

منى عبدالوهاب

Master Thesis Mona Abdulwahab .docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:338056490

Submission Date

Dec 18, 2025, 1:50 AM GMT+2

Download Date

Dec 18, 2025, 8:54 AM GMT+2

File Name

Master Thesis Mona Abdulwahab .docx

File Size

5.6 MB

55 Pages

12,366 Words

74,753 Characters

50% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

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.

Detection Groups



51 AI-generated only 45%

Likely AI-generated text from a large-language model.



6 AI-generated text that was AI-paraphrased 4%

Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

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 the Almighty **God** first and foremost. His blessings direction and compassion that accompanied me every step of this voyage. In the most challenging moments of my existence. He granted me the fortitude to continue. In periods of tiredness, He filled my core with tolerance and optimism. Without His Favor, this accomplishment would not have been feasible. Attaining a master's degree has been a long-held aim. The route to this instant was not simple; it was replete with hurdles, late evenings, and instances of uncertainty. Yet God furnished me with individuals who mentored me when I felt swamped. Their backing transformed those hardships into manageable phases.

I express my thank and gratitude profound thankfulness and esteem to **Dr. Hagar Saleh**, whose direction, endurance, and steadfast backing were crucial in finalizing this investigation and she motivated me to surmount difficulties I presumed I could not reach her good will and commitment will perpetually remain imprinted on my remembrance.

I am thankful to **Mid-Ocean University** for affording me the chance to attain my scholarly aims. The nurturing setting and facilities and counsel I obtained from the university played a vital part in this success. I extend my heartfelt thanks and regard to all the faculty and advisors whose counsel and prodding proved priceless throughout my academic pursuit.

I am exceedingly appreciative of **my family**, whose affection, petitions, and steadfast aid provided me the power to persist. You stood beside me in every trial, uplifting my spirits whenever I felt worn out or disheartened from the bottom of my heart, thank you.

Thank you for being patient and strong during the long study sessions, my dear **Joory, Abdul Malik, and Lana**. You are the stars that guided me, the light that showed me the way, and the reason I keep going on this journey. Your support and genuine love helped me get through times when I was tired. You were the reason I reached my goals.

Thank you for your love and support, **my dear husband**. Your faith in me helped me get through tough times. Being with you made every moment more meaningful and brighter. This success is not just mine; it belongs to everyone who helped me, believed in me, and walked this path with me. I will always be grateful to all of you.

The journey goes on because my goals are not yet met.

**Abstract:**

Artificial intelligence (AI) is becoming more popular and is being used in more and more places. 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.

In the subsequent stage I will utilize broader. More varied medical records to enhance the model's precision and dependability in genuine situations and employ more advanced frameworks that provide better outcomes and clearer explanations. My forthcoming investigation will centre on solidifying the connection between ailment forecasting, therapy suggestions to render the system more potent in clinical practice.



Table of Contents

• List of Abbreviations.....	vii
• List of Figures.....	viii
• List of Tables.....	viii

Chapter 1: General Introduction

• Introduction.....	1
• 1.1 Research Question	1
• 1.2 Problem Definition	2
• 1.3 Thesis Objectives.....	2
• 1.4 Thesis Contribution.....	3
• 1.5 Thesis Organization.....	4

Chapter 2: Overview

• Introduction.....	6
• 2.1 Artificial Intelligent (AI).....	6
• 2.2 Machin Learning (ML).....	7
• 2.3 Deep Learning (DL).....	9
• 2.4 Natural Language Processing (NLP).....	11
• 2.5 Transformer Architecture.....	12
• 2.6 TF-IDF.....	13
• 2.7 Evaluation Matrices.....	15

Chapter 3: Related Work

• Introduction.....	16
• 3.1 History	16
• 3.2 Table 1: summary of Related work.....	21
• 3.3 Summary of Related work.....	24



Chapter 4: Proposed Approach

- Introduction.....26
- 4.1 Data collecting.....27
- 4.2 Pre-processing.....28
- 4.2.1 Data cleaning.....28
- 4.2.2 Feature Extraction.....28
- 4.3 Data splitting.....29
- 4.4 Training model.....30
- 4.4.1 Machine Learning model.....30
- 4.4.2 Deep Learning model.....32
- 4.5 Testing.....33
- 4.6 Evaluation.....33
- 4.7 Summary.....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.....39
- 5.4 Confusion Matrix clinical BERT40
- 5.5 Discussion42
- 5.6 Strength and limitations of the proposed system.....44
- 5.7 System performance.....45

Chapter 6: Conclusion and Future Work

- 6.1 Conclusion.....46
- 6.2 Limitations.....46
- 6.3 Future Work.....47
- References.....48



List of Abbreviations:

- A. AI: Artificial Intelligence
- B. BERT: Bidirectional Encoder Representations from Transformers
- C. Bio BERT: Biomedical Bidirectional Encoder Representations from Transformers
- D. Clinical BERT: Clinical Bidirectional Encoder Representations from Transformers
- E. CNB: Complement Naïve Bayes
- F. DL: Deep Learning
- G. LR: Logistic Regression
- H. ML: Machine Learning.
- I. NB: Naïve Bayes
- J. NLP: Natural Language Processing
- K. SVM: Support Vector Machine
- L. TF-IDF: Term Frequency–Inverse Document Frequency

List of Figures:

- Fig.1 Artificial Intelligent (AI)7
- Fig.2 Machin Learning (ML).....9
- Fig.3 Deep learning (DL).....10
- Fig.4 Natural language processing (NLP).....12
- Fig.5 Transformer Architecture.....13
- Fig.6 Work flow..... 27
- Fig.6 Confusion Matrix of SVM.....37
- Fig.7 Confusion Matrix of Clinical BERT.....40

List of Tables:

- 1. Table.1Summary of Related work.....21
- 2. Table. 2 Machin learning model result.....36
- 3. Table.3 Deep learning model result.....39
- 4. Table.4 Reasons for the similarity in result DL.....42
- 5. Table.5 Summary of Research and their answer.....42





Chapter 1: General Introduction

Introduction:

The medical industry is weak and needs to grow. Keeping up with and maintaining new technologies. There has been a lot of progress in the use of artificial intelligence (AI) techniques in this field over the past few years, especially when it comes to analysing structured medical data. This has made diagnostic decisions better, improved the speed of detection, and made it easier to get health information faster. As the number and types of health records keep growing, narrative complaints are one of the most important things doctors look at when they first check on a patient's health.

Medical platforms still have a lot of trouble with non-structured narratives, even though this area has come a long way. Patients write about their illnesses in different ways and with different levels of skill. These texts often use vague language, repeat themselves, or use imprecise phrasing, which makes it hard for automated systems to check them. This makes it harder for current systems to accurately find conceptual frameworks that are related to different situations.

This study investigates the constraints of unstructured data and its precise association with potential diseases. A multitude of previous studies have focused on either traditional machine learning systems or advanced learning architectures without a comprehensive comparison. This study aims to conduct a fair comparison of the systems, determine the most appropriate one, and focus specifically on unstructured datasets. This system will use natural language processing (NLP) along with machine learning models and deep learning that is based on transformers.

Research Questions:

A main question that is similar to the study's goal guides this investigation. The suggested framework's goal is to look at written complaints and identify the most likely illness. The study question is how synthetic intelligence can be used to achieve these goals in a consistent and reliable way.

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?



1.1 Problem Definition:

The medical field faces significant challenges, including accurately diagnosing illnesses based on symptoms reported by patients. With the growing number of health statistics and digital patient files, it's important to use this information to make a smart system that can look at medical records and predict 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. It accurately predicts the most likely disease. Solving this problem is a crucial step in keeping pace with development and leveraging artificial intelligence technologies.

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

1. Collect medical symptom data from reliable sources and prepare it for AI processing.
2. 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



3. 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

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

- BERT
- Bio BERT
- Clinical BERT

5. 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

6. 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. Creating an integrated AI-based system able, to interpret open-text symptom narratives. Forecasting the highest probability ailment, drawing from a reliable medical repository.

2. Performing an extensive comparison, among six varied Machine Learning architectures, incorporating standard approaches (Linear SVM, Logistic Regression, Naïve Bayes) and sophisticated transformer-based designs (BERT, Bio BERT, Clinical BERT), to ascertain the premier model for medical text categorization.

3. Presenting a full pipeline for medical text, examination encompassing data readiness and text encoding utilizing TF-IDF and tokenizing. Subsequent to model instruction and evaluation, concluding with therapy suggestions.



4.Emphasizing the confirmation of transformer models, particularly Clinical BERT, in grasping plain language (NL), and attaining superior precision in ailment forecasting.

5.Furnishing a functional prototype that may serve as a decision-support, aid by connecting symptom comprehension with clinical direction.

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

This chapter delves into considerable depth regarding how this project was executed. It clarifies how the information was collected and refined, what frameworks were utilized and how they were educated and validated, how the apparatus functions to predict ailments, and the primary performance indicators employed.

Chapter 5: Experimental Results and Discussion

This part talks about what the different models used in the research found. It looks at the results of each model, compares its pros and cons, checks how well it can predict diseases, and lists the gaps that were filled and the goals that were met.

Chapter 6: Conclusion and Future Work

The last chapter sums up the main points of the dissertation and talks about what the whole study added. It talks about what was done and gives suggestions for future research. It also gives ideas on how the equipment can be improved, made bigger, or used in real medical settings.



Chapter 2: Overview:

2.1 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.

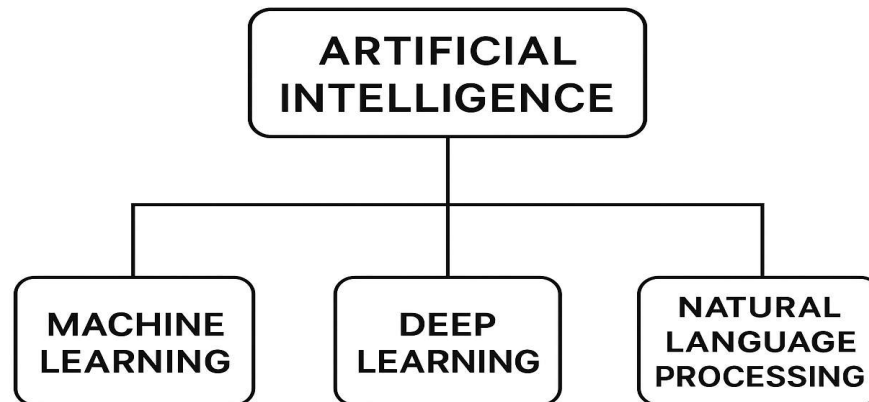


FIG.1 Artificial Intelligence (AI)

2.3 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.

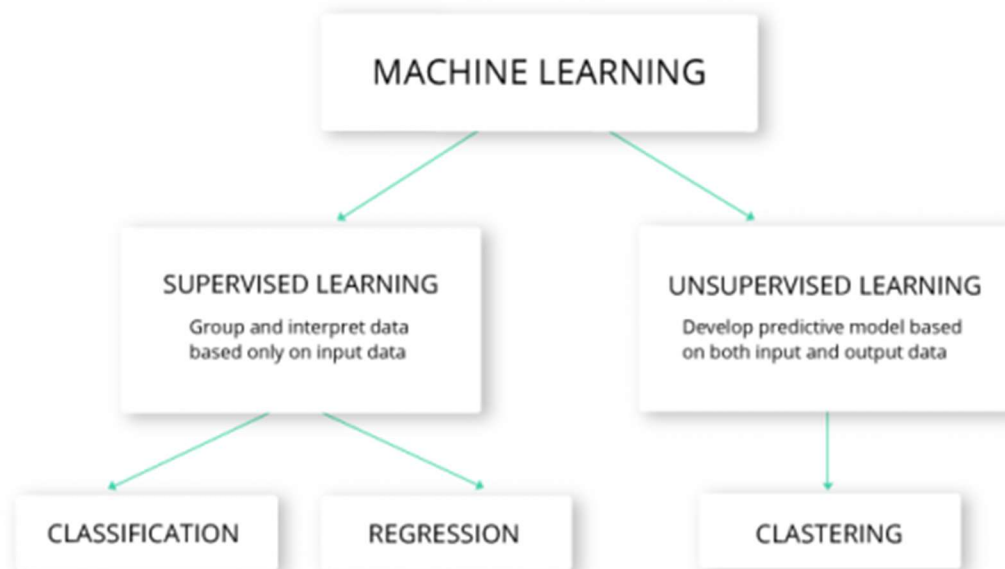


FIG.2 Machina learning (ML):

2.4 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]

Why Deep Learning is Important in Medicine:

1. When it looks at medical images like X-rays, CT scans, and MRIs, it is very accurate.
2. Reads medical literature to find important medical information.



3. Can see complex structures that the human eye can't see.
4. Creates automated medical evaluations and treatment plans.
5. Helps healthcare providers make better decisions and cuts down on medical mistakes.

Deep Learning technology is used by healthcare professionals for a number of important tasks:

1. Finds tumours in pictures taken with radiology.
 2. They use text-based symptoms to find illnesses by putting them into groups.
 3. The system can keep an eye on patients' health with portable devices.
 4. Makes guesses about how well the treatment will work.
 5. Gets medical information from electronic health records (EHRs).
- Deep Learning is the foundation for smart medical systems that let machines understand things like people do, so they can do complex diagnostic tasks quickly.

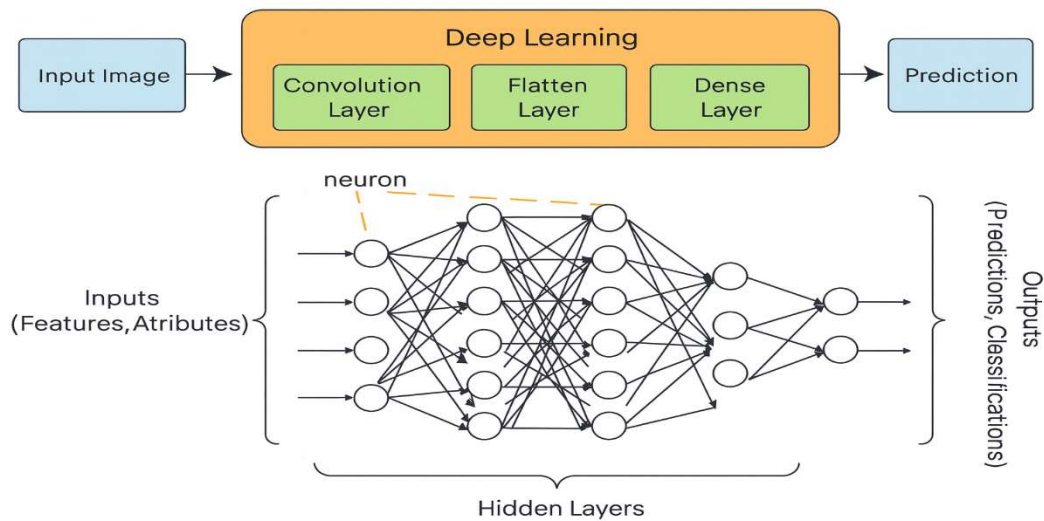


FIG.3 Deep Learning (DL)

2.5. 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.
5. The system enables patients to interact with AI medical assistants through communication channels.

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.
5. The system generates concise versions of medical research documents.

NLP technology enables computers to move past basic keyword searches by understanding human language context and medical significance which makes it essential for developing



intelligent healthcare systems.

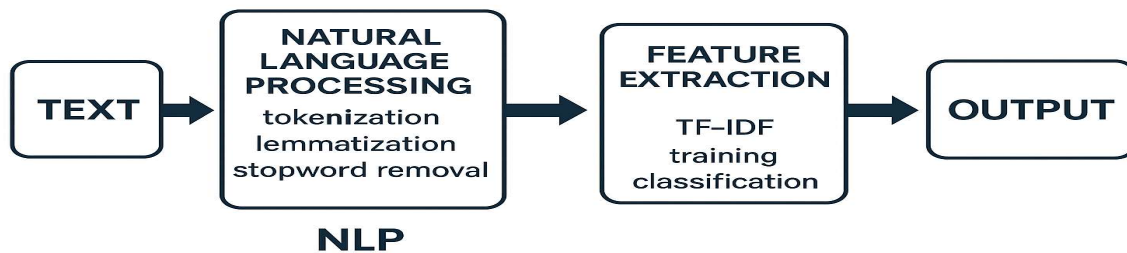


FIG.4 Natural Language Processing (NLP)

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.



- 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.

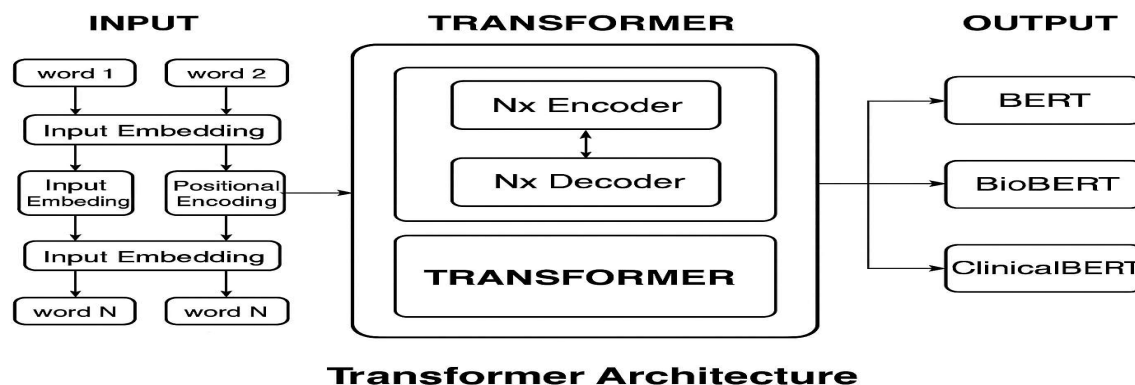


FIG.5 Transformer Architecture

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

TF-IDF is a numerical representation used in Natural Language Processing (NLP) to measure how important a word is within a document relative to an entire 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).

This helps highlight **important medical terms** in symptom descriptions and reduce the weight of common words. [6]

Tokenization The procedure of splitting raw text into smaller significant segments called tokens.

These tokens might be words, subwords, or characters, contingent on the model and the tokenizer employed. Tokenization is a core preparatory phase in Natural Language Processing



(NLP) because machine-learning and deep-learning systems cannot comprehend raw text directly; they necessitate text to be divided and transformed into numerical formats.

In medical text analysis, tokenization helps transform symptom descriptions into structured units that models such as BERT, Bio BERT, and Clinical BERT can process. Modern transformer models use **subword tokenization**, which allows them to handle rare medical terms, abbreviations, and variations in clinical language.[7]



2.8 Evaluation Metrics (Accuracy, Precision, Recall, F1-score, Confusion Matrix):

In any machine learning classification undertaking, the assessment phase is vital to confirm the model can function competently on new data and is reliable enough for use in actual medical environments. Unlike regression problems, classification tasks demand metrics that evaluate not just the total accuracy of forecasts but also how adeptly the model separates between distinct illness groups.

The most frequently employed evaluation figures for classification comprise Accuracy, Precision, Recall, and F1-score Confusion Matrix.

These readings give a full picture of the model results from different points of view.

Accuracy calculates the overall percentage of accurate predictions, whereas.

Precision tells you what percentage of the predicted good cases are actually correct.

Recall checks how well a model can find all the real positive cases. This is very important in medicine, where missing a condition could have serious consequences.

F1-score combines Precision and Recall into one number, giving you a more balanced view, especially when working with datasets that aren't evenly spread out.

Confusion Matrix is a simple tool used to see how well a classification model works. It shows how many correct and incorrect predictions there are by comparing the true class labels with the predicted ones. This helps you understand how well the model works and where it goes wrong.

These appraisal metrics are set by comparing the estimated labels to the baseline truth labels. They are widely used in classification tasks, such as predicting diseases based on symptoms reported by patients.[8]



Chapter 3: Related Work

3.1 Introduction:

Our time is defined by change and speed. AI is now a useful tool for looking at data and finding patterns that were hard to find with older methods. Researchers have been looking into new ways to help people make better decisions, especially in healthcare, by making predictions more accurate. This is possible because AI technology and models are becoming more complex.

Before starting an experiment, it's important to read about past studies on the same topic. They can help you with your work and fill in any gaps. The previous study talks about the methods used, the results that were found, and the limitations that still exist. This background sets the stage for further research and makes sure that scientific knowledge keeps growing by filling in the gaps that already exist. This was the first step in putting the system together.

3.2 History

Ferhi et al. (2025) [9] The investigator conducted multiple experiments utilising machine learning and crowd learning techniques to enhance characteristic-based health assessment tools formatted in tabular structures. The study did not provide any numerical results; it merely selected the best model. The study utilised three publicly available, symptom-oriented diagnostic repositories: the Reference Collection, the Cough Differential Diagnosis (Cough-DDX) collection, and the Cough Differential Diagnosis (Cough-DDX2) collection. These collections focus on breathing symptoms and contain structured information that makes it possible to repeat and confirm the results of the procedures.

Several modelling techniques were employed, including Decision Trees (DT), Support Vector Machines (SVM), and Logistic Regression (LR). Moreover, ensemble strategies based on grouping, such as cumulative and sequential grouping, were examined. The experimental findings indicated that grouping techniques achieved superior efficacy compared to standalone classifiers. Cumulative grouping combined with DT and SVM produced the optimal results across all evaluation metrics, including accuracy, precision, sensitivity, and F1-score.

Sogandi et al. (2024) [10] This research is especially relevant to current studies, as it showcases the merits of machine learning approaches in structured diagnostic environments. Conducted a study that combined supervised and unsupervised machine learning techniques. They used a publicly available dataset from the Kaggle platform, containing symptoms and diseases. The dataset consisted of structured data, where symptoms were represented by predefined



characteristics, facilitating feature review and model training. The research did not offer any numeric findings; it simply chose the optimal model.

The study aimed to identify the most important symptoms and establish a comprehensive set of diagnostic criteria. The prior method was used to generate correlation rules. Several predictive models were then tested, including stepwise regression, support vector machine (SVM), bootstrap forest, and booster trees. The experimental results showed that stepwise regression outperformed the tested models, demonstrating significant effectiveness in accurately classifying diseases.

Hassan et. 2024[11] researchers used advanced linguistic models to improve disease prediction from symptom descriptions. Their study focused on comparing traditional deep learning methods, such as BiLSTM, with specialised transformer models, such as MCN-BERT. The researchers used two datasets: one containing 1,200 text pairs linking symptoms to specific diseases, and the other containing over 23,000 symptoms associated with adverse drug reactions. The results showed that MCN-BERT consistently performed better, especially when trained using the Damp enhancer. The accuracy of the first dataset was approximately 99.6%, while the second dataset achieved approximately 96% accuracy. These results indicate that specialised transformers are capable of detecting subtle patterns in clinical language, which could significantly contribute to automating disease classification and early diagnosis

Al-Qarni et al. (2025) proposed a deep learning-based framework for sickness prediction utilising textual symptom descriptions articulated in colloquial language. The researchers built two unique collections. The first one had symptom descriptions for 24 diseases, while the second one had more diseases, bringing the total to 41. The scientists put these collections together by hand using symptom information from public ally available sources, but they were not made available to the public as archives.

The study included natural language processing (NLP) techniques such as tokenisation and TF-IDF representation, in conjunction with deep learning architectures including Long Short-Term Memory (LSTM), CNN-LSTM, and Gated Recurrent Unit (GRU). The testing results showed that the CNN-LSTM model with tokenisation had the highest accuracy of 99.9% on the 41-disease collection. The LSTM model with TF-IDF had an accuracy of 98.8% on the 24-disease collection. These findings demonstrate the effectiveness of deep learning architectures in modelling textual symptom data, while also highlighting the limitations related to dataset accessibility and reproducibility.



Serigne'd et. 2024. [13] The scientists carried out a study entitled "Symptom-Based Disease Prediction: A Machine Learning Approach" to improve medical diagnosis by using data on symptoms. The research employed a publicly accessible Kaggle dataset of 132 indicators linked to 41 disorders, organised in a systematic manner.

We looked at a few different machine learning frameworks, like Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). The trial findings demonstrated that XGBoost surpassed all other models, attaining superior diagnostic accuracy in predicting diseases from symptom inputs.

Finkelstein et al. 2022[14] conducted research. The primary objective was to facilitate timely medical decision-making by forecasting symptom progression through self-reported patient data. The researchers used two deep learning architectures, Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), using a long-term collection of daily symptom recordings kept by 208 cancer patients during their chemotherapy. These records had information about weariness, illness, pain, and other related problems. The structures were evaluated across different durations "3 to 7 days" to predict the occurrence of symptom exacerbation. The results showed that the CNN design with a 3-day prior view was better than the LSTM, with an accuracy of 81%, a precision of 87%, a recall of 80%, and an F1-score of 83%. The LSTM system got results that were slightly worse but still similar. The study found that CNN is a strong candidate for real-time symptom tracking in digital wellness systems. The data set utilised in this study consists of genuine, clinic-derived, patient-reported symptom assessments and is available upon request.

Abhinav et al. 2023[15] proposed a machine learning-based architecture to predict many chronic diseases, including breast cancer, cardiovascular disease, and diabetes. The device is meant to be simple and easy to use. It has a web interface made using Python, Django, and Stream lit that lets users enter their health data and get predictions about illnesses. The research used Decision Tree (DT) and Support Vector Machine (SVM) methods. The Decision Tree algorithm had the best performance in every case, with 97.98% accuracy for breast cancer, 92.62% for heart disease, and 91.55% for diabetes.

Afrin et. 2021 [16] The study titled "Supervised Machine Learning–Based Liver Disease Prediction Approach with LASSO Feature Selection" aimed to develop an effective technique for the early diagnosis of liver diseases using supervised machine learning techniques. The researchers utilised the Indian Liver Patient Dataset (ILPD) as the primary data source.

We looked at a wide range of classifiers, including Logistic Regression, Decision Tree, Random Forest, AdaBoost, K-Nearest Neighbours (KNN), Linear Discriminant Analysis (LDA), Gradient Boosting, and Support Vector Machine (SVM). The Least Absolute Shrinkage and Selection Operator (LASSO) was used for feature pruning to improve both the system's results and its



clarity. This made it possible to find the most important medical aspects related to liver pathology.

We used 10-fold cross-validation and common evaluation criteria like correctness, exactness, responsiveness, and F1-measure to test the systems. The results showed that the Decision Tree structure combined with LASSO did better than the other methods, with a correctness of 94.29%, an exactness of 92%, a responsiveness of 99%, and an F1-measure of 96%. These findings underscore the effectiveness of integrating tree-based frameworks with robust feature selection methodologies for reliable and timely liver disease prediction.

Chintan M. et al. 2023. [17] The researchers proposed a machine learning framework to predict cardiovascular diseases with considerable accuracy. The goal was to make it easier to quickly and accurately identify heart problems using automated methods to reduce the chances of making mistakes in clinical decisions.

The study utilised various supervised Machine Learning (ML) methods, including Decision Tree (DT), Random Forest (RF), XGBoost (XGB), and Multilayer Perceptron (MLP). We used a k-modes clustering approach with Huang initialisation during preparation to improve categorisation performance, and we used GridSearchCV to fine-tune the parameters.

The models were trained and tested using a real dataset with 70,000 patient records from Kaggle. The Multilayer Perceptron (MLP) model achieved superior results, with an accuracy of 87.28% and an AUC of 0.95, outperforming the other models evaluated.

Umar et al. (2022) [18] proposed various machine learning-based techniques for the rapid detection of cardiac abnormalities. The study presented four primary structures. The first one used a weighted Naïve Bayes system, and the second one used a combination method that brought together Support Vector Machine (SVM) and XGBoost, using time-domain, frequency-domain, and information-theoretic features. The third structure showed a better SVM model that was fine-tuned using duality refinement. The fourth structure combined a clinical decision-support workflow that used DBSCAN to get rid of anomalies, SMOTE-ENN to handle class disparity, and XGBoost to predict diseases in the end.

We used standard criteria including accuracy, precision, recall, and F1-score to see how well the system worked. XGBoost consistently achieved superior accuracy and robustness across all test circumstances, surpassing the other offered methodologies. The researchers stressed how machine learning could help doctors get accurate and quick diagnoses of heart disease. However, the study did not clearly explain how the data was collected or make a publicly available collection available, which suggests that private or site-specific patient information may have been used.

Abdur et al. (2022) [19] performed a study to forecast COVID-19 infection outcomes utilising the Light Gradient Boosting Machine (LGBM) methodology. The main goal was to create a quick, simple, and accurate predictive framework that could be used in medical advising apps.



To accomplish this, the researchers employed a structured dataset that included patient-reported indicators such as fever, coughing, and exhaustion, together with diagnostic evaluation outcomes and demographic factors. LGBM was our choice since it can process a lot of structured data with very little computing power. The testing showed that the LGBM framework was quite accurate and didn't take long to learn. This makes it great for situations when you need to quickly test for COVID-19 and sort patients. The dataset used in this study is called the COVID-19 Symptom and Patient Information. It gave us a complete picture of how well the framework works in real-life medical settings.

Lee et al. (2023) [20] utilised the MIMIC-III dataset to develop a personalised medicine recommendation framework leveraging Graph BERT, a graph-enhanced transformer architecture. The study sought to provide individualised treatment suggestions by accurately modelling patients' clinical histories and elucidating intricate interactions among medical entities.

The experimental results showed that Graph BERT did far better than traditional baseline models. This was mostly because it could show how complex relationships were in clinical data. This capacity helped provide pharmaceutical suggestions that were more accurate and useful in a clinical setting. The use of the MIMIC-III dataset, which is based on real hospital records, made the evaluation more reliable and made the suggested approach more useful in the real world.

This study is particularly useful as it underscores the significance of integrating relational and contextual information in the design of medical recommendation systems. Nonetheless, its emphasis on treatment prescription over symptom-based disease prediction indicates a potential avenue for further study that amalgamates both diagnostic and therapeutic elements within a cohesive clinical decision-support framework.

Prabhakar et al. (2025) [21] performed a study focused on enhancing the detection of patient characteristics from unstructured clinical notes through the application of advanced deep learning methodologies. The researchers used the MIMIC-III dataset because it has detailed and realistic clinical narratives that are good for testing phenotype extraction in real-life medical settings.

We tested the suggested system using standard performance criteria and found that it worked better than traditional models like vanilla BERT, BioBERT, and basic CNN-based architectures. The model got an AUROC of 0.80, an F1-score of 0.72, and a Hamming loss of 0.09, which shows that it worked well for multi-label classification tasks that are often needed in clinical phenotyping.

Kumar et al. (2023) [22] conducted a study to develop a prediction framework for diabetes diagnosis using established machine learning techniques. The researchers employed the PIMA



Indians Diabetes Dataset, a widely utilised and publicly available compilation of structured numerical clinical data.

The purpose was to see how effectively several classifications models, like Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT), worked for separating diabetes and non-diabetic cases. Among the models examined, SVM had the best accuracy at 89%. This indicates its efficacy for medical prediction tasks utilising structured numerical features.

This study shows that traditional machine learning models can give reliable results when used on well-organised clinical datasets. But it can't be utilised in real life because it uses numbers instead of words to express symptoms. In real life, patients usually explain their symptoms in free text. The results are still important since they show us how well conventional machine learning approaches work in controlled diagnostic conditions. They may also be used as a standard for more advanced models.

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	Reference, Cough-DDX-DDX2	DT, SVM, Bagging, Stacking	Bagging +SVM	High accuracy and recall



2	2024	Kaggle “symptom-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 symptom-disease pairs. Dataset -2: - 23,516 “ADR vs non-ADR”.	BiLSTM, CNN-LSTM, BERT, BioBERT, MCN-BERT	MCN-BERT A damp optimizer.	Dataset-1: - Accuracy 99.58%, Precision 99.46%, Recall 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 herapy "self- reported , daily, custom dataset"	"CNN" LSTM	CNN "3-day window"	Accuracy: - 81% Precision: - 87% Recall: - 80% 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 sympto ms-to- disease mapping	Decision Tree, K- Nearest Neighbours', Naïve Bayes, Random Forest	Random Forest	Accuracy 90 % Random Forest
----	------	---	---	------------------	---

3.4 Summary:

The examined studies display the methods employed by the investigators and their foremost discoveries. Investigators evaluated a wide selection of tactics, ranging from basic machine learning predictors to sophisticated grouping techniques and transformer architectures. Numerous articles indicated sound results. For instance, integrating a Support Vector Machine (SVM) with particular feature-scanning strategies enhanced the prognosis of pulmonary ailments, while sequential regression pinpointed the most pertinent indicators. Other investigation employed CNN-LSTM frameworks to categorize illnesses based on text and XGBoost to detect long-term conditions. Furthermore, several works depended on clinical records like MIMIC-III to reveal patient behavior, track symptom development, and suggest treatments using transformer mechanisms. The path of Artificial Intelligence scrutiny in medical text evaluation is clear in these outcomes. Most earlier reports utilized established machine learning forecasters, such as SVM, proportional estimation, and Naïve Bayes, while others explored grouping processes like bagging and XGBoost, paired with merged deep learning setups combining CNN and LSTM. Current exploration indicates a growing inclination towards transformer-based designs, specifically BERT, BioBERT, and ClinicalBERT, due to their superior capacity to grasp medical vocabulary, meaning, and contextual links throughout distributed patient sign accounts.

Despite the successes of these inquiries and the encouraging evidence they produced, several constraints persist. A number of studies still depend on organized data sets, whereas actual clinical information, notably patient-provided metrics, frequently appears as unrestricted, unstructured prose containing ambiguous phrases and uncommon medical terminology. Only a small fraction of research concentrates on this intricate type of data. Moreover, although transformer systems surpass older machine learning approaches in managing unstructured text, further deficiencies involve the constrained access to extensive medical repositories and an absence of authentic-world validation, rendering the replication of research findings or their deployment in healthcare settings more challenging. Owing to these deficiencies, there is a vital



requirement for a system able to interpret open-format sign narratives and anticipating diseases utilizing both machine learning and transformer frameworks. Closing this research gap would render diagnostic assistance tools more advantageous, precise, and quicker in reaching determinations and sound assessments.

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 analysing 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. Each model learns patterns that link symptoms to the most likely diseases. After training, the models are tested on previously unseen data to ensure the system has learnt and not simply memorised. Their performance is evaluated using accuracy, conciseness, recall, the F1 score, and the confusion matrix. These metrics measure the consistency and accuracy of each model in predicting disease outcomes and its effectiveness in distinguishing between similar cases.

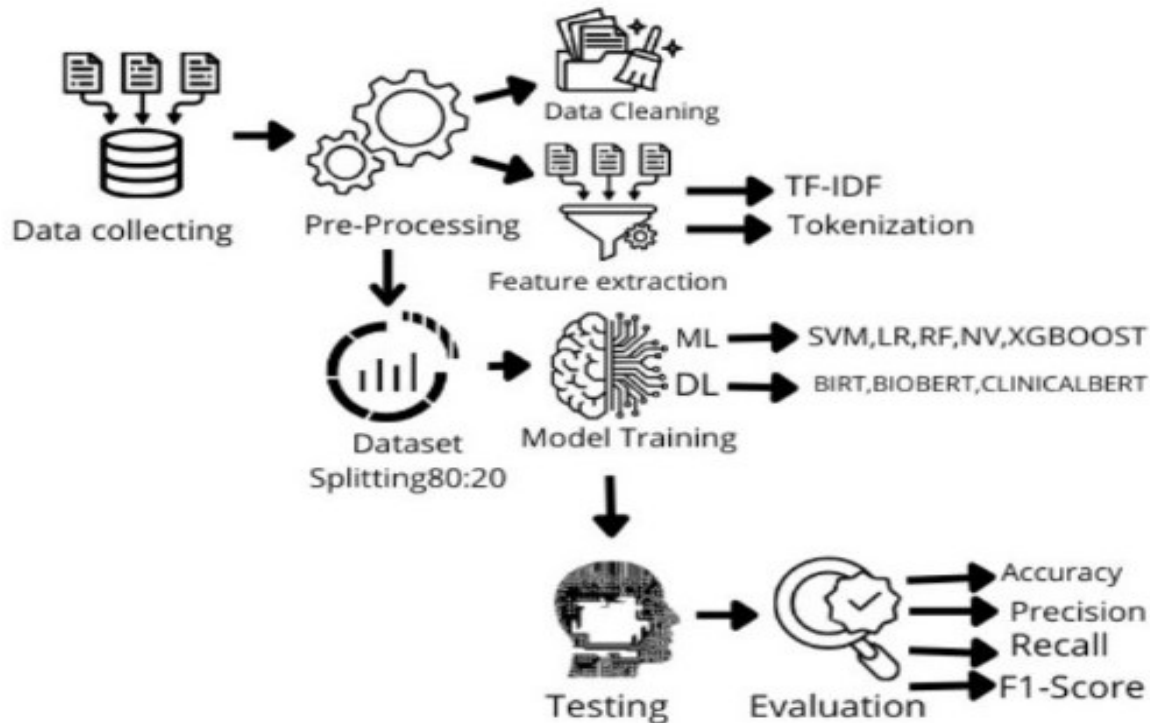


FIG.6 Workflow

4.1 Data collecting:

The first step was to find a dataset suitable for my research, and I obtained one from the Kaggle platform. The goal of this process was to collect unstructured medical records describing diseases and their symptoms in a textual format, which would then be used to implement the objectives of this study.

I obtained an unstructured text dataset called **Symptom2Disease** [24]. This dataset includes 24 structured diseases, each with 50 unstructured symptom descriptions written in textual format, totalling 1200 descriptions. The data shows that people talk about their problems in different ways, from short phrases to long sentences. This is crucial for enhancing natural language understanding systems.

This renders it appropriate for training machine learning and transformer-rooted deep learning mechanisms for multi-class categorization assignments.



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 to perform best. This is a vital phase in changing preprocessed text characteristics into numerical arrays that allocate significance to key terms.

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. Term-Frequency Inverse Document Frequency (TF-IDF) is used to prepare data for SVM classifiers, logistic regression, random forests, XGBoost, and Naïve Bayes. 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]. Term frequency (TF): This is a common term that means the number of times the same word appears in the same sentence, so it depends on the frequency.



Frequency (IDF): This scale measures word rarity by focusing on distinctive words that have not appeared much and their distribution within texts, giving them higher weights. [42].

4.3 Dataset Splitting:

The dataset contained approximately 1200 symptom entries related to 24 diseases, evenly distributed across 50 symptoms per disease. This ensures a balanced database.

To guarantee an accurate and objective assessment of the models' capabilities, the data was divided using a method commonly employed in research: an 80/20 split. This technique is widely used in machine learning and deep learning applications such as BERT, BioBERT, and ClinicalBERT. The data is divided into two separate sections: one for model training (80%) and the other for testing (20%).

To further enhance the integrity and reliability of the assessment, stratification was implemented. This method ensures that each disease category is equally represented in both sets of data. This division prevents overlearning and makes the assessment more reliable.

Consequently, the teaching set (80%) comprises 960 entries utilized to educate the systems and calibrate their settings, whereas the validation set (20%) includes 240 wholly distinct entries employed to objectively gauge the final system's effectiveness on novel 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:

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. This is why it's called "naive," as this premise is often unrealistic, but it works well in practice. The model looks at past data and learns how often each feature appears with each category (e.g., disease). It's widely used for classifying texts and medical data due to its simplicity and speed. 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. Each tree yields a decision.



The final decision is based on a majority vote. It is used because it provides accurate predictions. It is a crowd learning algorithm that builds many decision trees during the training phase. This algorithm combines its predictions to arrive at a more reliable final result. [30]

XG Boost Extreme Gradient Boosting:

XGBoost is an advanced machine learning model and algorithm belonging to the Ensemble Learning category. It relies on the concept of gradient boosting to build robust and accurate models. Instead of training a single model, XGBoost: creates a series of decision trees, each focusing on correcting errors from the previous tree. All the trees are then combined to obtain a final, highly accurate model, known for its success in prediction tasks. It progressively reduces errors, handles structured data efficiently, is very fast due to software optimizations, incorporates built-in techniques to prevent overfitting, and can effectively handle large datasets.[31]

MNB Multinomial Naive Bayes:

MNB is a type of Naive Bayes algorithm. It's a probabilistic model that assumes features are independent of each other. Based on probability theory, it relies on the assumption of feature independence. This model is particularly suitable when dealing with countable data, such as the frequency of words in text or word repetitions within documents. For this reason, it's frequently used in text classification and symptom analysis. The basic idea is to consider how often each word appears and the probability that it belongs to a specific category, then select the category with the highest probability, which is the disease. For this reason, MNB has been widely adopted in text classification tasks and has demonstrated strong effectiveness in Natural Language Processing applications, especially when working with feature representations like term frequencies and TF-IDF. [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. Bio BERT 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:

This step is very important: after completing the training phase, the system must be tested on data it has never seen before to ensure that the system has understood and been trained properly. This step checks how well the models work on data they have never seen before. To do this, 20% of the dataset was kept for testing only. Using new data helps show if the model can really understand symptoms and predict the correct disease. In the testing phase, each symptom entry is given to the trained model.

The model predicts a disease label. 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. It gauges the model's capability to detect all the positive instances. Low recall implies that the system is "overlooking" ill patients ("False Negatives"), which could turn out to be dangerous in a clinical environment.

4. F1-Score:

The harmonic mean of Precision and Recall. This metric is required by the study because it penalizes model operation when either precision or recall figures drop abruptly, which assists in judging system stability across diverse illness types.

5. Confusion Matrix:

It is a fundamental evaluation tool used to measure the performance of classification models in machine learning and transformer-based deep learning. Its core concept is a table that illustrates the model's predictions compared to the actual values in the data. This allows us to determine how many cases were correctly classified, how many were incorrect, and the type of error used to identify the type of error, not just the number of errors. It helps reduce misdiagnosis and is used to extract important evaluation metrics, particularly in medical fields.



4.8 Summary:

This chapter discusses the methodology and steps I took to implement the code. The methodology developed a system for predicting diseases based on textual symptoms using structured workflows. A comprehensive comparison was made between traditional machine learning models and transformer-based deep learning models. I began by processing a medical dataset containing 1200 textual symptom rows and 24 diseases. A thorough preprocessing process was applied to ensure the highest data quality, including cleaning, normalisation, text encryption, and TF-IDF feature extraction for the machine learning models. Transformer-based deep learning models did not require external assistance; they are sophisticated models capable of understanding the meaning of text and words from context. The machine learning phase involved training six different models: Supporting Vector Machine (SVM), Logistic Regression, Random Forest (RF), XGBoost, and a simple Bayesian model. Transformer-based deep learning models were trained using BERT, BioBERT, and Clinical BERT models, pre-trained on medical data, which facilitated their ability to capture symptoms and predict diseases. Following training, a comprehensive evaluation was conducted using accuracy, adjustment, recall, F1 score, and confusion matrix to measure the system's ability to identify diseases efficiently.



Chapter 5: Experimental Results and Discussion:

Introduction:

This chapter will discuss the key results achieved by the system framework for disease prediction using unstructured, text-based data. A database containing 1200 symptom rows associated with 24 diseases was used, with each disease being associated with 50 symptom rows. Six machine learning and transformer-based deep learning algorithms were employed: Supporting Vector Machines (SVMs), logistic regression, simple Bayesian regression, random forest, XGBoost, BERT, BioBERT, and ClinicalBERT.

The aim of this study is to conduct a comprehensive and fair comparison highlighting the strengths and weaknesses of each model.

Each model was evaluated individually to assess its performance and determine the best model. The results indicate that transformer-based deep learning models are superior in disease prediction and natural language processing.

5.1 Results of Traditional Machine Learning Models:

The most important results obtained from the practical application of the code: First, I will present the results related to machine learning. Six models were evaluated, all of which were robust classification models, using TF-IDF. The technique is a method that assists machine learning models, assigns weights to important words, and enables the models to extract symptoms. These models included the Supporting Linear Vector Machine (SVM), logistic regression, the complementary simple Bayesian algorithm, the polynomial simple Bayesian algorithm, random forest, and the XGBoost algorithm. Each algorithm demonstrated unique advantages in handling textual symptom data.

The SVM model achieved the best overall results with an accuracy of 0.97, demonstrating high accuracy and discrimination of the dataset. The application of TF-IDF, along with the quality and size of the data, led to these excellent results.

Logistic regression also made predictions that were consistent and reliable, with an accuracy rate of 0.95. It performed impressively across all metrics. The Complement Naïve Bayes and Multinomial Naïve Bayes models performed better than expected for simple probability predictors. While Complement NB achieved an accuracy of 0.91, Multinomial NB achieved a slightly higher accuracy of 0.92. Both frameworks were fast and efficient and handled the unbalanced dataset appropriately.



The Random Forest algorithm achieved a good accuracy of 0.95, demonstrating its ability to extract nonlinear relationships within the data.

The XGBoost algorithm achieved the lowest score among the six algorithms with an accuracy of 0.88, although it was able to identify significant structures in the input indicators.

The results suggest that all known machine learning algorithms have the potential to understand important relationships between symptoms and diseases.

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:

We used the confusion matrix, which is a basic tool for measuring how well classification models work in machine learning and transformer-based deep learning models. The main idea is to compare the predictions made by the model with the actual values in the data. This lets you see how many cases were correctly classified and how many were not. This is done with a graph that shows the number of correct cases on the same line and the number of incorrect cases not on the main line. It also shows the type of error, not just the number of errors. This tool is very important in medicine because it helps doctors avoid making mistakes when diagnosing patients. It is also used to get important evaluation metrics like accuracy, precision, recall, and F1 score. The confusion matrix was used to test the SVM because it got the best score of all the models. This shows that the model can correctly tell the difference between symptoms and link them to the right diseases.

Most of the predictions were on the main line, but a few were not. These mistakes mostly happened in diseases that have similar symptoms, like the common cold and the flu or a migraine and a tension headache. This is normal in real-world medical datasets where symptoms can overlap. The linear SVM model did a great job of classifying things, as shown by the confusion matrix.

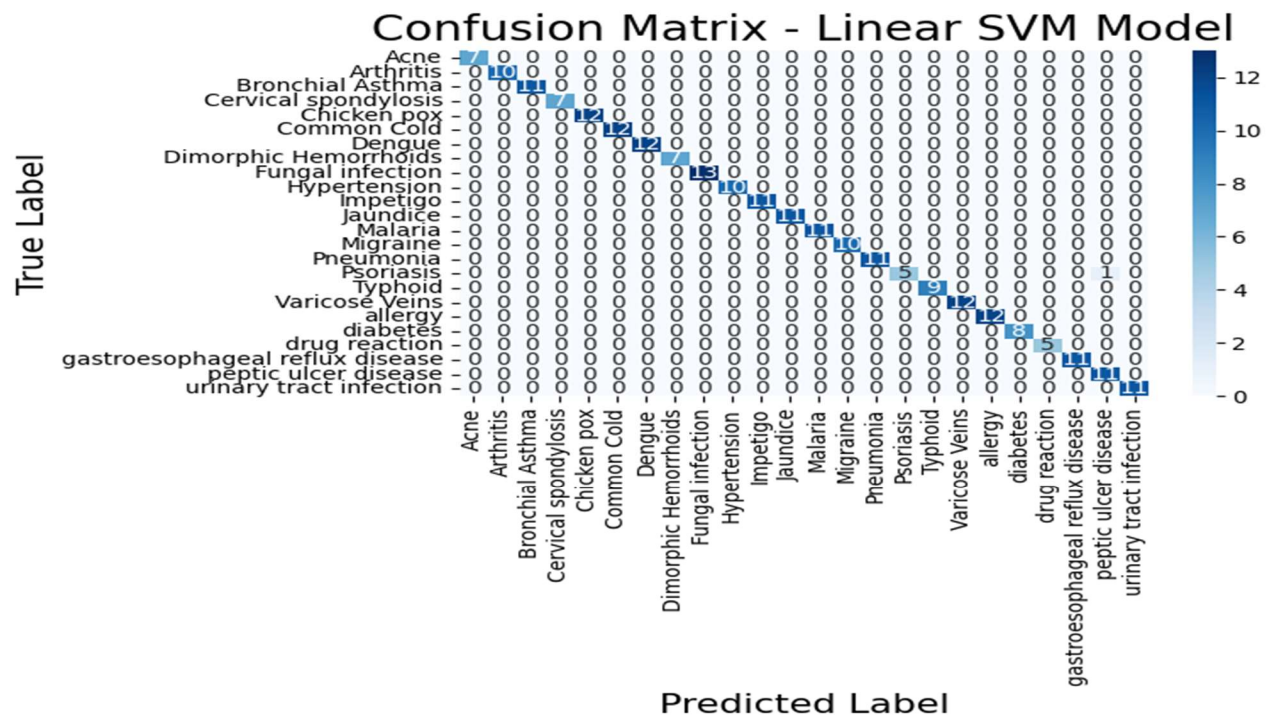


Fig. 7 Confusion Matrix SVM



5.3 Results of Deep Learning and Transformer-Based Models:

I began by using three transformer models. I chose models pre-trained on medical data to ensure good results. Three advanced deep learning models based on transformers were evaluated for their ability to understand medical texts and classify diseases with significantly higher accuracy compared to traditional machine learning methods. These systems include BERT, BioBERT, and Clinical BERT, all of which use transformers designed to extract semantics and contextual links from textual information.

The Clinical BERT architecture achieved the best overall results, with an accuracy of 97.5% and an F1 score of 97%. Thanks to its initial training on medical and healthcare terminology, Clinical BERT demonstrated high performance due to its pre-training and the quality of its data after cleaning.

The BioBERT architecture also achieved strong and consistent results, with an accuracy of 0.96. Using biomedical databases (such as PubMed and PMC) in training the system allowed it to efficiently decode disease-related terminology, although its effectiveness was slightly lower than that of Clinical BERT.

The standard BERT design achieved an accuracy of 0.95 across all metrics. Although not trained exclusively on health data, its transformer-based framework helped extract contextual patterns from textual symptom reports.

In summary, the results indicate that all transformer-based deep learning systems outperformed traditional machine learning approaches in understanding and analysing medical texts. Among these systems, Clinical BERT emerged as the most efficient, delivering highly reliable disease predictions due to its ability to understand context and its prior training on medical data.

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 confusion matrix was used because it reveals errors, not just the error rate; it is an effective tool, especially in the medical field. The clinical BERT model gave the highest accuracy among all models. The idea behind this is that most of the correct values are concentrated along the diagonal line, indicating that the model accurately predicted almost all disease categories with very few off-line misclassifications. This confirms the ability of the clinical BERT model to understand the medical context and effectively differentiate between similar symptoms.

This diagram illustrates that the numbers on the main diagonal in dark blue are the correct predictions, and the numbers outside the main diagonal are the incorrect ones.

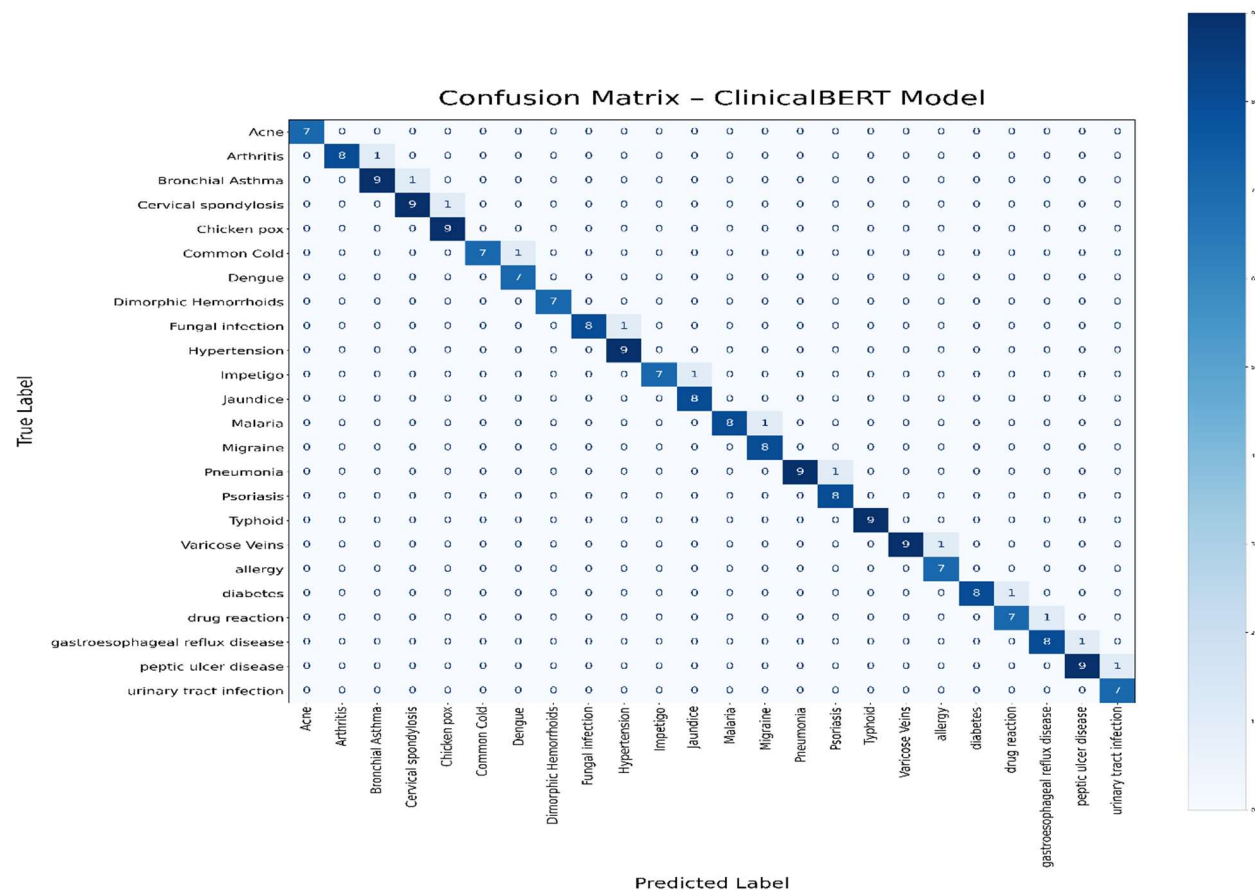




FIG.8 Confusion Matrix Clinical BERT

5.5 Discussion:

After evaluating all models, the transformer-based deep learning models demonstrated superior performance in symptom analysis and disease prediction, yielding good and logical results. The results were confirmed to be non-overlearning through a 50-epoch trial with early stop-training monitoring (validation loss) and discontinuation of training if performance did not improve after 5 consecutive epochs. Following an objective evaluation of the models, the study indicated that transformer-based deep learning models outperformed traditional machine learning models in predicting diseases based on textual symptom descriptions. This superiority is attributed to the quality and volume of the data, as well as the fact that they were pre-trained on medical data.

Performance of Machine Learning Models:

The machine learning models performed well, understanding the symptoms from the texts and yielding good results in predicting the disease. This is likely due to the use of TF-IDF, which helped the models identify symptoms from the texts. The moderate data size also played a significant role. Furthermore, thorough data cleaning resulted in high data quality.

Performance of Deep Learning / Transformer Models:

Transformer-based deep learning models have shown excellent results, due to their ability to understand contextual meaning and their prior training on clinical medical data. The volume and quality of the data also contributed to this outcome.

Reasons for Similar Performance Among DL Models:

The transformer models achieved similar performance with a slight ClinicalBERT advantage, and this table illustrates the main reasons for the similarity in results.

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?	NLP techniques were successfully implemented using TF-IDF for machine learning models and contextual tokenization for transformer-based models. These methods enabled the extraction of clinically relevant features and preserved the meaning of medical terminology.
b. Can existing machine learning algorithms (SVM, LR, NB) and advanced transformer-based learning models (BERT, BioBERT, ClinicalBERT) be used for disease prediction?	ML models provided strong baseline performance, with Linear SVM performing best among them. However, transformer-based models achieved higher accuracy overall, with Clinical BERT outperforming all other models due to its medical domain pretraining.

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 and Limitations of the Proposed System:

Points of Strength:

The suggested framework displays several benefits that allowed it to attain excellent results in ailment categorization using written symptom accounts. **Mainly, employing a fusion of established machine learning methods and transformer-based deep learning structures permitted a thorough assessment across varied architectures, enabling the precise selection of the best model. Furthermore, relying on advanced frameworks like ClinicalBERT and BioBERT provided a strengthened comprehension of medical terminology and handling of clinical expressions. Also, the quality of the cleansed data and the clear relationship between metrics and conditions significantly assisted in improving system performance.**

**Limitations:**

All models achieved strong results, but the transformers ultimately outperformed them. However, some limitations are worth noting. The dataset was of medium size, containing 1200 rows, which is small for sophisticated transformer-based deep learning models. It only included textual symptoms, omitting age and chronic diseases, making the model limited and impractical. The database included only 24 diseases, a very small number.



Chapter 6:

6.1 Conclusion:

With the emergence of artificial intelligence and its advanced technologies that have transformed the world and brought about a qualitative leap, it became necessary to leverage these technologies in many systems, including healthcare systems.

Healthcare organizations need AI-based systems that help solve the problem of written patient symptoms, as existing systems only work with structured and organized data. This study aims to address this problem by using techniques that help analyze written symptoms and accurately predict the correct disease. The study followed a clear methodology and utilized an open database from Kaggle. Machine learning models were trained using Supporting Linear Vector Machines (SVM), logistic regression (LR), Naive Bayes((NB), random forest (RF), and XGBoost, as well as transformer-based deep learning models including BERT, BioBERT, and Clinical BERT. Each model was evaluated individually, and the best model was selected.

Clinical BERT achieved the highest accuracy rate of 0.97% which is expected since it was pre-trained on medical data. This is where the importance of this study lies, as it has demonstrated the effectiveness of Transformers in the medical field and their ability to understand meaning from context and predict disease. This will benefit both physicians, reduce diagnostic errors, and improve the efficiency of healthcare services.

6.2 Limitations:



The system achieved favourable results, but this does not negate the limitations of this study. Among the most significant limitations was the database size; it was of moderate size, containing 1200 rows related to 24 diseases. This size is considered limited for transformer models trained on large datasets. Another limitation was the tiny number of diseases included, and the database lacked any vital indicators such as age, sex, and chronic diseases. These limitations do not indicate a weakness in the system; on the contrary, the system excelled and yielded satisfactory results. However, to implement such a system in healthcare facilities, the database must be expanded.

The system should be tested under medical supervision to confirm its effectiveness. This system does not replace doctors but rather complements their work.

6.3Future Work:

This research experience was a wonderful and enriching one from which I learned a great deal. This research will not stop here; I will continue my work by expanding the database and experimenting with other advanced models. I will also work on predicting the appropriate treatment to create a comprehensive symptom-based treatment system.



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.

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

