

منى عبدالوهاب

Mona Mohamad Abdulwahab.docx

- 1- Final Thesis (Word Format) Submission
- Thesis_(2)_AI901_Dr_Hager_Saleh_Third Trimester 2025 (Moodle PP)
- Midocean University

Document Details

Submission ID

trn:oid::1:3450557291

Submission Date

Dec 25, 2025, 1:45 AM GMT+4

Download Date

Dec 25, 2025, 11:11 PM GMT+4

File Name

xcod.302240369_401621_bahawludbA_damahoM_anoM_منى_عبدالوهاب_17671

File Size

5.6 MB

45 Pages

11,818 Words

70,572 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.





AI-Based Medical Decision Support Framework for Disease Prediction Using Natural Language Symptom Descriptions

Prepared Name: **Mona Mohamad Abdulwahab**

ID No.: 233000426

Submitted in fulfillment of the requirements for the degree Master of Science in
Artificial Intelligence

At Collage of Informatics Midocean University UAE

Supervisor Name: **Hager Saleh**

Collage of Informatics, Midocean University, Comoros

Email: 233000426@st.midocean.edu.km

Acknowledgements:

I thank God for his guidance which led me to achieve this achievement. Praise be to God first and last. Every step I took and every piece of knowledge I learned comes from his source of all credit. I thank God because His divine blessing allowed me to achieve this success.

I don't know how or where to begin, for words cannot express my gratitude to this remarkable woman. I want to express my sincerest appreciation to **Dr. Hager Saleh** who provided me with complete academic support throughout my entire educational journey and offered me vital assistance during my most difficult times.

Your positive feedback together with your religious support through prayer brings me great value. I am grateful for your affection and backing which you combine with your knowledge of the situation. Your presence in our lives has given us strength because God's grace together with your support has sustained us through every challenge.

All my love and appreciation to my **beautiful family** who gave me love and support. Through his prayers my father succeeded in opening doors which had been locked while my mother provided me with the support to keep going. Thank you, my family.

My children Joory Abdul Malik and Lana brought me strength because they were still young. I want to express my gratitude for your support which you provided during our entire extended study period. Your presence in my life serves as both my guiding light and my inner power which drives me to keep moving forward. Your genuine backing and love enabled me to defeat my exhaustion. You were the reason I achieved my goals and dreams.

I want to show my appreciation to **my husband** because he has given me his love and full support throughout my academic years. Thank you for your trust.

Midocean University will always be in my mind because it let me keep studying, which gave me the opportunity to continue my studies. Thank you to everyone involved, and thank you to all the professors and supervisors. Thank you for your efforts.

My achievement exists because of all people who supported me throughout my journey. I will always be grateful to all of you. The journey goes on because my goals are not yet met.

Abstract:

Artificial intelligence (AI) has become more popular because it now functions in multiple new domains which extend its original operational range. This shows that it can transform things. Healthcare systems have problems understanding unstructured text-based medical data, which slows down diagnosis and makes it hard to make judgements. This study develops an AI-driven medical decision-making system capable of interpreting symptom descriptions articulated in layman's terms and utilising a database of unstructured symptoms alongside their organised disease descriptions to infer potential disorders.

The research aims to address this issue. Employs a comprehensive experimental approach by training various machine learning and deep learning models. These models use XGBoost, random forest algorithms, simple Bayesian algorithms, support vector machines, logistic regression, and simple supplemental Bayesian algorithms. We deployed three advanced machine learning models based on transformers: BERT, BioBERT, and ClinicalBERT. These models helped us process unstructured clinical documents, get contextual representations, and make illness classification more accurate.

The clinical BERT model was the most accurate, with an accuracy of 97.5%. This means that it understands complicated medical concepts better than the other models; therefore, it is the best choice for assessing clinical materials that aren't in an organised format. Because of this, the clinical BERT model is considered the basic model in the proposed framework. This research addresses significant shortcomings in previous studies by utilising unstructured medical data in the form of clinical texts. It looks at how transformer-based deep learning models are different from standard machine learning models. The computerised system that comes out of this can aid both doctors and patients by speeding up diagnosis and making better medical decisions.

Keywords: artificial intelligence, disease prediction, medical NLP, Clinical BERT, machine learning, deep learning, medical decision support.

List of Publications:

I formally submitted my study to Midoceans for journal. The research underwent rigorous examination by the editors and was officially approved for release. This is a significant benchmark in my academic path and an input to the arena of medical artificial intelligence and release will also permit healthcare practitioners to access and employ the outcomes in creating the framework for clinical uses.

My future work will include expanding the database to include more cases, as well as testing other, more advanced models to choose the best one, and my dream is to become a harbinger of disease prediction leading to appropriate treatment.

Table of Contents

- List of Abbreviations..... 5
- List of Figures.....5
- List of Tables..... 6

Chapter 1: General Introduction

- Introduction.....7
- 1.1 Research Question7
- 1.2 Problem Definition8
- 1.3 Thesis Objectives.....8
- 1.4 Thesis Contribution.....9
- 1.5 Thesis Organization.....10

Chapter 2: Overview

- Introduction.....12
- 2.1 Artificial Intelligent (AI).....13
- 2.2 Machin Learning (ML)..... 13
- 2.3 Deep Learning (DL).....14
- 2.4 Natural Language Processing (NLP).....14
- 2.5 Machine learning terms and definition.....15
- 2.6 Transformer Architecture16
- 2.7 Transformer terminology and definition17

Chapter 3: Related Work

- Introduction20
- 3.1 History20
- 3.2 Table 1: Summary of Related Work...24

Chapter 4: Proposed Approach

- Introduction.....28
- 4.1 Data collecting.....29
- 4.2 Pre-processing.....30
- 4.2.1 Data cleaning.....30
- 4.2.2 Feature Extraction.....30
- 4.3 Data splitting.....31

- 4.4 Model Training31
- 4.4.1 Machine Learning model.....31
- 4.4.2 Deep Learning model.....33
- 4.5 Testing.....34
- 4.6 Evaluation.....34

Chapter 5: Experimental Results and Discussion

- Introduction.....36
- 5.1 Results of Traditional Machine learning Models.....36
- 5.2 Confusion Matrix SVM38
- 5.3 Results of Deep learning Models.....38
- 5.4 Confusion Matrix clinical BERT39
- 5.5 Discussion41
- 5.6 Strength and limitations of the proposed system.....42

Chapter 6: Conclusion and Future Work

- 6.1 Conclusion.....43
- 6.2 Limitations.....43
- 6.3 Future Work.....44
- References.....45

List of Abbreviations:

- AI: Artificial Intelligence
- BERT: Bidirectional Encoder Representations from Transformers
- Bio BERT: Biomedical Bidirectional Encoder Representations from Transformers
- Clinical BERT: Clinical Bidirectional Encoder Representations from Transformers
- CNB: Complement Naïve Bayes
- DL: Deep Learning
- LR: Logistic Regression
- ML: Machine Learning.
- NB: Naïve Bayes
- NLP: Natural Language Processing
- SVM: Support Vector Machine
- TF-IDF: Term Frequency–Inverse Document Frequency

List of Figures:

- Fig.1 Artificial Intelligent (AI)12
- Fig.2 Machin Learning (ML).....13
- Fig.3 Deep learning (DL).....14

- Fig.4 Natural language processing (NLP).....15
- Fig.5 Transformer Architecture.....17
- Fig.6 Work flow..... 29
- Fig.6 Confusion Matrix of SVM.....38
- Fig.7 Confusion Matrix of Clinical BERT.....40

List of Tables:

- Table.1Summary of Related work.....24
- Table. 2 Machin learning model result.....35
- Table.3 Deep learning model result.....39
- Table.4 Reasons for the similarity in result DL.....41
- Table.5 Summary of Research and their answer.....42

Chapter 1: Introduction

Introduction:

As the world begins to change old concepts and focus on technological development, artificial intelligence is no longer just a useful tool; it has become a crucial component in reshaping many vital sectors, and in every sector, it enters, it drives radical changes, with the healthcare sector at the forefront. With the increasing reliance on intelligent systems, the need has arisen to develop solutions capable of understanding human language as naturally used by people, because patients actually describe their symptoms in their own individual ways and styles, especially when describing health-related symptoms, which are often unorganized, varied in style, and somewhat ambiguous. Medical platforms still face significant challenges in handling unstructured complaints. Patients write about their ailments in different ways and with varying skill levels. Often these texts use vague language, are repeat.

This study examines the limitations of unstructured data and its specific correlation with potential diseases. Research studies have studied either basic machine learning systems or complex learning systems but they did not perform a complete evaluation between these two platforms. The research establishes a method to evaluate these systems for their fairness while selecting the best system which operates on unorganized data sets. This system will use natural language processing (NLP) along with machine learning models and deep learning that is based on transformers.

1.1 Research Questions:

A main question that is similar to the study's goal guides this investigation. The proposed framework functions to analyse written complaints which enables the identification of the most probable medical condition. The research investigates methods to achieve these objectives through dependable standardized operations of synthetic intelligence. The main questions for the research are:

- Might a natural language processing system be created to aid in forecasting illness?
- Can we use current machine learning methods like SVM, LR, and NB, as well as more advanced transformer-based learning models like BERT, BioBERT, and ClinicalBERT, to predict diseases?

1.2 Problem Definition:

One of the biggest problems in medicine is accurately diagnosing diseases based on the symptoms that patients report. As the number of health statistics and digital patient files grows, it's important to use this data to build a smart system that can look at medical records and figure out what might be wrong.

While contemporary machine learning and deep learning methods have appeared that can process and examine language data, present studies center mostly on organized datasets, restricting the complete capability of these methods. Consequently, the research difficulty resides in the lack of a unified platform based on artificial intelligence and current techniques that permits:

1. It analyzes written symptoms in the patient's own language and understands the context from the meaning.
2. The system correctly identifies which disease will most probably occur.

1.3 Thesis Objectives:

Main Objective of the Study:

To develop an intelligent AI-based medical diagnostic system capable of interpreting patient-reported symptom descriptions, predicting the most likely disease,

Sub-Objectives:

1. Data Collection and Pre-Processing

Collect medical symptom data from reliable sources and prepare it for AI processing.

Perform text pre-processing steps, including:

- Cleaning
- Tokenization
- TF-IDF feature extraction

These steps convert raw symptom descriptions into structured input suitable for ML and DL models.

2. Training and Evaluating Models

Train and evaluate six Machine Learning (ML) models:

- Linear SVM
- Logistic Regression (LR)
- Complement Naïve Bayes (CNB)
- Multinomial Naïve Bayes (MNB)
- Random Forest (RF)
- XGBoost

Investigate the performance of three transformer-based Deep Learning (DL) models:

- BERT
- Bio BERT
- Clinical BERT

Compare the performance of all ML and DL models to identify the most effective approach for predicting diseases from natural-language symptom descriptions.

3. System Evaluation

Assess the ability of the proposed system to function as a practical medical decision-support tool in real clinical environments

1.4 Thesis Contribution:

This study presents a set of contributions aimed, at enhancing the use of artificial intelligence techniques in disease prediction based on textual symptom descriptions. The main contributions of this work can be summarized as follows:

1. Develop an AI-based and natural language processing system capable of analysing symptom descriptions in texts and predicting the most likely diseases from an unstructured database.
2. A comprehensive comparison was conducted among six different models, including traditional machine learning approaches Linear Support Vector Machine (Linear SVM), Logistic Regression (LR), and Naïve Bayes (NB) as well as advanced transformer-based models, namely BERT (Bidirectional Encoder Representations from Transformers), BioBERT (Biomedical BERT), and ClinicalBERT, to identify the most effective model for medical text classification.
3. Pre-process the data before training to obtain high-quality data, including data readiness checks, text encoding using TF-IDF, and hashing. After model training and evaluation, treatment recommendations are provided.
4. Train the transformer models and ensure they understand the clinical medical meaning from the context.
5. Develop a preliminary system to assist clinicians in decision-making and expedite diagnosis.
- 6- Providing an adaptable and facilitating structure for research, enabling subsequent scientists to build upon it. And broadening the scope of its applications to encompass different areas related to medical text analysis.

7- Validating transformer models for medical natural language processing applications. And demonstrating their capability to generate plausible treatment advice based on expected patient outcomes.

1.5 Thesis Organization:

This thesis is organized into six chapters, each guiding through a different stage of the study in a clear and coherent manner.

Chapter 1: Introduction

This chapter explains the primary point of the research. It makes clear how important it is to make an intelligent system that can anticipate diseases based on what patients say in writing. It also talks more about the issue that led to my inquiry, the aims I wish to reach, and the primary questions that motivated this work.

Chapter 2: Overview

This chapter explains the main words used in this study, which are meant to define and explain the algorithms, models, and measurement tools used.

Chapter 3: Related Work

This chapter examines prior research on illness prediction that utilises both structured and unstructured data, the models employed in each study, and the key findings. It puts the studies and their results into a table so that they are easy to find. Finally, it talks about the research's flaws and limitations, lists the study's goals, and explains how those flaws were fixed.

Chapter 4: Proposed Methodology

In this chapter, I discussed in detail how to implement this project. I started by collecting and refining the data, cleaning it, and preparing it for training, as well as the training models used, the results, and the metrics that were relied upon.

Chapter 5: Experimental Results and Discussion

This section is considered one of the most important sections because it presents the key results obtained, discusses the reasons behind these results, the objectives achieved, and the gaps that have been addressed.

Chapter 6: Conclusion and Future Work

This chapter reviews the conclusions reached, the main limitations we faced, and future work.

Chapter 2: Overview:

Introduction:

In this section of the thesis, I will discuss the most important concepts related to artificial intelligence and its branches, the terminology used, the models on which the data was trained, and the evaluation methods employed. This section provides a brief overview of the key points related to this work.

2.2 Artificial Intelligence (AI):

Artificial intelligence is defined as the ability of machines to learn and perform tasks similar to humans in intelligence, decision-making, and problem-solving. Advanced AI technologies have emerged that are capable of analyzing inputs and understanding meaning from context. [1]AI

provides assistance with:

- 1The system works with huge amounts of medical data, some of which is very complicated.
2. It helps doctors and nurses with their diagnostic tests and decisions.
3. These cuts down on medical mistakes and speeds up the delivery of medical care.
4. AI technology makes telehealth services better.

Medical applications of AI technology include:

1. Use the symptoms of patients to make models for predicting diseases.
2. The system looks at medical images to find tumours and broken bones.
3. Uses AI to find new drugs and make treatment plans.
4. AI applications include virtual assistants and medical chatbots.

AI technology turns traditional healthcare systems into smart platforms that use data to give patients faster and more accurate medical care.

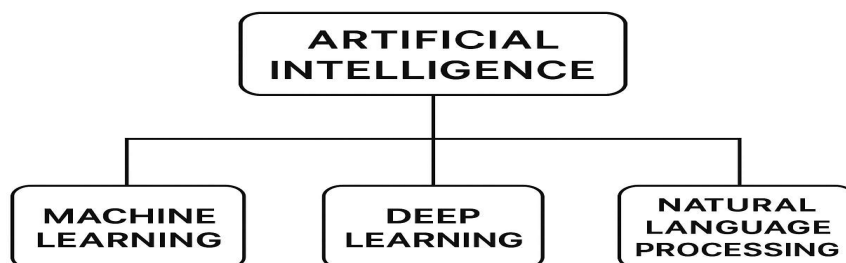


FIG.1 Artificial Intelligence (AI)

2.2 Machine learning (ML):

Machine learning is a part of artificial intelligence that works on making algorithms and statistical models that let computers find patterns in data, guess what will happen, and make decisions without being told what to do. Machine learning configurations naturally improve their skills over time, allowing them to find hidden patterns in data sets and adapt to new, unknown inputs.[2]

Medical applications of ML depend on its capability to:

1. Quickly go through big medical information databases
2. See things that the human eye can't see
3. Help doctors with their diagnostic evaluations;
4. Gives better results while making fewer mistakes in its output.
5. Helps doctors find illnesses in their early stages.

Examples of ML applications in healthcare:

1. They use medical indicators and patient data to guess how a disease will get worse.
2. Looks at X-ray pictures along with MRI and CT scan data.
3. Puts medical papers and clinical records into groups.
4. Finds new medical treatments and drug discoveries.
5. Keeps an eye on the health of patients in real time.

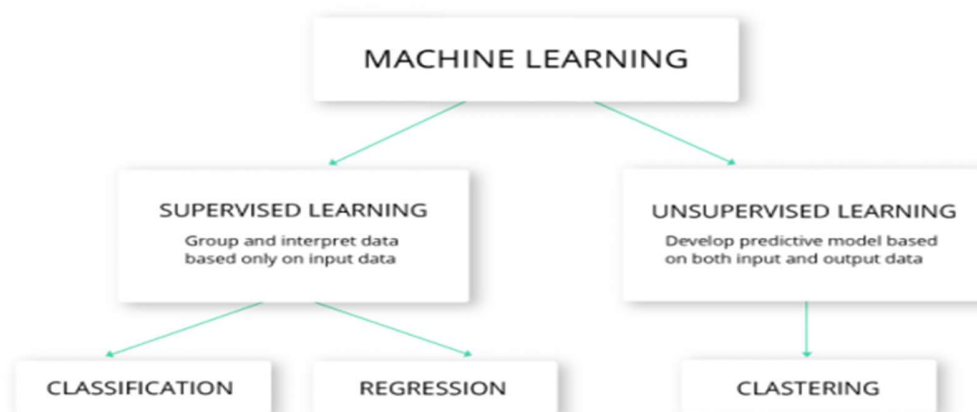


FIG.2 Machina learning (ML):

2.3 Deep Learning (DL):

Deep learning (DL) is the newest segment of machine learning (ML), employing multi-layered neural networks (deep networks) to glean insights from data. These deep networks allow the model to discern intricate patterns. Because of this skill, deep learning can handle complex tasks like recognising images, processing natural language, and analysing medical texts.

When deep learning models are trained on large datasets and have a lot of computing power, they become more accurate. Because of this, deep learning has produced some of the best results in many current AI applications. [3]

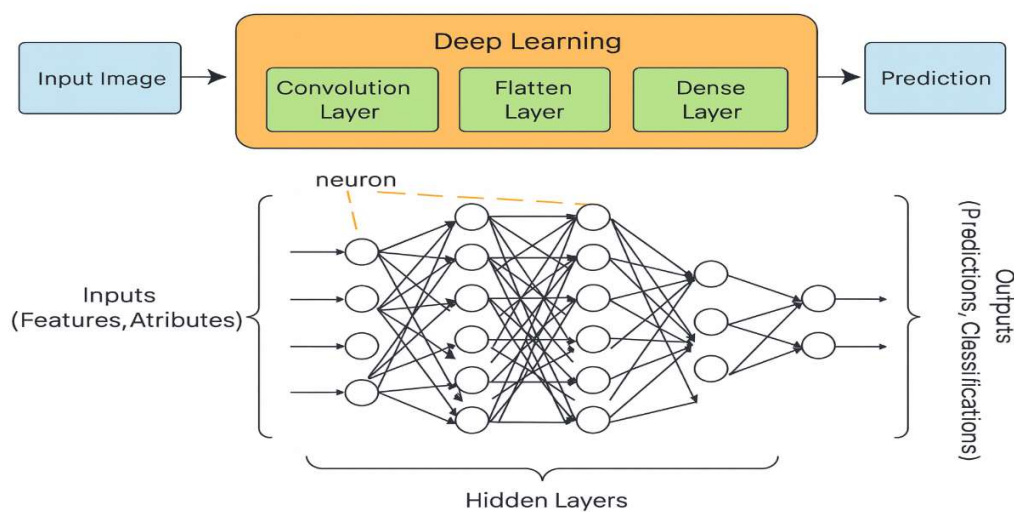


FIG.3 Deep Learning (DL)

2.4 Natural Language Processing (NLP):

Natural Language Processing “NLP” is a field of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. It allows systems to analyze medical text, extract symptom information, and convert unstructured text into meaningful features.[4]

The medical field depends on NLP to analyse data from patient symptom reports and electronic health records and clinical reports and medical research publications. The transformation of unstructured medical text into machine-readable information becomes possible through NLP because most healthcare data exist in natural language format.

Why NLP Is Important in Healthcare:

1. The system identifies medical terms and symptoms which patients write down during their descriptions.
2. The system converts unorganized clinical data into numerical values which it understands.
3. The system enables doctors to make diagnoses automatically while helping them make clinical choices.
4. The system enables fast analysis of extensive medical literature collections.

Examples of NLP Applications:

1. The system identifies essential for extracting symptoms from patient-provided information.
2. The system uses text descriptions to perform disease classification tasks.
3. Medical chatbots and virtual health assistants operate through NLP technology.
4. The system analyses medical reports and patient records for clinical data extraction.

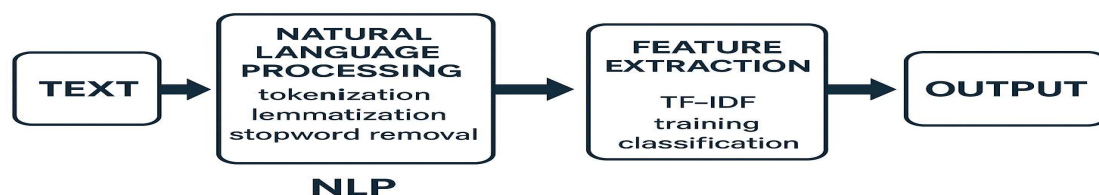


FIG.4 Natural Language Processing (NLP)

2.5 Machine Learning Terms and Definitions

TF-IDF (Term Frequency - Inverse Document Frequency):

The numerical representation TF-IDF in Natural Language Processing (NLP) determines word importance by combining two elements which measure both word importance and their relative value to the complete dataset. It increases with the number of times a word appears in a document (TF) but decreases with how commonly the word appears across all documents (IDF). The system detects vital medical terms which appear in patient symptoms but it reduces the influence of common words found in the descriptions. [6] Tokenization in medical text analysis enables symptom descriptions to become structured units which BERT and its variants Bio BERT and Clinical BERT can analyze. The current transformer models employ subword tokenization as a method to process infrequent medical terminology along with medical abbreviations and clinical language differences.[7]

Dataset: A set of records or examples containing information used to train and evaluate the model.

Features: The characteristics or variables that are input into the model, and which it relies on in the prediction or classification process.

Label / Class: The correct result associated with each record in the data, such as the disease name or target group.

Model: An algorithm or mathematical equation that learns the relationship between inputs (Features) and outputs (Labels).

Supervised Learning: A type of learning in which the model is trained using data that contains known inputs and outputs.

Unsupervised Learning: In this method, the model is trained on unlabelled data with the goal of discovering patterns or clusters.

Semi-supervised Learning: It combines labelled and unlabelled data to improve model performance.

2.6 Transformer Architecture:

Transformers are highly advanced Deep Learning (DL) models and thus a subset of ML designed to handle sequential data using a self-attention mechanism. They process entire sentences at once, allowing them to capture contextual meaning more effectively than traditional RNN-based models.[5]

The Transformer model uses self-attention to analyse word relationships across any sentence position which enables better context understanding than RNN-based models.

The following advantages make Transformers outperform RNNs in their operations:

1. The model handles complete sentence input at once instead of processing words sequentially.
2. The model demonstrates superior ability to detect distant word connections throughout sentences.
3. The model trains more efficiently when working with extensive datasets.
4. The system operates at increased speed when running on GPU hardware.
5. The models based on Transformers achieve the highest possible results in Natural Language Processing applications.

The medical field benefits from Transformer-based models including BERT and Bio BERT and Clinical BERT because they achieve exceptional results in medical text analysis and disease forecasting and clinical decision-making systems.

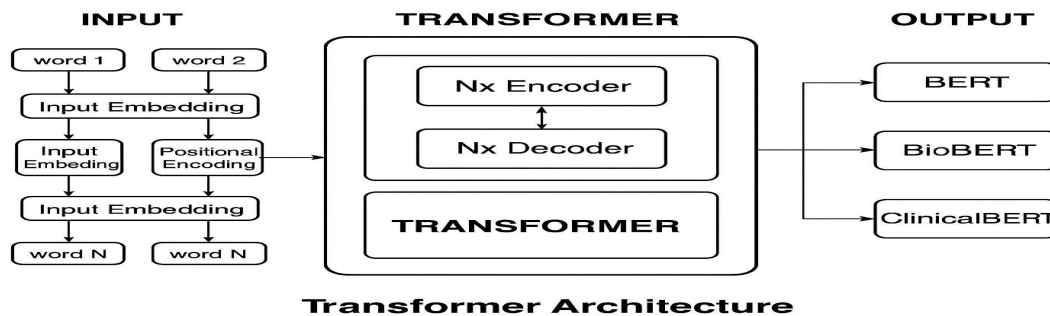


FIG.5 Transformer Architecture

2.7 Transformer terminology and definition:[7]

Encoder: The part responsible for understanding the text and extracting the semantic representation of the sentence.

Decoder: The part responsible for generating outputs (such as translation or new text).

Embedding: Converting words into digital vectors that represent their meaning

Token: A text unit (a word or part of a word) is entered into the form.

Attention Mechanism: A mechanism that allows the model to identify the most important words when processing a given word.

Self-Attention: A mechanism that links each word to all other words in the same sentence to understand the context.

Query (Q): Like the current word for which we are looking for context.

Key (K): It represents the words we compare.

Value (V): Information is extracted based on the degree of attention.

Multi-Head Attention: Using multiple attention heads in parallel to capture different linguistic relationships at the same time.

Head: An independent attention channel that focuses on a particular type of relationship between words.

Positional Encoding: A technique that adds information about word order because the transformer does not automatically understand order.

Masked Attention: A type of attention that prevents the model from seeing future words during prediction.

Dropout: A technique that reduces overlearning by randomly disabling some modules.

Bidirectional: Processing text from both the right and left simultaneously.

Pre-training: The model was trained on large public datasets.

Fine-Tuning: Retraining the model on a specific task (such as classifying diseases).

CLS Token: A special symbol used to represent the entire sentence in classification tasks.

Attention Weights: The values that determine how much the model focuses on each word.

Contextual Embeddings: Representing words according to context, not in a fixed way.

Parallel Processing: Processing all words at the same time speeds up the training.

BERT: A deep learning model based on the Transformer architecture (Encoder only), designed to understand texts in a bidirectional way, i.e., from right and left at the same time.

Bidirectional: This means that the model understands the word based on the words before and after it at the same time, giving a more accurate understanding of the context.

Encoder-only Architecture: BERT uses only the Encoder part of the Transformer, because it focuses on understanding the text and not generating it.

Tokenization (Word Piece): Breaking words into smaller parts is necessary to deal with rare or complex medical terms.

Token: A word or part of a word is entered into the form after the splitting.

[CLS] Token: A special symbol is added at the beginning of the sentence and is used to represent the entire sentence in classification tasks.

[SEP] Token: A symbol used to separate two sentences or to end a sentence.

[PAD] Token: A symbol used to equalize sentence lengths within a batch.

Contextual Embeddings: The representation of a word changes according to the context (the same word may have a different representation in different sentences).

Attention Weights: The values that determine how much the model focuses on each word during comprehension.

Fine-Tuned BERT Model: The BERT model is tailored for a specific task such as medical NLP.

Domain-Specific BERT: A version of BERT trained in a specific field, such as: (BioBERT, ClinicalBERT)

Chapter 3: Related Work

3.1 Introduction:

The world exists in its present condition because everything changes at a fast rate. AI technology allows users to find concealed data patterns which traditional methods were unable to detect. Scientists continue their search for new methods which will improve healthcare decision-making through better prediction results. The implementation of this becomes possible because AI systems and technological infrastructure show continuous development in their complexity.

Before starting any project, it is important to read about other studies on the same topic. The research will enable me to grasp the subject matter better while I detect gaps in previous studies and their weaknesses which will guide my modern investigation through new models that have not been studied before.

3.2 History:

Ferhi et al. (2025) [9] The research used machine learning and ensemble learning methods which need data to be both structured and organized. The research used three diagnostic datasets which are available to the public for testing purposes. The research used three diagnostic datasets which included the Reference Set and two versions of Cough-DDX for testing purposes. The evaluation process included Decision Tree (DT) and Support Vector Machine (SVM) and Logistic Regression (LR) as well as ensemble learning methods which used aggregation and bagging approaches. The research results showed that ensemble learning-based methods outperformed individual classifiers although the study did not provide exact numerical results. The ensemble method which combined DT and SVM produced the highest performance results according to all standard evaluation criteria.

Sogandi et al. (2024) [10] The research applied supervised and unsupervised machine learning methods to a diagnostic system which operated under controlled conditions for disease identification through patient symptom assessment. The Kaggle platform provided us with a dataset which we used to achieve generalizable results through symptom transformation into predefined structured features. The study found out which symptoms had the biggest effect while creating a whole system for diagnosing illnesses. The first step involved using a prior probability-based method to create association rules. The research team conducted tests on multiple prediction models which included gradient descent regression and support vector machines (SVM) and random forests and gradient-boosted trees. The research study determined that gradient descent regression produced the best results for disease classification even though it did not provide any numerical performance data.

Hassan et al. (2024) [11] While searching, I found this study that discusses the use of advanced language models to improve disease prediction from textual descriptions of symptoms. The study compares traditional deep learning methods, such as BiLSTM, with specialized transformer-based models, particularly MCN-BERT. This is a great idea because it uses modern models. Two datasets were used: the first dataset contains 1,200 pairs of symptom and disease texts, while the second includes more than 23,000 descriptions of cases related to adverse drug interactions. The experimental results showed a strong ability of the MCN-BERT model to consistently outperform BiLSTM. The model achieved an accuracy of approximately 99.6% on the first symptom-disease dataset and about 96% on the adverse drug interaction dataset.

Al-Qarni et al. (2025) The research holds value for me because the author performed analysis through textual data and modern analytical methods. However, one of its drawbacks is that the researcher did not specify the database used. The research investigates deep learning methods which analyze unstructured symptom descriptions to predict diseases. A TF-IDF representation was used because deep learning models do not understand meaning from context, in order to measure the weights of meaningful words using deep learning models such as LSTM, CNN-LSTM, and GRU. The results showed that the CNN-LSTM model achieved the highest accuracy of 99.9% on the dataset covering 41 diseases, while the LSTM model achieved an accuracy of 98.8% on the dataset containing 24 diseases. These results demonstrate the great potential of deep learning models.

Serigne'd et al. (2024) [13] The researcher conducted this study through Kaggle structured data analysis with machine learning algorithms which he used to create disease prediction models from patient symptoms. The research employed traditional models alongside an evaluation of four machine learning models which included Decision Tree (DT) and Support Vector Machine (SVM) and Random Forest (RF) and Extreme Gradient Boosting (XGBoost). The experimental results showed that XGBoost produced the best diagnostic results because it performed better than all other models when using symptoms for disease prediction. The researcher did not mention any results.

Finkelstein et al. (2022) [14] This study is new and used advanced models. The researchers launched a research initiative to speed up medical choice processes. The researchers conducted research to develop methods which would help doctors make their medical choices more quickly. The researchers started their research by creating two deep learning models which used Long Short-Term Memory (LSTM) networks and Convolutional Neural Network (CNN) architectures. The researchers analyzed data from 208 cancer patients who received chemotherapy through daily symptom reports which spanned a long period. The results demonstrated that the CNN model performed better than the LSTM model because it reached 81% accuracy and 87% precision and 80% recall and 83% F1 score. The results show that the CNN model is a good choice for real-time symptom monitoring, even though the LSTM model produced somewhat lower but comparable results.

Abhinav et al. (2023) [15] The research establishes a machine learning framework which uses data to predict three diseases that affect humans: breast cancer and heart disease and diabetes. The system included features which made it simple for users to operate and comprehend. The researcher worked with an unshared confidential database which he did not reveal to anyone.

The authors performed a study which compared decision tree (DT) models against support vector machine (SVM) models. The decision tree model always did better than the SVM model. The model was 97.98% accurate for breast cancer, 92.62% for heart disease, and 91.55% for diabetes. Tree-based models show their ability to forecast various diseases which makes them appropriate for healthcare applications in actual clinical practice.

Afrin et al. (2021) [16] The research establishes a supervised machine learning approach which will identify liver diseases at their initial stages. The Indian Liver Patient Dataset (ILPD) was used for testing, which is the main data source for researchers. The LASSO feature selection method allowed researchers to identify essential liver disease symptoms which produced improved model performance and easier model implementation. The models received their evaluation through the use of conventional metrics which included accuracy and precision and recall and F1 score. The decision tree model with LASSO produced the highest results because it

achieved 94.29% accuracy and 92% precision and 99% recall and 96% F1 score. This result shows that tree-based models with good feature selection methods work well for predicting liver diseases.

Chintan M. et al. (2023) [17] The research develops a machine learning system which uses symptom data to forecast cardiovascular diseases for fast and precise automated disease identification and improved medical choice accuracy. The research evaluated four supervised machine learning models which included Decision Tree (DT) and Random Forest (RF) and XGBoost (XGB) and Multilayer Perceptron (MLP). The preprocessing step used k-modes clustering with Huang's principle to boost classification results and GridSearchCV for model parameter optimization. The research analyzed 70,000 patient records from Kaggle which formed its large real-world dataset. The results indicated that the MLP model produced the best results because it reached 87.28% accuracy and 0.95 AUC.

Umar et al. (2022) [18] proposed implementing several techniques based on machine learning algorithms for the rapid detection of heart diseases. The study focuses on a weighted Naive Bayes model, a hybrid methodology that combines Support Vector Machines (SVM) and XGBoost through time-domain features, frequency-domain features, and information theory; an SVM model enhanced through dual optimization; and a comprehensive processing pipeline to facilitate clinical decision-making, which includes using DBSCAN to remove outliers, SMOTE-ENN to address class imbalance, and XGBoost for final disease prediction. The study demonstrated excellent results.

Abdur et al. (2022) [19] The researcher used the Light Gradient Boosting Machine (LGBM) technique to predict outcomes after infection with the novel coronavirus (COVID-19). The research established a fast method which generates exact medical predictions which healthcare professionals can use. The researchers used data from a structured database which contained patient demographic information along with their self-reported symptoms. LGBM was chosen because it can handle large structured datasets quickly and at a low cost.

Lee et al. (2023) [20] In this research, the investigator put out a framework for individualised medical advice with the Graph BERT model. The MIMIC-III dataset, which has real hospital records, was utilised to build accurate models of patients' medical histories and the complicated interactions between medical entities. The results of the experiment showed that Graph BERT did far better than the usual baseline models.

Prabhakar et al. (2025) [21] The researcher applied transformer-based deep learning techniques to extract patient characteristics from unstructured clinical notes during his experimental research. The researchers conducted their study using MIMIC-III dataset which represents a recognized medical database that includes actual hospital patient information for testing medical classification systems in actual healthcare environments. The researchers used the MIMIC-III dataset, which is a well-known medical dataset with real clinical records. The system functions well for conducting medical classification assessments which healthcare professionals use in their actual work environment.

The study used standard performance metrics to test the proposed model and found that it worked better than standard models like vanilla BERT, BioBERT, and basic CNN-based models. The AUROC for the system was 0.80, the F1-score was 0.72, and the Hamming loss was 0.09. The system achieved outstanding results in its medical information organization system because this system allows healthcare providers to find appropriate patient data.

Kumar et al. (2023) [22] The researcher used machine learning techniques to develop a method for diagnosing diabetes and utilized a well-organized digital clinical dataset for diabetes. The researcher tested three classification models which included Decision Tree (DT) and Logistic Regression (LR) and Support Vector Machine (SVM). SVM produced the highest accuracy at 89% among these models because it shows excellent performance in medical prediction tasks when numerical features are properly organized.

Sood et al. 2024. [23] The aim of this research was to develop an intelligent healthcare system that can predict illnesses based on symptoms reported by patients. The researchers employed a structured dataset containing over 5,000 entries of symptom-to-illness associations. This study used a number of machines learning algorithms, including Decision Tree (DT), K-Nearest Neighbours (KNN), Naïve Bayes, and Random Forest (RF). The Random Forest method had the highest accuracy of these, over 90%, making it the best model for the classification job. The system was also added to a web platform that could help with diagnosis right away. The investigators put together the data collection themselves, but it is still stored with the paper's release. The International Journal of Preventive Medicine and Health has all the data and the full investigation.

Table 1: Summary of Related Work:

Study	Year	Dataset	Models	Best Model	Result
1	2025	Referen ce, Cough- DDX- DDX2	DT, SVM, Bagging, Stacking	Bagging +SVM	High accuracy and recall
2	2024	Kaggle "sympto m- disease dataset	Prior + Stepwise Regression, SVM, Bootstrap Forest, Boosted Trees, Neural- Boosted Methods	Stepwise Regression	Achieved the most reliable performance and enabled extraction of useful decision rules
3	2024	Dataset -1: - 1200 sympto m-	BiLSTM, CNN- LSTM, BERT, Bio BERT, MCN- BERT	MCN-BERT A damp optimizer.	Dataset-1: - Accuracy 99.58%, Precision 99.46%, Recall

		disease pairs. Dataset -2: - 23,516 “ADR vs non-ADR”.			99.37%, F1 = 99.13%. Dataset-2: - Accuracy 96.15%
4	2025	Custom-built textual datasets “24 & 41 diseases” – symptom descriptions, not publicly available”	LSTM, CNN-LSTM, GRU with Tokenization & TF-IDF	CNN-LSTM “Tokenizer”	Accuracy: - 99.9% “41 diseases”, LSTM+TF-IDF: - 98.8% “24 diseases”
5	2022		Decision Tree “DT”, Support Vector Machine “SVM”, Random Forest “RF”, XGBoost	XGBoost	Achieved the highest accuracy in symptom–disease classification
6	2024	Symptom diaries from cancer patients on chemot	“CNN” LSTM	CNN “3-day window”	Accuracy: - 81% Precision: - 87% Recall: - 80%

		herapy “self- reported , daily, custom dataset”			F1-score: - 83%
7	2023		Decision Tree, SVM	Decision Tree,	Breast Cancer: - 97.98% heart disease: - 92.62% Diabetes: - 91.55%
8	2021	Indian Liver Patient Dataset “ILPD	DT, RF, SVM, KNN, AdaBoost, GB, LR, LDA	Decision Tree “LASSO”	Accuracy: - 94.29% Precision: - 92% Sensitivity: - 99% F1-score: - 96%
9	2023	Cardiovascular Disease Dataset from Kaggle	DT, RF, XGBoost, MLP, K-Modes + GridSearchCV	Multilayer Perceptron “MLP”	Accuracy: - 87.28%, AUC: - 0.95
10	2022	Clinical heart disease data “not specified	Naïve Bayes, SVM XGBoost, improved SVM, DBSCAN, SMOTE-ENN	XGBoost	Best overall performance in accuracy
11	2022	COVID-19 Symptom Dataset	LGBM	LGBM	Fast and accurate prediction

12	2023	MIMIC-III	Graph Transformer	Graph BERT	Effective medication recommendation
13	2025	MIMIC-III	Clinical BERT, Bio BERT, blue BERT, CNN, Vanilla BERT, and a custom hybrid model	Clinical BERT + Multi-head Attention + CNN	AUROC: - 0.80 F1-score: - 0.72 Hamming Loss: - 0.09
14	2023	PIMA Diabetes	SVM, LR, DT	SVM	89% accuracy
15	2024	dataset >5,000 symptoms-to-disease mapping	Decision Tree, K-Nearest Neighbours', Naïve Bayes, Random Forest	Random Forest	Accuracy 90 % Random Forest

Chapter: 4 Proposed Approach:

Introduction:

This chapter explains the practical part of this project. I will discuss how to design and build an AI-based system for analyzing patient-written textual symptoms, understanding and analysing contexts, and predicting the most likely diseases using an unstructured database. I will explain how this work was done step-by-step, starting with data collection using an unstructured Kaggle database containing clinical textual descriptions of symptoms. This was followed by a thorough preprocessing and feature preparation phase. This step involved cleaning the text by removing repetitions, punctuation, duplicate entries, and unnecessary content to produce high-quality data. The higher the data quality and the better the cleaning, the better the results. For the machine learning models, the TF-IDF method was used to extract relevant features and assign weight to each word, giving higher weight to important words from which symptoms could be extracted from the texts. In contrast, deep learning models rely on understanding meaning from context and do not depend on word repetition but rather understand each word within the context of the sentence. Every model learns how to link symptoms to the diseases that are most likely to be the cause. After the training, the models are evaluated on new data to make sure they genuinely comprehend the principles and aren't merely recalling them. To see how good they are, we use the F1 score and the confusion matrix to see how clear, precise, and easy to remember they are. These tests assess how well each model can predict what might happen with a sickness and how well it can recognise the difference between cases that are similar.

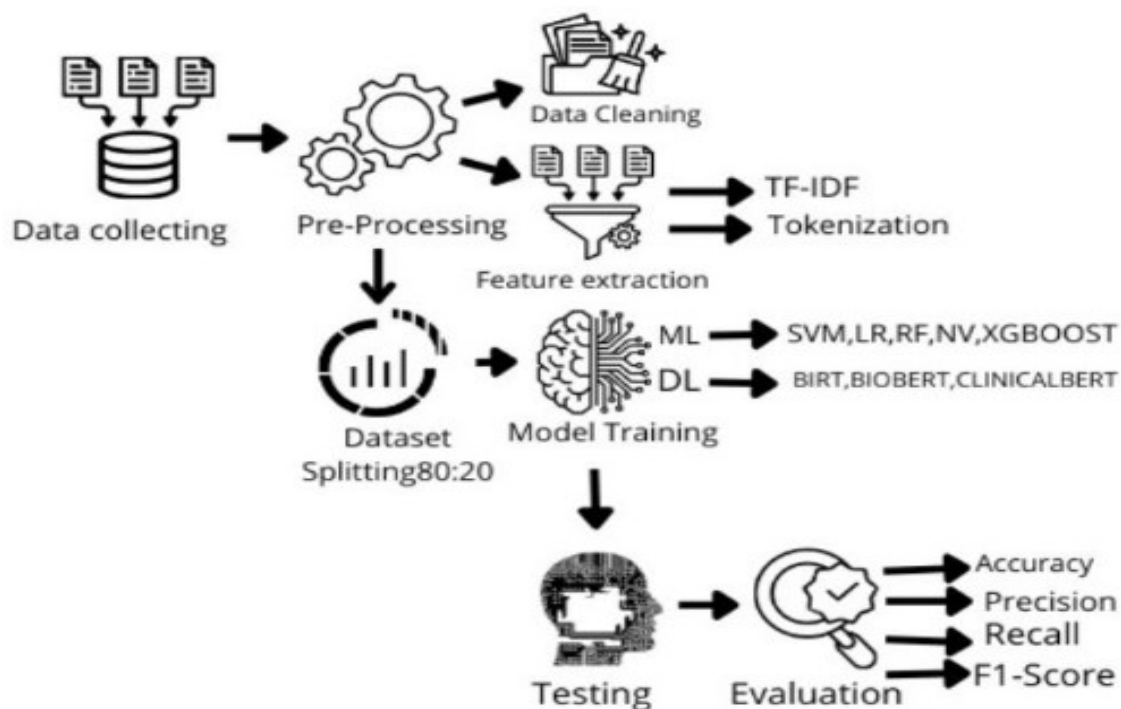


FIG.6 Workflow

4.1 Data collecting:

I got a good dataset for my research from the Kaggle platform. This was the first step. The goal of this method was to collect unstructured medical records that wrote down diseases and their symptoms. The records would then be used to meet the goals of this study.

I received a dataset of unstructured text called Symptom2Disease [24]. There are 24 structured diseases in this dataset, and each one has 50 unstructured symptom descriptions in text. There are 1200 descriptions in total. The data shows that people talk about their worries in many different ways, from short phrases to long sentences. This is very important for improving how systems that understand natural language work.

This makes it a good choice for training deep learning and machine learning systems that use transformers to do multi-class categorisation tasks.

4-2 Pre-Processing:

Clinical medical texts frequently have inconsistencies, formatting mistakes, and redundancies, which can adversely affect the performance of machine learning models and transformers.

Hence, a thorough preprocessing pathway was put into action to convert the disordered

symptom descriptions into a neat and standardized layout appropriate for examination, since data quality influences the outcomes.

4.2.1 Data Cleaning

The writing is tidied to ensure its suitability for training diverse systems. The purer and superior the information, the superior the assessment outcomes and the prevention of overfitting. Originally, the content was normalized by changing all terms to lowercase to stop a single word from being seen as multiple markers due to spelling differences. Next, unneeded symbols and marks were taken out because they did not add to the clinical sense of the writing, and their inclusion might negatively affect the learning steps. The input also went through thorough checking to find missing figures and repeats, and redundant entries were erased to guarantee the system stays impartial during instruction and keeps the dependability of the evaluation outputs. Frequent words with no direct clinical relevance were omitted, which aided in lessening textual clutter and concentrating more accurately on terms linked to indications.

4-2-2 Feature extraction:

Machine learning systems require feature extraction from data to reach their peak operational performance. The conversion of preprocessed text characteristics into numerical arrays which assign importance values to essential terms takes place during this critical stage.

This technique has been used with machine learning models, and TF-IDF feature extraction is a fundamental technique in natural language processing for modifying the appearance of text. The main idea is that it assigns a weight to each word, thus giving weight to important words and decreasing the weight of frequently occurring words, thereby enabling feature extraction from texts. The TF-IDF algorithm serves to transform data for SVM classifiers and logistic regression and random forests and XGBoost and Naïve Bayes models. TF-IDF It not only counts words but also indicates the importance of text within a data set or collection of text. It is not concerned with order but rather with the frequency of text. [41]. The term frequency (TF) shows the number of times a word appears in one sentence thus its value depends on word occurrence frequency.

The IDF scale determines word rarity through its analysis of unique words which appear infrequently in texts thus receiving greater weight in calculations. [42].

It is important to perform these steps before running TF-IDF to convert text to a numerical representation. This step is considered very important; before we can use these functions, we need to perform some basic tasks. The first step involves performing fundamental data processing operations to improve the data and verify that the feature extraction method operates properly. The process required me to eliminate all text elements which included punctuation marks and special symbols and numbers that did not belong to the written content. The prevention of meaning errors requires all texts to maintain identical formatting because this approach prevents the use of identical words which could have different interpretations. The

following step requires the removal of all punctuation symbols and special characters and numbers which do not belong to the text content. All texts need to follow the same format because different formats can create meaning errors. To prevent the same words from appearing in different forms, all letters were converted to lowercase. The document contains its text content through individual units which researchers call tokens. The system processes words one by one through this method. The process requires a complete elimination of all non-meaningful words which includes conjunctions and prepositions from the text. To achieve high accuracy, the focus is on words that have.

4.3 Dataset Splitting:

Any researcher must overcome the obstacle of finding appropriate data for their research. After research, a dataset containing approximately 1,200 records of symptoms related to 24 diseases was adopted, evenly distributed with 50 records per disease. The system maintains equal data distribution through this approach which prevents any form of bias from occurring.

The research distributed its data into two parts which consisted of 80% for training and 20% for testing. The method enables researchers to test model performance through unbiased testing procedures which most studies in this field employ.

The researcher applied Stratified sampling to maintain equal disease and symptom representation across all groups. The system enables all participants who enter the system to obtain equal access to system resources. The instruction set (80%) comprises 960 records that are used to train the systems and establish their variables. The testing subset (20%) has 240 records that are not at all like the instruction set. They are used to fairly test how well the finished system functions with new input.

4.4 Model Training:

This part establishes the foundation for this research, where the efforts were implemented in the Jupyter Notebook. Instruction took place utilizing the Scikit-Learn framework for conventional machine learning routines, incorporating SVM classifiers, logistic regression, random forests, XGBoost, and Naïve Bayes. The PyTorch and Hugging Face Transformers toolkits were employed to build and adjust transformer-based architectures, like BERT, BioBERT, and ClinicalBERT, inside a Python setting. These libraries furnished the required utilities for data preparation, embedding, model setting refinement, and effective instruction on complex textual attributes.

The algorithms were instructed to link high-dimensional attribute vectors either TF-IDF representations created by Scikit-Learn or token embeddings generated by Hugging Face transformers with one of 24 distinct ailment categories. This correlation procedure allows the system to transform initial symptom text into useful digital indicators that predictors can process and convert into precise ailment forecasts.

4.4.1 Machine Learning Models:

Support Vector Machine (SVM):

(SVM) is a supervised machine learning technique broadly employed in categorization jobs. This algorithm depends on locating the best boundary that divides data groups while attaining the biggest feasible gap between distinct data elements. SVM are very adept at managing high-dimensional information, rendering them especially proper for text categorization uses, such as medical writings, where textual aspects are abundant and intricate. [27]

Logistic Regression (LR):

Logistic regression (LR) is a supervised machine learning algorithm chosen as a training model because it is primarily used for classification tasks rather than predicting numerical values. It is used when the goal is to decide between two categories, such as sick/non-sick, hence the term binary classification. It calculates the probability of data belonging to a specific category using a logistic function. Despite its simplicity, this model has proven effective in classifying symptoms and predicting diseases. Its rapid training and ability to work efficiently with TF-IDF representations make it a useful choice for analysing textual data. [28]

Complement Naïve Bayes (CNB):

Naive Bayes is a supervised learning machine learning classification algorithm based on Bayesian probability theory. It assumes a very simple premise: that all features are independent of each other. For this reason, it is called the “Naïve” model, as it makes a simplified assumption that may not be entirely realistic, but it proves effective in practical application. The model looks at past data and learns how often each feature appears with each category (e.g., disease). It is one of the common models for classifying medical texts and data, due to its ease of application and speed of processing. The traditional naive Bayes classifier was developed by CNB to improve its ability to handle unbalanced text classification tasks. [29]

Random Forest (RF):

It is a supervised machine learning algorithm used for classification and prediction. It relies on the idea of combining a large number of decision trees to obtain more accurate and stable results. The decision process generates one result from each tree. The group determines its choice by voting which results in the majority deciding the final answer. It is used because it provides accurate predictions. Random Forest uses multiple decision trees which it combines through their collective decisions to produce the final outcome. The trees in the forest present their own perspectives until the forest selects the most powerful viewpoint. [30]

XG Boost Extreme Gradient Boosting (XG Boost):

XGBoost (Extreme Gradient Boosting) is a very powerful machine learning model based on decision trees. The boosting method is widely used in competitions and medical applications due to its high accuracy and speed. The basic idea is to train independent trees (like a random forest). XGBoost trains the trees sequentially; each new tree attempts to correct the errors of the previous one. The first tree is built, and then a simple decision tree is trained on the data. It gives initial predictions and calculates residuals. The difference between the actual values and the predicted values is calculated.[31]

Multinomial Naive Bayes (MNB):

It is a machine learning classification algorithm based on probability theory (Bayes' theorem). It assumes that features are independent of each other, hence the name Naive. The basic idea of the algorithm is to calculate the probability that data belongs to a specific category based on the available features, and then select the category with the highest probability. Naive Bayes calculates the probability of each category; for example, disease A, disease B, and disease C. The probability of each disease in the data is calculated. The method functions as an organizational system which helps users evaluate their symptoms. The research uses word frequency analysis together with category probability assessment to determine the main concepts. Select the disease category because it offers the highest probability of being correct. MNB has become a popular choice for text classification work because it shows excellent results in Natural Language Processing applications which use term frequencies and TF-IDF as feature representations. [32]

4.4.2 Deep Learning Models:

During the second stage of instruction, we employed transformer-based deep learning architectures. These systems grasp immediate contextual meanings from the text. Three distinct transformer frameworks, initially trained on health data, were adjusted to assist in ailment categorization.

1. BERT (Bidirectional Encoder Representations from Transformers):

Employed as the foundational model in this research, BERT is constructed upon a multi-layer bidirectional Transformer encoder and underwent pre-training using extensive general text resources, such as the Books Corpus and English Wikipedia. Although BERT exhibits robust capabilities across numerous natural language processing assignments. [33]

2. BioBERT(Biomedical BERT) :

BioBERT was utilized in this study. While it shares the same base structure as BERT, BioBERT was additionally pre-trained on vast biomedical textual information, including PubMed abstracts

and full articles from PubMed Central (PMC). How does it work? It reads text from right to left simultaneously. It understands the meaning of words within the medical context. It picks up on relationships between symptoms, diseases, and medications. [34]

3. Clinical BERT The Core Model:

ClinicalBERT is a deep learning model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, but specifically trained on real-world clinical texts such as physician notes and medical reports. How does it work?

It relies on self-attention. It reads text from right to left simultaneously to understand the full context. It understands medical abbreviations, grammatical errors, and the informal style found in clinical notes and unstructured clinical texts filled with abbreviations and written in a non-academic style. ClinicalBERT is able to handle these challenges with high accuracy.

4.5 Testing:

The process of validation stands as a vital requirement because it enables the model to learn from unknown data which it has not seen during training. The system proves its ability to learn and apply knowledge through this demonstration. The evaluation process enables researchers to determine how well models predict outcomes from data that has not been seen before. To do this, 20% of the dataset was kept for testing only. The model shows its ability to detect medical signs and find proper illnesses through its results when testing new data. In the testing phase, each symptom entry is given to the trained model.

The model generates a disease prediction output. These predictions are then compared with the true labels.

4.6 Evaluation:

To determine the efficacy of the suggested framework, a thorough assessment was conducted utilizing five metrics to guarantee precise outcomes. Depending solely on correctness is insufficient; thus, we investigated the ensuing criteria to confirm clinical dependability [36]:

1. Accuracy:

Accuracy denotes the percentage of correctly forecasted events among all observations. While it is a frequently employed assessment measure, accuracy alone might convey a deceptive view of model performance when the data is imbalanced, as high accuracy can still emerge even if infrequent categories are poorly estimated.

2. Precision (Positive Predictive Value):

Precision assesses the trustworthiness of the model when forecasting a specific ailment. Within the scope of the proposed system, a high precision figure signifies that when the AI model identifies a condition, the forecast is highly likely to be sound. This is especially vital in healthcare uses, as it

helps diminish erroneous positive cases and eases needless patient worry arising from wrong disease predictions.

3. Recall (Sensitivity):

This is arguably the paramount measure in medical AI. This is a metric that indicates the model's ability to detect all valid cases of a disease. The higher the value, the fewer cases the system fails to detect (false negatives), which is very important in medical applications to avoid missing diagnoses.

4. F1-Score:

It is the average between precision and recall. The F1 score is used when it is important not to miss genuine patients while simultaneously minimizing misdiagnosis.

5. Confusion Matrix:

The confusion matrix operates on a main diagonal such that values on the line represent the correctness and values outside it represent the error.

Chapter 5: Experimental Results and Discussion:

Introduction:

The research section of this thesis presents the project's worth along with its beneficial and detrimental elements and its most significant findings. This chapter will demonstrate the most important outcomes that the system framework for exploiting unstructured text data to predict diseases has gotten. The database comprised 1200 symptom rows, and each disease had 50 of those rows. We employed six machine learning and transformer-based deep learning algorithms: Supporting Vector Machines (SVMs), logistic regression, simple Bayesian regression, random forest, XGBoost, BERT, BioBERT, and ClinicalBERT.

The research contains experimental work and investigative studies which face potential success or failure in their results. The following section presents essential results and metrics which stem from the code execution for discussion purposes while I will analyse each result independently for complete comparison purposes.

5.1 Results of Traditional Machine Learning Models:

Now I will look at how well machine learning models that used TF-IDF technology did. This technique helps machine learning models by giving important phrases weights, which helps find symptoms. This, in turn, helped the models figure out what the text meant and how to analyse it. The models I used were easy to grasp, simple, and quick to train. The Supporting Linear Vector Machine (SVM), logistic regression, simple Bayes supplementary algorithm, simple polynomial Bayes algorithm, random forests, and the XGBoost algorithm are all examples of these models. When working with text-based symptom data, each algorithm showed its strengths.

The Supporting Linear Vector Machine (SVM) was the best, even though it was easy to understand. It got an overall accuracy of 0.97, which shows that it was quite accurate and knew a lot about the dataset. The TF-IDF technique and the fact that there was a lot of accurate data available helped us get these great results.

Logistic regression consistently got the right answer with an accuracy of 0.95, showing that it worked well in all areas. The Complement Naïve Bayes and Multinomial Naïve Bayes models did better than predicted at basic probability predictions, even though they are simple. The Complement NB got an accuracy of 0.91, and the Multinomial NB had an accuracy of 0.92, which was a little better. Both models were fast and efficient in working with unbalanced datasets. The Random Forest model was able to find nonlinear relationships in the data with an accuracy of 0.95. The XGBoost algorithm was the least accurate of the six, with an accuracy of 0.88. It could find important patterns in the input indications, but it couldn't figure out what the disease was. The results show that all machine learning algorithms can look at text and figure out what disease it is, as long as the dataset is big and smart enough.

Table 2: Machine Learning ML Model Performance Results:

	Model	Accuracy	Precision	Recall	F1-Score
1	Linear SVM	0.97	0.97	0.97	0.97
2	Logistic Regression	0.95	0.96	0.96	0.95
3	Complement NB	0.91	0.93	0.91	0.91
4	Multinomial NB	0.92	0.93	0.92	0.92
5	Random Forest	0.95	0.96	0.95	0.95
6	XGBoost	0.88	0.89	0.88	0.88

5.2 confusion matrix SVM:

The medical nature of this field requires precise error location identification so researchers employed a confusion matrix for their analysis. The tool serves as a basic assessment tool which helps determine how well classification models perform their tasks. The confusion matrix serves as a performance evaluation tool which shows how predicted results match their actual values in the dataset. The method shows us both right and wrong classifications along with the exact spot where each mistake took place. This is done through a graph showing the number of correct cases on the main line and the number of incorrect cases off it. The graph also shows the type of error, not just the number of errors. The medical field requires this tool because it enables clinicians to prevent diagnostic mistakes which occur during patient treatment. The system uses this method to calculate essential evaluation metrics which include accuracy and precision and recall and the F1 score. The confusion matrix was used to test the SVM model because it achieved the best score among all the models. The model shows its ability to identify medical symptoms correctly while it links these symptoms to their corresponding medical conditions.

Most of the predictions were on the main line. These errors mostly occurred in diseases with similar symptoms, such as the common cold and influenza, or migraine and tension headache. Medical datasets from real-world settings show this pattern because different symptoms tend to appear together in patients. The linear SVM model achieved outstanding results in case classification according to the confusion matrix.

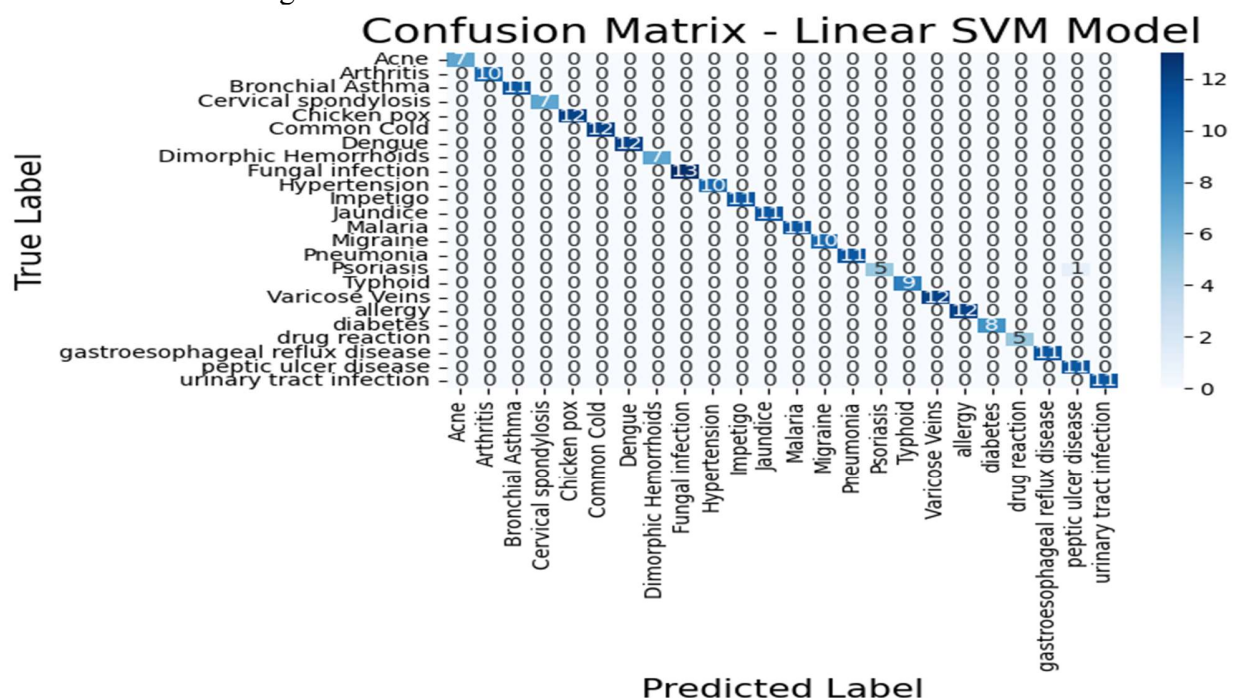


Fig. 7 Confusion Matrix SVM

5.3 Results of Deep Learning and Transformer-Based Models:

Transformers are relatively new and sophisticated models, and there is limited research on them. The research establishes its primary goal to evaluate different models which exist in the field. The model employed transformer models for its operations. These models were much better at understanding medical literature and grouping diseases than regular machine learning methods because they were trained on medical data before they were used. The research used three systems which researchers named BERT and BioBERT and Clinical BERT. They all know what something means based on the context. Medical models serve as an effective tool for students to learn medical terminology because they provide visual support.

The Clinical BERT model achieved the best results because it learned from clinical data which resulted in 97.5% accuracy and 97% F1 score. The BioBERT model achieved an accuracy of 0.96 in its performance. The model received its training data from biomedical databases which included PubMed and PMC but it did not use clinical information which resulted in its reduced performance. The standard BERT model produced results with 0.95 accuracy for all assessment points. The model employed its transformer structure to process health information from text data although its training material included only non-medical content.

The results indicate that all transformer-based deep learning systems outperformed traditional machine learning methods in understanding and analysing medical texts. The systems produced Clinical BERT as their most efficient solution because it used its contextual abilities and medical data experience to achieve reliable disease prediction results..

Table 3: Deep Learning DL Model Performance Results:

Model	Accuracy	Precision	Recall	F1-Score
Clinical BERT	0.975	0.97	0.974	0.97
Bio BERT	0.96	0.96	0.96	0.96
BERT	0.95	0.95	0.95	0.95

5.4 Confusion matrix ClinicalBERT:

The medical field requires exact assessment methods which is why the confusion matrix served as the assessment tool. The system shows all mistakes which it detects in addition to its overall mistake percentage so it functions as an effective system. The clinical BERT model produced the best results among all models tested. This is because most of the correct values were concentrated along the diagonal line, indicating that the model accurately predicted almost all disease categories with very few misclassifications outside the line. The clinical BERT model demonstrates its ability to recognize medical information which enables it to identify between two symptoms that share many characteristics.

This graph shows that the numbers on the main diagonal in dark blue represent correct predictions, while the numbers outside the main diagonal represent incorrect predictions.

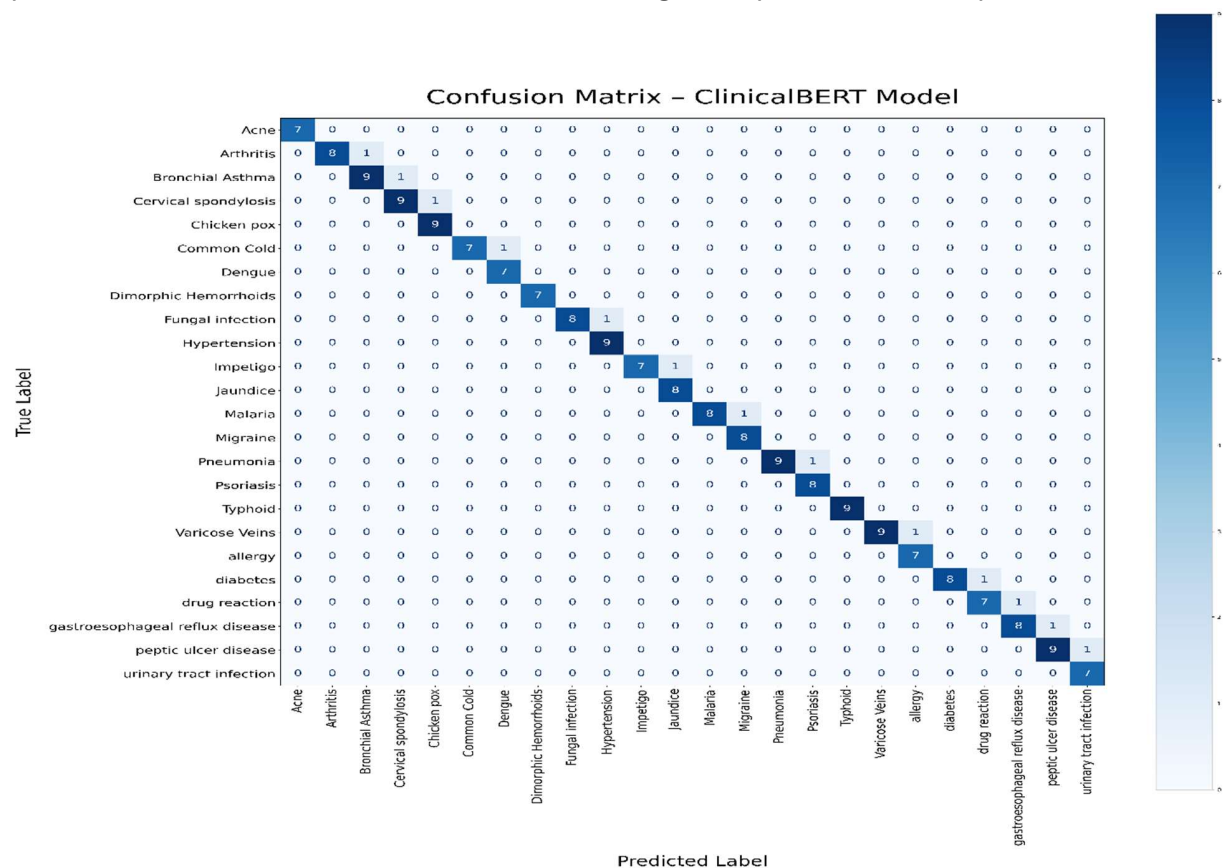


FIG.8 Confusion Matrix Clinical BERT

5.5 Discussion:

The experimental results showed that the system maintained reliable operation through all assessment parameters. Two types of models were used to ensure the best results and to select the optimal model. Multiple cleaning and processing operations needed to be performed on the data until it reached suitable quality standards. The TF-IDF technique was used alongside deep learning models to help understand the meaning. The results were verified through five different methods which included accuracy and recall and precision and the F1-score and the confusion matrix. You have to just look at one number. The deep learning models gained ability to interpret medical signs and symptoms through the implementation of Transformers. The system determines input data meaning through its transformer architecture which depends on context information. The training lasted for 50 epochs, but it was early stopped if the results didn't get better. The model avoided excessive fitting to the training data through this approach.

How well it works of Machine Learning Models:

The machine learning models were able to correctly guess the disease based on the symptoms in the texts.

This is probably because TF-IDF helped the models find symptoms in the texts. The data size which fell between small and large proved to be essential for the research. The data cleaning process enabled us to reach higher standards of data quality throughout the entire process.

How well it works of Deep Learning / Transformer Models:

Transformer-based deep learning models have done very well because they can understand the meaning of things in context and have been trained on clinical medical data before. The amount and quality of the data also played a role in this result.

Why transformer models work the same way:

The transformer models did about the same thing, with ClinicalBERT having a small edge. This table shows the main reasons why the results were so similar.

Table 4: Reasons for the similarity in result DL:

Factor	Explanation
Limited Dataset Size (~12,00 samples)	The dataset was sufficient for training but not large enough to reveal major differences between models. Larger datasets generally highlight stronger variations between architectures.
Well-Structured and Clean Data	Minimal noise and ambiguity improved learning performance, reducing variability between model outputs.

Table 5: Summary of Research Questions and Their Answers

Research Question	Answer & Findings
a. Might a natural language processing system be created to aid in forecasting illness?	The research was able to answer the questions by successfully applying natural language processing techniques using TF-IDF machine learning models, and transformer models that can understand meaning from context.
b. Can existing machine learning algorithms (SVM, LR, NB) and advanced transformer-based learning models (BERT, BioBERT, ClinicalBERT) be used for disease prediction?	The transformer-based models achieved higher accuracy overall, with the clinical BERT model outperforming all others thanks to its prior training in the medical field.

Comparison with Previous Studies and Research Gap Contribution

Many studies were reviewed before this research began, some using structured data and neglecting unstructured data, and others focusing solely on machine learning or deep learning models. These studies consistently left gaps. This research attempts to bridge these gaps by using an unstructured dataset written as text. Machine learning and deep learning models based on transformers were then applied, compared, and the best model selected. This distinguishes my study from previous research. I utilized the open database Kaggle to allow other researchers to build upon my findings and experiment with different models.

5.6 Strengths of the Proposed System:

The suggested framework displays several benefits that allowed it to attain excellent results in ailment categorization using written symptom accounts. The good findings were attributable in part to the great quality of the data, which was cleaned well and of the right size. Also, using machine learning models helps provide weight to the right terms. Using transformer-based models that have already been trained on medical data helped with understanding the context. The system worked well because of all of these things.

Chapter 6:

6.1 Conclusion:

The domain of Natural Language Processing (NLP) examines human writing and speech patterns. Numerous disciplines will see profound transformations when integrated. Using machine learning and transformer-based deep learning models. The healthcare sector is one such domain. Notwithstanding these modifications, the healthcare system remains incapable of comprehending patient communications. This extensive data is not being exploited effectively. Owing to the challenges associated in creating computers that can understand and analyse speech. This thesis sought to provide an alternative methodology and address the deficiencies of prior research that depended exclusively on structured data. The project commenced with a Cagle database with 1,200 entries of documented symptoms in textual format, corresponding to 24 disorders. The symptoms were uniformly distributed throughout the disorders. Machine learning and transformer-based deep learning models were trained. We picked a Kaggle database that was easy to get to because it was based on the ideas of openness and generalisability. We trained machine learning models using linear support vector machines (SVMs), random forests (RFs), basic Bayesian approaches (NBs), logistic regression (LRs), and XGBoost. We also used transformer-based deep learning models such as BERT, BioBERT, and Clinical BERT. All models were evaluated using the same methods, and the best was selected. The Clinical BERT model was the best, with an accuracy rate of 97%.

6.2 Limitations:

I did not expect the system to achieve such impressive success and operate so smoothly; however, this does not negate the existence of some limitations in this study.

- The available data contains restricted numbers of samples which do not represent all possible diseases.

The dataset contains 1,200 records which span 24 diseases but this amount of data is insufficient for transformer models that need extensive training on big datasets.

- Missing key clinical information:

The dataset lacks vital clinical data about patient age and gender and their chronic disease status which would prevent the system from operating in real medical environments.

- Lack of real clinical validation:

The system underwent testing without hospital-based medical supervision while the assessment relied on the data that was available.

- Experimental nature of the study:

The system was implemented for research and experimental purposes only and has not yet been deployed as an operational system in healthcare facilities.

6.3 Future Work:

I am very pleased with this achievement and intend to further develop this work in the future.

- The first step will be expanding the database to enhance the model's generalizability.
- I also plan to work on predicting appropriate treatments to build a comprehensive, symptom-based treatment system.

- Add support for more languages to facilitate wider adoption.
- I will work largest possible number of research in international journals so that the largest possible number of researches can benefit it.

References:

Afrin, S., et al. (2021). Liver disease prediction using LASSO-based feature selection and machine learning. *Bulletin of Electrical Engineering and Informatics*, 10(6), 3369–3376.
<https://doi.org/10.11591/eei.v10i6.3242>

- Ahmed, R., et al.** (2023). Decision Tree and SVM for breast cancer, heart disease, and diabetes prediction. *IEEE Access*. <https://ieeexplore.ieee.org/abstract/document/10167400/>
- Aissaoui Farhi, L., Ben Amar, M., Massoudi, A., & Bouallague, R.** (2025). Improving symptom-based medical diagnosis using ensemble learning approaches. *Systems Research & Behavioral Science*. <https://doi.org/10.1002/sres.3139>
- Al-Qarni, S. S., & Agarín, A.** (2025). Disease prediction from symptom descriptions using LSTM and CNN. *International Journal of Advanced Computer Science and Applications (IJACSA)*. https://thesai.org/Downloads/Volume16No5/Paper_41-Disease_Prediction_from_Symptom_Descriptions.pdf
- Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M.** (2019). Clinical BERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.
- Bioengineering Team.** (2024). Prediction of chemotherapy symptom progression using CNN and LSTM. *Bioengineering*, 11(11), 1172. <https://www.mdpi.com/2306-5354/11/11/1172>
- Bishop, C. M.** (2006). *Pattern recognition and machine learning*. Springer.
- Cortes, C., & Vapnik, V.** (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- COVID-19 symptom-based prediction using LGBM.** (2023). *Springer Proceedings*. https://doi.org/10.1007/978-981-97-7190-5_6
- ddrbcn.** (2023). FDA drug labeling (medical treatment information) [Data set]. *Kaggle*. <https://www.kaggle.com/datasets/ddrbcn/openfda-drug-labeling>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).
- Diabetes prediction using machine learning models (SVM, LR, DT).** (2023). *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-023-16745-4>
- Gerona, A.** (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.)*. O'Reilly Media.
- Goldberg, Y.** (2017). *Neural network methods for natural language processing*. Morgan & Claypool Publishers.
- Goodfellow, I., Bengio, Y., & Courville, A.** (2016). *Deep learning*. MIT Press.
- Han, J., Camber, M., & Pei, J.** (2011). *Data mining: Concepts and techniques (3rd ed.)*. Morgan Kaufmann.

Hassan, E., et al. (2024). Application of transformer-based models for disease prediction using symptom–disease datasets. *Scientific Reports*. <https://doi.org/10.1038/s41598-024-51615-5>

Hosmer, D. W., Lemshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression (3rd ed.)*. Wiley.

Hybrid transformer architecture for patient phenotyping using Clinical BERT and CNN. (2025). *IEEE Journal on Biomedical and Health Informatics*. <https://ieeexplore.ieee.org/abstract/document/9751198>

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing (3rd ed. draft)*. <https://web.stanford.edu/~jurafsky/slp3/>

Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4), 485–525.

Kumar, S., Sharma, R., & Gupta, A. (2023). Prediction of cardiovascular disease using MLP and XGBoost. *Applied Sciences*, 16(2), 88. <https://doi.org/10.3390/a16020088>

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

Lee, J., Kim, H., & Park, S. (2023). Graph BERT: A transformer-based model for medication recommendation using MIMIC-III dataset. *arXiv preprint arXiv:1906.00346*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.

Machine learning approaches for heart disease classification using DBSCAN and SMOTE-ENN. (2022). *Journal of Healthcare Engineering*. <https://doi.org/10.1155/2022/7351061>

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

niyarbarman. (2024). Symptom-to-disease mapping dataset [Data set]. *Kaggle*. <https://www.kaggle.com/datasets/niyarbarman/symptom2disease>

Rennie, J. D., Shih, L., Trevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive Bayes text classifiers. In *Proceedings of ICML-03* (pp. 616–623).

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach (4th ed.)*. Pearson.

Sugandh, F. (2024). Identifying diseases through symptom analysis using supervised and unsupervised machine learning methods. *Scientific Reports*. <https://doi.org/10.1038/s41598-024-69029-8>

Sokolova, M., & Lapalme, G. (2009). Performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.

Symptom-based disease classification using DT, SVM, RF, and XGBoost models. (2024). *ResearchGate*. <https://www.researchgate.net/publication/379650846>

Symptom-to-disease mapping using Random Forest and KNN models. (2024). *IEEE Conference Publication*. <https://ieeexplore.ieee.org/abstract/document/9388603/>

Vaswani, A., Shazeer, N., Parmar, N., Ozokerite, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

U.S. Food and Drug Administration. (2023). *Open FDA Drug Labelling Database*. Retrieved from <https://open.fda.gov/apis/drug/label/>

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019).

BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 1, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019).

Publicly available clinical BERT embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. <https://doi.org/10.18653/v1/W19-1909>