Titre: Title:	Virtual Medical Assistant Model Using Natural Language Processing in Healthcare System
Auteur: Author:	Ali Elahi Naraqi
Date:	2024
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Elahi Naraqi, A. (2024). Virtual Medical Assistant Model Using Natural Language Processing in Healthcare System [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/61859/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/61859/
Directeurs de recherche: Advisors:	Samuel Pierre
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Virtual Medical Assistant Model using Natural Language Processing in Healthcare System

ALI ELAHI NARAQI

Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie informatique

Décembre 2024

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Virtual Medical Assistant Model using Natural Language Processing in Healthcare System

présenté par **Ali ELAHI NARAQI**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Martine BELLAÏCHE, présidente

Samuel PIERRE, membre et directeur de recherche

Ranwa AL MALLAH, membre

DEDICATION

This thesis is dedicated to my parents, my brother and his wife. I would not have been able to finish my graduate studies without them unending support and encouragement.

ACKNOWLEDGEMENTS

I would first like to express my sincere gratitude to my research director, Professor Samuel Pierre of Polytechnique Montreal and director of LARIM, for providing me an opportunity to obtain my master's degree in his lab and for his trust, support, and direction during this incredible journey. I consider myself incredibly fortunate to have been given an opportunity to join your esteemed laboratory.

Furthermore, I extend my deepest appreciation to Dr. Franjieh El Khoury, LARIM’s research associate and coordinator and supervisor of the APSEC project. I am endlessly grateful for Dr. Franjieh El Khoury's consistent encouragement, support, and patience throughout this journey. Her mentorship, advice, and expertise far surpassed my expectations throughout my master’s degree. I am also deeply thankful for her continued assistance and critical review of my dissertation. It was a tremendous honor to learn under such an experienced and capable supervisor.

In loving memory of my grandparents, whose presence and guidance continue to inspire me even though they passed away a few years ago. Their values and strength remain deeply embedded in my journey, and I hold their memory close to my heart.

I would also like to personally thank my family for their constant encouragement throughout the years and for supporting my confidence when I needed it. I am particularly grateful to my parents, who are the most admirable people in the world and have always supported me during my most challenging moments.

A special thanks flows out to my brother Morteza and his wife Elin for their unwavering support during this journey. Their encouragement and understanding have been immensely valuable, and I am truly grateful for their unwavering belief in me.

Finally, I'm really appreciate of all the support and friendship from everyone in LARIM.

RÉSUMÉ

L'importance des soins de santé dans nos vies est indéniable, et les progrès rapides continuent de transformer ce domaine quotidiennement. Un domaine clé dans cette évolution est l'intégration de l'intelligence artificielle et de l'apprentissage automatique pour améliorer les services de santé. De nos jours, les systèmes de santé peuvent utiliser ces modèles basés sur l'intelligence artificielle pour répondre aux besoins des patients, améliorer l'accessibilité à l'information médicale et accroître l'efficacité opérationnelle des prestataires de soins de santé. Cependant, en parallèle de ces avancées, de nouveaux défis sont apparus, notamment en matière de confidentialité des données, de précision des modèles, d'évolutivité et d'adaptabilité aux scénarios du monde réel.

Dans ce mémoire, nous nous concentrerons sur le développement d'un système en deux parties pour améliorer la prestation des services de santé. Nous mettrons en œuvre un modèle de mémoire à long terme bidirectionnelle (BiLSTM) pour la détection des symptômes, en utilisant l'analyse de texte pour fournir des recommandations de services médicaux précises, et un modèle KNN pour localiser les centres médicaux les plus proches à l'aide de données géographiques telles que les codes postaux et les villes. La performance des deux modèles sera évaluée à l'aide des courbes caractéristiques de fonctionnement du récepteur (ROC) afin de garantir une précision de prédiction élevée.

Les modèles seront implémentés sur un MacBook Pro fonctionnant sous macOS Monterey. Les résultats devraient montrer des améliorations prometteuses en termes d'efficacité et de précision. Les résultats de notre modèle proposé démontrent une grande précision et efficacité. En conclusion, la mise en œuvre réussie de ces modèles améliore considérablement l'efficacité de notre système de soutien de santé, validant son potentiel à améliorer l'accès aux services médicaux et à fournir une assistance rapide basée sur l'identification des symptômes.

ABSTRACT

The importance of healthcare in our lives is undeniable, and rapid advancements continue to transform this field daily. A key area within this evolution is the integration of artificial intelligence and machine learning to enhance healthcare services. Today, healthcare systems can use these AI-driven models to address patient needs, improve accessibility to medical information, and increase the operational efficiency of healthcare providers. However, along-side these advancements, new challenges have arisen, including data privacy, model accuracy, scalability, and adaptability in real-world scenarios.

In this dissertation, we will focus on developing a two-part system to improve healthcare service delivery. We will implement a Bidirectional Long Short-Term Memory (BiLSTM) model for symptom detection, using text analysis to provide accurate medical service recommendations, and a K-Nearest Neighbors (KNN) model to locate the nearest medical centers using geographical data such as postal codes and cities. The performance of both models will be evaluated using Receiver-operating characteristic curves (ROC), to ensure high predictive accuracy.

The models will be implemented on a MacBook Pro running macOS Monterey. The results are expected to show promising improvements in terms of efficiency and precision. The results of our proposed model demonstrate high accuracy (i.e., 98.61% for the training set, 98.55% for the validation set and 98.50% for the test set). In conclusion, the successful implementation of these models significantly enhances the efficiency of our healthcare support system, validating its potential to improve access to medical services and provide timely assistance based on symptom identification.

TABLE OF CONTENTS

LIST OF TABLES

.

LIST OF FIGURES

LISTE OF SYMBOLS AND ABBREVIATIONS

AI	Artificial Intelligence
ADASYN	Adaptive Synthetic Sampling
BI LSTM	Bidirectional Long Short-Term Memory Networks
CNN	Convolutional Neural Networks
DL	Deep Learning
DNS	Domain Name System
ECG	Electrocardiogram
EHRs	Electronic Health Record
IoT	Internet of Things
KNN	K-Nearest Neighbor
LSTM	Long Short-Term Memory Networks
mHealth.	Mobile Health
ML	Machine Learning
NLG	Natural Language Generation
NLP	Natural Language Processing
RNN	Recurrent Neural Network
ROC	Receiver-Operating Characteristic Curve
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
UI	User Interface
UIMA	Unstructured Information Management Architecture
VMA	Virtual Medical Assistance

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

Natural Language Processing (NLP) is an important technological advancement that improves communication within healthcare systems at a time when technology has a significant impact on healthcare. This dissertation explores the integration of NLP in creating a Virtual Medical Assistant Model to improve the healthcare system. Despite these advancements, many healthcare systems are still having difficulty achieving satisfactory patient communication and administrative efficiency. Virtual medical assistants, powered by NLP, promise to address these issues by automating tasks and facilitating accurate patient-provider dialogue. The primary objective of this research is to propose an effective and accurate model for healthcare chatbots that enhances patients' knowledge about their health, ensuring that patients not only receive timely care but also gain a better understanding of their health conditions. In this dissertation, we propose a model based on NLP to enhance the healthcare system. Then we implement and evaluate the performance of the proposed model, method and strategy in terms of responding time and precision in the answers of the chatbots. In this chapter, we first define the basic concepts and significance of NLP in healthcare; secondly, we detail the statement of the problems that we will seek to address, thirdly, we present the objectives of this dissertation; and fourthly, we illustrate the plan of the dissertation.

1.1 Basic definitions and concepts

In this section, we review the basic definitions and main concepts related to NLP in healthcare systems, the role of virtual medical assistants, and the technologies that support these systems, including the AI frameworks and data processing methods that will be used in our proposed models. Moreover, we discuss the type of health data that will be analyzed, how NLP can enhance the management and interaction with this data, and the proposed methods for ensuring the accuracy and efficiency of these interactions.

1.1.1 Natural Language Processing (NLP)

In recent years, significant advancements in research and technology have led to the development of Natural Language Processing (NLP) to enhance healthcare system operations. As a key component of the digital transformation of healthcare systems, natural language processing (NLP) aims to address the communication gap between human languages and machine understanding. Natural Language Processing (NLP) is defined as a crucial subfield within computer science and artificial intelligence, which deals with the interaction between computers and human natural languages. The fundamental aim of NLP is to enable computers not only to understand and interpret human language but to generate it as well, thereby facilitating meaningful and effective communication [1]. Natural language processing (NLP) is revolutionizing the healthcare sector by automating the gathering and analysis of medical data, which is mainly unstructured and includes patient feedback, prescription information, and clinical notes. The application of NLP technologies enables healthcare systems to process vast amounts of data efficiently, helping to improve diagnostic accuracy, patient care, and operational efficiencies [2]. Techniques like machine learning and deep learning within NLP have further enhanced its effectiveness, allowing for a more nuanced understanding and generation of human language. This capability is pivotal in developing systems that can interact naturally with users, providing them with reliable and accessible healthcare information and services [3].

1.1.2 Virtual Medical Assistants

Virtual Medical Assistants (VMAs) are AI-driven platforms designed to simulate human conversation and enhance healthcare services. They are programmed to assist patients by providing timely health-related information, scheduling appointments, and facilitating seamless communication between patients and healthcare providers. These assistants use advanced algorithms from the field of NLP to interpret and generate human-like responses, making interactions as natural and helpful as possible. VMAs engage patients by providing instant responses to their inquiries, which is crucial for maintaining continuous communication, especially for chronic condition management. Rathore *et al*. [4] emphasize how virtual assistants enhance patient engagement by offering personalized interaction and easy access to healthcare information, which improves patient satisfaction and adherence to treatment protocols. Administrative efficiency in healthcare settings is markedly improved through the automation of routine tasks such as appointment scheduling and patient data management. Prof. D. L. Falak [5] highlights that generative AI, utilized by these virtual assistants, helps in automating form-filling tasks, collecting patient data for various uses, and integrating seamlessly with Electronic Health Records (EHRs), thereby streamlining administrative processes and reducing the workload in the healthcare sector.

Virtual Medical Assistants are integral to modern healthcare systems, utilizing AI to automate and optimize a variety of tasks traditionally performed by human staff. This includes:

1. Information Provision: VMAs provide patients with accurate and personalized health information, helping them understand their health conditions and the necessary steps for care.
2. Appointment Scheduling: These assistants automate the scheduling process, allowing patients to book, reschedule, or cancel appointments without human intervention, thus improving accessibility and efficiency.
3. Communication Facilitation: VMAs ensure that patient inquiries are responded to accurately and quickly, addressing the communication gap, which is crucial for patient engagement and satisfaction.

The implementation of Virtual Medical Assistants in healthcare settings revolutionizes the patient care model by making healthcare more accessible. Moreover, VMAs enhance the patient experience by providing 24/7 support, reducing wait times for assistance, and personalizing patient interactions based on previous data and interactions. Virtual Medical Assistants are a testament to the advances in AI and NLP within the healthcare industry. They are continuously enhanced with the latest in machine learning algorithms to improve their ability to understand and process complex medical dialogue. As these technologies progress, the capability of VMAs to offer advanced services, such as predicted health monitoring and customized health management advice, highlights their increasing significance in the healthcare industry. Integrating Virtual Medical Assistants into healthcare systems enables facilities to optimize operations and enhance the quality of patient care, proving the revolutionary influence of AI in healthcare.

1.1.3 Healthcare Systems

Healthcare systems consist of a wide-ranging network of organizations, institutions, and professionals responsible for providing medical services to patients. These systems increasingly rely on digital technologies to streamline operations and communication within the healthcare industry. Digital healthcare systems integrate various components, such as electronic health records (EHRs) and patient management systems, facilitating more effective coordination and data management.

NLP is applied to many tasks that involve functions as simple as appointment scheduling or as complicated as patient flow and clinical recording. AI and NLP are two technologies that assist in handling big data and enhance individual data processing by making it more efficient. This has a positive impact both on the efficiency of healthcare services and on patients.

- Operational Efficiency: AI technologies, including NLP, transform healthcare by automating tasks that traditionally require human intervention, thus reducing errors and saving time. For example, AI applications in medical imaging and interactive response systems are enhancing diagnostics and patient care, showing a notable impact on healthcare accessibility and the quality of care provided [6].
- Patient Data Management: AI and NLP significantly contribute to the management of patient data by improving the accuracy and efficiency of electronic health record systems. These technologies enable sophisticated data analysis capabilities, such as sentiment analysis and emotion categorization, which support clinical decision-making and foster more empathetic and personalized care plans [7].
- Enhancing Patient Care: The application of AI and NLP in healthcare goes beyond administrative efficiency, impacting patient care directly by facilitating enhanced communication tools, predictive health analytics and personalized treatment options. This not only improves the delivery of healthcare services but also ensures that patients receive more tailored and effective treatments [8].

1.1.4 Overview of Machine Learning (ML) Techniques in NLP and Related Concepts

Machine Learning (ML) plays a fundamental role in the field of NLP by enabling systems to automatically learn and improve from experience without being explicitly programmed. ML techniques are crucial for the development of models that can process, analyze, and generate language in ways that mimic human abilities [9].

Core ML Techniques in NLP:

- Supervised Learning: This involves training a model on a labeled dataset, where the model learns to predict outputs from input data. It's widely used in NLP for tasks like text classification and sentiment analysis [9].
- Unsupervised Learning: This model learns to identify patterns and structure from unlabeled data, which is useful in NLP for clustering and topic modeling [10].

- Reinforcement Learning: This involves models that learn to make sequences of decisions by receiving feedback on their actions, applicable in dialogue systems and interactive NLP applications [11].

Regarding all the concepts mentioned earlier, we can conclude that virtual medical assistants powered by Natural Language Processing (NLP) are pivotal in transforming healthcare communication and management. They are the main focus of this dissertation. For this purpose, we propose a model based on a combination of simulation and health data to evaluate the effectiveness of NLP techniques in enhancing healthcare systems. Our approach involves the use of advanced NLP tools and machine learning frameworks to simulate interactions between patients and virtual assistants, in order to improve the accuracy and responsiveness of these systems. NLP and machine learning technologies allow for the automation of healthcare administrative tasks and improve patient-provider communications. These technologies provide the backbone for our testbed, which includes various NLP-driven scenarios to assess how virtual assistants can enhance patient engagement and streamline healthcare operations. The testbed will use popular NLP frameworks such as TensorFlow and Spacy, which support the implementation of complex NLP models and are widely recognized for their efficiency and scalability [12]. TensorFlow and Spacy are integral to the development of advanced NLP applications within healthcare systems. TensorFlow, a robust machine learning framework developed by Google, enables the design and training of deep learning models that can effectively process and analyze large datasets. Spacy, on the other hand, is optimized for high-performance NLP tasks and is widely recognized for its efficiency in processing and understanding human language. TensorFlow and Spacy provide the computational power and sophisticated algorithms necessary for developing virtual medical assistants. These tools support the implementation of complex NLP models that can interpret clinical dialogue, extract medical information from unstructured data, and automate interactions between patients and healthcare systems. The integration of these technologies enhances the responsiveness and accuracy of virtual assistants, ensuring they can handle the nuances of medical communication effectively [13]. For instance, TensorFlow's ability to manage large-scale neural networks makes it suitable for tasks that require analyzing patterns in vast amounts of health data, such as predictive analytics and patient monitoring. Spacy contributes by offering powerful syntactic analysis, helping parse patient input to extract relevant medical entities and information swiftly [14].

1.2 Problem Statement

While evaluating the literature within the context of healthcare systems, it was revealed that Virtual Medical Assistants supported by NLP experienced certain issues, and it was determined that the current methods had certain drawbacks. Addressing these challenges is crucial for improving the effectiveness and utility of VMAs in providing patient care and administrative support. However, we can take into account these critical issues:

- Database Limitations: A major challenge is the insufficient datasets available for training the model, which limits their ability to perform accurately and handle diverse medical inquiries. Enhancing the database with more comprehensive and varied health data can significantly improve the VMA's understanding and response accuracy [15][16].
- Limited Patient Education: VMAs often fail to provide comprehensive and understandable health information to patients, which is crucial for patient education and compliance. This could be improved by integrating more patient-friendly communication modules and educational materials that are easy to understand and follow [15-17].
- Inefficient Administrative Task Handling: VMAs struggle with handling administrative tasks efficiently, which often leads to a lack of availability for direct patient care. Optimizing these systems to handle administrative functions more swiftly and effectively can free up valuable time for healthcare providers to focus more on patient care [19][20].
- Multilingual Support and Cultural Sensitivity: The lack of support for multiple languages and cultural sensitivities restricts access to VMAs for non-English speakers. Expanding the linguistic capabilities of VMAs to include multilingual support can enhance accessibility and usability for a broader patient base [16][17].
- Scalability Issues: As the demand for VMAs grows within healthcare systems, the need for scalable solutions becomes increasingly critical. Existing systems often struggle with handling large volumes of interactions efficiently, which leads to delays and reduced functionality. Our research will explore methods to enhance the scalability of VMAs to ensure they can handle growing user bases without decreasing in performance [21][22].

Considering these problems, our main research question is the following: Which model can be used to improve healthcare chatbots in healthcare? From this primary question, derive the following two secondary questions:

1- Which method can be used for healthcare chatbots while helping healthcare providers with administrative tasks to have more time and resources available for directly caring for patients at an acceptable cost?
2- What challenges do both patients and doctors face when using healthcare chatbots, and how may these challenges be dealt with to maximize the advantages of using chatbots in healthcare?

1.3 Research Objectives

The main objective of this dissertation is to propose an effective and accurate model for healthcare chatbots to enhance patients' knowledge about their health.

More specifically, this research aims to:

1- Propose a method used for healthcare chatbots while helping healthcare providers with administrative tasks to have more time and resources available for directly caring for patients at an acceptable cost;
2- Design an architecture for addressing the challenges of using healthcare chatbots by patients and doctors, and maximizing their benefits in healthcare;
3- Implement the proposed method and architecture;
4- Evaluate the performance of the proposed method and architecture in terms of responding time and precision in the answers of the chatbots.

1.4 Dissertation plan

The remainder of this dissertation is organized as follows. Chapter 2 presents the literature review regarding the virtual medical assistant systems, discussing their functionalities and identifying challenges (i.e., data management, patient engagement and multilingual support), as well as exploring recent advancements in NLP and AI that could potentially enhance VMA systems. Chapter 3 details our proposed model for healthcare chatbot, describes the data collection processes, NLP techniques and algorithms employed in developing the virtual medical assistant systems, and discusses the architecture of the proposed model and the testbed setup. Chapter 4 discusses the results and the evaluation of the performance of the proposed model. Finally, Chapter

5 summarizes the research work, outlines its limitations, and offers some indications for future direction.

CHAPTER 2 LITERATURE REVIEW

In this chapter, we will present a literature review of healthcare systems, healthcare chatbots and research issues and challenges. Then, we will investigate healthcare chatbots in depth, to evaluate their potential as innovative instruments for change in the health care industry. Our research emphasizes the numerous aspects of healthcare chatbots, including their functionality, and impact on patient care and engagement. We start our investigation with an in-depth review of existing literature on healthcare chatbots, examining their developmental history, features, and applications in the field of medical service. We aim to provide an advanced overview of the current state of healthcare chatbot technology by precisely integrating previous research works and scientific discussions. Moreover, we examine the potential advantages and limitations of integrating healthcare chatbots into medical processes and patient care systems. We aim to identify the impact of healthcare chatbots on improving access to medical knowledge, increasing patient involvement, and optimizing healthcare service delivery through practical studies and contextual analysis.

2.1 Healthcare System

There is a considerable amount of related work on healthcare systems, including key developments and pursuing study. The objective of this section is to perform an in-depth review of relevant works, with a focus on healthcare service models, especially on significant results and contributions in healthcare chatbots. The obvious importance of healthcare in our lives highlights the field's continuous advancement. Healthcare systems are at the heart of providing important services for individuals all over the world, with a diverse range of healthcare delivery strategies. These models range from traditional hospital-based care to new community health initiatives, each with distinctive advantages and challenges. Furthermore, the integration of advanced technologies and information systems has revolutionized healthcare infrastructure, facilitating the efficient management of patient records, enhancing communication between healthcare providers, and improving clinical decision-making processes. However, with these advancements come new challenges, including ensuring the accuracy of healthcare system responses, facilitating seamless booking of appointments, managing administrative tasks efficiently, and ensuring user-friendly interfaces for both healthcare providers and patients. Despite these limitations, integrating technology into healthcare systems creates new opportunities for development and innovation.

Healthcare systems can improve overall healthcare outcomes by adopting new technologies like artificial intelligence [23].

2.2 Natural Language Processing (NLP) in Healthcare

Natural Language Processing is a field of study that combines computer science, artificial intelligence (AI), and language. The main emphasis is on the natural language exchanges that take place between computers and humans. It aims to make it possible for computers to generate, comprehend, and analyze human language in a beneficial and important way [24]. NLP is useful in a broad range of computer and artificial intelligence applications, from basic tasks like keyword searching and spell checking to complex procedures like sentiment analysis, language translation, and conversational agent creation. In addition to reducing the gap between human and machine learning, NLP improves the capacity of computers to process large volumes of textual input, allowing for the extraction of insights and the automation of answers.

The development of NLP from a theoretical idea to a fundamental technology is a history of ambitious objectives, creative discoveries, and profound changes in the ways that language is understood. Rule-based models characterized the early stages of NLP research in the 1950s and 1960s, as scientists manually programmed language rules into computers. The earliest machine translation systems and the initial investigations in computational languages were developed at this time. Nevertheless, rule-based techniques became more difficult and restricted in their use due to the complexity and diversity of human language. A new era for NLP was welcomed by the development of statistical techniques in the late 1980s and early 1990s, which used mathematical models to infer language patterns from massive datasets. The possibilities and uses of NLP were greatly enhanced by this move toward data-driven methods. NLP gained more momentum in AI research with the advent of machine learning algorithms, which allowed for advanced vocabulary generation and comprehension based on context and intent [25].

The development of deep learning techniques has revolutionized NLP during the past ten years, leading to the creation of models that can process and produce human language with never-before-seen accuracy and fluency. Natural language processing is becoming more and more integrated across many industries as a result of technologies like Google's BERT and OpenAI's GPT series, which have shown impressive gains in language generation and understanding. NLP has become a critical technology in the healthcare industry, transforming operations and services. NLP has opened up fresh opportunities in patient care, research, and administration, from analyzing clinical notes for meaningful information to powering conversational agents for patient support. NLP is essential to the continuous development of the healthcare industry because of its capacity to sort through and comprehend vast amounts of medical literature and patient data, which is enables improvements in diagnosis, treatment, and patient engagement [26].

2.2.1 Core technologies and algorithms in NLP

In the exploration of core Natural Language Processing technologies and algorithms, significant contributions have been made, underlining the essential role of machine learning and deep learning in advancing NLP applications, particularly in healthcare. The literature has extensively explored the potential of deep learning algorithms, like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to execute delicate NLP tasks [27]. For instance, Authors provided an overview of popular deep learning methods including CNNs, RNNs, Restricted Boltzmann Machine, and Autoencoders, emphasizing their significance in the field of NLP. Deep learning approaches have revolutionized machine learning of human language, leading to significant improvements in domains like sentiment analysis, language translation, and automated summarization. These techniques are recognized for their depth of processing and pattern recognition skills. Deep learning offers improved learning from massive data sets using complex neural networks, which makes it more useful than traditional algorithms [28].

Researchers discussed a range of deep learning algorithms, including CNN, LSTM, and others, highlighting their diverse applications not just in healthcare but across various domains. These algorithms have been demonstrated to be incredibly successful in analyzing delicate medical data, which has made an important contribution to patient care plans and healthcare forecasting models.

Moreover, the advancements made in deep learning have brought about innovative uses in silent audio technology, indicating the broad impact of deep learning in several fields [29]. Authors illustrate the application of deep learning in Silent Sound Technology, proposing its use to improve communication and diagnostic procedures within healthcare. These new perspectives on NLP technologies and algorithms emphasize how active the field is, with continuous research and development pushing the limits of machine learning and deep learning applications in natural language processing and healthcare [30].

2.3 NLP Applications in Healthcare

The integration of Natural Language Processing in healthcare has facilitated a transformative shift in how clinical data is processed, analyzed, and utilized for enhancing patient care, research, and operational efficiency. Automated clinical documentation, a cornerstone application of NLP, streamlines the process of converting doctor-patient interactions into structured electronic health records (EHRs).

Wang *et al.* [22] emphasized that NLP could streamline clinical note transcription and classification, greatly reducing the administrative burden on healthcare providers and allowing them to concentrate more on patient care instead of paperwork. It aims to provide specialized NLP methods for clinical use, demonstrate NLP's utility in clinical research at organizations like the University of Minnesota and Mayo Clinic, encourage collaboration among diverse NLP researchers to advance clinical NLP research, and simplify clinical environments for those without medical backgrounds to broaden understanding and participation. The methodology includes a structured approach to comprehensively address clinical NLP. It establishes a Big Data Infrastructure for analyzing large datasets and explores recent NLP advancements through case studies like family history extraction. Additionally, it uses tools like UIMA and cTAKES for Clinical Information Extraction and examines methods for identifying patient groups using electronic health records, all demonstrated through practical case studies. The study includes several limitations that warrant acknowledgment., there is the limitation of restricted access to EHR data, which may provide challenges for researchers in the broader natural language processing field because of privacy concerns and the complexity of medical language. Furthermore, the tutorial's comprehensive covering in a brief period of time could lead to a general overview instead of a thorough investigation of specific methods or tools. NLP-based patient interaction analysis makes it possible to glean insightful information from patient conversations, supporting the provision of individualized care and treatment optimization.

Aramaki *et al.* [31] illustrated how NLP technologies can analyze patient feedback, queries, and interactions across clinical settings, providing a deeper understanding of patient needs and experiences. The main goals of this study are to explore and use NLP technologies in various healthcare settings. This includes improving patient care by better understanding clinical information and developing new NLP tools for healthcare. The focus is on making these tools accurate, scalable, and useful for real-world applications. The study also explores the literature concerning potential NLP products and services applied to a wide range of medical, clinical, and healthcare areas. It delves into various clinical applications of NLP across different departments, such as internal medicine, surgery (pre- and post-), oncology, radiology, pathology, psychiatry, rehabilitation, and obstetrics and gynecology. In this study, data from electronic health records and clinical notes are collected and analyzed. Researchers then develop NLP tools and algorithms, like machine learning models, to analyze and extract useful information from the data. Real-world case studies are also conducted to show how these NLP technologies can be used in healthcare. The application of NLP in healthcare faces several challenges, including data security and privacy concerns, interoperability issues between healthcare systems, and the complexity of medical language. There are also difficulties integrating NLP into the daily workflows of healthcare providers and the need for NLP systems to accurately handle the specificity of clinical terminology. Another crucial application for NLP is literature mining, which allows researchers to find patterns, gaps, and possibilities in the massive amounts of medical literature and databases they discover.

Uzuner *et al.* [32] discussed in extensive detail Biomedical/Clinical NLP in their review, covering a wide range of topics including data sources, research topics, difficulties in managing biomedical texts, particular word meaning research interests, dataset availability, annotation procedures, and primary processing techniques. This thorough summary offers insights into the field's fundamentals, procedures, and applications. The authors offered a thorough review of biomedical/clinical NLP in the methodology part, including data sources, research areas, and difficulties that arise when working with biomedical and clinical literature. They explored the kinds of queries that scholars in various fields, such as linguistics and semantics, propose. To provide clarity and context, they described the three main methods for processing biomedical and clinical text: rule-based, statistical, and hybrid techniques. These methods are supported by examples from current research projects. The authors emphasized difficulties in Biomedical/Clinical NLP when discussing limitations. These include data accessibility and privacy restrictions, complicated language structures, methodological variety, and ongoing research requirements for continued development and innovation in the sector.

The impact of NLP technology on patient care, accuracy, and efficiency is demonstrated by case studies of successful deployments in the medical field. Névéol and Zweigenbaum [12] indicated how NLP may be used to improve healthcare delivery systems by providing an introduction to the fundamental techniques that enable effective healthcare using NLP. These techniques include applications in diagnostic support, treatment planning, and patient tracking. However, Névéol and Zweigenbaum in [33] highlighted that clinical NLP research is evolving towards complex text analysis and showcases five studies using clinical NLP on diverse text sources like online forums, encyclopedias, and EHRs aimed at practical outcomes. It covers advancements in word sense disambiguation, negation detection, and temporal analysis, which help address critical clinical issues such as cancer patient triage. Additionally, it explores effective health communication strategies through research on clinically relevant online There are still many difficult NLP challenges to be addressed, such as accurately understanding clinical texts from EHRs and extracting useful insights. Additionally, the requirement for robust de-identification procedures ensures patient confidentiality while enabling enhanced access. The integration of NLP in healthcare applications indicates significant advances in the application of AI and machine learning technologies to enhance patient care, clinical procedures, and medical research. In addition, healthcare is becoming more accurate, data-driven, and efficient because of NLP's automated clinical paperwork, patient interaction analysis, and literature mining.

2.3.1 Healthcare chatbot

As mentioned before, healthcare chatbots are one of the most popular sectors in health services, with many articles on the subject. In this section, we are going to investigate some of the articles in this domain. Healthcare Chatbots use Artificial Intelligence that can make a human-system interaction to resolve basic queries regarding health parameters before consulting a doctor. The actual purpose behind this work is to work on the user's symptoms and to provide medical suggestions according to it to reduce the time and cost required for the process. The chatbot works based on the user's input. It takes phrase keywords, generates decisions to solve the user's inquiry, and responds correctly. This chatbot can be used by regular humans in any kind of emergency case, where it can play the role of medical consultant to people about primary care before contacting a doctor, or it can work as a doctor for small and short-term health conditions like cold, headache, etc.

The goal of the authors in this research work [34] was to transform the relationship between technology and healthcare. The primary objective of this project is to develop an advanced healthcare chatbot that facilitates seamless human-system interactions. Users can utilize the chatbot to address their basic health questions, achieving the objective of resolving these concerns prior to seeking professional medical consultation. The platform aims not only to use the user's symptoms to offer unique medical recommendations but also to significantly reduce the time and financial costs typically connected with the healthcare procedure. At the heart of this system lies the strategic employment of NLP, which ensures effective communication by enabling the chatbot to interpret user inputs based on sentence keywords, thereby providing accurate and relevant responses. Furthermore, the system uses complex machine learning algorithms, such as TF-IDF, stemming and n-grams, to analyze and understand user inputs comprehensively. This technological foundation supports a range of critical functionalities, including disease detection based on symptoms, provision of medical term explanations, doctor recommendations, treatment schedules, and comprehensive health parameter tracking and monitoring. The healthcare chatbot, designed for both offline and online use, highlights system versatility and accessibility. It was developed with a comprehensive approach using advanced technologies, including NLP, to ensure its effectiveness and usability. This technology enables natural and consistent chats with patients, regardless of language, expanding the chatbot's accessibility to a larger, broader user base. Machine learning techniques form the backbone of the chatbot's analytical capabilities. Techniques such as TF-IDF analysis are utilized for understanding the importance of words in the user's input, while stemming algorithms reduce words to their root form. The system uses n-grams for word sequence prediction and cosine similarity for sentence assessment to deliver relevant and precise chatbot responses. Moreover, it supports critical functions like creating user profiles for personalized medical suggestions, doctor recommendations, and medication reminders, enhancing user experience. To verify the system's quality and performance, comprehensive tests were carried out, followed by statistical analysis to ensure dependability and efficacy. The proposed approach is just showing how this chatbot could change healthcare, being a low-cost, available, and easy-to-use tool while improving the accuracy of the classification of adults. However, critiques highlight issues with the chatbot's question accuracy and its lack of coverage for a wide range of health issues. Additionally, the system's limitation in not catering to users under 18, excluding youngster’s considerations from the system's range is an important issue. Future updates need to make the chatbot more inclusive, providing accurate medical advice for all ages, expanding disease coverage, and enhancing Q&A functions for greater reliability and user diversity [34].

Goel *et al.* [35] presented a neural network-driven medical chatbot that can provide essential information, diagnose diseases, and advise whether one needs to see a medical consultation. Advanced NLP techniques used in the chatbot facilitate smooth expression of health problems by the customers and allow for accurate responses. It will also analyze the sentiments of the customers to improve its conversation. This would render a straightforward, user-friendly, 24/7 health advice platform that would reduce web searches and physical consultations. Such an interface targets being available instantly and hassle-free across the globe, employing sentiment analysis to make the response from the chatbot even more empathetic in nature. One of the pivotal advantages of employing chatbots in healthcare is their ability to offer continuous, 24/7 support. Moreover, the authors proposed a comprehensive plan to develop a medical chatbot that leverages the latest advancements in Neural Networks and NLP. The intended outcome is to improve patient engagement, streamline the information dissemination process, and ultimately, transform the way individuals seek and receive medical advice. The NLP used here enhances the system's understanding of user queries so that the chatbot can reply naturally and intuitively. Data preparation and sentiment analysis make the response much more accurate and empathetic; hence, the chatbot will meet the user's needs with effectiveness. A friendly user interface will support users' signup, login, and interaction easily for rapid access to health information. Text preprocessing: Text is broken into words, punctuation and stop words are removed to make the data simpler. It provides fast feedback given by an expert system for an efficient and accessible GUI using a python tkinter module. Medical chatbots are yet to overcome some challenges regarding accuracy and reliability, especially on diagnosis precision and quality of training data. These are just but some of the challenges that proof intricacy in increasing health chatbots [35].

Rajesh *et al*. [19] presented a comparative study of different ML algorithms for intent classification in chatbots, aiming to evaluate the performance of four ML algorithms: logistic regression, random forest, support vector machines (SVM), and deep learning (DL) neural networks. In their study, a comparison analysis is conducted to assess the effectiveness of several machine learning algorithms for intent classification in chatbots. The focus lies on assessing the performance of logistic regression, random forest, SVM, and deep learning neural networks. Emphasis is placed on investigating the influence of data preprocessing techniques and feature engineering on the models' performance to identify optimal strategies for enhancing chatbot intent recognition. Additionally, an innovative approach is proposed for constructing healthcare-oriented chatbots utilizing SVM.

This approach demonstrates SVM's proficiency in accurately classifying and predicting user intent from textual input, thereby assisting chatbot responsiveness and accuracy, particularly within healthcare contexts. Furthermore, it explored how to integrate them with backend systems for integrations, including electronic health records and appointment-scheduling software, to provide more relevant information to the users. The study also explored into the challenges and opportunities inherent in healthcare chatbot development, encompassing ethical considerations, technological obstacles, and user acceptance.

Lastly, the technical facets of chatbot construction, including NLP, ML algorithms, and chatbot architectures, are thoroughly examined to provide a comprehensive understanding of the landscape.

The methodology for developing a chatbot typically involves four key steps [15]:

1. Dataset Preparation encompasses collecting, cleaning, and preprocessing data, often sourced from real-world interactions or publicly available datasets.
2. Feature Engineering extracts relevant features from the text data using techniques like tokenization and TF-IDF. Following this, Model Training involves utilizing various machine learning models such as logistic regression, random forests, SVM, and deep learning neural networks, with data divided into training and test sets.
3. Model Evaluation assesses model performance using metrics like accuracy and F1 score. Optimization techniques like hyperparameter tuning are then applied based on evaluation results, leading to the selection of the best-performing model.

Integration and deployment involve integrating the chosen model into the chatbot framework and deploying it, considering factors like interfacing with backend systems and designing user interfaces. Additionally, the chatbot development process entails identifying its purpose and goals, understanding the target audience and domain, developing conversation flows, and implementing technical aspects such as NLP and chatbot architecture to ensure accurate information provision.

Rajesh *et al*. [19] identified several key challenges surrounding the development and deployment of chatbots in healthcare as follows:

Despite advancements in AI, the dependence on human oversight persists as a significant limitation. The chatbot often demands regular monitoring and intervention to handle complex queries, manage errors, and incorporate updates in response to evolving medical knowledge or user feedback.

The accuracy and effectiveness of machine learning models are influenced by the limitations inherent in the datasets used for training and testing. Factors such as dataset size, quality, and diversity, as well as biases and representational gaps, can impact the model's performance in real-world applications.

The complexity of the models employed presents challenges in terms of computational resources, scalability, and efficiency in deployment environments.

The dynamic nature of language presents a continuous challenge, as models may struggle to adapt to new phrases, and expressions without ongoing updates to the training data. Addressing these challenges is crucial for enhancing the chatbot's friendliness, accuracy, and helpfulness, ensuring a more effective and reliable tool for assisting patients while acknowledging that not all questions may be accurately addressed.

Menon *et al.* [36] developed a comprehensive system leveraging deep learning techniques for the transcription and summarization of clinical conversations between healthcare professionals and patients. This involves accurately converting spoken dialogue into text form and reducing it to brief overviews that contain crucial details such as diagnoses, treatments, and patient concerns. These might also involve automating such processes that could further enhance efficiency and accuracy in clinical documentation, apart from reducing the administrative burden on health professionals. Noise reduction techniques will be applied to improve the quality of audio recording and transcription. Additionally, the authors investigated the efficacy of various deep learning models, including convolutional neural networks and support vector machines, to identify optimal approaches for transcription and summarization tasks. Ultimately, the implementation of AI and NLP-driven automation seeks to streamline administrative tasks during doctor-patient interactions, allowing healthcare professionals to devote more time and attention to delivering quality patient care. Moreover, the development of a noise suppression mechanism aims to ensure that only essential information is retained from recorded clinical conversations, further enhancing the system's utility and effectiveness. The methodology outlined involves a systematic process for enhancing the transcription and summarization of clinical conversations.

The dialogue between doctor and patient is recorded and stored in an appropriate format. Noise suppression techniques are applied using a Convolutional Autoencoder architecture, comprising multiple convolutional and deconvolutional layers, effectively reducing unwanted sounds.

Subsequently, the cleaned audio file undergoes conversion into text via the Google Speech-to-Text API, ensuring rapid and accurate transcription. The transcribed text is then utilized as input for a summarization model, which generates a tabular summary highlighting essential medical terms and details.

Further refinement is achieved through the implementation of a supervised deep learning technique, utilizing a convolutional network for text categorization based on provided tags and word frequency. This categorization aids in creating a condensed, tabular summary of the conversation. Additionally, significant words are filtered from the transcribed conversations, resulting in a refined dataset suitable for various applications, thus enhancing the utility of the transcribed data for further analysis and decision-making processes.

Menon *et al*. [36] indicate several limitations, identifying areas for future research and enhancement. Despite leveraging advanced speech-to-text technologies, the transcription process encounters difficulties in accurately capturing certain words or phrases, particularly with diverse accents or different languages. This underscores the need to enhance the adaptability and accuracy of the speech recognition model across various speech patterns. While extracting main ideas, may not fully capture the complexities of medical conversations; advanced NLP could improve this. Transcription and summarization depend on diverse, unbiased training data to avoid inaccuracies. Additionally, effectiveness varies by medical specialty, requiring adjustments for specific terminologies and patterns.

Garg [37] is dedicated to crafting a drug recommendation system to address prevalent challenges in the healthcare sector, notably the scarcity of specialists and the escalating trend of self-medication due to limited access to medical consultation. Key objectives include easing the burden on healthcare professionals by offering a system capable of recommending drugs based on sentiment analysis of patient reviews. Additionally, the system elevated patient care standards by utilizing machine learning to accurately predict sentiments from drug reviews and suggest the most suitable medication for specific conditions. A pivotal aspect of the research involves advancing the utilization of various machine learning classifiers and vectorization techniques, such as Bag of Words, TF-IDF, Word2Vec, and Manual Feature Analysis, to investigate patient reviews and forecast sentiments. This attempt highlights the transformative potential of machine learning in revolutionizing drug recommendation processes within the healthcare domain. The methodology includes analyzing sentiment from patient reviews and building a drug recommendation system. It starts with data preparation like cleaning and visualization. Text data is processed with vectorization techniques such as Bag of Words and TF-IDF. Machine learning classifiers like LinearSVC and Logistic Regression are used for sentiment prediction, with their performance evaluated by metrics like precision and recall. The recommendation system uses the best models and considers review 'useful count' to recommend drugs based on sentiment scores from reviews. The methodology also involves text processing techniques like tokenization and lemmatization, and machine learning algorithms like Perceptron for binary sentiment classification, ensuring robustness through various evaluation metrics. The research identifies several limitations that pave the way for future improvements and exploration. The inaccuracies in performance metrics between positive and negative classes indicate data imbalance within the training set. This imbalance can distort sentiment predictions, favoring the overrepresented class. Techniques like SMOTE, ADASYN, are recommended to rectify this imbalance and enhance model performance. Garg 's models [37] could benefit from meticulous hyperparameter optimization to fine-tune their parameters and support prediction accuracy. The absence of such optimization may have constrained the performance of the sentiment analysis classification algorithms. The current recommendation framework aggregates the best-predicted result from each method directly. However, for more robust and complex results, a thorough ensemble of different predicted outcomes is imperative. This could accumulate the strengths of individual models to elevate the overall accuracy and dependability of recommendations.

Zhou *et al*. [38] explored a comprehensive review of existing studies on NLP within the smart healthcare domain, focusing on both technical methodologies and practical applications. It defines various NLP approaches and the tailored NLP pipeline specific to smart healthcare, underscoring NLP's pivotal role in analyzing and comprehending human language. Furthermore, the paper introduces representative smart healthcare scenarios where NLP techniques find application, covering clinical practice, hospital management, personal care, public health, and drug development. Additionally, the paper explores the significant impact of NLP-driven smart healthcare in addressing two pressing medical issues: the COVID-19 pandemic and mental health challenges, highlighting the transformative influence of NLP in these critical areas. Moreover, the paper categorizes NLP methods into rule-based, statistical, and neural approaches, accompanied by a detailed comparative analysis that elucidates the strengths and limitations of each approach.

This comprehensive examination aims to showcase the diverse applications of NLP in healthcare settings, emphasizing its relevance and potential in addressing multifaceted challenges across various healthcare domains. Zhou *et al.* [38] adopt a thorough review methodology, examining and contrasting various Natural Language Processing (NLP) approaches, including rule-based, statistical, and neural NLP, alongside their applications in healthcare settings. It explores the technical intricacies of NLP, encompassing feature extraction, modeling, and the NLP pipeline, highlighting their significance within smart healthcare contexts. The NLP pipeline is deconstructed into three pivotal components: preprocessing, feature extraction, and modeling. Preprocessing techniques such as tokenization, stemming, and lemmatization are elaborated upon extensively, underscoring their crucial role in text data preparation for NLP tasks within healthcare domains. The authors highlighted several notable limitations within the realm of NLP in healthcare as follows:

1- It emphasizes the complex nature and ambiguity of human language, which present hurdles in accurately interpreting biomedical texts.
2- It raises concerns regarding the interpretability of neural NLP models, particularly crucial in clinical settings where transparency and reliability are paramount for decision support systems.
3- The practical implementation of NLP-driven applications in healthcare encounters various challenges, including computing power requirements, training costs, and ethical considerations surrounding patient privacy. These limitations underscore the imperative for further research aimed at overcoming these obstacles and fully harnessing NLP's potential for smart healthcare applications.

While Zhou *et al.* [38] comprehensively cover a wide range of applications in smart healthcare, it may benefit from deeper exploration into specific subfields within the healthcare domain. Exploring the user experience aspects of NLP-driven healthcare applications could provide valuable insights into enhancing their effectiveness and usability in real-world scenarios. Thus, a comprehensive approach that addresses these considerations is vital for maximizing the efficiency and applicability of NLP in healthcare settings.

Kamra *et al*. [39] proposed a cognitive disease prediction model focusing on psychological disorders like bipolar disorder, obsessive-compulsive disorder, and schizophrenia. It aims to provide early diagnosis through NLP and machine learning techniques, leveraging audio files of user symptoms for prediction. This approach intends to save time and resources for both healthcare professionals and patients, emphasizing the potential of AI and NLP in enhancing diagnostic processes and treatment planning in mental health care. The study's methodology utilizes NLP and ML to create a cognitive disease prediction model via speech recognition for predicting psychological disorders. It involves feature extraction, then dictionary creation, and finally AWD-LSTM neural network classification using TensorFlow. This approach focused on early detection of conditions like bipolar disorder, OCD, and schizophrenia by analyzing speech patterns. The process begins with preprocessing the dataset and audio inputs into a standardized CSV format. Then, model development was started, and for that, AWD-LSTM was used as the core, where the dataset was split 70/30 for training and testing. This, of course, goes hand in hand with the intensive training and testing of the model on the performance-reliability trade-off. The model predicts cognitive ailments with audio input and gives a percentage accuracy of the prediction. However, limited dataset diversity and lack of prediction interpretability may impact its effectiveness. Expanding the dataset, testing in real-world settings, and developing a user-friendly interface are crucial. These improvements could significantly refine the model's applicability and trustworthiness in clinical practices

Chkirbene *et al*. [40] proposed several ambitious objectives intended to transform the offering of healthcare services by incorporating modern technology. The study's main goal is to create an advanced intelligent healthcare model that uses machine learning methods to precisely identify aberrant electrocardiogram (ECG) signals. To improve the scalability of healthcare services and reduce wait times in emergency rooms, this project is essential for providing prompt and effective patient care. In addition, the study provided an entirely novel feature selection process intended to maximize machine learning model performance. The system presented how to accelerate the process of evaluating complicated health information by locating and identifying the most important elements within ECG data for indexing and classification. Moreover, the authors [20] suggested developing a data-dependent indexing method. This innovative method streamlines the exchange and storage of medical data by converting high-dimensional ECG data into a distinct index code. The essay predicts a transformational impact on healthcare infrastructures through these aims, allowing them to maintain high standards of service and operational efficiency while managing an increasing quantity of patient data. In addition, the authors employed advanced techniques to manage healthcare data, using the PTBDB ECG dataset for initial data structuring. A feature selection algorithm extracts key ECG characteristics, essential for data simplification. Various machine learning methods, including SVM, are used for ECG classification, with SVM selected for its accuracy. A new indexing method is introduced for efficient data storage and retrieval, enhancing cloud database management. These methods combine edge computing and machine learning to improve healthcare, especially for urgent care where quick accurate decisions are crucial. Limitations include the need for dataset balancing, hyperparameter optimization, and potentially better ensemble methods for performance enhancement. Despite these issues, the system shows promise in detecting irregular heartbeats and lowering data collision rates, underscoring the transformative potential of merging edge computing with machine learning in healthcare. Future research is called for to refine the system for everyday healthcare use.

Agarwal *et al*. [41] developed a user-friendly system that engages with individuals to address basic health inquiries before seeking professional medical consultation. The chatbot would use NLP techniques to quickly sort out the symptoms of the users and offer them medically related suggestions, saving them both time and healthcare costs. This interactive platform processes user inputs to resolve queries, offering tailored answers and recommendations, including suggesting appropriate doctors and providing dosage reminders. Primarily designed for emergency situations or minor health concerns, the chatbot serves as a valuable advisor or temporary healthcare provider. Furthermore, the objectives of this endeavor encompass various goals as follows:

1- By providing an affordable alternative to conventional medical consultations, authors hope to reduce the number of people who disregard their health.
2- By incorporating advanced machine learning models into the system, illness diagnosis will be more accurate and efficient.
3- This will promote a better way of living and positively affect general social healthcare standards. The suggested chatbot hopes to transform how people interact with their health and well-being by achieving these interconnected goals, which would ultimately lead to a society that is healthier and better informed.

The research begins with data collection from an online repository, followed by pre-processing to ensure data quality. It evaluates the performance of three machine learning algorithms: Naïve

Bayes, Decision Tree, and Random Forest for disease prediction. The study uses these algorithms to predict diseases from symptoms entered into a chatbot interface, which processes inputs using NLP techniques. Classification models predict diseases, while association models suggest treatments. The goal is to determine the most effective prediction model and advance the development of predictive healthcare solutions. The authors lack detailed information about the dataset used for training and testing the machine learning model, affecting the validity of its conclusions. Accurate healthcare predictions require a high-quality, inclusive dataset, which the article fails to address. Despite reporting high accuracy rates, the model's generalizability to other diseases and diverse populations is not discussed. The article also omits clinical validation of the model's predictions, a crucial step for real-world application. Comparing the model's suggestions with medical professionals' input is necessary for reliable performance. Additionally, improving the chatbot's communication skills will enhance user satisfaction and trust.

Gupta *et a*l. [42] developed a comprehensive model for helping patients in selecting the most suitable doctor or hospital for their diagnosed condition. The system intends to provide customized recommendations by utilizing user requirements and insights obtained from prior patient assessments. Factors such as location convenience, affordability, doctor quality, and healthcare facility standards are meticulously considered in the recommendation process. The model collects real patient reviews from many reliable sources and uses sentiment analysis to determine how positive or negative the reviews are as individuals. Employing the decision tree algorithm, the system effectively maps user preferences to deliver personalized recommendations. Additionally, the platform facilitates user engagement by allowing individuals to share their own hospital and doctor experiences, contributing to a collaborative healthcare community. To keep the system flexible and responsive to changing user demands and preferences, it is necessary to continuously learn from new data and modify the model. The aim is to provide optimal recommendations that are customized to address the specific healthcare needs of each patient by integrating NLP, data analysis techniques, and machine learning algorithms. This ultimately improves the quality of healthcare received by patients. The study employs a methodology that involves various steps.

1- Data is gathered to generate a large dataset of patient evaluations from various websites and questionnaires. The obtained data is then cleaned and prepped for analysis through preprocessing, which includes standardizing text formats and removing unneeded information.

2- Methods for sentiment analysis are used to determine the content of patient feedback and categorize them as positive, neutral, or negative. Feature extraction is then employed to identify significant aspects of patient satisfaction or dissatisfaction from the text.
3- A recommendation algorithm is developed to rank hospitals based on patient requirements, considering factors like location and disease type.
4- A user-friendly web interface is created to facilitate patient input and provide hospital recommendations based on the system's analysis. Various NLP models are used in this process, including Bag of Words, TF-IDF, lexicon-based models, machine learning classifiers like SVM and Naive Bayes, and deep learning models such as CNNs and RNNs for sentiment analysis.

The limitations in [42] include factors like the system's scalability as the volume of reviews increases, potential biases in patient reviews, the reliance on publicly available patient reviews that might not cover all healthcare providers equally, and the challenge of accurately interpreting sentiments from text due to language nuances. Furthermore, the quantity and quality of evaluations, which could differ significantly between platforms and healthcare providers, are an essential requirement for the success of the system.

Jeidy and Melanie [43] focused on the development of a medical chatbot leveraging machine learning techniques to identify illnesses and offer preliminary healthcare advice. The chatbot leverages speech recognition for informal pre-consultations to make healthcare more accessible and cost-efficient. It records interactions for further analysis, facilitating continuous improvement. Advanced techniques like TF-IDF and N-gram enhance text classification and similarity evaluation, while a referral system directs users to specialized software for complex queries. A proposed framework combines AI, Machine Learning, and R scripting for efficient chatbot conversations. Additionally, the chatbot engages college students, offering basic health information and assistance. This proposed methodology involves developing chatbots that use advanced machine learning algorithms to understand and respond to speech inputs. Users must register to ensure personalized interactions. If the chatbot lacks immediate answers, it employs optimization methods to enhance responsiveness. Additionally, incorporating Domain Name System (DNS) experts broadens the chatbot's knowledge, providing users with detailed and accurate information.

The methodology includes Morpheme-based Analysis, using morphemes to assess data contributions and N-grams to calculate high similarity percentages, enhancing the chatbot's response relevance. An expert system modifies statements after analyzing data for accuracy in interactions. Google API supports smooth text or video chats, providing user-specific responses. SVM Classification refines data categorization, while the Porter Method removes superfluous elements to optimize processing. Dataset verification involves n-grams, TF-IDF, and Fourier analysis for similarity scoring. Performance testing of the chatbot evaluates its accuracy, relevance, and engagement against set standards, ensuring its reliability and effectiveness. The work's limitations cover several important areas.

- **Difficulty in Gathering Feedback**: A major obstacle is the difficulty in collecting enough feedback regarding customer satisfaction or user experiences. Insufficient data on user interactions and satisfaction levels makes it hard to evaluate and enhance the chatbot's performance efficiently.
- **Learning via User Input**: Whereas a reinforced feedback system allows the chatbot to learn from user input in an ideal world, real-world implementation challenges make this technique less successful. The complexities associated with designing and implementing such a system pose significant challenges, potentially limiting the chatbot's learning capabilities and overall performance.
- **Lack of Explicability**: An additional significant constraint involves the absence of clarification in the chatbot's decision-making technique. Developing insight into the chatbot's information processing and decision-making is essential for successfully addressing specific problems. Nevertheless, when the chatbot issues or performs inadequately, the fundamental opacity in its decision-making processes makes it more difficult to identify and address performance-related problems. This lack of transparency undermines the chatbot's reliability and effectiveness in real-world scenarios, necessitating further research and development efforts to enhance explicability and performance.

2.3.2 Administrative Task in Healthcare Chatbot System

The efficient fulfillment of administrative duties is essential to maintaining efficiency in operations and improving patient care in the complex healthcare ecosystem. These duties, which have an impact on everything from patient admission to post-care follow-up, are vital to the operation of healthcare facilities, even if they are not directly related to clinical treatment. This section explores the fundamental administrative responsibilities found in healthcare environments and emphasizes how critical it is to handle these responsibilities well.

An outline of typical administrative duties in healthcare settings is given in this overview. These are important and varied tasks that are necessary for the efficient operation of healthcare services. Among these is appointment scheduling, which entails arranging for patient appointments and controlling the schedules of healthcare providers to guarantee prompt treatment. Furthermore, patient data management which includes managing medical records and other patient data to support well-informed clinical decisions is essential.

Effective administrative task management is important for a number of reasons. First off, by reducing procedures, it improves the delivery of healthcare by allowing healthcare professionals to concentrate more on patient care, which eventually improves health outcomes and enhances patient satisfaction. In addition, it improves the patient experience since efficient administrative procedures provide patients with a positive view and increase their confidence in medical institutions. Furthermore, effective task management is a vital component of achieving operational excellence. It improves resource allocation, lowers costs, and minimizes errors, all of which raise the standard of healthcare services offered overall.

Sowmya *et al.* [44] proposed a combination of assistive robots and healthcare chatbots to address the absence of medical services in highly populated areas. These robots reduce healthcare providers of some of their daily duties by quickly and accurately diagnosing illnesses. The recommended healthcare chatbot answers users' questions about health by offering Q&A sessions where they may enter symptoms and get advice on possible illnesses and how to get medical attention. The chatbot, which is powered by machine learning and natural language processing, outperforms current models with a remarkable 93% accuracy rate when corresponding symptoms to known illnesses. The medical profession aims to gain from the potential for improved patient access and healthcare delivery through the integration of precise healthcare chatbots with assistive robots. The research focuses on developing these systems, which combine NLP and AI algorithms with mechanical interfaces for interactive patient support. Challenges include integrating with existing healthcare systems, accurately interpreting complex health queries, and ensuring the reliability of automated advice. The approach utilizes advanced NLP methods like CNNs and LSTMs, alongside classification algorithms such as decision trees, SVMs, and ensemble methods, to enhance diagnostic accuracy and predict potential health issues. The study's limitations include the need for continuous data updates to maintain the chatbot's accuracy, potential challenges in understanding complex medical terminology, and ensuring user privacy and data security in interactions with the chatbot and assistive robots. This innovative approach underscores the potential of integrating NLP-powered chatbots with assistive robots to revolutionize medical assistance, especially in resource-constrained environments, by providing efficient, accurate, and accessible healthcare solutions.

Dammavalam *et al*. [45] implemented a comprehensive healthcare management system powered by an AI chatbot. This chatbot is designed to interact organically with users, addressing their queries using a knowledge base pre-processed into a JSON format and analyzed using bag-of-words for text and speech input. Their objectives were to improve healthcare accessibility by providing navigation, symptom diagnosis, doctor recommendations, diagnostic information, and appointment booking within a single application. In order to analyze queries and provide responses, the system uses LSTM models. These models have been supported by a sequential model that uses Relearning activation functions and a softmax output layer that is optimized for the least cost function using random gradient descent. This methodology provides an interface that is convenient to use for hospital administration duties. The limitations of this work include accessibility and user-friendliness. The fact that not everyone is able to use the chatbots suggests that they are not widely available. This limitation could result from a number of things, such as device compatibility, internet access, or technological expertise. Although efforts have been made to make the project more user-friendly, it has been acknowledged that some users, especially those who are not comfortable with digital tools may find it difficult to communicate with the chatbot.

Madhavan [46] proposed an application "Medoo" an application for smartphones that aims to improve patient-hospital interactions by facilitating continuous medical observation. It suggests streamlining administrative tasks like appointment scheduling and health services by combining chatbots, big data, and biometric tracking devices. The application is being developed using Internet of Things (IoT) technology for real-time patient health monitoring. Chatbots and big data are utilized for data analysis and communication. Google APIs assist in location services for finding nearby hospitals. However, limitations include concerns regarding data security and privacy, dependency on internet connectivity, challenges in integrating with existing healthcare systems, and user adaptability. Addressing these limitations is crucial for the successful implementation and acceptance of the application, representing a significant advancement in healthcare technology.

2.4 The Role of NLP in Healthcare Chatbots

To improve healthcare chatbots' usability and efficacy and to make these digital assistants more like human healthcare providers in terms of recognizing and processing patient inputs, NLP performs a critical role. Nowadays, researchers have a better understanding of how NLP technologies have advanced healthcare chatbots due to recent studies.

Christopherjames *et al*. [47] proposed the development of an assisted conversational agent that gathers and analyzes patient health data using NLP. One of the goals is to give patients recommendations along with relevant health information depending on their input. Techniques include building a conversational agent that can provide personalized healthcare support and using NLP techniques to process natural language inputs. The authors used Dialog flow, an API developed by Google, as its main NLP tool. Dialog flow is well known for its powerful NLP features, which enable the development of conversational interfaces that can comprehend and interpret natural language user requests. The chatbot is programmed to understand the purpose behind the user's inquiries, ensuring that their text input is correctly linked to the most suitable response that the chatbot should provide. Besides intent recognition, the system also recognizes important entities from the user's queries, including symptoms or particular health problems, which are essential for providing accurate health advice. By utilizing the identified intents and entities, Dialogflow enables simpler to provide appropriate responses that provide users with details about illness symptoms, treatments, and suggestions for further activities. However, there are limitations, including the potential that the conversational agent's efficacy may be impacted by the complexity and accuracy of the NLP model, as well as its dependency on patients' potentially variable capacity to accurately explain their symptoms in natural language. It is vital to address these limitations to ensure the reliability and quality of medical support given by the conversational agent.

Desai *et al.* [48] proposed a method to utilize NLP techniques in a medical chatbot to comprehend patient requests and provide relevant answers. To evaluate and correctly answer to patient inquiries, methodologies include using machine learning and NLP algorithms. Furthermore, ongoing learning from patient interactions is used to improve the chatbot's responses. Nevertheless, there are certain limitations, such as the chatbot's dependence on the quality and amount of training data, which might limit its capacity to respond to a range of medical queries. The methodology of this work highlights the utilization of NLP algorithms for the chatbot to understand patient inquiries effectively and provide relevant responses. This involves text classification to categorize inputs into predefined categories like symptoms, diseases, or treatments. Additionally, entity recognition extracts medical entities such as symptoms or diseases from the patient's text. Sentiment analysis can be used to assess the emotional states of patients, and intent recognition can be used to identify the purpose of queries and assist the chatbot in providing relevant information. Important challenges also exist in obtaining precise medical diagnoses and advice in the absence of human supervision. It is vital to address these constraints to ensure the efficacy and dependability of the medical chatbot in assisting patients with their medical requirements.

Sharma [49] proposed developing an intelligent chatbot that uses NLP to identify illnesses and provide users with information before to a consultation. The proposed methodology focuses on the deployment of an artificial intelligence (AI)-powered chatbot system. The core of the AI system includes ML and NLP, which enable the chat-bot to understand health-related data patterns and enhance its diagnostic capabilities over time through user interactions. Python is a fundamental programming language used in the rule-based text-to-text implementation of the chatbot. It engages users in conversations about their health issues and delivers personalized diagnoses based on their inputs. The chatbot functions as a virtual assistant and uses text, voice, and video communication technologies to diagnose a variety of illnesses and offer necessary information, such as referrals to doctors and appointment scheduling. NLP techniques are employed to instruct the chat-bot with datasets containing conversational dialogs, allowing it to construct grammatically correct sentences and interpret the meaning and structure of user inputs effectively. The system's performance is largely dependent on the accuracy and detail of the user's responses; consequently, this work has issues relating to its demand for user inputs. Inaccuracies or misunderstandings in user inputs may result in incorrect diagnoses or treatment recommendations. The challenge of making medical diagnoses is another significant issue. Although the chatbot can provide initial health assessments and recommendations, it might not be able to deal with complex medical situations requiring for a more complex comprehension and interpretation than what it can provide.

Coşkun [28] provided an extensive architecture for a healthcare system powered by AI, which includes NLP to facilitate user interaction and medical diagnosis. The system's primary function is to allow users to interact with chatbots via a simplified user interface. The conversation is handled by a server that employs NLP. The architecture integrates several kinds of elements, such as servers for handling database management and user interactions, NLP modules for producing and comprehending language, and connectors to external data sources. With a tool that may be able to identify health problems, help users navigate healthcare procedures, and connect them with medical specialists when needed, this setup attempts to make getting medical help easier and more engaging for users. The technology is designed to make health advice more accessible, particularly for people who live in faraway places or are unable to physically visit healthcare institutions. Their proposed architecture is depicted in Figure 2.1.

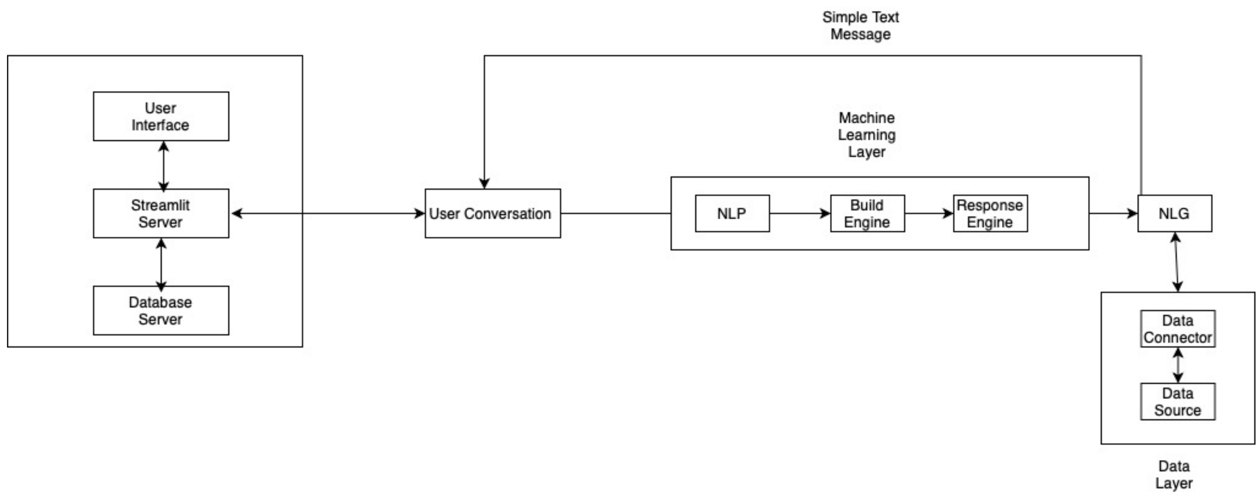


Figure 2.1 Architecture of proposed system

The provided architecture of the proposed system outlines an AI-based healthcare system utilizing NLP.

This architecture is composed of the following components:

1- User Interface:

- This is the front end that users interact with. It's designed to be intuitive, allowing users to input their queries and receive responses.

2- Streamlit Server:

- Streamlit is an open-source app framework for Machine Learning and Data Science teams. In this context, it likely serves as the middleware facilitating user interaction with the system, managing the flow of data between the user interface and backend services.

3- Database Server:

- A database server stores and manages health-related data, possibly including patient information, medical records, and other relevant datasets that the chatbot would need to access.

4- User Conversation:

- This component manages the dialogue between the user and the chatbot, likely using a dialogue management system to keep track of the conversation's state and context.

5- NLP Module:

- The core of the chatbot's understanding capabilities, this module processes the natural language input from users, interprets their intent, and extracts relevant information.

6- Build Engine:

- This engine likely constructs the response logic based on the NLP's output and predefined rules or models. It's part of the backend processing that determines the appropriate action or response.

7- Response Engine:

- Once the Build Engine has determined the appropriate action, the Response Engine generates a response that will be sent back to the user.

8- NLG (Natural Language Generation):

- This component translates the system's actions or responses into natural language text that the user can easily understand.

9- Data Connector:

- The Data Connector bridges the system with external data sources or APIs that can provide additional information needed by the chatbot to formulate its responses.

10- Data Source:

- External databases or information repositories that the chatbot can query to retrieve information not stored in its primary database.

2.5 Research Challenges and Issues

This section highlights the main challenges and obstacles that develop throughout the development and implementation of chatbots for healthcare. and we categorize them based on their weaknesses and similarities in Table 2.1. According to these findings, the primary limitations would be enhancing patient knowledge, assisting healthcare providers with administrative tasks, user-friendly Interface.

2.5.1 Enhancing Response Accuracy

The accuracy of responses indicates diversity, especially in intricate medical situations where an in-depth understanding is essential for accurate diagnosis and assistance. Evolving NLP models and algorithms are commonly used in systems, much like in many other systems, to continuously improve understanding and response accuracy over time. Accuracy of health information provided, ability to process complex medical questions, and personalization of health recommendations, as mentioned previously (cf. 2.3.1, 2.4) The key issues in this area include:

- **Complexity of Medical Language:** Chatbots must explore a vast array of medical terms and patient expressions, often specific to individuals' health conditions**.**
- **Continuous Learning**: Chatbot accuracy requires constant updating and learning as medical knowledge improves and patient needs change.
- **Personalization**: Effective care requires making individualized health recommendations that adequately take into account each patient's individual medical history and symptoms.

2.5.2 Assisting Healthcare Providers with Administrative Tasks

The inefficiency of appointment booking processes is impacted by restrictions on real-time scheduling capabilities caused by integration issues with current healthcare systems. There is an obvious move toward scheduling technology to reduce administrative costs, although challenges with interoperability still exist. We highlighted this point previously (cf. 2.3.1, 2.3.2).Chatbots offer an opportunity to revolutionize this aspect of healthcare by:

- **System Integration**: One major technical issue is ensuring chatbots interact accurately with different healthcare provider platforms.
- **Dynamic Scheduling**: Developing chatbots that can manage complex scheduling scenarios, such as emergency cancellations and modifications.

- **Scalability**: It is important to solve the issue of chatbots' capacity to scale up and manage large amounts of administrative queries while patient loads change.

2.5.3 User-Friendly Interface

User experience inequalities, particularly for those who require more user-friendly interfaces or have low knowledge of technology**.** Various studies have been done to enhance chatbot interfaces by implementing adaptive and increasingly conversational user interface designs. (cf. 2.4, 2.4.1) Current research is directed towards:

- **Technological Accessibility**: It is essential to ensure that the chatbot is user-friendly for anyone with different degrees of technical ability.
- **Inclusivity**: One of the primary issues is addressing the various demands of various user populations, especially those who suffer from disabilities.
- **Engagement**: Users must be engaged by the interface design in a way that encourages continuous and meaningful interaction with the chatbot.

In addition to being necessary for the successful implementation of healthcare chatbots, addressing these issues is also necessary for achieving the potential of these technologies to improve patient care and streamline healthcare delivery.

Table 2.1 Summary of existing healthcare chatbots

Proposed method	Limitations
NLP for accurate patient information extraction [19],[28],[29],[33],[36],[39],[44]	Response accuracy can be compromised by complex query interpretation and may lack personalization.
NLP-driven appointment scheduling systems [19],[46],[45],[49]	Challenges with system integration and scalability, and potential rigidity in handling nuanced patient-provider interactions.
Conversational AI for patient interaction [31],[34],[36],[48]	Requires continuous learning to improve response quality; Maintaining a natural dialogue flow
Sentiment analysis for patient feedback and support [28],[29],[33],[41]	Difficulty in accurately capturing emotional nuance; Reliant on diverse and nuanced training datasets
Machine learning models for symptom checking and diagnosis [33],[41],[45],[47],[48],[49]	Potential for misdiagnosis due to oversimplified models; Need for constant updating with medical advances
Chatbot UI/UX design enhanced by user-centric NLP [33],[38],[41],[42]	Maintaining accessibility for all user groups while establishing a balance between technical language and user-friendly interactions

CHAPTER 3 Proposed Optimized Natural Language Processing Architecture for Chatbot

This chapter describes the methodology and proposed optimized architecture used to optimize and evaluate a Natural Language Processing (NLP)-based Virtual Medical Assistant (VMA) model for the healthcare system. The focus will be on the system architecture, integration of various components, and the overall workflow from data input to output. This chapter provides a comprehensive understanding of how the system functions and the methodologies used to ensure its efficiency and accuracy**.** The goal is to ensure a robust and reliable approach to achieving the research objectives, thereby contributing to advancements in healthcare technology and patient care.

3.1 System Architecture

The system architecture of the virtual medical assistant is designed to ensure efficiency, scalability, and accuracy in healthcare services. This architecture is divided into three layers, as shown in Figure 3.1: User Interface (UI), Natural language generation (NLG), preprocessing and processing.

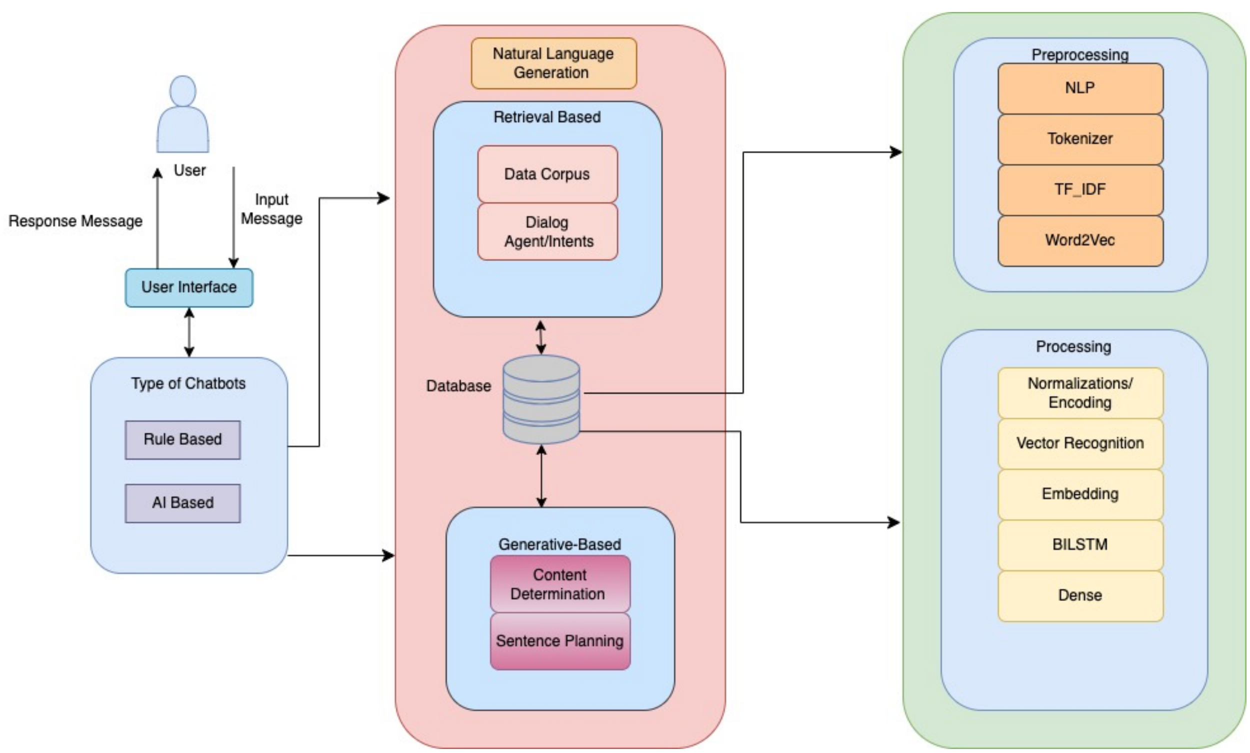


Figure 3.1 Architecture of the proposed model

In this section, we proposed two models for the VMA designed to provide an efficient and responsive healthcare chatbot system, integrating advanced NLP techniques. The model can handle both rule-based and AI-based approaches for generating responses to user queries, ensuring flexibility and accuracy in addressing a wide variety of medical questions and tasks. The user begins by interacting with the system through a user-friendly interface, either via web or mobile platforms. This interface captures the user's input message and directs it to the appropriate backend component based on the complexity of the query [51]. The user engages with the system by interacting on online or mobile platforms. This interface receives the user's input message and sends it to the appropriate backend component based on the complexity of the inquiry. The system includes two categories of chatbots, namely Rule-Based and AI-Based. Rule-based chatbots are used in scenarios where the tasks are specific to predefined frameworks of conversations. In contrast, the AI-based chatbot uses advanced algorithms in machine learning to answer more complex questions, which require a deep understanding of its context. Natural Language Generation (NLG) could be categorized into two main models: retrieval-based and generative-based approaches [6][23]. The Retrieval-Based approach relies on a data corpus and predefined dialog agents or intents to retrieve the most relevant response to the user's query. In contrast, the Generative-Based approach uses content determination and sentence planning algorithms to generate responses dynamically based on the context and input, making it highly flexible for complex or novel queries.

3.2 Natural Language Generation Layer

A database is a fundamental storage place that centers and keeps all information from the used corpus in retrieval-based approaches to their interaction history [53]. The system continuously interacts with the database to retrieve or store information, ensuring the user receives the most relevant response [19]. Based on the user query, several preprocessing steps were carried out in order to prepare the input for further analysis. The preprocessing steps include NLP techniques such as tokenization, TF-IDF (Term Frequency-Inverse Document Frequency), and Word2Vec, all of which transform the raw input text into a format suitable for machine learning models. This makes sure the input data is standardized and ready for analysis [6] [23][24]. In this section, we describe the feature extraction and training phases of our proposed model for symptom detection and medical service location identification. The feature extraction conducts symptoms data preparation for our BiLSTM model training. The first step in this is to prepare the symptom-related training data for our BiLSTM model. This phase is crucial for ensuring that the model effectively learns from the dataset, allowing for accurate symptom classification. For this research, the primary dataset utilized to develop and evaluate the VMA model using NLP in healthcare systems is the Medical Speech, Transcription, and Intent Dataset [52]. This dataset, collected by Paul Mooney and published on Kaggle in 2020, includes 8.5 hours of audio and text data related to 25 common medical symptoms, such as "knee pain" and "headache." Each audio file is paired with a transcription provided by contributors who describe specific symptoms, ensuring alignment between audio and text data. These phrases were used to generate corresponding audio recordings, ensuring alignment between text and audio data. This dataset includes 25 different symptoms [52]. The whole symptoms are illustrated in Table 3.1.

Acne	Knee pain	Muscle pain	Open wound	Internal pain
Emotional pain	Joint pain	Neck pain	Skin issue	Hard to breath
Foot ache	Headache	Infected wound	Hair falling out	Cough
Shoulder pain	Feeling cold	Earache	Feeling dizzy	Blurry vision
Body feels weak	Back pain	Injury	Stomachache	Heart hurts

Table 3.1 List of symptoms for Kaggle dataset

Additionally, in this dataset, we have many phrases each related to a specific type of symptoms, as shown in Figure 3.2. The variety of phrases relating to back pain is demonstrated by the different back pain phrases found in each row (APPENDIX A).

▼ Back pain	I can't stand up or sit down I have a pain in my back that annoys me
	I feel a pain in my back
	I feel a pain in my back when I sit on a chair for long.
	I feel back pain when I carry heavy things
	i feel pain in my back
	I feel pain in the lower back
	I have a back pain since I fell on the floor.
	I have a back pain since I turned 70 years old.
	I have a dull ache in my lower back which makes it difficult to move
	i have a problem in my back i cannot extend it
	I have shooting pains up and down my back.
	I love to garden but I get a terrible twinge in my lower back when I lean over.
	I think I overdid it when I carried all that lumber from the yard. My lower back is killing me.
	I used alot of pain killer to get better but i still feel the same back pain
	longitudinal burning line across back with hard respiratory movements
	My back hurts a lot when I bend
	my back hurts me a lot
	My back hurts me and i can't bend it or walk
	My back hurts so much I can't bend down to tie my shoelaces.
	My back is hurting so much.
	My upper back has been sore for a week.
	standing less than five minutes and my back start to ache
	The pain in my back dwwls like a sharp knife in it
	When I bend over I get a shooting pain down my back
	When I carry heavy things i feel like breaking my back
	When I play sports I have some burning sensation in my spine

Figure 3.2 List of phrases for specific symptom

3.2.1 Retrieval Based Module

In this module, we have two phases including data corpus and dialog agents. In the data corpus, we gathered and preprocessed data from medical symptom datasets. This dataset was selected for the training of the BiLSTM model [59]. In addition, text preprocessing techniques like tokenization and embedding were applied to the symptom descriptions to standardize the input data. Moreover, we consider another dataset about the details of medical centers like location, code postal, and so on for the classification model. In the training phase, the extracted features were conducted into the BiLSTM model to classify the symptoms. In addition, we implement this phase using the Keras library for optimizing the model's hyperparameters to gain high accuracy. The data is processed using NLP and speech recognition to convert it into CSV format, enabling neural network analysis for cognitive disease prediction.

We use a dataset for the development of our KNN algorithms specifically related to medical centers. This dataset provides essential location information for healthcare services, enabling our system to match user input with the nearest relevant healthcare providers effectively. By utilizing this dataset, we aim to optimize the accuracy and reliability of our model in locating medical services based on user needs.

3.2.2 Generative Based Module

The generative-based module is responsible for creating meaningful responses that are tailored to user queries. It primarily involves two key components:

1. Content Determination: This step involves deciding what information or content should be included in the response. It filters and selects the most relevant data based on the user's input and the database content.
2. Sentence Planning: After determining the content, this part structures the response into coherent and natural-sounding sentences. Sentence planning ensures the response is grammatically correct and logically organized, allowing for better user interaction.

3.3 Preprocessing and Processing Layer

The next stages are preprocessing and processing, which involve multiple advanced steps. The normalized text is transformed into vectors through vector recognition, enabling the system to capture semantic meaning (APPENDIX B). These vectors are then passed through an embedding layer to produce a dense, low-dimensional representation of the text, allowing the model to recognize deeper contextual relationships [4]. A BiLSTM (Bidirectional Long Short-Term Memory) network processes the input in both forward and backward directions, ensuring the system captures both past and future context within the text, which is essential for accurate understanding in medical conversations. Finally, the dense layers improve the model's output and help make more accurate predictions or generate responses. Finally, the system sends the generated response back to the user through the user interface, ensuring the conversation is fluid, contextually appropriate, and tailored to the user's needs. The entire workflow is designed to improve the user experience by providing accurate, timely, and meaningful responses, ultimately improving both patient knowledge and healthcare efficiency. This model illustrates the integration of sophisticated

NLP techniques, deep learning networks, and flexible chatbot mechanisms to create an effective virtual medical assistant for healthcare systems [54].

3.3.1 Preprocessing

The preprocessing contains four steps as follows:

- **NLP:** All these preprocessing steps involve applying basic NLP techniques. NLP refers to the methods whereby computers are able to understand and modify natural human languages. In this step, a basic level of processing of the user's input message is made such that when the system tries to break down the message, it can be interpreted properly. NLP techniques include the identification and segmentation of units of meaning from language such as words and sentences, for example. In this regard, NLP provides a foundation for some high-end processes like tokenization and word embedding [19][54].
- **Tokenization:** The tokenization is a process that splits input text into separable units called tokens. These tokens come in the form of words or even characters as small, depending on the strategy of tokenization being employed. Therefore, the process of tokenization allows turning the input text into components more manipulated by the system. For instance, in a sentence like "I have a headache", tokenization would break this down into: This step enables the system to treat each word or token as a separate entity [32].
    -

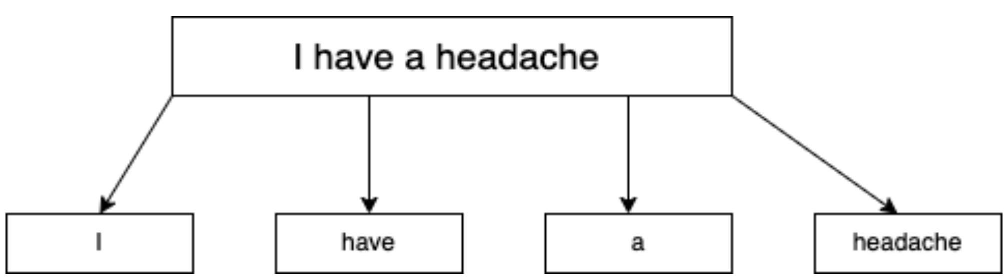


Figure 3.3 Tokenization of sentence

- TF-IDF (Term Frequency-Inverse Document Frequency): While the text has been tokenized, the system calculates TF-IDF (3.3) values for each word in the query. TF-IDF is a statistical measure used to evaluate how important a word is in a given document or dataset. It estimates the frequency within the given document, regarding its general frequency in the whole corpus, represented as IDF [16][55][58].

$$\text{TF (t,d)} = \frac{\text{number of times (t)} \, appears\ in\ document(d)}{total\ numbers\ of\ terms\ in\ d} \qquad (3.1)$$

$$\text{IDF(t)}= \log \frac{N}{1+df} \qquad (3.2)$$

$$\text{TF} - \text{IDF} = \text{TF(t,d)} * \text{IDF(t)} \qquad (3.3)$$

- **Word2Vec:** As a vector space representation of words has to be produced, in which the semantic relations between words should stand out, a continuous way of doing this is Word2Vec. It creates vectors driven by the use of words within big data. During the second step after TF-IDF, these words are again transformed into vectors, expressing their meaning and relationship to other words [58].

3.3.2 Processing

After the initial preparation steps are completed with the input data being preprocessed, the information then moves into the processing stage. This is where normalization and encoding methods are utilized, followed by applying vector recognition to translate text-based data into machine formats known as vectors.

- **Normalization/Encoding**: The input text is standardized by transforming it into a uniform structure and encoding it into numerical values to facilitate processing [51-53].
- **Vector Recognition**: Vectorization refers to a classic approach of converting input data from its raw format, like text, into vectors of real numbers, which is the format that ML models support [51][63].
- The BILSTM is a **recurrent neural network** used primarily on natural language processing. Unlike standard LSTM, the input flows in both directions, and it's capable of utilizing information from both sides. The model is able to better understand the relationship between sequences, like knowing the following and preceding words in a sentence [59].

- **Dense Layers**: There are fully connected layers that further refine the output from the BILSTM, helping the system make final predictions and prepare data for response generation [64].

3.4 LSTM (Long Short-Term Memory)

LSTM is one of the RNN architectures that has been successfully implemented in representing sequential information, for example, text or time series data. While standard RNNs handle long-range dependencies poorly, at least in part due to the vanishing gradients and other optimization problems, LSTMs were designed to hold information for more time steps [56].

An LSTM cell includes three key gates [56][57]:

1- **Forget Gate**: The first gate is going to decide which previous information is going to be forgotten. It takes the current input and the hidden state from a previous step, returning a value in the range 0-1 for every piece of information, where 0 tells the system to forget and 1 tells it to keep.

- $f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f$ (3.4)
- f_t is the forget gate vector at time step t.
- σ is the sigmoid activation function, which makes the output values range between 0 and 1. Thus, it does the critical function of deciding what proportion of information to retain and what proportion to discard.
- W_f is the weight matrix for the forget gate.
- $[h_{t-1}, x_t]$ is the concatenation of the previous hidden state (h_{t-1}) and the current input (x_t). This combines the information from the previous time step and the current input.
- b_f is the bias term for the forget gate.

2- **Input Gate**: This gate will decide what new information needs to be let into the cell state. This input gate will update the cell state with the new information regarding its relevancy from the current input and from the hidden state at the previous time step.

- $i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i$ (3.5)
- $C_t = tanh(W_c * [h_{t-1}, x_t] + b_f$ (3.6)

3- **Output Gate**: The output gate decides which piece of information of the present cell state is to be kept and carried forward to the next hidden state by passing filtered and preprocessing data to form the present output.

- $\sigma_t = \sigma(W_o * [h_{t-1}, x_t] + b_o$ (3.7)
- $h_t = O_t$ * tanh (C_t) (3.8)

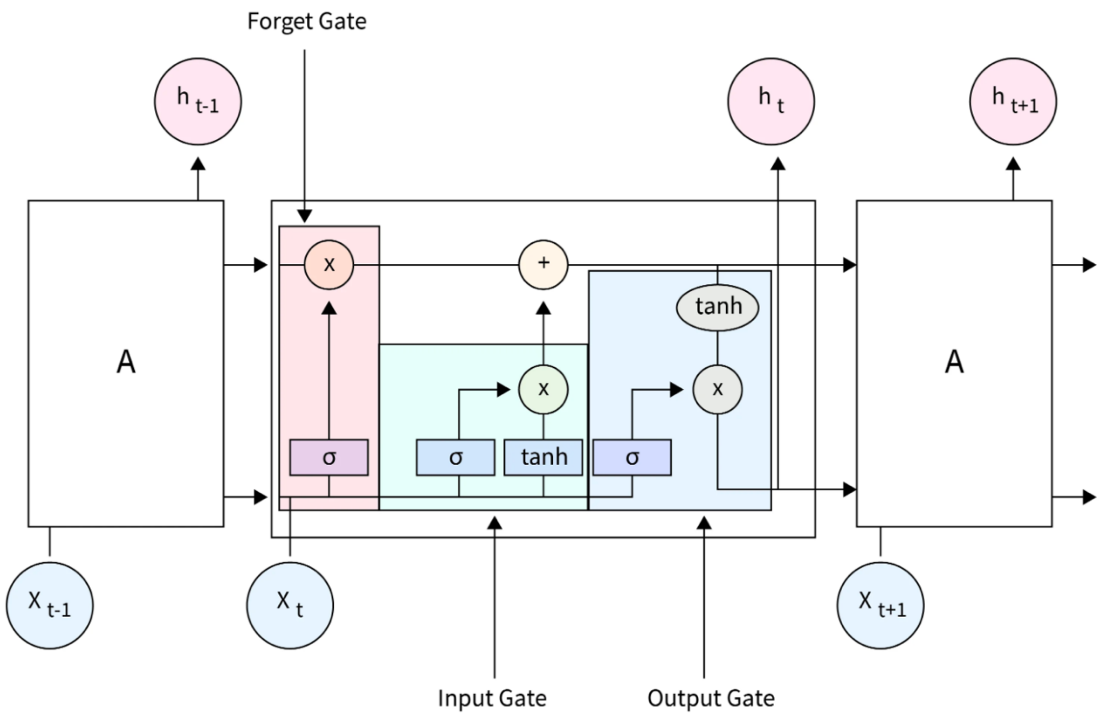


Figure 3.4 LSTM architecture

4- Cell State: The forget gate and the input gate update the cell state. The output of the forget gate is multiplied by the previous cell state, and the product is added to the output of the input gate. This sum will be used in computing the hidden state in the output gate.

- $C_t = f_t * C_{t-1}, i_t * \overline{C}_t$ (3.9)

The LSTM architecture is similar to RNN, but instead of the feedback loop has an LSTM cell. The sequence of LSTM cells in each layer is fed with the output of the last cell. This enables the cell to get the previous inputs and sequence information. A cyclic set of steps happens in each LSTM cell [56][57]:

- The Forget gate is computed.
- The Input gate value is computed.
- The Cell state is updated using the above two outputs.

The output (hidden state) is computed using the output gate.

These series of steps occur in every LSTM cell. The intuition behind LSTM is that the cell and hidden states carry the previous information and pass it on to future time steps. The cell state is aggregated with all the past data information and is the long-term information retainer. The hidden state carries the output of the last cell (i.e. short-term memory). This combination of long-term and short-term memory techniques enables LSTM's to perform well in time series and sequence data.

3.5 BiLSTM (Bidirectional Long Short-Term Memory)

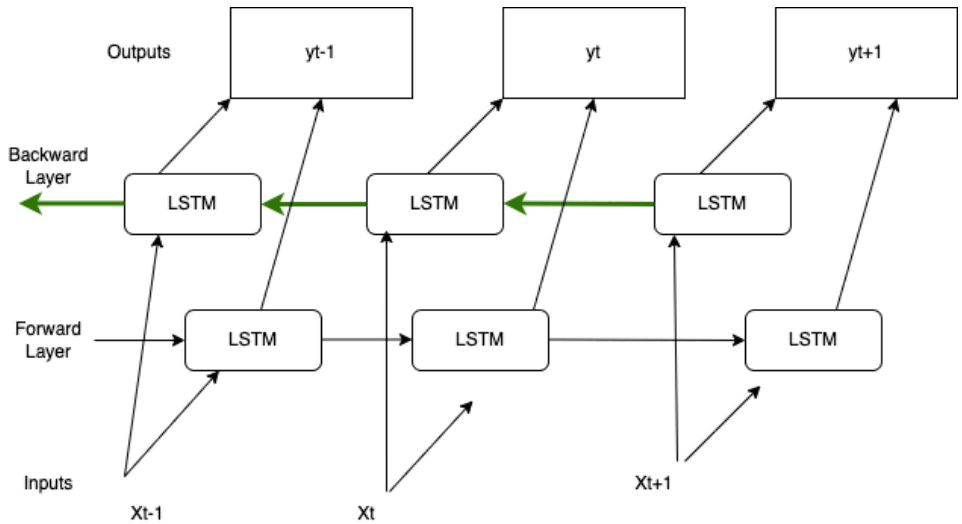
We proposed BiLSTM model for the symptom dataset. This model is an extension of LSTM, which processes the input sequence in both directions: from past to future (forward) and from future to past (backward). A bi-directional architecture provides the model with context from both preceding words and following words, making it particularly useful in applications that benefit from the entire context of a sequence.

In a BiLSTM network [59][60][62]:

1- **Forward LSTM**: A single LSTM layer receives the input sequence sequentially, in the forward direction, that is, from the first token to the last token of the entire sequence.

2- **Backward LSTM**: A separate LSTM layer processes the same input in reverse sequential order, from the last token towards the first token. This implementation allows the model to take future context into account, which may influence current action.

3- **Outputs**: The outputs from both the forward and backward LSTM layers are then merged. This kind of integration of the entire input sequence is what allows the model to look not only at the going words of the sentence but also back to the ones preceding the token where it is.



•

Figure 3.5 BiLSTM Architecture

3.6 Dense Layer

A dense layer known as a fully connected layer, is a core building block in neural networks. It connects each neuron in the layer to every neuron in the previous and next layers, allowing the model to learn complex patterns in the data [64].

- $y = \sigma(W * x + b),$ (3.10)

where x is the input vector, W is the weight matrix that connects the input neurons to the output neurons,b is the bias vector, and σ is the activation function, such as Relu or sigmoid, which introduces non-linearity into the model.

The activation function introduces non-linearity, enabling the model to capture complex, non-linear relationships between inputs and outputs

3.7 K-Nearest Neighbors (KNN) Algorithm

The K-Nearest Neighbors (KNN) algorithm [65] is a simple and supervised machine learning algorithm used for both classification and regression tasks. KNN works by comparing the distance between data points and making predictions based on the nearest neighbors of the input. We used this algorithm to find the closest medical center based on user input, such as a postal code or other location data. The KNN algorithm includes two phases [65]: Training and prediction.

1- Training Phase: Unlike many other algorithms, KNN doesn't involve any explicit training. Instead, it simply stores the entire dataset during training.

2- Prediction Phase: When a new data point is provided, KNN calculates the distance between this point and all the other points in the training dataset. The most common distance metric used is Euclidean Distance, though others like Ottawa or Toronto distance can also be applied. The Euclidean distance formula between two points $A(x_1, x_2)$ and $B(y_1, y_2)$ [65].

- $d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ (3.11)

3- After calculating the distances, the algorithm selects the K nearest neighbors (where K is a user-defined parameter). These neighbors are the training data points closest to the new input based on the chosen distance metric.

3.7.1 Receiver Operating Characteristic (ROC) Curve

After using KNN algorithms for finding the closet neighbors in the neighborhood based on postal code, we should use ROC to evaluate the performance of the previous algorithms. ROC curve is a graphical representation that evaluates the performance of a classification model. The ROC curve is a plot of the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis for different classification thresholds. This is a single method of adjusting the sensitivity setting for a classifier. A higher threshold will result in fewer false positives but may miss true positives, while a lower threshold will capture more positives but may also increase false positives. TPR is known as sensitivity or recall, this measures the proportion of actual positives that the model correctly identifies [66-68].

$$True\ Positive\ Rate = \frac{True\ Positive}{True\ Positive + True\ Negative} \qquad (3.12)$$

FPR represents the proportion of actual negatives that the model incorrectly classifies as positives.

$$False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive + False\ Negative} \qquad (3.13)$$

3.8 Summary

This chapter presented our proposed NLP-based Virtual Medical Assistant (VMA) model for detecting symptoms and locating the nearest medical services within the healthcare system. The architecture includes three layers with the names of UI, NLG, and Processing.

Tokenization, TF-IDF, and Word2Vec embedding were applied in the process of feature extraction to investigate symptom descriptions in order to prepare the input data for the model.

In the training phase, we applied the BiLSTM network to organize symptoms with high accuracy.

We utilized the K-Nearest Neighbors (KNN) algorithm, to determine the closest medical service in the vicinity by measuring distances between the user's location and nearby medical centers. Our model can easily handle more symptom types and medical service locations, making it scalable.

CHAPTER 4 IMPLIMENTATION AND RESULTS

In Chapter 3, we discussed the design of our proposed optimized Natural Language Processing model towards effective virtual medical assistance in the healthcare system by pointing out its three layers: User Interface, Natural Language Generation, and Processing. In this chapter, we will present the implementation of the model to assess its performance. We present our testbed set up and configuration of testing environment. In addition, we provide our findings and results related to the processing layer of the proposed architecture, such as BiLSTM network efficiency for symptom's detection and K-Nearest Neighbors (KNN) for retrieval of nearest medical service by user request.

4.1 Testbed

We provide our testbed on one machine by using Visual Studio Code (VSCode) [70]. VSCode was selected because of its versatility and support for extensive machine learning and NLP libraries [70][71], making it well-suited for handling the data-intensive tasks required by this project, as follow:

- Development environment: VSCode platform was conducted with important extensions for Python development, such as Jupyter Notebook [71], that facilitated for interactive programming and real-time visualization of data processing and model training results [70].
- Software Specifications:
  a) Programming Language: We choose Python language [71] because of its large and several type of Natural Language Processing libraries [71][72]: NLTK, SpaCy and Tensorflow, which were crucial components for the language function of the chatbot.
  b) Operating System: We used a macOS system which offered an environment where the NLP libraries run stable, and VSCode with extensions to interface with the other functionalities provided for model development and testing.
    - Testbed environment for all testing was carried out on a MacBook Pro ((15-inch, 2016) with macOS Monterey (Version 12.7.6). We ran on macOS to run NLP libraries in a stable environment and smoothly combined VSCode extensions for model development and testing. The system's specifications are as follows:

- **Processor**: Our processor was a 2.7 GHz Intel Core i7, which provides the processing capacity required to run the NLP task and model assessments.
- **Memory**: Our memory had a 16 GB 2133 MHz LPDDR3, which supported operation during data processing and facilitated operations in model training.
- **Storage**: Our storage had a 512 GB SSD, which enabled fast data access and run operations, which are essential for handling large datasets and quick model iterations.

We implement NLP architecture by using Python code in VSCode locally on a macOS system, which is associated with the processing of medical data and aggregation for a healthcare chatbot like symptom detection and finding a nearby medical service center. We use an NLP pipeline, which includes tokenization, vectorization, and embedding layers to get a good understanding of medical symptoms to get high accuracy.

4.2 Natural Language Generation Layer

In Chapter 3 (cf. Chap. 3, §3.2.1), we explained that the first module in the Natural Language Layer Module is the Retrieval based module, which is used to prepare the data in our proposed model. This section presents our selected Medical text dataset [52] and Medical service dataset [53] for Generative-based module. These data will be analyzed by using our proposed BILSTM model. There are plenty of popular public medical datasets that are used for detecting symptoms regarding the literature review in Chapter 2 (cf. Chap. 2, §2.4). Among all of them, we will work on a combination of two public datasets from the Kaggle datasets and Government datasets [53].

4.2.1 Retrieval Based Module Implementation

In this section, we describe our dataset for developments for each module. In addition,we used Python to implement these models (APPENDIX A).

4.2.2 Medical Datasets Implementation

In Chapter 3 (cf. Chap. 3.3.1), we explained that the first model for the dataset that is related to medical symptoms is the BiLSTM. This section presents our selected medical dataset for the BiLSTM model for detection. This data will be normalized and cleaned using the scikit-learn framework. There are plenty of popular public datasets that exist for training and evaluating healthcare chatbots, as outlined in the literature review (cf. Chap. 2, § 2.3.1). We utilized two large datasets, including the Medical Speech, Transcription for detection of symptoms [52], and the Intent Dataset and the Ministry of Health Service Provider Locations (MOHSERLO) dataset for finding the nearest medical services in the neighborhood [53]. These datasets cover various medical symptoms and healthcare provider information.

- Medical Speech, Transcription, and Intent [52]: It has 8.5 hours of audio utterances paired with text for common medical symptoms. This data contains thousands of audio utterances for common medical symptoms like “knee pain” or “headache” (Table 3.1), totaling more than 8 hours in aggregate. Each audio was created by individual human contributors based on a given symptom (APPENDIX A). These audio snippets can be used to train conversational agents in the medical field. These phrases, as shown in Figure 3.2, were used to generate corresponding audio recordings, ensuring alignment between text and audio data. This dataset includes 25 different symptoms [52].
- Ministry of Health Service Provider Locations (MOHSERLO) dataset [53]: This dataset contains 19 types of health service providers and consists of a total of 11,606 records, with the most recent update in May 2023. Each health service provider has a variety of attributes, such as the type of service provider and specific details related to the service provider, including address information, which encompasses fields like Address Line 1, Community, and Postal Code [53].

4.3 Generative Based Module Implementation

In this section, we implemented the Generative-Based module for the first model, which is a good model for the Medical Speech, Transcription, and Intent [52] dataset for detecting prompts and phrases. First of all, we found that NLP and ML algorithms work with numerical data, so we converted the data into numerical form using ML algorithms, which we will explain in detail. Each word and prompt in this dataset has a specific code, known as the word index and encoded prompt, we can also see the count of words, as shown in Figure 4.1.

```
{'i': 1,
 'my': 2,
 'a': 3,
 'in': 4,
 'have': 5,
 'pain': 6,
 'feel': 7,
 'and': 8,
 'the': 9,
 'when': 10,
 'is': 11,
 'to': 12,
 'it': 13,
 'of': 14,
 'on': 15,
 "can't": 16,
 'up': 17,
 'like': 18,
 'hurts': 19,
 'get': 20,
 'with': 21,
 'that': 22,
 'back': 23,
 'me': 24,
 'after': 25,
...
 'bottoms': 997,
 'arches': 998,
 'heels': 999,
 'knocked': 1000,
 ...}
```

```
[('my', 4691),
 ('in', 1983),
 ('have', 1617),
 ('pain', 1604),
 ('feel', 1533),
 ('and', 1528),
 ('the', 1478),
 ('when', 1424),
 ('is', 1240),
 ('it', 1077),
 ('to', 1009),
 ('of', 804),
 ('can', 650),
 ('on', 591),
 ('up', 510),
 ('like', 507),
 ('hurts', 460),
 ('get', 409),
 ('that', 408),
 ('with', 404),
 ('back', 387),
 ('me', 371),
 ('after', 350),
 ('was', 346),
 ('there', 337),
...
 ('neck', 272),
 ('skin', 257),
 ('ache', 255),
 ('stomach', 253),
 ('not', 242)]
```

	prompt	prompt_encoded
0	Internal pain	16
1	Stomach ache	24
2	Emotional pain	6
3	Body feels weak	3
4	Heart hurts	13
5	Infected wound	14
6	Injury from sports	15
7	Open wound	21
8	Back pain	1
9	Foot ache	9
10	Head ache	12
11	Joint pain	17
12	Knee pain	18
13	Muscle pain	19
14	Neck pain	20
15	Shoulder pain	22
16	Cough	4
17	Feeling cold	7
18	Hard to breath	11
19	Blurry vision	2
20	Ear ache	5
21	Feeling dizzy	8
22	Acne	0
23	Hair falling out	10
24	Skin issue	23

Figure 4.1 Index of words and prompt

After that, before using the BiLSTM model for training, we tokenized and converted phrases into sequences and encoded the prompts, then converted them into padded sequences and placed them into a vector with a fixed size (APPENDIX B). We choose to use TF-IDF to process the phrases in the dataset before feeding them into our model [55][56]. TF-IDF is a technique used to evaluate the importance of a word within a sentence relative to a dataset. This way, common words that appear frequently across all phrases, like pain or have been given less weight, while less frequent and potentially more informative words like sharp or stomach, are emphasized, as shown in Figure 4.2.

	prompt	phrase	aches	aching	and	between	body	but	can	chest	...	on	pain	place	sharp	shoulders
0	Internal pain	i feel a sharp pain in my stomach	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.325137	0.000000	0.587857	0.000000
1	Internal pain	i have a pain internal	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.351725	0.000000	0.000000	0.000000
2	Internal pain	i have pain but i can't specify place	0.000000	0.0	0.000000	0.000000	0.000000	0.329977	0.267809	0.0	...	0.000000	0.198047	0.609164	0.000000	0.000000
3	Internal pain	my body aches on the inside between my hips an...	0.350292	0.0	0.169629	0.406423	0.273344	0.000000	0.000000	0.0	...	0.225847	0.000000	0.000000	0.000000	0.397926
4	Internal pain	i have a pain internal	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.351725	0.000000	0.000000	0.000000

Figure 4.2 TF-IDF weights for symptom detection phrases

We applied the TF-IDF to the text data, and then we created and trained a Bidirectional LSTM model [57][58]. We set 100 as the length of the vector to define the maximum sequence length for padding, which ensures that all sequences have a consistent length (APPENDIX C). We took 70% of the original data for the training set, 12% of the original data for validation, and 18% of the original data for the testing set (APPENDIX C). We used validation during training to assess the model's performance and to help prevent overfitting [61]. Our proposed BiLSTM model started with an embedding layer that transforms input sequences of word indices into dense vectors of size 128, which helped the model recognize the relationships between words, as shown in Figure 4.2. It included three BiLSTM layers with L2 regularization to prevent overfitting. The next two layers are both 64 units and have a dropout rate set to 0.5 for enhanced regularization. The third layer has 32 units and generates a single vector with a reduced 25% dropout to retain more information for the dense layers. Then, we applied 128 units with a Relu dense layer to capture the non-linear relationships and employ L2 regularization at the expense of the maximum dropout rate of 60% dropout rate for robustness. Finally, the output layer employed softmax activation function to produce probabilities across the phrases (APPENDIX C).

Model: "functional"

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 100)	0
embedding (Embedding)	(None, 100, 128)	640,000
bidirectional (Bidirectional)	(None, 100, 128)	98,816
dropout (Dropout)	(None, 100, 128)	0
bidirectional_1 (Bidirectional)	(None, 100, 128)	98,816
dropout_1 (Dropout)	(None, 100, 128)	0
bidirectional_2 (Bidirectional)	(None, 64)	41,216
dropout_2 (Dropout)	(None, 64)	0
dense (Dense)	(None, 128)	8,320
dropout_3 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 25)	3,225

Figure 4.3 Summary of BiLSTM Model

Our proposed model shows promising results, as illustrated in Figure 4.4. As we can observe, in the chart on the left, the rate of accuracy for training and validation has increased, which indicates the model is gradually improving, and in the right chart, the rate of loss for both training and validation decreased steadily, which means that the model is learning effectively while maintaining stability (APPENDIX D).

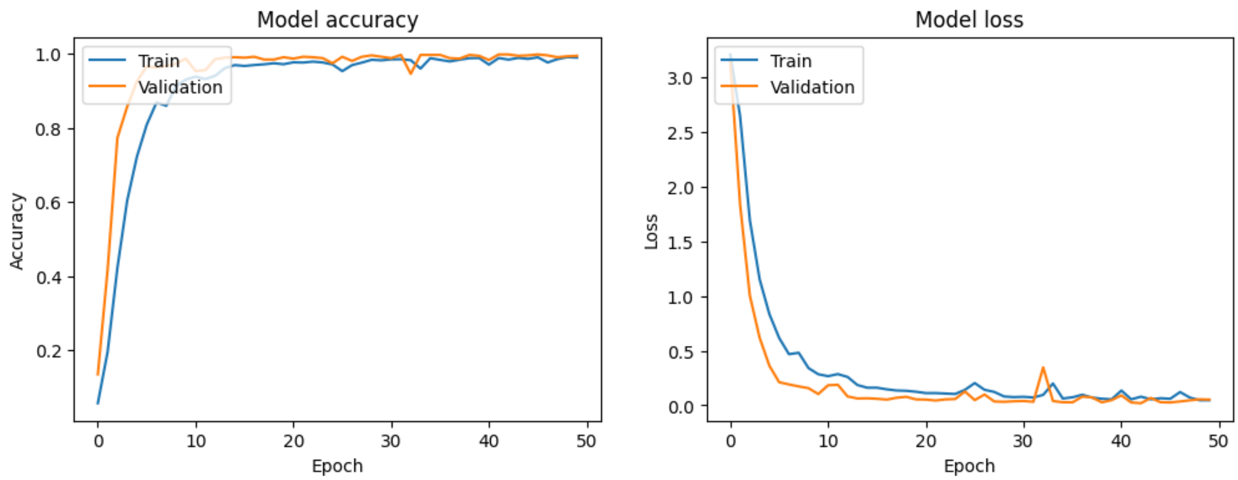


Figure 4.4 Accuracy and loss of model

4.3.1 Monitoring of BiLSTM Model

We first found the most appropriate and comprehensive dataset that is related to symptoms, in which the voice data has already been converted to text data, and then we applied the processing steps, including cleaning and normalizing the dataset. After that, we implemented the TF-IDF technique to find out the weight of each word within these medical symptoms-related phrases. We implemented a Bidirectional LSTM model in order to successfully detect symptom phrases in the text.The architecture of our proposed model consists of three bidirectional LSTM layers with L2 regularization applied for each layer to reduce the chance of overfitting. Moreover, we used dropout to learn better for getting better accuracy. The dense layer before the output used ReLU activation to capture non-linear relationships in the data, followed by a softmax output layer for multi-class classification. We utilized early stopping based on validation loss to ensure that the model did not overfit and preserved optimal performance. We considered Adam as an optimizer with a reduced learning rate that was used for efficient training (APPENDIX C).

Table 4.1 Accuracy of proposed model

Dataset	Accuracy
Train Data	98.50
Validation Data	98.55
Test Data	98.61

4.4 Implementation of Ministry of Health Service Provider Locations

In this section, we applied some process with another dataset [53], which contains 19 types of health service providers and consists of a total of 11,606 records, with the most recent update in May 2023. Each health service provider has a variety of attributes, such as the type of service provider and specific details related to the service provider including address information, which encompasses fields like Address Line 1, Community, and Postal Code. Our purpose was to find the appropriate algorithm for this dataset to find a closet medical center in the neighborhood. We analyzed this dataset to get meaningful insights into the types and distribution of healthcare services across various regions. We represented the distribution of various service types provided by the Ministry of Health as shown in Figure 4.5. This dataset includes a wide range of healthcare-related services. As we can observe, pharmacies have the largest rate in the distribution of service types, with over 5,000 centers across regions. These charts represented the lowest counts of service types, which should be increased ((APPENDIX E).

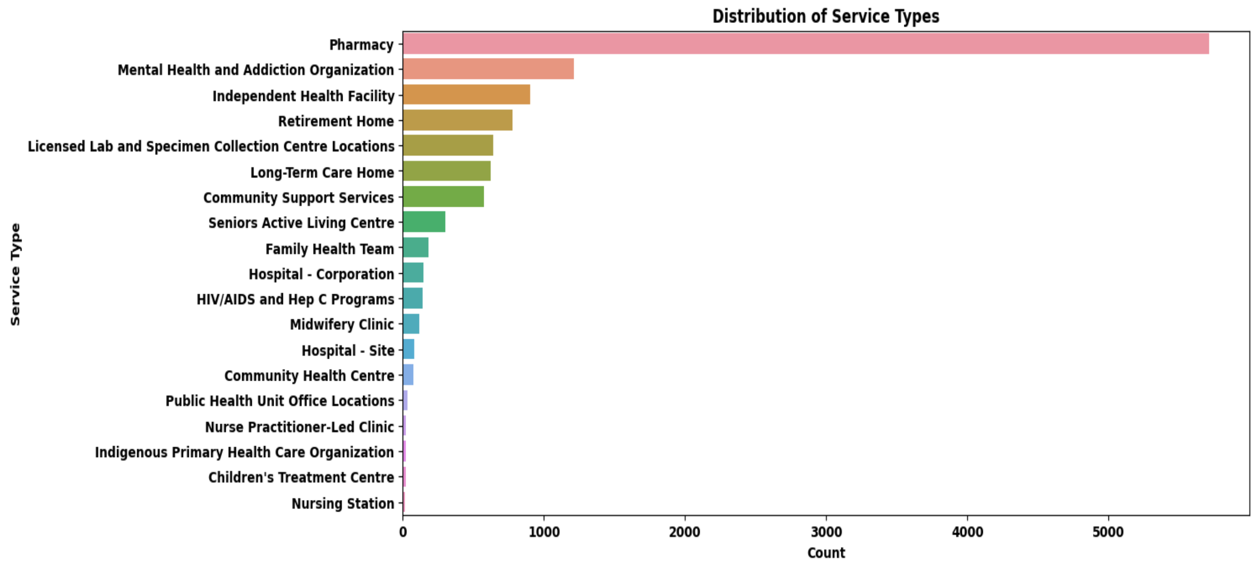


Figure 4.5 The distribution of service types

The second chart presents the distribution of healthcare services across the top 20 communities, as shown in Figure 4.6. Toronto is one of the communities with the most services available; Ottawa and Hamilton have high numbers. The graph indicates that bigger cities have more places offering healthcare services, showing their population sizes and the need to support a larger number of residents (APPENDIX F). This distribution gives insight into how healthcare services are spread geographically and helps identify potential areas that might benefit from increased healthcare access, as shown in Figure 4.6.

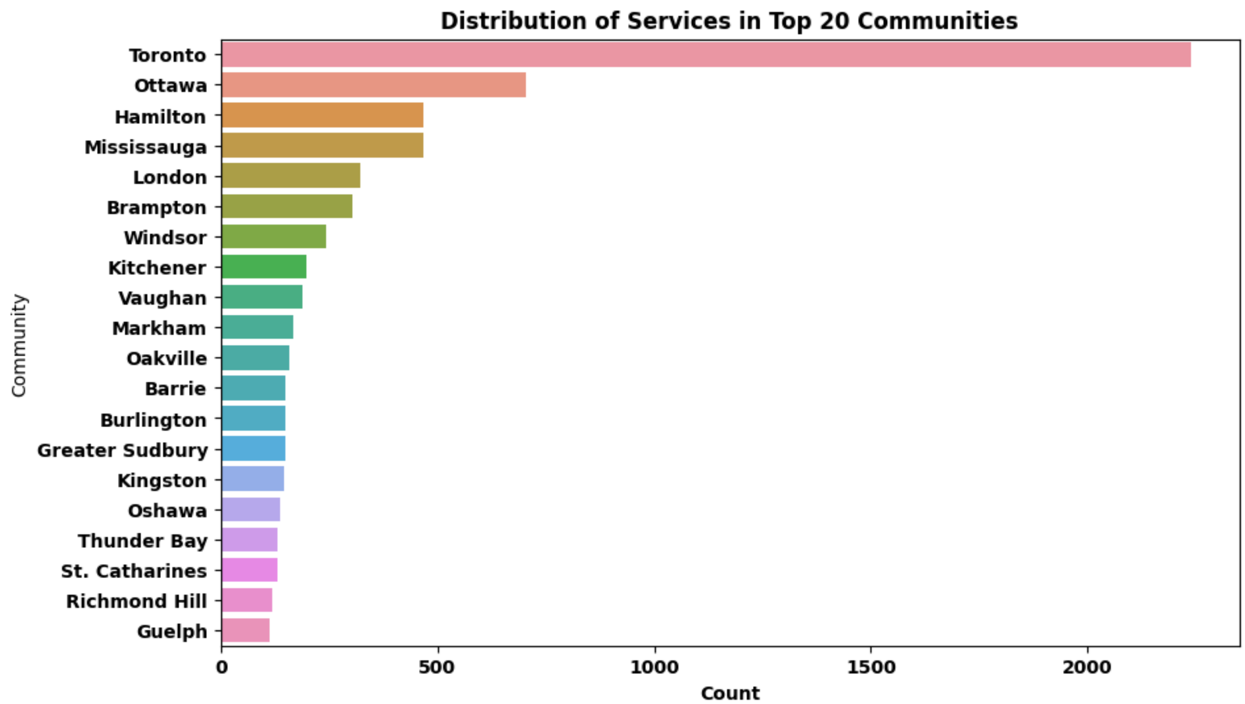


Figure 4.6 Distribution of services in top 20 communities

We combined the previous two charts in a stacked bar chart to obtain more insight into the distribution of healthcare services between communities, as shown in Figure 4.7. This visualization presents insights into the availability and variety of healthcare services within the top 20 communities. Toronto has the highest diversity and quantity of services, indicating extensive healthcare infrastructure. As we can see, Thunder Bay and Ottawa also have a significant service of pharmacies, mental health services, and long-term care homes. We may identify significant shortcomings and regions where certain healthcare services are concentrated simply according to this combined visualization. This knowledge is helpful for organizing and maximizing the distribution of healthcare resources among various regions.

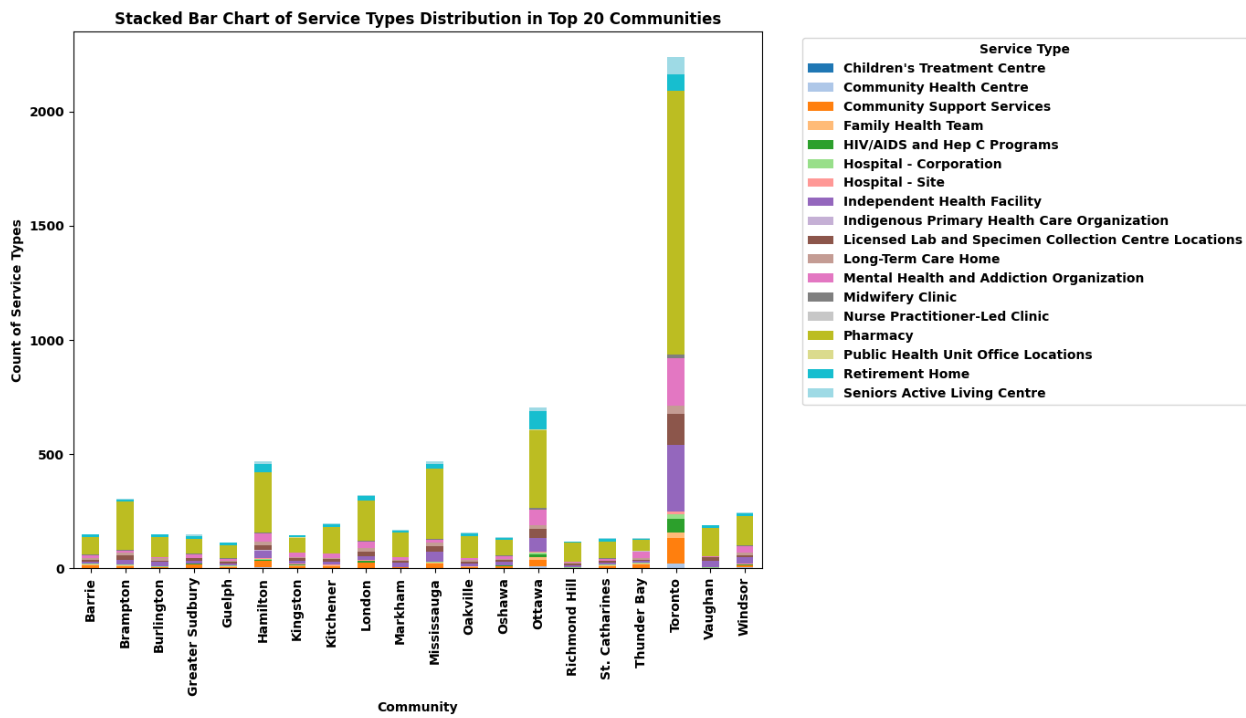


Figure 4.7 Distribution of service types in top 20 communities

In this analysis, we sorted and cleaned the dataset to focus on healthcare services available near specific postal codes, as shown in Figure 4.8. We identified the types of services available and counted the total number of service types within the area for all of them (APPENDIX G).

	POSTAL_COD	Community	Service_Types	Service_Type_Count
0	K0A1A0	Mississippi Mills	[Licensed Lab and Specimen Collection Centre L...	9
1	K0A1E0	Clarence-Rockland	[Pharmacy, Long-Term Care Home]	2
2	K0A1L0	Ottawa	[Licensed Lab and Specimen Collection Centre L...	6
3	K0A1M0	Casselman	[Independent Health Facility, Pharmacy, Mental...	4
4	K0A1N0	Clarence Creek	[Pharmacy, Long-Term Care Home]	2
5	K0A1R0	Berwick	[Pharmacy]	1
6	K0A1W0	Embrun	[Midwifery Clinic, Retirement Home, Pharmacy, ...	5
7	K0A1W1	Embrun	[Pharmacy, Mental Health and Addiction Organiz...	2
8	K0A2A0	Clarence-Rockland	[Retirement Home]	1
9	K0A2M0	Limoges	[Pharmacy, Independent Health Facility, Long-T...	4

Figure 4.8 Distribution of postal code and community

During the data processing phase, each postal code and community was known by a specific numerical code to facilitate analysis, as shown in Figure 4.9. After encoding postal code and community, we can more easily perform statistical analyses, visualize patterns, and manage data related to the distribution of healthcare services across different regions. This approach ensures that our processing steps are efficient and scalable, particularly when dealing with datasets as large as this one, which contains over 11,600 rows (APPENDIX G).

	COMMUNITY	POSTAL_COD	COMMUNITY_ENCODED	POSTAL_COD_ENCODED	Service_Type_Count
0	Geraldton	P0T1M0	189	6894	8
1	Manitouwadge	P0T2C0	298	6898	8
2	Schreiber	P0T2S0	445	6902	3
3	Nipigon	P7C1A7	357	7346	1
4	Toronto	M5T1Z1	523	4824	1
...	...	...	...	...	...
11601	Mississauga	L5V2X5	326	2698	3
11602	Scarborough	M1S5T7	443	4057	2
11603	Scarborough	M1V5L3	443	4095	3
11604	Maple	L6A2L1	300	2717	2
11605	Newmarket	L3Y1J6	352	1914	3

11606 rows × 5 columns

Figure 4.9 Encoded of postal code and community

4.4.1 Implementation of KNN algorithms for Medical Service Provider

In this section, we implemented the KNN algorithm for the medical service provider dataset to find the closest medical center in the neighborhood. The KNN algorithms is a non-parametric and instance-based learning algorithm that stores the entire training dataset and makes predictions [63][64]. These algorithms assign the label that is most frequent among the K nearest neighbors They can also be useful to classify and regress. They are applied in use cases such as image recognition, recommendation systems or medical diagnosis (APPENDIX H). In our project, we used KNN based on city and postal code, achieving an accuracy of 0.9156, as depicted in Figure 4.10.

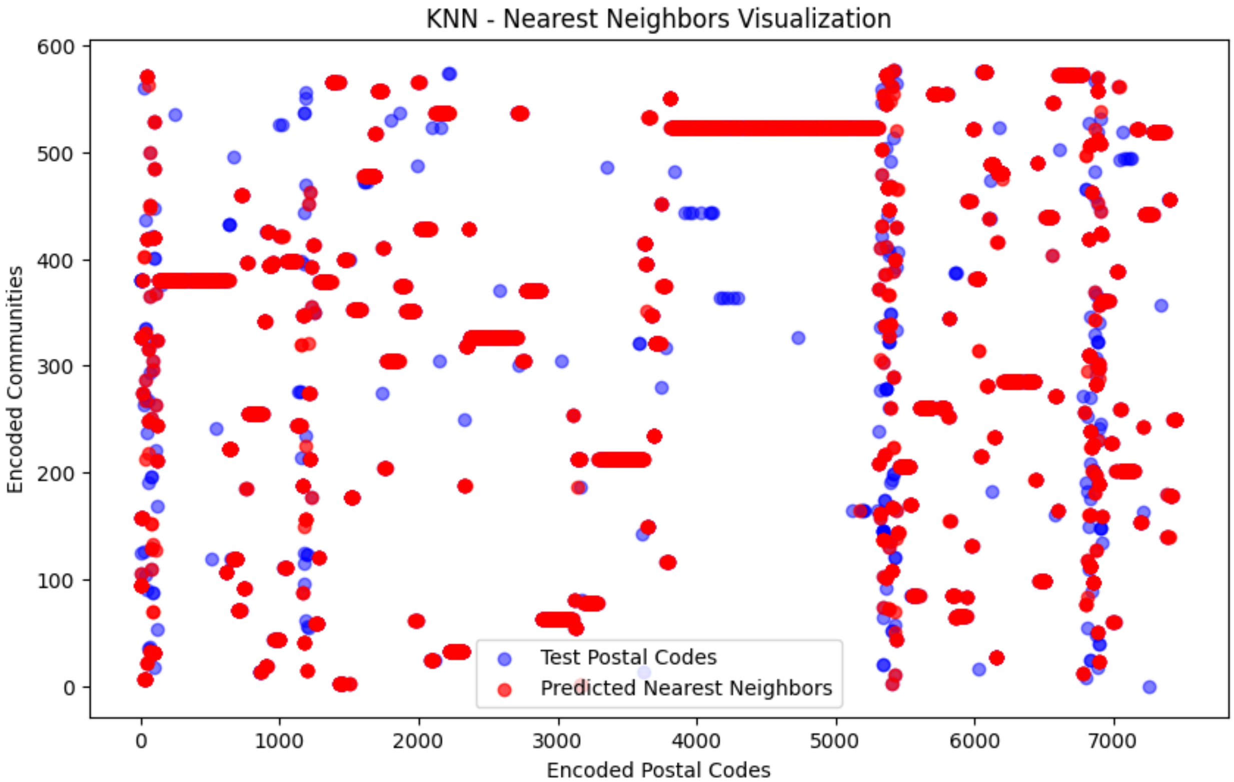


Figure 4.10 KNN - Nearest neighbors visualization

4.4.2 Monitoring of KNN algorithms with ROC for Generative based Module

In this section, after implementing the KNN algorithm, we evaluated the KNN algorithm with an ROC curve, which is a graphical representation used to evaluate the performance of a classification model [45][46]. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings (APPENDIX I). It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The chart of ROC is illustrated in Figure 4.11.

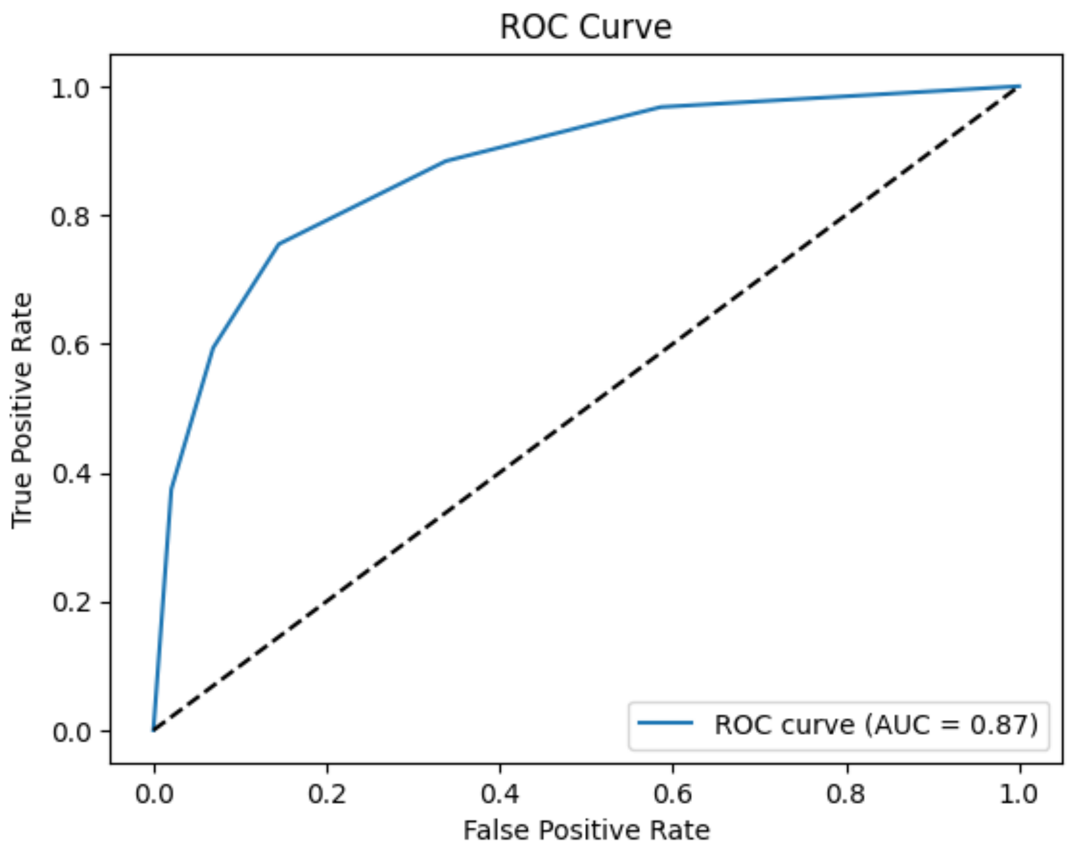


Figure 4.11 ROC (Receiver Operating Characteristic) curve

4.5 Summary

This chapter presented the implementation of our proposed model. First, we established our testbed in one machine by using Visual Studio Code (VSCode) [71]. Then, we discussed the implementation of the two modules for each dataset. We processed the data using TF-IDF to emphasize important terms in medical phrases [53]. In our experiment, we implemented the Bidirectional LSTM layers as a framework and used L2 norm for regularization. Data was divided into training dataset, validation dataset and evaluation to robust model training. Moreover, we analyzed the distribution of healthcare services with the Ministry of Health dataset, in order to improve our perception of the availability of services in various communities. The performance of KNN algorithm was validated using ROC curves for the healthcare services dataset (APPENDIX I). Finally, we discussed our implementation to provide a good model for our project, which is done by Python languages and NLP libraries.

Table 4.1 provides an overview of the proposed model, the technologies and the enhancements depending on each layer in our model based on both chapters 3 and 4.

Table 4.2 Summary of the proposed architecture for healthcare chatbot

Layers	Module	Technologies/ Dataset	Duty	Achievements
UI	User Interface	Rule Based Chatbot AI Based Chatbot	Collect user input and deliver response and choose appropriate chatbot type	Manages interaction between user and system
Natural Language Generation	Retrieval Based	Public health dataset, TF-IDF, Python libraries	Process user query, retrieve and generate response	Normalizing and cleaning the data
Natural Language Generation	Generative Based	Public health dataset, TF-IDF, Python libraries,	Detection of Symptomps	Enhance model reliability
Preprocessing and Processing	Retrieval Based Generative Based	NLP Libraries (e.g., NLTK, SpaCy), Machine Learning Models (e.g., KNN, BiLSTM)	Prepare input data for models, Encode and classify text data	Analyzes and classifies data for accurate input detection
Monitoring	BILSTM Model And KNN	ROC Curves, scikit-learn, Python	Evaluate model performance, calculate accuracy and AUC	Enhance model reliability, improve classification accuracy

CHAPTER 5 CONCLUSION

This dissertation presented our model for enhancing healthcare chatbots through NLP and ML for symptom detection and a KNN-based system for medical center locations. This chapter concludes the dissertation. The first section provides a summary of our work. The second section details the limitations of our work. Finally, the third section discusses future works.

5.1 Summary of Works

In this research, we developed a two-part system to address the growing need for efficient healthcare service delivery. The BILSTM model was built for symptom detection, utilizing text analysis to provide accurate medical service recommendations. The KNN model was employed to find the closest medical centers, factoring in geographical data like postal codes and cities. The performance was evaluated using ROC curves, ensuring high precision. The models were implemented on MacBook Pro (15-inch, 2016) with macOS Monterey (Version 12.7.6).

By conducting the literature review on healthcare chatbots, we identify several weaknesses and shortcomings in existing systems, including a lack of high accuracy for detecting symptoms, difficulties in finding the medical center to book an appointment, lack of data interoperability, and lack of scalability. Therefore, several questions come to mind. Which method can be used for healthcare chatbots while helping healthcare providers with administrative tasks to have more time and resources available for directly caring for patients at an acceptable cost? What challenges do both patients and doctors face when using healthcare chatbots, and how may these challenges be dealt with to maximize the advantages of using chatbots in healthcare?

Regarding all the weaknesses mentioned above, this dissertation proposes an effective and accurate model for healthcare chatbots to enhance patients' knowledge about their health. In addition, we propose a method used for healthcare chatbots while helping healthcare providers with administrative tasks to have more time and resources available for directly caring for patients at an acceptable cost. In this model, we aim to design an architecture for addressing the challenges of using healthcare chatbots by patients and doctors and maximizing their benefits in healthcare and we aim to implement and evaluate the performance of the proposed model, method, and strategy in terms of responding time and precision in the answers of the chatbots.

To accomplish our objectives, we developed BILSTM for one of the phases of our project, which is related to the detection of symptoms, and KNN for the other phase, which is related to finding the closest medical center in the neighborhood. We considered Adam as an optimizer of the model, and we evaluated KNN with the ROC curve. The outcomes of the proposed model for healthcare chatbots are promising, and in the following paragraphs, we will discuss each component of the system, its results, and how it contributes to achieving our objectives.

Our proposed model for the healthcare chatbot system consists of several key components that work together to deliver efficient and accurate responses to user queries. The system is designed to process input messages, detect symptoms, and recommend medical services by leveraging both rule-based and AI-based approaches. The proposed system lets the user communicate through a user interface based on input messages. Formulating the responses that need to be delivered, in turn, via the same interface, the system integrates two kinds of bots: one with rule-based approaches, where the developed modules follow a set of predefined rules and patterns while answering the user queries, and another is AI-based, using machine learning techniques for dynamic and more context-aware response generation. The core component of the system is a natural language generation mechanism, powered by retrieval-based methods. It has a data corpus component that stores relevant information for retrieval and a dialog agent identifying user intents to guide the responses. [4][6][23]

In the third component of this architecture, we have the preprocessing and processing layer. The preprocessing layer applies a variety of NLP techniques for parsing input data in preparation. With tools like tokenization, TF-IDF, and Word2Vec, the input is transformed into numerical formats that can be analyzed. The processing layer normalizes and encodes the data to ensure consistency, further handling with vector recognition, embedding techniques, and a Bi-LSTM model in handling sequential data for precise predictions. Therefore, it applies generative-based methodologies of content determination and sentence planning in formulating an appropriate and relevant response from the input pre-processed by the user to guarantee a meaningful interaction [6][23].

For our BILSTM model, we developed and trained a model in which the sequence length for padding was set to 100, thus guaranteeing coherence with regard to the input provided. The data is then divided into 70% for training, 12% for validation, and 18% for testing. Validation during training was crucial; this would monitor the performance and avoid overfitting [37]. The architecture of our model also contains three BiLSTM layers with L2 regularization to reduce overfitting. The first two layers consisted of 64 units each with a 0.5 dropout rate to regularize the neural network, while the third one was made up of 32 units and a dropout rate of 25% as a way to retain information useful for the dense layers. As results, we obtain an accuracy of 98.61% for the training set, 98.55% for the validation set and 98.50% for the test set, indicating our performance for the BILSTM proposed model.

We processed another dataset consisting of 11,606 records of 19 types of health service providers, whose last update was in May 2023 [52]. The attributes included in each record are the provider type, home address, and detailed address (Address Line 1, Community, and Postal Code) [53]. The objective of this effort was to determine which is the best algorithm for finding the nearest medical center to a given neighborhood. This dissertation analyzed this dataset in search of any useful insight related to the distribution and types of healthcare services across various regions. Then, we applied the K-Nearest Neighbors to get the closest medical center in the neighborhood of the Medical Service Provider dataset. KNN is considered to be the most basic type of non-parametric and instance-based learning; it holds every example from the training dataset and makes a prediction via the most frequent label amongst the k-nearest neighbors [43][44]. As for the project at hand, this ensues from city and postal code-based for KNN with 0.9156.

Finally, after implementing the KNN algorithm, we plotted the performance using an ROC curve, which is usually a graphical performance measure presentation for our classification models. An ROC curve is a plot between TPR and FPR at different threshold levels, which presents a visual of a model's performance [68][69].

5.2 Limitations

Our proposed model has several limitations during the creation and testing stages, which have revealed multiple areas of possible improvement. Among the major limitations was the quality and diversity of the dataset used for training the models. The dataset, although large in the data from health service providers and medical records, was nevertheless limited concerning geographical reach and a wide range of various conditions. The fact is, the models might not generalize well across different regions or when certain medical services or neighborhoods are under-represented in the dataset, considering the application of the KNN algorithm for finding the nearest medical centers. Thirdly, there is the availability of recent data; though the dataset was updated in May

2023, updates will need to be made periodically to keep the system current and accurate. Another limitation is the scalability of the models. Whereas KNN and the NLP models performed well within the objectives of the project but may not perform as efficiently on large datasets or real-world healthcare applications that are run concurrently with hundreds of users. The KNN algorithm may suffer from high memory usage since it stores the entire dataset for instance-based learning, which could become an overload in a system that needs to process real-time data efficiently. Moreover, the integration of models was relatively simplistic and might not capture the complex interactions between symptoms, medical services, and geographical locations in the real world. For example, symptom detection and suggestions by an NLP model were based on a pre-set structure and might fail when symptoms consisted of more serious medical conditions that require more sophisticated understanding and context. The chatbot interface was itself dependent upon the input a user would give it and could be restricted by the accuracy and clarity of information being put forth by the users, which sometimes is not very reliable within the scope of the clinical domain.

Lastly, the models were tested in a controlled environment using the MacBook Pro; however, the testing environment was not representative of real-world conditions with variables like network latency, concurrent user queries, and hardware limitations in lower-resource environments that are likely to reduce the performance of the model. Moreover, the use of certain pre-processing techniques like TF-IDF and Word2Vec in the NLP part is effective but may not be precisely efficient in any other healthcare context where real-time decision-making or highly personalized medical services are needed. In conclusion, our project presented strong results in terms of accuracy and usability; the limitations related to the dataset, scalability, model integration, and real-world testing conditions suggest areas for future work to improve the robustness and applicability of the healthcare chatbot system.

5.3 Future Work

The first improvement in our proposed model would be expanding diversification and extension of the training dataset. Further work on this system might include data from more medical service providers and larger coverage of geographical areas to make the KNN model more adaptable and accurate for different regions. Moreover, incorporating more up-to-date, diverse medical datasets that include real-time updating or even patient-generated data would enhance the relevance and robustness of the recommendations the system could provide. Further development of this would enable the chatbot to give more accurate and personalized suggestions of medical centers and services even in less represented areas or rare medical conditions.

Optimization of the computational performance of the KNN algorithm may be done in further developments, or other models of machine learning more real-time friendly may be considered. Some of the approaches that could make the KNN much faster while shrinking memory overhead can be KD-Tree or Ball Tree algorithms [51][52]. It would maybe have led other algorithms, like random forests or gradient-boosted trees, through a performance test in finding a location for a medical center that may yield similar or higher performances but with better computational efficiency.

Further sophistication can be added in the integration of the NLP model for the symptom detection process. For example, the technique, as it stands, relies on conventional techniques such as TF-IDF and Word2Vec. Some deep learning advanced techniques, for example, transformers or BERT (Bidirectional Encoder Representations from Transformers), can add value to it [74]. These models will be particularly powerful for context-based understanding and may make the chatbot better at interpreting complex medical symptoms as well as language variations, thus bringing forth even more precise recommendations. Additionally, fine-tuning them on domain-specific healthcare data might enhance their detection of rare or nuanced medical conditions to a great extent [75].

One of the most promising factors for future work involves the integration of real-time data processing. By integrating real-time processing APIs from healthcare providers, the system will be more up-to-date with regard to the availability of services, waiting time estimates, or even personalized advice regarding the user's current health data. In such a way, this bot can be practical in real situations where response timing is important, like emergency medical services.

Another suggestion would be to enhance the user interface and user interaction features. As it is, this chatbot depends on a user who can input the symptoms or queries properly, which may not always be the case in a healthcare scenario. Speech recognition or multi-modal interfaces can enable users to interact through voice or even through pictures, such as sending pictures of symptoms. This would make the system more accessible and user-friendly for those people who might have difficulties in typing or describing their symptoms with precision.

Lastly, future work could explore the integration of personalized medicine approaches. The patient's history and preference, even genetic information, could be integrated into a system providing recommendations on health that are most personal. This will enable the strategy to provide advice that best fits the health profile for each individual; therefore, the advice will be more relevant and effective. These enhancements are going to enable the system to perform highly accurate, personalized healthcare services.

In the future, work on a healthcare chatbot should include dataset expansion, scalability of the model, deep learning techniques, real-time data processing, advanced user interaction interfaces, and integrating personalized recommendations for healthcare. These developments would greatly enhance the system's effectiveness, making it a more robust and versatile tool in the healthcare sector.

REFERENCES

[1] TYAGI, N., & Bhushan, B. (2023). Demystifying the Role of Natural Language Processing (NLP) in Smart City Applications: Background, Motivation, Recent Advances, and Future Research Directions. Wireless Personal Communications. https://doi.org/10.1007/s11277-023-10312-8

[2] Vashishtha, E., & Kapoor, H. (2023). Enhancing Patient Experience by Automating and Transforming Free Text into Actionable Consumer Insights: A Natural Language Processing (NLP) Approach. International Journal of Health Sciences and Research, 13(10), 275–288. https://doi.org/10.52403/ijhsr.20231038

[3] Vincent Quoc-Nam Trinh, Zhang, S., Kovoor, J. G., Gupta, A. K., Weng Onn Chan, Gilbert, T., & Bacchi, S. (2023). The Use of Natural Language Processing (NLP) in Detecting and Predicting Falls Within the Healthcare Setting: A Systematic Review. International Journal for Quality in Health Care, 35(4). https://doi.org/10.1093/intqhc/mzad077

[4] Rathore, Y., Vandana Mishra Chaturvedi, Khadilkar Sujay madhukar, Vijay Suresh Karwande, Anil Haribhau Rokade, & Yogesh Nagargoje. (2023). Patient Engagement and Satisfaction in Ai-Enhanced Healthcare Management. https://doi.org/10.1109/icaiihi57871.2023.10489712

[5] Prof. D. L. Falak. (2023). A Survey on Healthcare Virtual Assistant Using Generative AI. International Journal for Research in Applied Science and Engineering Technology, 11(11), 130–134. https://doi.org/10.22214/ijraset.2023.56449

[6] Sathe, N., Vaibhav Deodhe, Sharma, Y., & Shinde, A. (2023). A Comprehensive Review of AI in Healthcare: Exploring Neural Networks in Medical Imaging, LLM-Based Interactive Response Systems, NLP-Based EHR Systems, Ethics, and Beyond. https://doi.org/10.1109/icacctech61146.2023.00108

[7] Prashant Kumar Nag, Bhagat, A., R. Vishnu Priya, & Deepak Kumar Khare. (2023). Emotional Intelligence Through Artificial Intelligence: NLP and Deep Learning in the Analysis of Healthcare Texts. https://doi.org/10.1109/icaiihi57871.2023.10489117

[8] Iqbal, S. (2023). Artificial Intelligence Tools and Applications for Elderly Healthcare - Review. https://doi.org/10.1145/3594315.3594347

[9] Shetty, S. H., Shetty, S., Singh, C., & Rao, A. (2022). Supervised Machine Learning: Algorithms and Applications. Fundamentals and Methods of Machine and Deep Learning, 1–16. https://doi.org/10.1002/9781119821908.

[10] Jin, W. (2020). Research on Machine Learning and Its Algorithms and Development. Journal of Physics: Conference Series, 1544(1), 012003. https://doi.org/10.1088/1742-6596/1544/1/012003

[11] In Institution of Engineering and Technology eBooks. Institution of Engineering and Technology. https://doi.org/10.1049/pbte081e

[12] Malamas, N., Papangelou, K., & Symeonidis, A. L. (2022). Upon Improving the Performance of Localized Healthcare Virtual Assistants. Healthcare, 10(1), 99. https://doi.org/10.3390/healthcare10010099

[13] Liu, S., Liu, Y., Xu, X., Chen, R., Liang, D., Jin, Q., Liu, H., Chen, G., & Zhu, Y. (2023). Accelerated cardiac diffusion tensor imaging using deep neural network. Physics in Medicine & Biology/Physics in Medicine and Biology, 68(2), 025008–025008. https://doi.org/10.1088/1361-6560/acaa86

[14] Chaithra, Fernandes, R., Rodrigues, A. P., & Venkatesh, N. (2020). An Approach Toward Stateless Chatbots with the Benefit of Tensorflow Over Spacy Pipeline. Advances in Intelligent Systems and Computing, 483–496. https://doi.org/10.1007/978-981-15-3514-7_37

[15] Shinde, N. V., Akhade, A., Bagad, P., Bhavsar, H., Wagh, S. K., & Kamble, A. (2021, June 1). Healthcare Chatbot System using Artificial Intelligence. IEEE Xplore. https://doi.org/10.1109/ICOEI51242.2021.9452902

[16] Badlani, S., Aditya, T., Dave, M., & Chaudhari, S. (2021, May 1). Multilingual Healthcare Chatbot Using Machine Learning. IEEE Xplore. https://doi.org/10.1109/INCET51464.2021.9456304

[17] Rathi, N. D., & Deorankar, A. V. (2023). Overview and an Approach to AI based Healthcare System by using NLP. https://doi.org/10.1109/icaect57570.2023.10117865

[18] Rajasrikar Punugoti, Duggar, R., Risha Ranganath Dhargalkar, & Neha Bhati. (2023). Intelligent Healthcare: Using NLP and ML to Power Chatbots for Improved Assistance.

https://doi.org/10.1109/icicat57735.2023.10263708or Multidisciplinary Research. (n.d.). www.ijfmr.com. https://www.ijfmr.com

[19] Rajesh, V., Perumal, B., Ganesh, U., Rajkumar, V., Kumar, D., & Kumar, M. (2023). Building Customer Support Chatbots With Intent Recognition. https://doi.org/10.1109/vitecon58111.2023.10157329

[20] Jeidy Panduro-Ramirez, None MelanieLourens, None AnitaGehlot, Durga Prasad Gangodkar, None TanushreeSarkar, & None SivakumarSattanathan. (2023). Artificial Intelligence's Function In Chatbots That Use NLP And SVM Algorithms. https://doi.org/10.1109/icacite57410.2023.10182854

[21] Thakre, K., Rothe, Dr. P. R., Kukade, S., Shinde, P., & Madame, K. (2022). Health Care Chatbot Using NLP and Flask. International Journal for Research in Applied Science and Engineering Technology, 10(5), 2632–2654. https://doi.org/10.22214/ijraset.2022.42854

[22] Wang, Y., Tafti, A., Sohn, S., & Zhang, R. (2019, June 1). Applications of Natural Language Processing in Clinical Research and Practice. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-5006

[23] Ms. Shruti Desai, & Mr. Rahul Erulan. (2023). Medical Chatbot in Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 11(4), 2101–2103. https://doi.org/10.22214/ijraset.2023.50502

[24] DALE, R. (2015). NLP meets the cloud. Natural Language Engineering, 21(4), 653–659. https://doi.org/10.1017/s1351324915000200

[25] M. Bahja, "Natural Language Processing Applications in Business," *E-Business*, May 2020, doi: https://doi.org/10.5772/intechopen.92203.M. Bahja, "Natural Language Processing Applications in Business," *E-Business*, May 2020, doi: https://doi.org/10.5772/intechopen.92203.

[26] Crema, C., Attardi, G., Sartiano, D., & Redolfi, A. (2022). Natural language processing in clinical neuroscience and psychiatry: A review. Frontiers in Psychiatry, 13. https://doi.org/10.3389/fpsyt.2022.946387

[27] Zhong, J., & Yi, X. (2020). Categorizing Patient Disease into

ICD-10 with Deep Learning for Semantic Text Classification. Recent Trends in Computational Intelligence. https://doi.org/10.5772/intechopen.91292

[28] COŞKUN, M., YILDIRIM, Ö., UÇAR, A., & DEMIR, Y. (2017). AN OVERVIEW OF POPULAR DEEP LEARNING METHODS. *European Journal of Technic*, *7*(2), 165–176. https://doi.org/10.23884/ejt.2017.7.2.11

[29] Chen, X., Xie, H., Wang, F. L., Liu, Z., Xu, J., & Hao, T. (2018). analysis of natural language processing in medical research. *BMC Medical Informatics and Decision Making*, *18*(S1). https://doi.org/10.1186/s12911-018-0594-x

[30] Wu, H., Wang, M., Wu, J., Francis, F., Chang, Y.-H., Shavick, A., Dong, H., Poon, M. T. C., Fitzpatrick, N., Levine, A. P., Slater, L. T., Handy, A., Karwath, A., Gkoutos, G. V., Chelala, C., Shah, A. D., Stewart, R., Collier, N., Alex, B., & Whiteley, W. (2022). A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *Npj Digital Medicine*, *5*(1), 1–15. https://doi.org/10.1038/s41746-022-00730-6

[31] Aramaki, E., Wakamiya, S., Yada, S., & Nakamura, Y. (2022). Natural Language Processing: from Bedside to Everywhere. *Yearbook of Medical Informatics*, *31*(01), 243–253. https://doi.org/10.1055/s-0042-1742510

[32] Uzuner, O., Yetişgen, M., & Stubbs, A. (2014, August 1). *Biomedical/Clinical NLP* (Q. Liu & F. Xia, Eds.). ACLWeb; Dublin City University and Association for Computational Linguistics. https://aclanthology.org/C14-30018

[33] Névéol, A., & Zweigenbaum, P. (2015). Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare. *Yearbook of Medical Informatics*, *24*(01), 194–198. https://doi.org/10.15265/iy-2015-035

[34] Shinde, N. V., Akhade, A., Bagad, P., Bhavsar, H., Wagh, S. K., & Kamble, A. (2021, June *Healthcare Chatbot System using Artificial Intelligence*. IEEE Xplore. https://doi.org/10.1109/ICOEI51242.2021.9452902

[35] Goel, R., Ratnesh Puri Goswami, Somesh Totlani, Arora, P., Bansal, R., & Vij, D. (2022). Machine Learning Based Healthcare Chatbot. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. https://doi.org/10.1109/icacite53722.2022.9823901

[36] Menon, N. G., Shrivastava, A., Bhavana, N. D., & Simon, J. (2021). Deep Learning based Transcribing and Summarizing Clinical Conversations. *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. https://doi.org/10.1109/i-smac52330.2021.9640683

[37] Garg, S. (2021). Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. https://doi.org/10.1109/confluence51648.2021.9377188

[38] Zhou, B., Yang, G., Shi, Z., & Ma, S. (2022). Natural Language Processing for Smart Healthcare. *IEEE Reviews in Biomedical Engineering*, 1–17. https://doi.org/10.1109/RBME.2022.3210270

[39] Kamra, V., Kumar, P., & Mohammadian, M. (2021, May 1). Natural Language Processing Enabled Cognitive Disease Prediction Model for Varied Medical Records Implemented over ML Techniques. IEEE Xplore. https://doi.org/10.1109/ICSPC51351.2021.9451785

[40] Chkirbene, Z., Mohamed, A., Aiman Erbad, & Mohsen Guizani. (2020). *Smart Edge Healthcare Data Sharing System*. https://doi.org/10.1109/iwcmc48107.2020.9148195

[41] Agarwal, J., Kumar, M., & Srivastava, A. K. (2021, December 1). *Symptoms Based Disease Diagnosis and Treatment Recommendation*. IEEE Xplore. https://doi.org/10.1109/ComPE53109.2021.9751805

[42] Agrawal, R., Bajaj, H., Gupta, R., & Asma. (2021). Healthcare recommendation system using patients' review. *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*. https://doi.org/10.1109/icoei51242.2021.9452960

[43] Jeidy Panduro-Ramirez, None MelanieLourens, None AnitaGehlot, Durga Prasad Gangodkar, None TanushreeSarkar, & None SivakumarSattanathan. (2023). *Artificial Intelligence's Function In Chatbots That Use NLP And SVM Algorithms*. https://doi.org/10.1109/icacite57410.2023.10182854

[44] Daram Sowmya, Kousthubha Debbata, Gunda Sai Vignesh, Dasari, Anudeep Reddy Bavanthula, & Lanke Pallavi. (2023). *Emerging Role of Healthcare Chatbots in Improving Medical Assistance*. https://doi.org/10.1109/asiancon58793.2023.10270725

[45] Dammavalam, S. R., Chandana, N., Rao, T. R., Lahari, A., & Aparna, B. (2022, November 1). *AI Based Chatbot for Hospital Management System*. IEEE Xplore. https://doi.org/10.1109/ICAN56228.2022.10007105

[46] A, P., S, M., S, M., M, S., & Madhavan, R. (2021, May 1). *Mobile Application for Hospital Management System*. IEEE Xplore. https://doi.org/10.1109/ICICCS51141.2021.9432286

[47] Christopherjames, J. E., Saravanan, M., Thiyam, D. B., S, P. A. S., Sahib, M. Y. B., Ganapathi, M. V., & Milton, A. (2021, August 1). *Natural Language Processing based Human Assistive Health Conversational Agent for Multi-Users*. IEEE Xplore. https://doi.org/10.1109/ICESC51422.2021.9532913

[48] Ms. Shruti Desai, & Mr. Rahul Erulan. (2023). Medical Chatbot in Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *11*(4), 2101–2103. https://doi.org/10.22214/ijraset.2023.50502

[49] Sharma, Y. (2022). Intelligent Conversational Healthcare Chat-Bot System. *INTERANTIONAL JOURNAL of SCIENTIFIC RESEARCH in ENGINEERING and MANAGEMENT*, *06*(05). https://doi.org/10.55041/ijsrem12864

[50] Vincent Quoc-Nam Trinh, Zhang, S., Kovoor, J. G., Gupta, A. K., Weng Onn Chan, Gilbert, T., & Bacchi, S. (2023). The Use of Natural Language Processing (NLP) in Detecting and Predicting Falls Within the Healthcare Setting: A Systematic Review. International Journal for Quality in Health Care, 35(4). https://doi.org/10.1093/intqhc/mzad077

[51] A. Singh, S. Joshi, and M. Domb, “Embedded Conversational AI, Chatbots, and NLP to Improve Healthcare Administration and Practices,” Dec. 2023, doi: https://doi.org/10.1109/icacrs58579.2023.10404985.

[52] Medical Speech, Transcription, and Intent,” www.kaggle.com. https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent

[53] locations,” geohub.lio.gov.on.ca. https://geohub.lio.gov.on.ca/datasets/lio::ministry-of-health-service-provider-locations/about (accessed Mar. 01, 2024).

[54] N. D. Rathi and A. V. Deorankar, "Overview and an Approach to AI based Healthcare System by using NLP," Jan. 2023, doi: https://doi.org/10.1109/icaect57570.2023.10117865

[55] S. Badlani, T. Aditya, M. Dave, and S. Chaudhari, "Multilingual Healthcare Chatbot Using Machine Learning," IEEE Xplore, May 01, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9456304

[56] B. Das and S. Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation," *arXiv (Cornell University)*, Jun. 2018, doi: https://doi.org/10.48550/arxiv.1806.06407.

[57] Q. Chen et al., "A Survey on an Emerging Area: Deep Learning for Smart City Data," vol. 3, no. 5, pp. 392–410, May 2019, doi: https://doi.org/10.1109/tetci.2019.2907718.,

[58] N. Malingan, "What is LSTM?," *Scaler Topics*, Apr. 18, 2023. https://www.scaler.com/topics/deep-learning/lstm/

LSTM-chapter 3N. Malingan, "What is LSTM?," *Scaler Topics*, Apr. 18, 2023. https://www.scaler.com/topics/deep-learning/lstm/

LSTM-chapter 3

[59] B. K. AlSaidi, S. K. AlMamari, and F. H. Mohideen, "A survey on supervised and unsupervised NLP algorithms for mental health detection applications," Jan. 2022, doi: https://doi.org/10.1049/icp.2023.0390.

[60] Bidirectional Long Short-Term Memory Network - an overview | ScienceDirect Topics," *www.sciencedirect.com*. https://www.sciencedirect.com/topics/computer-science/bidirectional-long-short-term-memory-network

[61] S. A. AP, Cinehaa. M, Brintha. D, Keerthana. T, and Maheswari. J, "Advanced Chatbots for Home Patients using AI," Apr. 2023, doi: https://doi.org/10.1109/icoei56765.2023.10125886.

[62] S. A. AP, Cinehaa. M, Brintha. D, Keerthana. T, and Maheswari. J, "Advanced Chatbots for Home Patients using AI," Apr. 2023, doi: https://doi.org/10.1109/icoei56765.2023.10125886.

[63] Srishti Gureja, "LSTMs and Bi-LSTM in PyTorch- Scaler Topics," *Scaler Topics*, Apr. 25, 2023. https://www.scaler.com/topics/pytorch/lstm-pytorch/ (accessed Jun. 06, 2024).

[64] C. Aki, “NLP — Vector Models - The Deep Hub - Medium,” *Medium*, Feb. 14, 2024. https://medium.com/thedeephub/nlp-vector-models-ca13f25aa450 (accessed Sep. 06, 2024).

[65] Densely Connected Convolutional Networks,” *ar5iv*. https://ar5iv.labs.arxiv.org/html/1608.06993

[66] B. K. AlSaidi, S. K. AlMamari, and F. H. Mohideen, “A survey on supervised and unsupervised NLP algorithms for mental health detection applications,” Jan. 2022, doi: https://doi.org/10.1049/icp.2023.0390.

[67] K. Soreide, “Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research,” *Journal of Clinical Pathology*, vol. 62, no. 1, pp. 1–5, Sep. 2008, doi: https://doi.org/10.1136/jcp.2008.061010

[68] F. Nazari and W. Yan, “Convolutional versus dense neural networks: comparing the two neural networks’ performance in predicting building operational energy use based on the building shape,” Building Simulation Conference proceedings, Sep. 2021, doi: https://doi.org/10.26868/25222708.2021.30735.

[69] M. Biswas, R. Sikdar, R. Sarkar, and M. Kundu, “A CNN model with pseudo dense layers: some case studies on medical image classification,” Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 13, no. 1, Aug. 2024, doi: https://doi.org/10.1007/s13721-024-00474-1.

[70] Microsoft, “Visual Studio Code,” *Visualstudio.com*, Nov. 03, 2021. https://code.visualstudio.com/download

[71] Python, “Download Python,” *Python.org*, 2019. https://www.python.org/downloads/

[72] -“PyPI · The Python Package Index,” *PyPI*. https://pypi.org/

[73] C. Cheng, Liu Xiaoli, W. Linfeng, L. Longxin, and Wu Xiaofeng, “Algorithm for k-anonymity based on ball-tree and projection area density partition,” Aug. 2019, doi: https://doi.org/10.1109/iccse.2019.8845384.

[74] Wang, “A fast exact k-nearest neighbors algorithm for high dimensional search using k-means clustering and triangle inequality,” Jul. 2011, doi: https://doi.org/10.1109/ijcnn.2011.6033373.

[75] Dini Adni Navastara, None Ihdiannaja, and Agus Zainal Arifin, “Bilingual Question Answering System Using Bidirectional Encoder Representations from Transformers and Best Matching Method,” Oct. 2021, doi: https://doi.org/10.1109/icts52701.2021.9608905.

APPENDIX A MEDICAL DATA

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
data=pd.read_csv("/Users/apple/Downloads/My Files/My File/Medical/overview-of-recordings.csv")
data_copy=pd.read_csv("/Users/apple/Downloads/My Files/My File/Medical/overview-of-recordings.csv")
```

```
data.head()
```

```
data.isna().sum()
```

APPENDIX B PREPROCESSING STEP OF MEDICAL DATA

```
grouped_data = data.groupby('prompt')['phrase'].apply(list).reset_index()
```

```
from sklearn.preprocessing import LabelEncoder,OneHotEncoder,StandardScaler


label_en = LabelEncoder()


columns_to_encode = ['audio_clipping', 'background_noise_audible', 'quiet_speaker']


for column in columns_to_encode:
    data[column] = label_en.fit_transform(data[column])
```

```python
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction.text import CountVectorizer
```

```python
label_encoder = LabelEncoder()
final_data['prompt_encoded'] = label_encoder.fit_transform(final_data['prompt'])


vectorizer = CountVectorizer()
phrase_tokens = vectorizer.fit_transform(final_data['phrase'])

phrase_tokens_df = pd.DataFrame(phrase_tokens.toarray(), columns=vectorizer.get_feature_names_out())

encoded_prompts = final_data[['prompt', 'prompt_encoded']].head()
tokenized_phrases = phrase_tokens_df.head()
```

```python
final_data[['prompt', 'prompt_encoded']]
```

```
#  Initialize a TfidfTransformer to compute TF-IDF values on phrase tokens.
# Apply TF-IDF transformation to phrase_tokens and store it in tfidf_matrix.

tfidf_transformer=TfidfTransformer()
tfidf_matrix =tfidf_transformer.fit_transform(phrase_tokens)

# Initialize a TfidfVectorizer with English stop words removal and lowercase conversion.
# Transform the 'phrase' column in final_data to a TF-IDF matrix and store it in tfidf_matrix_full.

tfidf_vectorizer_full = TfidfVectorizer(stop_words='english', lowercase=True)
tfidf_matrix_full = tfidf_vectorizer_full.fit_transform(final_data['phrase'])

# Convert the TF-IDF matrix into a DataFrame with feature names as column headers.

tfidf_df_full = pd.DataFrame(tfidf_matrix_full.toarray(), columns=tfidf_vectorizer_full.get_feature_names_out())
```

```
# Concatenate the original data with the TF-IDF DataFrame along columns to create a combined DataFrame.
# Display the first few rows of the resulting combined DataFrame.
combined_df_full = pd.concat([final_data[['prompt', 'phrase']], tfidf_df_full], axis=1)
combined_df_full.head()
```

APPENDIX C CREATING THE MODEL FOR MEDICAL DATA

```python
# Initialize a tokenizer to create a word index, considering only the top 'max_words' words.
tokenizer = Tokenizer(num_words=max_words)

# Fit the tokenizer on the 'phrase' column of final_data to build the vocabulary.
tokenizer.fit_on_texts(final_data['phrase'])

# Convert each phrase in final_data to a sequence of integer tokens based on the tokenizer's word index.
sequences = tokenizer.texts_to_sequences(final_data['phrase'])

# Pad or truncate sequences to ensure each has a length of 'max_len', padding/truncating at the end ('post').
padded_sequences = pad_sequences(sequences, maxlen=max_len, padding='post', truncating='post')

# Define the features (X) as the padded sequences and the labels (y) as the encoded prompts.
X=padded_sequences
y = label_encoder.fit_transform(final_data['prompt'])

# Split the data into training (70%) and evaluation (30%) sets.
X_train,X_eval,y_train,y_eval=train_test_split(X,y,test_size=0.3,random_state=42)


# Further split the evaluation set into validation (40%) and test (60%) sets.
X_val,X_test,y_val,y_test=train_test_split(X_eval,y_eval,test_size=0.6,random_state=42)
```

```
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, LSTM, Dense, Bidirectional, Dropout,Input
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.models import Model
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import classification_report

max_len = 100  #  Adjusting this based on your data
# X_padded = pad_sequences(X, maxlen=max_len, padding='post', truncating='post')
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Build the Bidirectional LSTM model
input_layer = Input(shape=(max_len,))
# Embedding layer
embedding_layer = Embedding(input_dim=5000, output_dim=128)(input_layer)
# First Bidirectional LSTM layer
lstm1 = Bidirectional(LSTM(64, return_sequences=True))(embedding_layer)
dropout1 = Dropout(0.5)(lstm1)
# Second Bidirectional LSTM layer
lstm2 = Bidirectional(LSTM(64, return_sequences=True))(dropout1)
dropout2 = Dropout(0.5)(lstm2)
# Third Bidirectional LSTM layer
lstm3 = Bidirectional(LSTM(32, return_sequences=False))(dropout2)
dropout3 = Dropout(0.5)(lstm3)
# Dense layer before output
dense_layer = Dense(128, activation='relu')(dropout3)
dropout4 = Dropout(0.5)(dense_layer)
# Output layer
output_layer = Dense(len(label_encoder.classes_), activation='softmax')(dropout4)
# Defining the model
model = Model(inputs=input_layer, outputs=output_layer)
# Compiling the model
optimizer = Adam(learning_rate=1e-3)
model.compile(optimizer=optimizer, loss='sparse_categorical_crossentropy', metrics=['accuracy'])
model.summary()
```

APPENDIX D PLOT OF BI LSTM MODEL

```
import matplotlib.pyplot as plt

plt.figure(figsize=(12, 4))
plt.subplot(1, 2, 1)
plt.plot(history.history['accuracy'])
plt.plot(history.history['val_accuracy'])
plt.title('Model accuracy')
plt.ylabel('Accuracy')
plt.xlabel('Epoch')
plt.legend(['Train', 'Validation'], loc='upper left')

# Plot training & validation loss values
plt.subplot(1, 2, 2)
plt.plot(history.history['loss'])
plt.plot(history.history['val_loss'])
plt.title('Model loss')
plt.ylabel('Loss')
plt.xlabel('Epoch')
plt.legend(['Train', 'Validation'], loc='upper left')
plt.show()
```

APPENDIX E DISTRIBUTION OF SERVICE TYPE

```
import matplotlib.pyplot as plt
import seaborn as sns

# Distribution of service types
plt.figure(figsize=(13, 6))
sns.countplot(y='SERVICE_TY', data=Hdata, order=Hdata['SERVICE_TY'].value_counts().index)
plt.title('Distribution of Service Types',fontweight='bold')
plt.xlabel('Count',fontweight='bold')
plt.ylabel('Service Type',fontweight='bold')
plt.xticks(fontweight='bold')
plt.yticks(fontweight='bold')
plt.show()

# Distribution of services over communities
plt.figure(figsize=(10, 6))
sns.countplot(y='COMMUNITY', data=Hdata, order=Hdata['COMMUNITY'].value_counts().index[:20])
plt.title('Distribution of Services in Top 20 Communities',fontweight='bold')
plt.xlabel('Count',fontweight='bold')
plt.ylabel('Community')
plt.xticks(fontweight='bold')
plt.yticks(fontweight='bold')
plt.show()
```

APPENDIX F SERVICE TYPES IN TOP COMMUNITIES

```
import matplotlib.pyplot as plt

# Define the top 20 communities and group data
top_20_communities = Hdata['COMMUNITY'].value_counts().index[:20]
grouped_data_top_20 = Hdata[Hdata['COMMUNITY'].isin(top_20_communities)].groupby(['COMMUNITY', 'SERVICE_TY']).size().unstack()

plt.figure(figsize=(14, 8))
grouped_data_top_20.plot(kind='bar', stacked=True, figsize=(14, 8), colormap='tab20',)

# Set title and labels
plt.title('Stacked Bar Chart of Service Types Distribution in Top 20 Communities',fontweight='bold')
plt.xlabel('Community',fontweight='bold')
plt.ylabel('Count of Service Types',fontweight='bold')

# Customize the legend
legend = plt.legend(title='Service Type', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.setp(legend.get_texts(), color='black', fontweight='bold')  # Change color and make text bold
plt.setp(legend.get_title(), color='black', fontweight='bold')  # Customize legend title as well

# Customize x-ticks and layout
plt.xticks(rotation=90,fontweight='bold')
plt.yticks(fontweight='bold')
plt.tight_layout()
plt.show()
```

APPENDIX G PREPROCESSING OF HOSPIITAL DATA

```
# Import the LabelEncoder for encoding categorical variables.
from sklearn.preprocessing import LabelEncoder

# Select the 'COMMUNITY' and 'POSTAL_COD' columns from Hdata and drop any rows with missing values.
community_data = Hdata[['COMMUNITY', 'POSTAL_COD']].dropna()

# Initialize separate LabelEncoders for 'COMMUNITY' and 'POSTAL_COD'.
community_encoder = LabelEncoder()
postal_code_encoder = LabelEncoder()

# Encode 'COMMUNITY' and 'POSTAL_COD' into numerical representations and add them as new columns.
community_data['COMMUNITY_ENCODED'] = community_encoder.fit_transform(community_data['COMMUNITY'])
community_data['POSTAL_COD_ENCODED'] = postal_code_encoder.fit_transform(community_data['POSTAL_COD'])

# import ace_tools as tools; tools.display_dataframe_to_user(name="Encoded Community and Postal Code Data", dataframe=community_data)

community_data.head(20)
```

```python
# Initialize separate Tokenizers for 'POSTAL_COD' and 'COMMUNITY' columns
tokenizer_postalcode = Tokenizer()
tokenizer_postalcode.fit_on_texts(community_data['POSTAL_COD'])

tokenizer_COMMUNITY = Tokenizer()
tokenizer_COMMUNITY.fit_on_texts(community_data['COMMUNITY'])

# Convert the texts to sequences
sequences_postalcode = tokenizer_postalcode.texts_to_sequences(community_data['POSTAL_COD'])
sequences_community = tokenizer_COMMUNITY.texts_to_sequences(community_data['COMMUNITY'])

# Pad the sequences based on max lengths
padded_sequences_postalcode = pad_sequences(sequences_postalcode, maxlen=max_code, padding='post', truncating='post')
padded_sequences_community = pad_sequences(sequences_community, maxlen=25, padding='post', truncating='post')
```

APPENDIX H PREPARING THE MODEL FOR HOSPITAL DATA

```python
from sklearn.neighbors import KNeighborsClassifier

# Initialize the KNN model
knn_model = KNeighborsClassifier(n_neighbors=5)


# Splitting data into features (X) and target (y)
X_df = df[['POSTAL_COD']]
y_df = df['COMMUNITY']

# Splitting dataset into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X_df, y_df, test_size=0.3, random_state=42)

# Train the KNN model on the postal codes and communities
knn_model.fit(X_train, y_train)

# Predicting the test set
y_pred_knn = knn_model.predict(X_test)

# Evaluating the performance with accuracy
accuracy_knn = accuracy_score(y_test, y_pred_knn)

accuracy_knn
```

APPENDIX I MONITORING OF KNN MODEL

```
import numpy as np
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import roc_curve, roc_auc_score
from sklearn.model_selection import train_test_split
from sklearn.datasets import make_classification
import matplotlib.pyplot as plt

# Generate a random dataset (binary classification)
X, y = make_classification(n_samples=1000, n_features=20, n_classes=2, random_state=42)

# Split into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Initialize and fit the KNN model
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)

# Get the predicted probabilities (for the positive class, which is 1)
y_probs = knn.predict_proba(X_test)[:, 1]

# Compute the ROC curve
fpr, tpr, thresholds = roc_curve(y_test, y_probs)

# Compute the AUC score
auc_score = roc_auc_score(y_test, y_probs)
print(f"AUC Score: {auc_score}")

# Plot the ROC curve
plt.figure()
plt.plot(fpr, tpr, label=f"ROC curve (AUC = {auc_score:.2f})")
plt.plot([0, 1], [0, 1], 'k--')  # Random classifier line
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.legend(loc="lower right")
plt.show()
```