

منى عبدالوهاب

Master Thesis Mona Abdulwahab .docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:338295600

Submission Date

Dec 19, 2025, 4:30 AM GMT+2

Download Date

Dec 20, 2025, 6:38 PM GMT+2

File Name

Master Thesis Mona Abdulwahab .docx

File Size

5.6 MB

53 Pages

11,808 Words

71,257 Characters

17% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **183 Not Cited or Quoted 16%**
Matches with neither in-text citation nor quotation marks
-  **10 Missing Quotations 1%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 11%  Internet sources
- 12%  Publications
- 11%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 183 Not Cited or Quoted 16%**
Matches with neither in-text citation nor quotation marks
- 10 Missing Quotations 1%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 11% Internet sources
- 12% Publications
- 11% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

- 1** **Internet**
doctorpenguin.com <1%
- 2** **Publication**
Arvind Dagur, Sohrit Agarwal, Dhirendra Kumar Shukla, Shabir Ali, Sandhya Sharm... <1%
- 3** **Publication**
Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Re... <1%
- 4** **Internet**
www.mdpi.com <1%
- 5** **Publication**
Nazmul Siddique, Mohammad Shamsul Arefin, K. M. Azharul Hasan, M. Shamim K... <1%
- 6** **Submitted works**
Midocean University on 2025-09-05 <1%
- 7** **Internet**
download.bibis.ir <1%
- 8** **Internet**
medium.com <1%
- 9** **Publication**
Dhirendra Kumar Shukla, Shabir Ali, Sandhya Sharma. "Artificial Intelligence and ... <1%
- 10** **Publication**
Prashant Pranav, Archana Patel, Sarika Jain. "Machine Learning in Healthcare an... <1%

11	Submitted works	University of Technology, Sydney on 2025-10-03	<1%
12	Internet	aclanthology.org	<1%
13	Internet	dspace.alquds.edu	<1%
14	Internet	www.wecreateproblems.com	<1%
15	Internet	researchmgt.monash.edu	<1%
16	Internet	www2.mdpi.com	<1%
17	Internet	doaj.org	<1%
18	Internet	repositorio.unini.edu.mx	<1%
19	Submitted works	University of San Diego on 2025-08-11	<1%
20	Internet	esource.dbs.ie	<1%
21	Submitted works	Islington College,Nepal on 2025-12-16	<1%
22	Submitted works	Strathmore University (Main Account) on 2025-11-27	<1%
23	Internet	www.frontiersin.org	<1%
24	Internet	www.lac.org.tw	<1%

25	Publication	Dinesh Goyal, Payal Bansal, Ritam Dutta, Madhav Sharma. "Intelligent Systems U...	<1%
26	Publication	Shalli Rani, Ayush Dogra, Ashu Taneja. "Smart Computing and Communication fo...	<1%
27	Internet	deepai.org	<1%
28	Internet	journal.esrgroups.org	<1%
29	Internet	openprairie.sdstate.edu	<1%
30	Submitted works	Munster Technological University (MTU) on 2025-08-20	<1%
31	Submitted works	Swiss School of Business and Management - SSBM on 2025-08-29	<1%
32	Submitted works	University of Surrey on 2025-01-01	<1%
33	Submitted works	University of Ulster on 2025-12-11	<1%
34	Internet	creativecommons.org	<1%
35	Internet	uu.diva-portal.org	<1%
36	Internet	www.ncbi.nlm.nih.gov	<1%
37	Publication	"Natural Language Processing and Chinese Computing", Springer Science and Bu...	<1%
38	Publication	Arvind Dagur, Sohith Agarwal, Dharendra Kumar Shukla, Shabir Ali, Sandhya Sharm...	<1%

39	Submitted works	University of Rwanda on 2025-09-14	<1%
40	Publication	Vangoor, Srujay Reddy. "Exploring Advances in Multi-Label Text Classification for ...	<1%
41	Internet	ijmrset.com	<1%
42	Publication	"Advances in Emerging Technologies and Computing Innovations", Springer Scie...	<1%
43	Publication	Salmah Saad Al-qarni, Abdulmohsen Algarni. "Disease Prediction from Symptom ...	<1%
44	Publication	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ...	<1%
45	Submitted works	Valencia College on 2025-01-21	<1%
46	Internet	computersciencejournal.org	<1%
47	Internet	inass.org	<1%
48	Internet	itegam-jetia.org	<1%
49	Internet	kth.diva-portal.org	<1%
50	Internet	publications.eai.eu	<1%
51	Internet	www.preprints.org	<1%
52	Submitted works	Alliance College of Engineering and Design on 2025-11-04	<1%

53	Publication	Debasis Chaudhuri, Jan Harm C Pretorius, Debashis Das, Sauvik Bal. "Internationa...	<1%
54	Publication	Shankar Babu, Mahesh Babu Kota. "Synergies in Smart and Virtual Systems using...	<1%
55	Internet	era.ed.ac.uk	<1%
56	Internet	www.ijert.org	<1%
57	Internet	www.jetir.org	<1%
58	Submitted works	Sydney Polytechnic Institute on 2025-05-26	<1%
59	Submitted works	UNICAF on 2025-07-20	<1%
60	Submitted works	University of Oxford on 2025-05-21	<1%
61	Internet	link.springer.com	<1%
62	Internet	www.semanticscholar.org	<1%
63	Internet	123dok.net	<1%
64	Publication	Durgesh Kumar Mishra, Nilanjan Dey, Bharat Singh Deora, Amit Joshi. "ICT for Co...	<1%
65	Publication	de Castro, Rute Salomé Pires. "Development of an Artificial Intelligence Breast Ca...	<1%
66	Internet	nano-ntp.com	<1%

67	Internet	ruor.uottawa.ca	<1%
68	Internet	www.cuantum.tech	<1%
69	Internet	www.ijitee.org	<1%
70	Publication	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dharendra Kumar Shukla. "Intelli...	<1%
71	Publication	Simon Elias Bibri. "Generative AI and Generative AI of Things for Sustainable Sma...	<1%
72	Submitted works	University of Hertfordshire on 2025-08-27	<1%
73	Submitted works	University of Stirling on 2025-08-16	<1%
74	Submitted works	University of Ulster on 2025-12-11	<1%
75	Internet	aircconline.com	<1%
76	Internet	arxiv.org	<1%
77	Internet	datahorizonresearch.com	<1%
78	Internet	inaoe.repositorioinstitucional.mx	<1%
79	Internet	ir.lib.uwo.ca	<1%
80	Internet	www.geeksforgeeks.org	<1%

81	Internet	www.medrxiv.org	<1%
82	Internet	www.theseus.fi	<1%
83	Submitted works	Fakultet elektrotehnike i računarstva / Faculty of Electrical Engineering and Com...	<1%
84	Publication	Lecture Notes in Computer Science, 2015.	<1%
85	Submitted works	Queen's University of Belfast on 2023-09-08	<1%
86	Submitted works	The University of the West of Scotland on 2025-12-08	<1%
87	Publication	Udara Yedukondalu, V Vijayasri Bolisetty. "Advancing Innovation through AI and ...	<1%
88	Submitted works	University College London on 2018-09-06	<1%
89	Submitted works	University of Leeds on 2004-11-01	<1%
90	Internet	d.docksci.com	<1%
91	Internet	dspace.ut.ee	<1%
92	Internet	lutpub.lut.fi	<1%
93	Internet	pmc.ncbi.nlm.nih.gov	<1%
94	Internet	riunet.upv.es	<1%

95	Internet	sciforum.net	<1%
96	Internet	thesis.unipd.it	<1%
97	Internet	torbath.ac.ir	<1%
98	Internet	www.diva-portal.org	<1%
99	Internet	www.researchsquare.com	<1%
100	Publication	Adamu Mohammed Mustapha, Peter Nimbe, Abdul Razak Nuhu. "Comparative an...	<1%
101	Submitted works	Adtalem Global Education on 2024-11-09	<1%
102	Publication	Botelho, Inês Alexandra Vieira. "Emotion Recognition Using Multimodal Time Seri...	<1%
103	Submitted works	Central Queensland University on 2024-06-16	<1%
104	Publication	Chukwuebuka Joseph Ejayi, Dongsheng Cai, Makuachukwu B. Ejayi, Ijeoma A. Chik...	<1%
105	Publication	Hussein Al-kaabi, Fuqdan Al-ibrahimi, Ali kadhim jasim, Zainab S. Idan. "BCGH-Ne...	<1%
106	Submitted works	ICS DI on 2025-12-03	<1%
107	Submitted works	ICS DI on 2025-12-03	<1%
108	Submitted works	Lahore College for Women University on 2025-11-21	<1%

109	Submitted works	Midocean University on 2025-09-05	<1%
110	Publication	Mujtaba, Ghulam. "Feature Engineering Techniques to Classify Cause of Death fro...	<1%
111	Publication	Rajeev Agrawal, Arun Pratap Srivastava, Akihiko Sugiyama. "International Confer...	<1%
112	Publication	Ramzi Haraty, Mira Chehade. "Transfer learning and sentiment analysis of lebane...	<1%
113	Publication	Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computati...	<1%
114	Submitted works	University of New York Tirana on 2025-12-05	<1%
115	Submitted works	University of Patras on 2022-11-04	<1%
116	Internet	assets-eu.researchsquare.com	<1%
117	Internet	assets.researchsquare.com	<1%
118	Internet	erepo.uef.fi	<1%
119	Internet	ijcsrr.org	<1%
120	Internet	jedm.educationaldatamining.org	<1%
121	Internet	mlcb.github.io	<1%
122	Internet	ojs.wiserpub.com	<1%

123	Internet	ray.yorks.ac.uk	<1%
124	Internet	worldwidescience.org	<1%
125	Internet	www.coursehero.com	<1%
126	Internet	www.gecekitapligi.com	<1%
127	Submitted works	Bournemouth University on 2023-05-12	<1%
128	Publication	"Proceedings of the Third International Conference on Cognitive and Intelligent ...	<1%
129	Publication	Deepak Gupta, Anuj Kumar Singh, Yu-Dong (Eugene) Zhang. "Security and Privacy...	<1%
130	Submitted works	Sheffield Hallam University on 2021-12-30	<1%
131	Publication	Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computati...	<1%
132	Submitted works	University of Hull on 2023-08-23	<1%



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.

My dear Joory, Abdul Malik, and Lana, thank you for being patient and strong during the long study sessions. You are the stars that led me, the light that showed me the way, and the reason I keep going on this journey. Your real love and support helped me get through times when I was tired. You were the reason I was able to reach my goals.

Thanks for your love and support, my **husband**. Your trust in me got me through hard 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 SVM	38
• 5.3 Results of Deep learning Models.....	39
• 5.4 Confusion Matrix clinical BERT	40
• 5.5 Discussion	42
• 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:

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. It accurately predicts the most likely disease.

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 extractionThese 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. Develop an AI-based and natural language processing system capable of analyzing 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

This chapter goes into a lot of detail on how this project was carried out. It explains



how the data was gathered and improved, what frameworks were used and how they were taught and checked for accuracy, how the system operates to forecast diseases, and the main performance indicators that were used.

Chapter 5: Experimental Results and Discussion

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

Chapter 6: Conclusion and Future Work

The last chapter goes over the main points of the dissertation and talks about what the whole study added. It talks about what was done and gives ideas for more research in the future. 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.

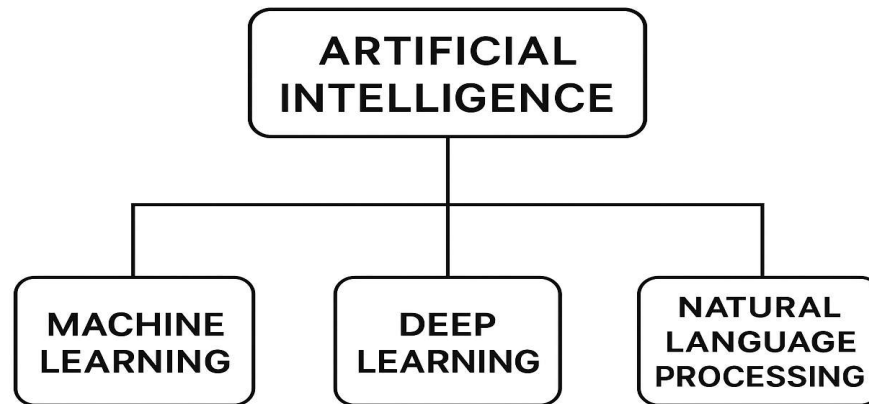


FIG.1 Artificial Intelligence (AI)

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.

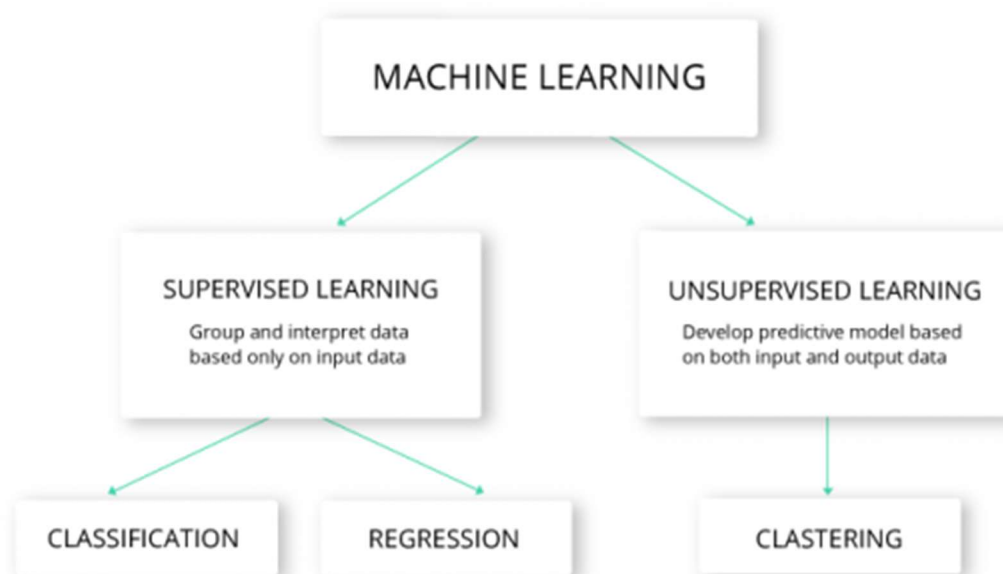


FIG.2 Machina learning (ML):

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]

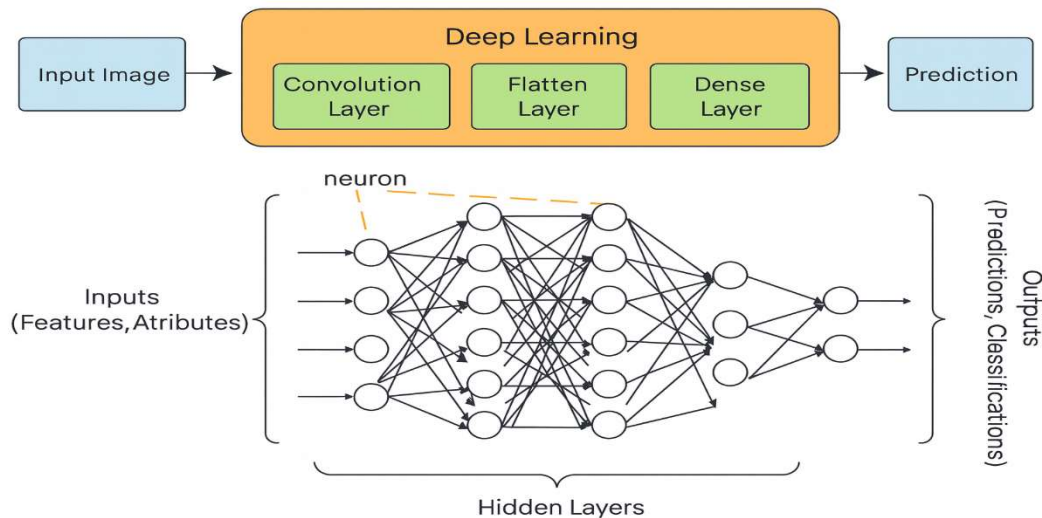


FIG.3 Deep Learning (DL)

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.

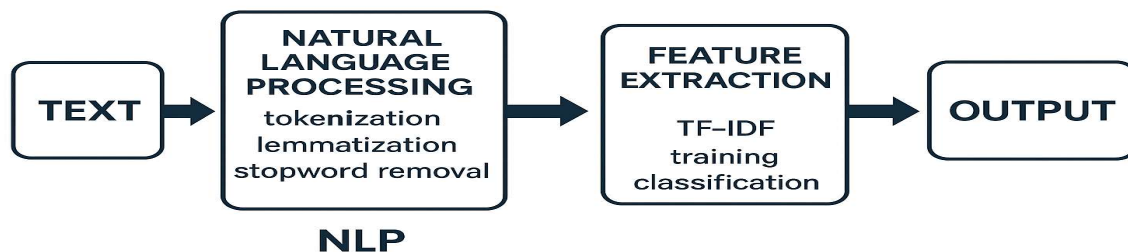


FIG.4 Natural Language Processing (NLP)

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.
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.

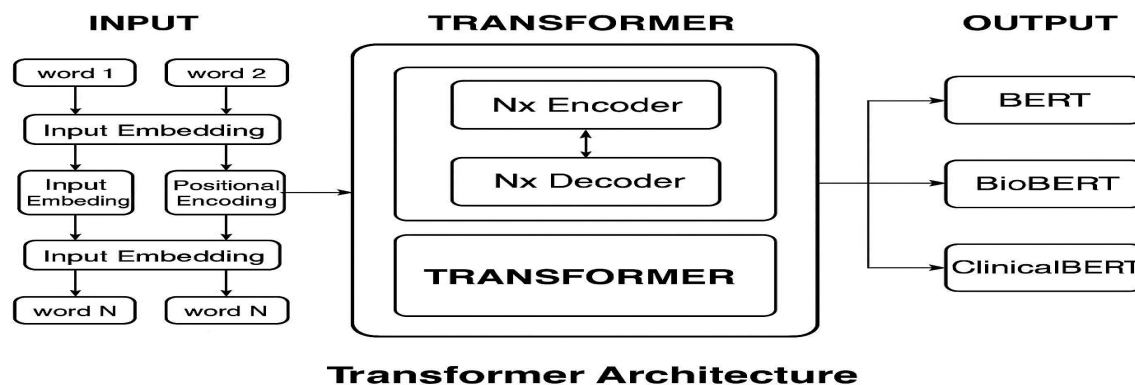


FIG.5 Transformer Architecture

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



72 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.

11 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).

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

80 **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.

77 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]



Chapter 3: Related Work

3.1 Introduction:

Our time is what it is because of change and speed. AI is a great tool for finding patterns in data that were hard to find before. Researchers have been looking for new ways to help people make better decisions, especially when it comes to health care, by making predictions more accurate. This is possible because AI models and technology are becoming more complex.

Before you start your own study, it's important to read about other studies on the same topic. They can help you with your work and fill in any gaps. The last study talks about the methods used, the results found, and the problems that still exist. This background sets the stage for more research and makes sure that scientific knowledge keeps growing by filling in the gaps that are already there. This was the first step in putting the system together.

3.2 History

Ferhi et al. (2025) [9] This study investigated the use of machine learning and crowd learning techniques to enhance symptom-based health assessment systems built on structured tabular data. The authors conducted multiple experiments using three publicly available diagnostic datasets: the **Reference Collection**, **Cough Differential Diagnosis (Cough-DDX)**, and **Cough-DDX2**, all of which focus on respiratory symptoms and provide reproducible structured



information.

Several classification models were evaluated, including **Decision Tree (DT)**, **Support Vector Machine (SVM)**, and **Logistic Regression (LR)**, alongside ensemble-based grouping strategies such as **cumulative and sequential grouping**. Although no detailed numerical results were reported, the findings indicated that **grouping-based ensemble approaches consistently outperformed individual classifiers**. In particular, **cumulative grouping combined with DT and SVM** achieved the best overall performance across standard evaluation metrics, including accuracy, precision, sensitivity, and F1-score.

Sogandi et al. (2024) [10] This study explored the use of both **supervised and unsupervised machine learning techniques** in a structured diagnostic setting to support disease classification based on symptom data. The authors utilised a **publicly available Kaggle dataset** in which symptoms were represented as **predefined structured features**, facilitating feature analysis and model training.

The primary objective was to identify the most influential symptoms and establish a comprehensive diagnostic rule set. Correlation rules were first generated using a prior-based approach, after which several predictive models were evaluated, including **stepwise regression**, **Support Vector Machine (SVM)**, **bootstrap forest**, and **booster trees**. Although no numerical performance metrics were reported, the study identified **stepwise regression as the most effective model** for accurate disease classification.

Hassan et al. (2024) [11] This study investigated the use of **advanced language models** to improve disease prediction from textual symptom descriptions. The authors compared traditional deep learning approaches, such as **BiLSTM**, with specialised transformer-based models, particularly **MCN-BERT**. Two datasets were utilised: the first contained **1,200 symptom–disease text pairs**, while the second included **over 23,000 symptom descriptions related to adverse drug reactions**.

The experimental results showed that **MCN-BERT consistently outperformed BiLSTM**, especially when enhanced using the **DAMP technique**. The model achieved an accuracy of approximately **99.6%** on the symptom–disease dataset and around **96%** on the adverse drug reaction dataset. These findings highlight the ability of specialised transformer models to capture subtle linguistic patterns in clinical text, supporting their effectiveness in automated disease classification and early diagnosis.

Al-Qarni et al. (2025) This study proposed a **deep learning–based framework** for disease prediction using **free-text symptom descriptions written in colloquial language**. The authors constructed two custom datasets: the first included symptom descriptions for **24 diseases**, while



the second expanded the scope to **41 diseases**. These datasets were manually compiled from publicly available sources; however, they were **not released publicly**, limiting external validation.

The proposed approach incorporated **natural language processing techniques**, including tokenisation and **TF-IDF representation**, alongside deep learning models such as **LSTM**, **CNN-LSTM**, and **GRU**. Experimental results showed that the **CNN-LSTM model with tokenisation achieved the highest accuracy of 99.9%** on the 41-disease dataset, while the **LSTM model with TF-IDF reached 98.8% accuracy** on the 24-disease dataset. These findings demonstrate the strong capability of deep learning models in handling textual symptom data, while also highlighting limitations related to **dataset availability and reproducibility**.

Serigne'd et al. (2024) [13] This study, aimed to enhance medical diagnosis by predicting diseases based on symptom data. The researchers utilised a **publicly available Kaggle dataset** consisting of **132 structured symptoms associated with 41 diseases**.

Several machine learning models were evaluated, including **Decision Tree (DT)**, **Support Vector Machine (SVM)**, **Random Forest (RF)**, and **Extreme Gradient Boosting (XGBoost)**. The experimental results showed that **XGBoost outperformed the other models**, achieving the **highest diagnostic accuracy in** disease prediction based on symptom inputs.

Finkelstein et al. (2022) [14] This study aimed to support timely medical decision-making by **predicting symptom progression using self-reported patient data**. The researchers applied two deep learning models, **Convolutional Neural Network (CNN)** and **Long Short-Term Memory (LSTM)**, on a longitudinal dataset consisting of **daily symptom reports from 208 cancer patients undergoing chemotherapy**. The recorded symptoms included fatigue, nausea, pain, and related conditions.

The models were evaluated over different prediction windows ranging from **3 to 7 days** to forecast symptom exacerbation. The results showed that the **CNN model with a 3-day prediction window outperformed LSTM**, achieving an **accuracy of 81%**, **precision of 87%**, **recall of 80%**, and an **F1-score of 83%**. Although LSTM achieved slightly lower but comparable results, the findings indicate that **CNN is a strong candidate for real-time symptom monitoring in digital health systems**. The dataset is based on real clinical patient-reported data and is available upon request.



Abhinav et al. (2023) [15] This study proposed a machine learning-based architecture for predicting multiple chronic diseases, including breast cancer, cardiovascular disease, and diabetes. The system was designed to be simple and user-friendly and was deployed through a web-based interface developed using Python, Django, and Streamlit, allowing users to input health-related data and receive disease predictions.

The authors evaluated Decision Tree (DT) and Support Vector Machine (SVM) models, with Decision Tree consistently achieving the best performance across all tasks. The model attained an accuracy of 97.98% for breast cancer, 92.62% for heart disease, and 91.55% for diabetes, demonstrating the effectiveness of tree-based models for multi-disease prediction in practical healthcare applications.

Afrin et al. (2021) [16] This study aimed to develop an effective supervised machine learning approach for the early prediction of liver diseases. The researchers utilised the Indian Liver Patient Dataset (ILPD) as the primary data source. A wide range of classifiers was evaluated, including Logistic Regression, Decision Tree, Random Forest, AdaBoost, KNN, LDA, Gradient Boosting, and SVM. To enhance model performance and interpretability, LASSO feature selection was applied to identify the most relevant clinical attributes associated with liver disease.

The models were evaluated using 10-fold cross-validation and standard metrics such as accuracy, precision, recall, and F1-score. The results demonstrated that the Decision Tree model combined with LASSO achieved the best performance, with an accuracy of 94.29%, precision of 92%, recall of 99%, and an F1-score of 96%, highlighting the effectiveness of integrating tree-based models with robust feature selection techniques for reliable liver disease prediction.

Chintan M. et al. (2023) [17] This study introduced a machine learning-based framework for cardiovascular disease prediction, aiming to support fast and accurate automated diagnosis while reducing the risk of clinical decision errors. Several supervised ML models were evaluated, including Decision Tree (DT), Random Forest (RF), XGBoost (XGB), and Multilayer Perceptron (MLP). To improve classification performance, k-modes clustering with Huang initialisation was applied during data preprocessing, and GridSearchCV was employed for hyperparameter tuning.

The experiments were conducted on a large real-world dataset consisting of 70,000 patient records from Kaggle. The results showed that the MLP model outperformed the other models, achieving an accuracy of 87.28% and an AUC of 0.95.



42 Umar et al. (2022) [18] This study proposed several **machine learning-based approaches** for the rapid detection of cardiac abnormalities. Four main frameworks were introduced, including a **weighted Naïve Bayes model**, a hybrid approach combining **SVM and XGBoost** with time-domain, frequency-domain, and information-theoretic features, an enhanced **SVM model refined through duality optimisation**, and a comprehensive clinical decision-support pipeline integrating **DBSCAN** for outlier removal, **SMOTE-ENN** for class imbalance handling, and **XGBoost** for final disease prediction.

47 The system was evaluated using standard metrics such as **accuracy, precision, recall, and F1-score**, with **XGBoost consistently achieving the highest accuracy and robustness** across
54 different experimental settings. The study highlights the potential of machine learning techniques
16 in enabling fast and reliable cardiac diagnosis. However, the lack of clarity regarding data
42 collection and the absence of a publicly available dataset limits the reproducibility and generalisability of the proposed approaches.

18 Abdur et al. (2022) [19] This study aimed to predict COVID-19 infection outcomes using the **Light Gradient Boosting Machine (LGBM)** algorithm. The primary objective was to develop a **fast, simple, and accurate** predictive framework suitable for medical advisory applications. The researchers utilised a **structured dataset** containing patient-reported symptoms such as fever, cough, and fatigue, along with diagnostic results and demographic information.
7 LGBM was selected due to its ability to efficiently handle large structured datasets with minimal
79 computational cost. Experimental results showed that the proposed framework achieved **high accuracy with low training time**, making it well suited for rapid COVID-19 screening and patient triage in real-world clinical settings. The use of the **COVID-19 Symptom and Patient Information dataset** further supported the practical applicability of the proposed approach.
2

39 Lee et al. (2023) [20] This study proposed a personalised medical recommendation framework using **Graph BERT**, a graph-based transformer model, to generate individualised treatment suggestions. The authors utilised the **MIMIC-III dataset**, which contains real hospital records, enabling accurate modelling of patients' clinical histories and complex relationships between medical entities.

121 The experimental results demonstrated that **Graph BERT significantly outperformed traditional baseline models**, primarily due to its ability to capture intricate relational and contextual information within clinical data. This led to more accurate and clinically meaningful



treatment recommendations.

While the study highlights the importance of integrating relational knowledge in medical recommendation systems, its focus on treatment recommendation rather than **symptom-based disease prediction** suggests an opportunity for future research that combines both diagnostic and therapeutic components within a unified clinical decision-support system.

Prabhakar et al. (2025) [21] This study focused on improving the extraction of patient characteristics from **unstructured clinical notes** using advanced deep learning techniques. The researchers employed the **MIMIC-III dataset**, which contains rich and realistic clinical narratives suitable for evaluating phenotyping tasks in real-world healthcare settings. The proposed model was evaluated using standard performance metrics and outperformed traditional approaches, including **vanilla BERT**, **BioBERT**, and basic **CNN-based models**. The system achieved an **AUROC of 0.80**, an **F1-score of 0.72**, and a **Hamming loss of 0.09**, demonstrating **strong performance in multi-label clinical classification**, which is essential for accurate patient phenotyping.

Kumar et al. (2023) [22]

This study aimed to develop a predictive framework for diabetes diagnosis using traditional machine learning techniques. The authors utilised the **PIMA Indians Diabetes Dataset**, which consists of structured numerical clinical data. Several classification models were evaluated, including **Support Vector Machine (SVM)**, **Logistic Regression (LR)**, and **Decision Tree (DT)**. Among these models, **SVM achieved the highest accuracy of 89%**, demonstrating its effectiveness in medical prediction tasks when working with well-structured numerical features. The findings indicate that conventional machine learning models can produce reliable results in controlled diagnostic settings using structured clinical data. However, the approach is limited in real-world applications, as patient symptoms are typically expressed in free-text form rather than numerical values. Despite this limitation, the study provides an important benchmark for evaluating more advanced models that incorporate natural language processing techniques.

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 machine 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	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%



		non-ADR”.			
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 chemotherapy “self-reported , daily,	“CNN” LSTM	CNN “3-day window”	Accuracy: - 81% Precision: - 87% Recall: - 80% F1-score: - 83%



		custom dataset"			
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

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. 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.

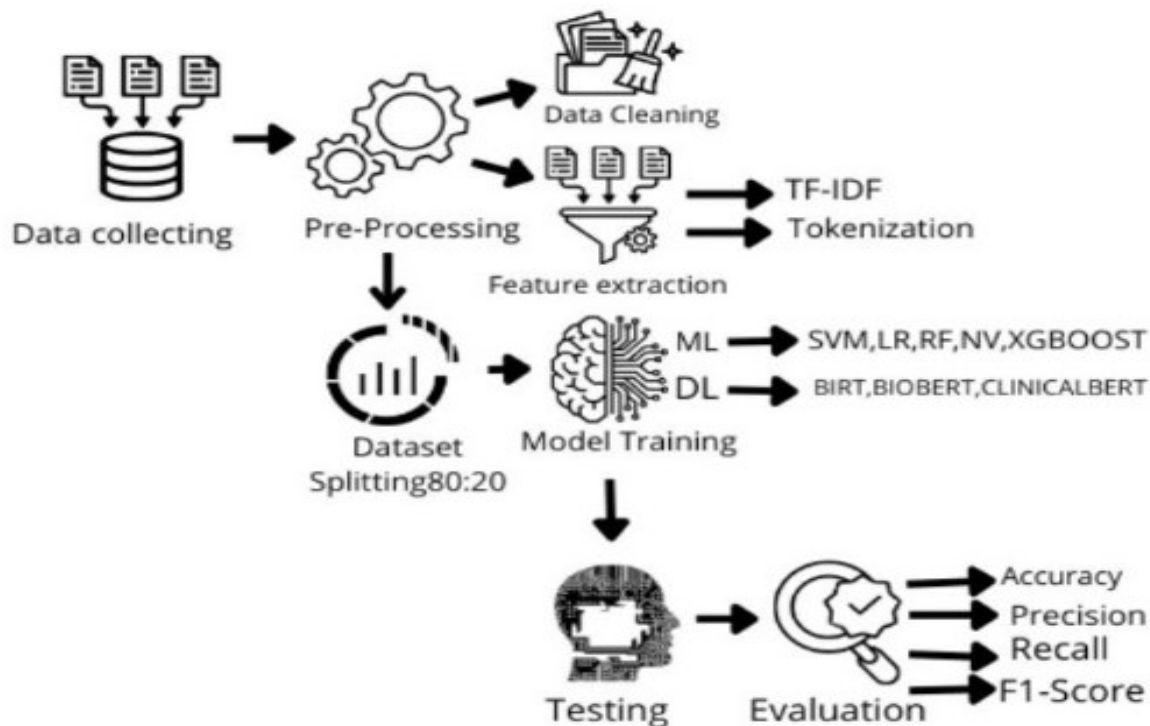


FIG.6 Workflow

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 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 keeps the repository in good shape.

We divided the material into two groups, 80% and 20%, which is a usual technique to study, to make sure that the models' abilities were tested fairly and accurately. BERT, BioBERT, and ClinicalBERT are all examples of deep learning and machine learning methods that employ this strategy a lot. There are two categories of information: one for training the system (80%) and the other for checking it (20%).

Stratification made the evaluation much stronger and more convincing. This technique makes sure that both data groups have the same number of cases for each illness type. This division keeps the test from overfitting and makes it more dependable.

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:

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. Each tree yields a decision.



44
21
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]

2 5 XG Boost Extreme Gradient Boosting:

It is a machine learning model used for prediction and decision-making. The idea behind it is to build a set of simple models that work together to gradually improve results. At each step, the model attempts to correct errors that appeared in previous predictions, making it very robust and accurate. Learning from its mistakes step by step, it delivers highly accurate results even with complex data. It is fast to train compared to other models and reduces the problem of overfitting.[31]

64 75 MNB Multinomial Naive Bayes:

MNB is a kind of Naive Bayes algorithm. It is a probabilistic model that assumes that features are not related to each other. It is based on probability theory and assumes that features are independent. This model works best with data that can be counted, like how often words appear in text or how often words are repeated in documents. Because of this, it is often used to sort text and look at symptoms. The main idea is to look at how often each word shows up and how likely it is to be in a certain category. Then, choose the category with the highest chance, 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.

76 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 important to make sure that the model can not only store data, but also deal with new data that it has never seen before. This shows its true ability to learn and generalise.

This helps make sure that the model has really learnt useful patterns and isn't just memorising the training data. 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. 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.

4.8 Summary:

It was just an idea, and here this idea was implemented, solutions were found for the problem, and the goals and contributions of this thesis were achieved.

The primary objective of this work was to predict diseases based on free-text symptom descriptions. To achieve this, the study started with a medical dataset consisting of 1,200 textual symptom records representing 24 disease categories. Since the quality of textual data plays a crucial role in model performance, a careful preprocessing stage was carried out. This included cleaning and normalization procedures, followed by TF-IDF feature extraction for machine



learning models, as this representation is well suited for handling symptom-based text in datasets of this size.

For transformer-based deep learning models, no manual feature extraction was required. These models were selected because they are inherently capable of capturing semantic and contextual relationships within text, which is particularly important when dealing with medical symptoms. In the machine learning stage, a set of classification models was tested to compare their performance on textual symptom data, including Support Vector Machine, Logistic Regression, Random Forest, XGBoost, and Naïve Bayes

In parallel, pretrained transformer models BERT, BioBERT, and ClinicalBERT were fine-tuned using the same dataset. Their prior exposure to large-scale biomedical and clinical text enabled them to better recognize symptom patterns and improve disease prediction performance. Among these models, ClinicalBERT showed a strong ability to interpret clinical language accurately.

Finally, the system was evaluated using commonly adopted performance metrics in medical classification tasks, such as accuracy, precision, recall, F1-score, and confusion matrices. This evaluation provided a clear indication of how effectively the proposed system can identify diseases and distinguish between different classes.

.



Chapter 5: Experimental Results and Discussion:

Introduction:

This segment of the thesis is very significant since it discusses the endeavour's worth, its advantages and drawbacks, and the most crucial 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.

This research aims to conduct a comprehensive and equitable comparison that elucidates the advantages and disadvantages of each model.

We examined each model individually to assess its performance and identify the optimal one.

The results show that deep learning models based on transformers are better at predicting diseases and understanding natural language.

5.1 Results of Traditional Machine Learning Models:

TF-IDF made classic machine learning models work really well. These models are easy to understand and have been tried out numerous times. This strategy assists machine learning models by providing important terms weights and helping the models detect symptoms. The Supporting Linear Vector Machine (SVM), logistic regression, the Simple Complementary Bayesian Algorithm, the Simple Polynomial Bayesian Algorithm, random forests, and the XGBoost algorithm were all examples of models. When it came to working with textual symptom data, each algorithm had its own strengths.

SVM was the best model, even though it was easy to understand. It had an overall accuracy of 0.97. This shows that it was quite accurate and knew the dataset well. TF-IDF and having a lot of solid data made these great outcomes possible.

Logistic regression always got the right answer, with an accuracy of 0.95. It did well in every way. The Complement Naïve Bayes and Multinomial Naïve Bayes models did better than expected for basic probability predictors, even though they were simple. The accuracy of Complement NB was 0.91, and the accuracy of Multinomial NB was 0.92, which was a little higher. Both models were quick and worked well with the data set that wasn't balanced.



4 With an accuracy of 0.95, the Random Forest can detect nonlinear correlations in the data.

The least accurate of the six algorithms was XGBoost, which had an accuracy of 0.88. It was able to discover crucial patterns in the input indications, but it couldn't guess the ailment.

The results reveal that all of the machine learning algorithms can look at text and identify what disease it is, as long as the data is big enough 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:

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 idea behind using a confusion matrix is to compare the model's predictions with the actual values in the data. This allows us to determine the number of cases correctly classified and the number of incorrect ones. This is done through a graph showing the number of correct cases on the main line and the number of incorrect cases outside this line. The graph also shows the type of error, not just the number of errors. This tool is crucial in the medical field because it helps clinicians avoid errors when diagnosing patients. It is also used to obtain important evaluation measures such as accuracy, precision, 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. This demonstrates the model's ability to accurately distinguish between symptoms and associate them with the correct diseases.

Most of the predictions were on the main line. These errors mostly occurred in diseases with similar symptoms, such as the common cold and the flu, or migraine and tension headache. This is normal in real-world medical datasets where symptoms can overlap. The linear SVM model did an excellent job of classifying things, as evidenced by the confusion matrix.

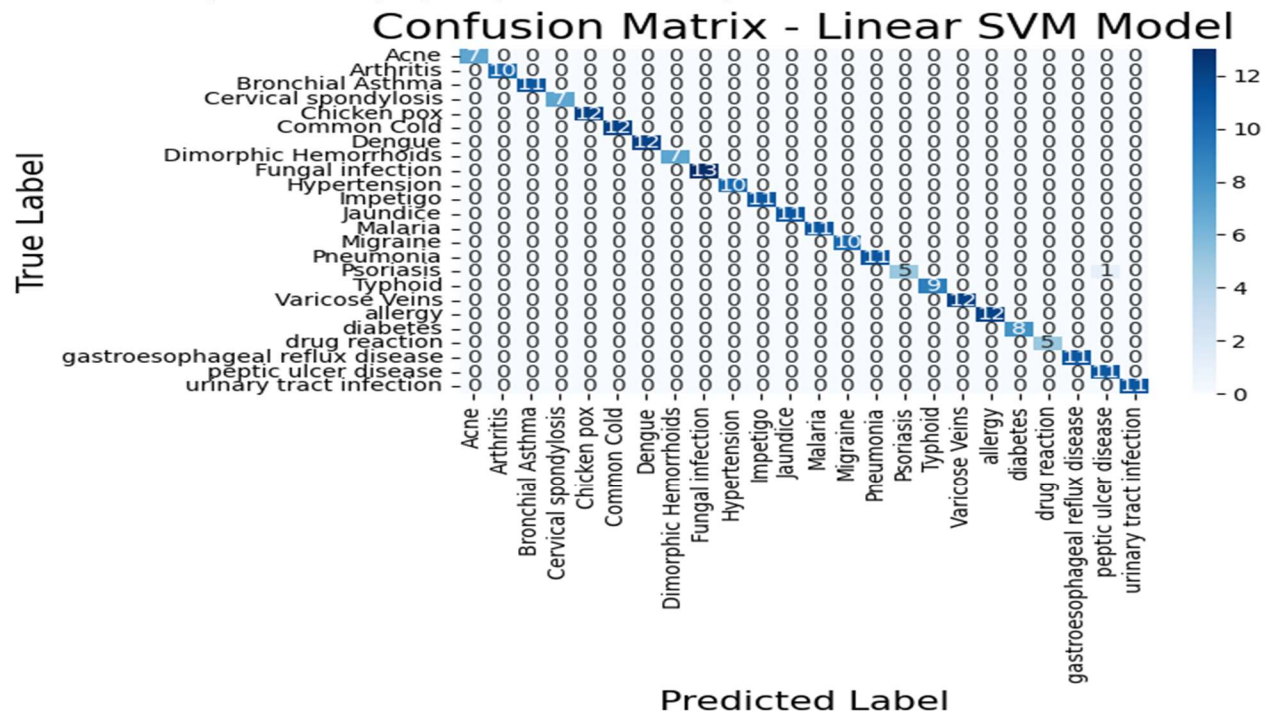


Fig. 7 Confusion Matrix SVM

5.3 Results of Deep Learning and Transformer-Based Models:



The primary objective of this study was to analyse textual symptoms and predict disease. The system successfully achieved this by employing three transformer models. These models, pre-trained on medical data, demonstrated a significantly higher accuracy in understanding medical texts and classifying diseases compared to traditional machine learning methods. The systems included BERT, BioBERT, and Clinical BERT, all of which utilize transformers designed to extract semantics and contextual links from textual information.

The Clinical BERT model 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 model also achieved strong and consistent results, with an accuracy of 0.96. However, it is trained on biomedical databases (such as PubMed and PMC) rather than clinical data, which explains its lower efficiency compared to Clinical BERT. The standard BERT design achieved an accuracy of 0.95 across all metrics. Although it wasn't trained exclusively on health data, its transformer-based framework helped extract contextual patterns from textual symptom reports.

In short, the results indicate that all transformer-based deep learning systems outperformed traditional machine learning methods in understanding and analysing medical texts. Among these systems, Clinical BERT stood out as the most efficient, delivering highly reliable disease predictions thanks to its contextual understanding and prior training on medical data. The idea is that with

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:

To ensure the accuracy of the assessment, a confusion matrix was used because it reveals errors, not just the error rate; it is a powerful tool, especially in the medical field. The clinical BERT model achieved the highest accuracy among all models. The reason 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 misclassifications outside the line. This confirms the clinical BERT model's ability to understand the medical context and effectively differentiate between similar symptoms.

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.

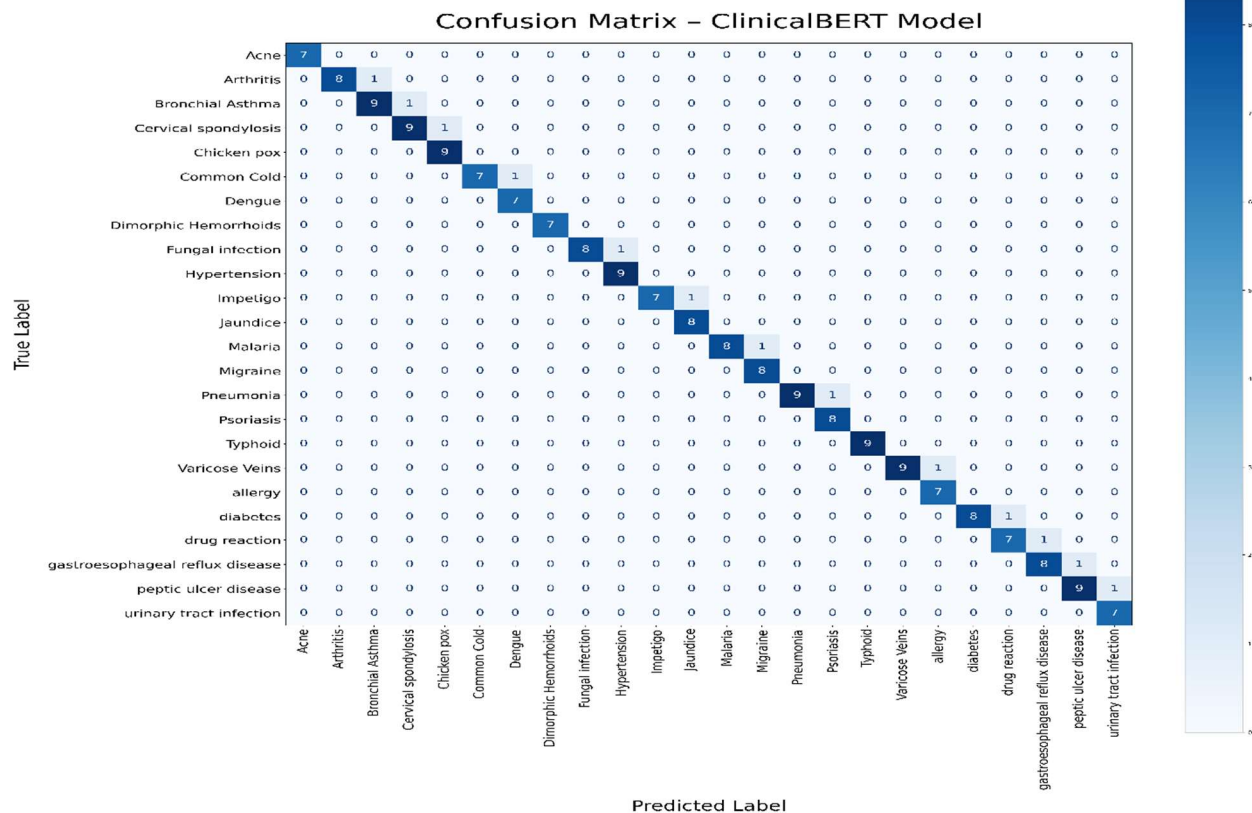




FIG.8 Confusion Matrix Clinical BERT

5.5 Discussion:

Five measures of accuracy, recall, Precision, recall, f1 score and confusion matrix were used to validate the results. Transformer-based deep learning models demonstrated superior performance in symptom analysis and disease prediction, achieving consistent and logical results. The absence of overfitting was confirmed through a 50-epoch training experiment with early stopping that monitored early training cessation (validation loss) and permanent training termination if performance did not improve after five consecutive cycles. 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 models' prior training on medical data.

Performance of Machine Learning Models:

The machine learning models did a good job of understanding the symptoms in the texts and making accurate predictions about 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 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. The research employed a defined technique and used an available database from Kaggle. Machine learning models were developed utilizing Supporting Linear Vector Machines (SVM), logistic regression (LR), Naive Bayes (NB), random forest (RF), and XGBoost, alongside transformer-based deep learning models such as BERT, BioBERT, and Clinical BERT. Each model underwent independent evaluation, and the optimal model was chosen. Clinical BERT attained the best accuracy rate of 0.97%. The favourable outcomes in both the machine learning and deep learning models are ascribed to the quality of the data post-cleaning. The application of TF-IDF technology in conjunction with machine learning models, as well as the efficacy of transformers in the medical domain, their proficiency in contextual meaning extraction, and their ability to forecast diseases. This will advantage professionals, diminish diagnostic inaccuracies, and enhance the efficiency of healthcare services.

6.2 Limitations:

I didn't expect the system to achieve such impressive success and operate so smoothly, but this doesn't negate the existence of some limitations in this study. One of the most significant limitations was the moderate size of the database, containing only 1200 records related to just 24 diseases. This size is considered limited for transformation models trained on large datasets. Another limitation was the database's lack of vital indicators such as age, sex, and chronic diseases. These limitations don't indicate a weakness in the system; on the contrary, the system excelled and achieved satisfactory results. However, implementing this system in healthcare facilities will require expanding the database.

The system should be tested under medical supervision to confirm its effectiveness. This system is not intended to replace physicians, but rather to complement their work.

6.3 Future Work:

I am very pleased with this achievement and will work on developing this work further in the future. I will begin by expanding the database to improve the model's generalizability. 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/>

