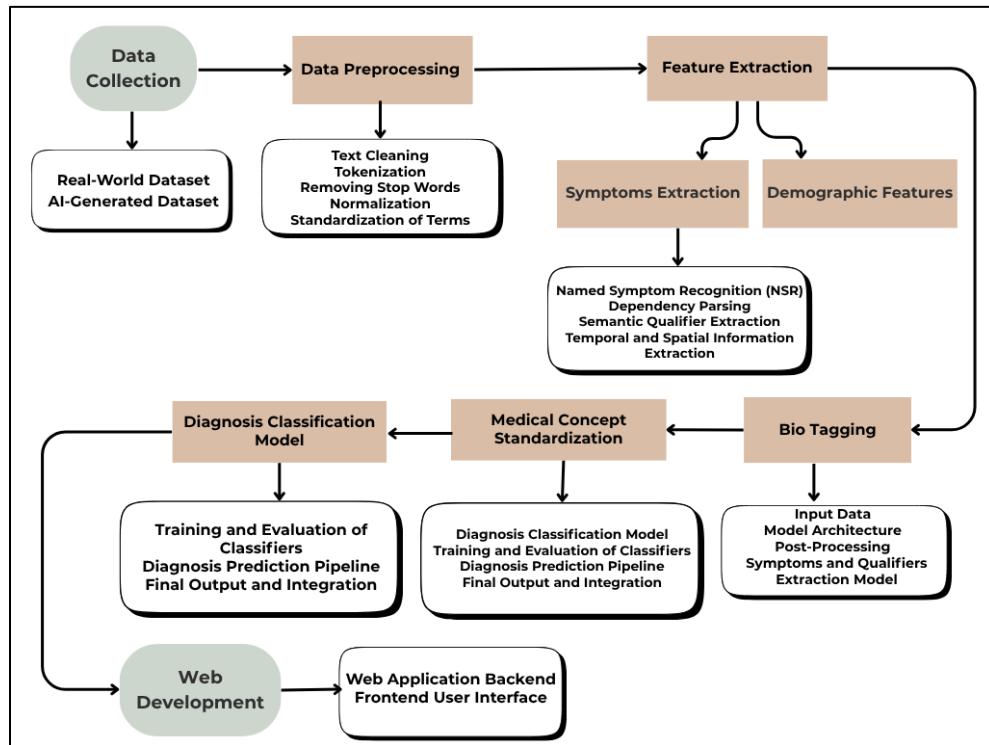


Figure 23: System flow chart

## **Chapter4 | Future Work and Current Limitations and Conclusion**

### **4.1 Possible Future Work**

#### **Expanding Data Collection**

Collaborate with healthcare institutions to collect larger datasets of real-world Arabic medical dialogues while ensuring data privacy and ethical standards.

#### **Enhancing Support for Rare Diseases**

Generate synthetic data for rare diseases and atypical symptoms to fill gaps in the dataset.

Incorporate expert-reviewed case studies to improve system performance on uncommon conditions.

#### **Extending Disease Coverage**

Expand the system's scope beyond heart-related diseases to include conditions like respiratory diseases, neurological disorders, endocrine diseases, gastrointestinal conditions, and infectious diseases.

#### **Multimodal Data Integration**

Incorporate data from multiple sources, such as audio recordings of consultations, diagnostic imaging, and wearable device data.

Use voice recognition technologies to process spoken Arabic dialogues.

#### **Enhancing Risk Assessment and Preventive Care**

We aim to enhance the system by including features that assess the severity of a patient's condition, providing clear risk indicators and personalized recommendations to help prevent disease progression. The system will track health trends over time, enabling early detection of worsening conditions, and will offer automated alerts for critical cases along with follow-up reminders to support proactive healthcare management.

This future work aims to enhance the system's functionality, scalability, and relevance while ensuring it adapts to evolving medical and technological advancements.

## **4.2 Limitations of the System:**

### **Difficulty in Collecting Real-World Data**

Gathering real-world data from doctors or patients poses a significant challenge. Medical cases often involve sensitive and private information, making individuals hesitant to share their data. This can result in a limited dataset, which impacts the system's ability to generalize across diverse medical scenarios.

### **Limited Resources in Arabic**

Compared to English and other widely spoken languages, Arabic lacks robust resources, including medical datasets, terminologies, and NLP tools. This limitation makes it more challenging to develop and fine-tune AI and NLP models for Arabic-speaking healthcare settings.

### **Insufficient Coverage for Rare Diseases**

Due to the scarcity of real-world and simulated data for rare diseases and atypical symptoms, the system might struggle to provide accurate diagnostic suggestions for less common medical conditions.

### **Accuracy of AI-Generated Data**

While AI-generated dialogues fill gaps in the dataset, they may not fully capture the nuances of real-world doctor-patient interactions. Over-reliance on synthetic data can lead to biases in the system's performance.

These limitations highlight areas that require focused attention to improve the system's performance, usability, and scalability in future iterations.

## **4.3 Conclusion**

This project demonstrated the successful development of a comprehensive AI-driven system for extracting symptoms and semantic qualifiers from Arabic medical dialogues and predicting the most likely cardiac diagnosis. By combining real clinical data with AI-generated dialogues, the

system was trained and evaluated across various settings using state-of-the-art models, including BiLSTM + CRF and AraBERT + BiLSTM + CRF.

The experimental results showed that integrating contextualized embeddings from AraBERT significantly improved extraction accuracy, particularly on structured AI-generated data. However, performance on real-world data revealed persistent challenges related to linguistic variability, rare entities, and limited annotated resources in Arabic. Despite these limitations, the system achieved strong token-level and entity-level results, particularly for common symptom attributes like location and severity.

In addition to the extraction component, the project included a diagnosis classification module, where Logistic Regression emerged as the most accurate classifier for heart-related conditions. A fully functional web interface was also built, enabling users to input free-text Arabic descriptions and receive structured, bilingual diagnostic outputs.

Ultimately, this work lays the foundation for practical, AI-supported clinical tools in Arabic-speaking healthcare environments. It bridges a critical gap in Arabic medical NLP by providing structured symptom interpretation and diagnosis support, while also offering opportunities for future expansion into other diseases and multimodal inputs. With further refinement, this system can play a pivotal role in improving diagnostic accuracy and communication between patients and clinicians across the region.

## References

- [1] M. C. Staff, "Heart disease: Symptoms and causes," [Online]. Available: <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>. [Accessed 7 Nov 2024].
- [2] J. S.-G. a. C. Sidey-Gibbons, "WHO issues first global report on Artificial Intelligence (AI) in health and six guiding principles for its design and use," 2019. [Online]. Available: <https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use>. [Accessed 7 Nov 2024].

- [3] C. S.-G. John Sidey-Gibbons, "BMC Medical Research Methodology," 2019. [Online]. Available: <https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0681-4>. [Accessed 7 Nov 2024].
- [4] S. S. & N. R. J. F. Milad Mirbabaie, "Artificial intelligence in disease diagnostics," 2021. [Online]. Available: <https://link.springer.com/article/10.1007/s12553-021-00555-5?form=MG0AV3>. [Accessed 7 Nov 2024].
- [5] C. Staff, "AI in Health Care: Applications, Benefits, and Examples," [Online]. Available: <https://www.coursera.org/articles/ai-in-health-care?form=MG0AV3>. [Accessed 10 Nov 2024].
- [6] I. Education, "AI healthcare benefits," 2023. [Online]. Available: <https://www.ibm.com/think/insights/ai-healthcare-benefits?form=MG0AV3>. [Accessed 12 Nov 2024].
- [7] A. S. J. H. a. C. L. Jing Li, "A Survey on Deep Learning for Named Entity Recognition," 2016.
- [8] H. L. Z. P. Y. & P. J. Somasundaram, "Deep Learning for Health Informatics," December 2016.
- [9] N. V. Otten, "Arabic NLP — How To Overcome Challenges, Tutorials In Python & 9 Tools/Resources Including Large Language Models (LLMs)," 2023. [Online]. Available: <https://spotintelligence.com/2023/10/29/arabic-nlp/?form=MG0AV3>. [Accessed 20 11 2024].
- [1] H. Gawbah, "AHD: Arabic Healthcare Dataset," 2024. [Online]. Available: 0] <https://data.mendeley.com/datasets/mgj29ndgrk/4?form=MG0AV3>. [Accessed 20 11 2024].
- [1] C. C. o. M. a. t. U. o. Iowa, "Master Clinician Program," [Online]. Available: 1] <https://medicine.uiowa.edu/internalmedicine/education/master-clinician-program/students/clinical-and-diagnostic-reasoning>. [Accessed 22 Nov 2024].

- [1] K. L. a. A. N, "Healthcare Technology Trends and News," 2024. [Online]. Available:  
2] <https://www.ominext.com/en/blog>. [Accessed 24 Nov 2024].
- [1] Z. C. Y. T. B. W. X. & X. H. Liu, "Automatic de-identification of electronic medical  
3] records using token-level and character-level conditional random fields. Journal of  
Biomedical Informatics," 2015. [Online]. Available:  
<https://www.sciencedirect.com/science/article/pii/S1532046415001197?via%3Dihub>.  
[Accessed 13 2025].
- [1] N. N. D. D. G. K. G. Naresh Kumar, "Efficient Automated Disease Diagnosis Using  
4] Machine Learning Models," 2021. [Online]. Available:  
<https://onlinelibrary.wiley.com/doi/10.1155/2021/9983652#data-availability..> [Accessed 24  
Nov 2024].
- [1] M. Ahmed, "arabic\_named\_entity," [Online]. Available:  
5] [https://github.com/MostafaAhmed98/arabic\\_named\\_entity/blob/main/model\\_implementatio](https://github.com/MostafaAhmed98/arabic_named_entity/blob/main/model_implementatio)n/arabic\_named\_entity\_arabert.ipynb. [Accessed 14 2025].
- [1] R. Agarwal, "MediCURE-Disease-Prediction-based-on-Symptoms," [Online]. Available:  
6] [https://github.com/rashi2911/MediCURE-Disease-Prediction-based-on-Symptoms?utm\\_source=chatgpt.com](https://github.com/rashi2911/MediCURE-Disease-Prediction-based-on-Symptoms?utm_source=chatgpt.com). [Accessed 14 6 2025].