University of Essex
Department of Mathematical Sciences

MA981: DISSERTATION

Predicting Cancer Types Based on Symptom Descriptions Using Natural Language Processing

Kalpesh Vijay Patil
2310628

Supervisor: Zoe Bartlett

September 17, 2024
Colchester

Contents

1 Abstract **6**

2 Introduction **7**

2.1 Introduction of Lung, Thyroid, and Colon Cancer 7

2.2 What is Artificial Intelligence and Why to use it ? 11

2.3 How to predict this Cancer Types Text by its symptom ? 11

3 Literature Review **12**

3.1 Symptom-Based Lung Cancer Prediction Using Machine Learning Techniques 13

3.2 Artificial Intelligence for Predicting Cancer Symptoms: Enhancing Early Diagnosis 14

3.3 Optimizing Disease Classification: Leveraging Language Models for Symptom Analysis 15

3.4 Enhancing Thyroid Cancer Diagnosis through NLP-Driven Symptom Analysis 15

3.5 Natural Language Processing in Clinical Notes: A Systematic Review of Colon Cancer Concept Extraction 17

3.6 Identifying and Predicting Cancer Patient Events Using Free Text in EHRs: A Trajectory Analysis 18

3.7 Application of Naive Bayesian Networks in Predicting Diseases: A Systematic Review of 2005-2015 Studies 19

3.8 Advancements in Multi-Disease Prediction: Leveraging LSTM Networks for Enhanced Clinical Decision Support 20

4 Methodology **21**

4.1 Introduction 21

4.2 Data Description: 22

4.3 Data Preprocessing and Exploration: . . . 23
4.3.1 Data Gathering, Cleaning and Preparation: . . . 23
4.3.2 Handling missing values and duplicate values: . . . 23
4.3.3 Removing unnecessary columns: . . . 24
4.3.4 Deriving the attributes related to Cancer_Type and Text_Description: . . . 24
4.3.5 Removing Outliers: . . . 25
4.3.6 Text Preprocessing and Cleaning: . . . 26
4.4 Data Visualization: . . . 27
4.5 Splitting of the dataset: . . . 32
4.6 Machine learning Models: . . . 33
4.7 Naive Bayes Classifier with a Pipeline for Efficient Workflow: . . . 37
4.8 Evaluation of Classification Model Performance Using Confusion Matrix and Classification Report: . . . 38
4.9 Text Classification Function for Predicting Cancer Types Using a Pre-trained Model: . . . 39
4.10 Building and Evaluating an Long Short-Term Memory(LSTM) Based Text Classification Model for Cancer Type Prediction: . . . 39
4.11 Text Classification Using BERT for Cancer Type Prediction: . . . 40

5 Results **41**
5.1 Machine learning models: . . . 41
5.2 Navies Bayes model: . . . 43
5.3 Evaluation of Classification Model Performance Using Confusion Matrix and Classification Report: . . . 43
5.4 Text Classification Function for Predicting Cancer Types Using a Pre-trained Model: . . . 44
5.5 Building and Evaluating an LSTM(Long short-term memory (LSTM))-Based Text Classification Model for Cancer Type Prediction(Neural Network): . . . 45
5.6 Text Classification Using BERT for Cancer Type Prediction: . . . 46

6 Conclusion: **48**
6.1 Future Scope: . . . 49

List of Figures

2.1 A Diagram for the causes of Lung cancer [31] 8
2.2 A Diagram for Thyroid cancer[32] 9
2.3 Identification of Colon cancer using text symptoms[33] 10

4.1 Steps followed for methodology. 21
4.2 Cancer Data 22
4.3 Shape of the given dataset [35] 23
4.4 Missing values of the given dataset [35]. 23
4.5 Duplicate values of the given dataset [35]. 24
4.6 Column name changed of the given dataset [35]. 24
4.7 Attributes that shows Cancer_Type and Text_Description [35]. 25
4.8 Different types of cancer 27
4.9 Distribution of the cancer types 28
4.10 Kernal Distribution of the number of the words 29
4.11 Distribution of Text Lengths 30
4.12 Distribution of word count 30
4.13 Word Cloud Visualization of Text Descriptions for Thyroid Cancer 31
4.14 Word Cloud Visualization of Text Descriptions for Colon Cancer 31
4.15 Word Cloud Visualization of Text Descriptions for Lung Cancer 32
4.16 Logistic Regression[48] 34
4.17 Random Forest[49] 35
4.18 Support Vector Machine(SVM)[50] 36

5.1 Visualization of Machine learing models 42
5.2 Visualization of Neural network models 45

List of Tables

4.1 Comparison of Outliers 25

5.1 Evaluation of Machine learning models. 41
5.2 Evaluation of Navies Bayes model. 43
5.3 Classification metrics for the models 44
5.4 Confusion Matrix diagram. 44
5.5 Long short-term memory (LSTM) - Neural Network model. 45
5.6 Evaluation of BERT Transformer. 46

CHAPTER 1

Abstract

Early and precise diagnosis is essential for both effective cancer therapy and better patient outcomes. In this dissertation, the application of Natural Language Processing (NLP) and Machine learning to the prediction of cancer kinds from symptom descriptions in unstructured text data is investigated. A thorough pipeline for data cleaning and preparation was put into place using a dataset of text-based cancer descriptions and symptoms. A number of classification models were assessed, such as Support Vector Machine(SVM), Random Forest, Decision Tree, Logistic Regression, Long short-term memory (LSTM) neural networks, and The Bidirectional Encoder Representations from Transformers (BERT).

The Bidirectional Encoder Representations from Transformers (BERT) model performed better than the others, demonstrating its capacity to grasp the subtleties of medical terminology. Techniques for data visualization were used to find trends that might help the models anticipate the future. This work adds to the expanding body of research on the application of Natural Language Processing (NLP) in cancer type prediction. The paper ends with suggestions for more research, such as creating AI(Artificial Intelligence) systems that can be understood by humans and incorporating more datasets.

CHAPTER 2

Introduction

2.1 Introduction of Lung, Thyroid, and Colon Cancer

In today's world, lung cancer causes more deaths from cancer-related causes than any other cancer variations like breast, prostate, and colon cancers. Early lung cancer diagnosis and treatments have justifiable consideration and positive impact on patient outcomes and mortality rates. The disease was frequently discovering at its advanced stages in many lung cancer cases since existing screening techniques are not always reliable. For the early diagnosis of lung cancer, machine-learning techniques present a viable approach [1]. Machine learning models may reliably forecast a patient's likelihood of acquiring lung cancer by identifying patterns and risk factors linked to the disease by evaluating vast databases of patient records. Using medical imaging data from CT scans and X-rays, machine-learning models have been created in recent years to help with the early identification of lung cancer. However, these techniques can be expensive and time-consuming. In this report, we explore the application of machine learning techniques to estimate the risk of lung cancer using text datasets. The model specifically looks at text description that include patient data such as demographics, clinical notes, and medical history.

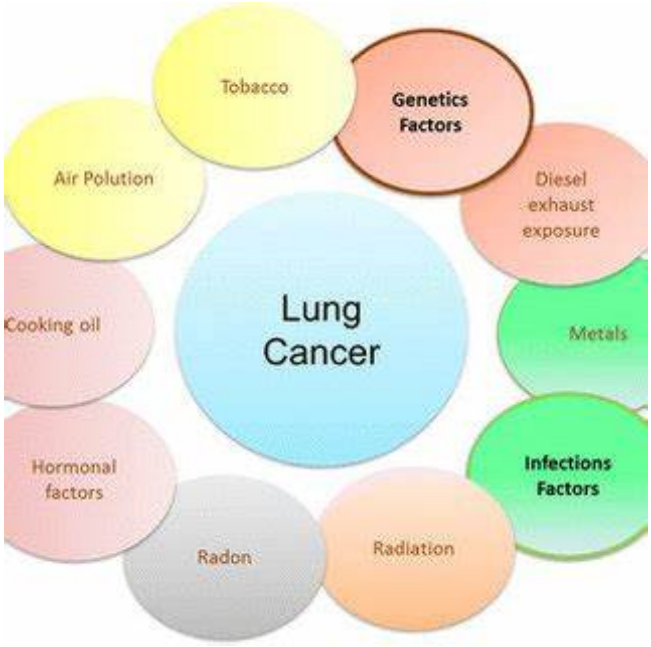

Figure 2.1: A Diagram for the causes of Lung cancer [31]

Figure 2.1 shows the causes of lung cancer, frequently brought in the human body by various life style habitats like smoking. Both smokers and the people who have been exposed to secondhand smoke also. A few wounds might happen in non-smokers, for example, individuals presented to radon gas, openness to cancer-causing agents, and a family background of cellular breakdown in the lungs.

In the United States, an estimated 20 million Americans have thyroid dysfunction [2]. The symptoms of hypothyroidism and hyperthyroidism are myriad and nonspecific, particularly in older individuals and thus testing is required to distinguish thyroid disease from other etiologies [3]. Prominent symptoms reported by patients include fatigue and weight change, with weight loss from thyroid overactivity and weight gain from underactivity [4, 5]. In order to understand the relationship between symptoms of fatigue and weight change in the outpatient setting and ordering patterns for thyroid testing in academic medical centers, we must be able to reliably identify documented symptoms in the text description. However, symptoms are often under-coded using billing codes and documented in the clinical free-texts.

Naive Bayes and Decision Tree, as illustrated in Figure 2.2, are two machine learning approaches that are used to train and evaluate the thyroid dataset in order to address the problem of thyroid cancer. These models will identify the thyroid cancer symptoms and the patient's stage of illness. There are three stages of thyroid cancer: minor, major, and critical.

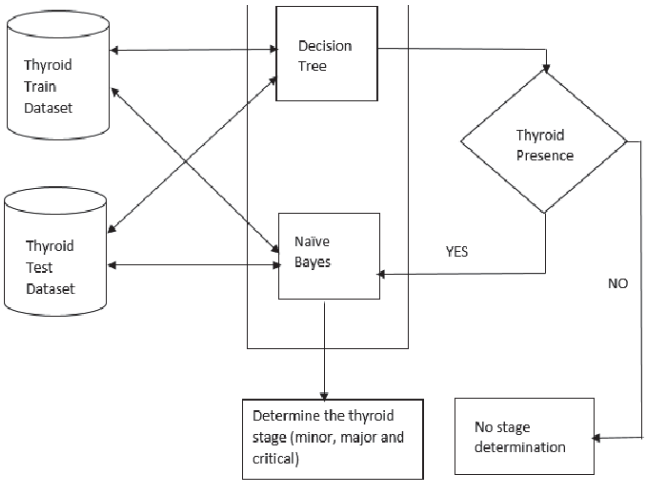

Figure 2.2: A Diagram for Thyroid cancer[32]

Colon cancer is the third most common malignancy worldwide (10 percent) and the second most common cause of cancer-related deaths (9.4 percent). With the ongoing research on molecular mechanisms of cancer and the joint development of various omics studies, an increasing number of treatment options are now available for local lesions and advanced diseases, thereby improving individualized diagnosis, treatment, and precision medicine. Natural language processing (NLP) is an essential branch of AI technology that aims to convert natural language into a commutable digital form to achieve text-level understanding and calculation. In the medical domain, the mainstream of NLP focuses on extracting clinically meaningful entities or classifying subgroups using Electronic Health Record (EHR), radiology reports, or drug instructions. Moreover, studies utilizing deep learning from free text to predict patient outcomes deserve further attention. Predictive models trained by machine learning may allow more personalized predictions by using many features of a patient's particular characteristics and disease and have been shown to outperform the prediction of treating oncologists. Some models developed to date utilize structured data that is, data that are processed into specific features such as the presence of genetic markers, demographics or specific aspects of clinical history. This may limit the widespread use of such models, as data availability varies among cancer treatment centers and between patients. It also limits what data can be used for a model, as not all clinical data are easily coded or categorized for extraction and analysis. The use of unstructured data such as text within medical documents, may address these drawbacks. Almost all patients receiving treatment for cancer have an initial consultation document from their oncologist. Such documents

generally have many details relevant to survival. For example, tobacco use or marital status, even if the clinic does not routinely store such data in structured data sets. The use of machine learning with documents, known as natural language processing (NLP), has increasingly been applied throughout medicine.

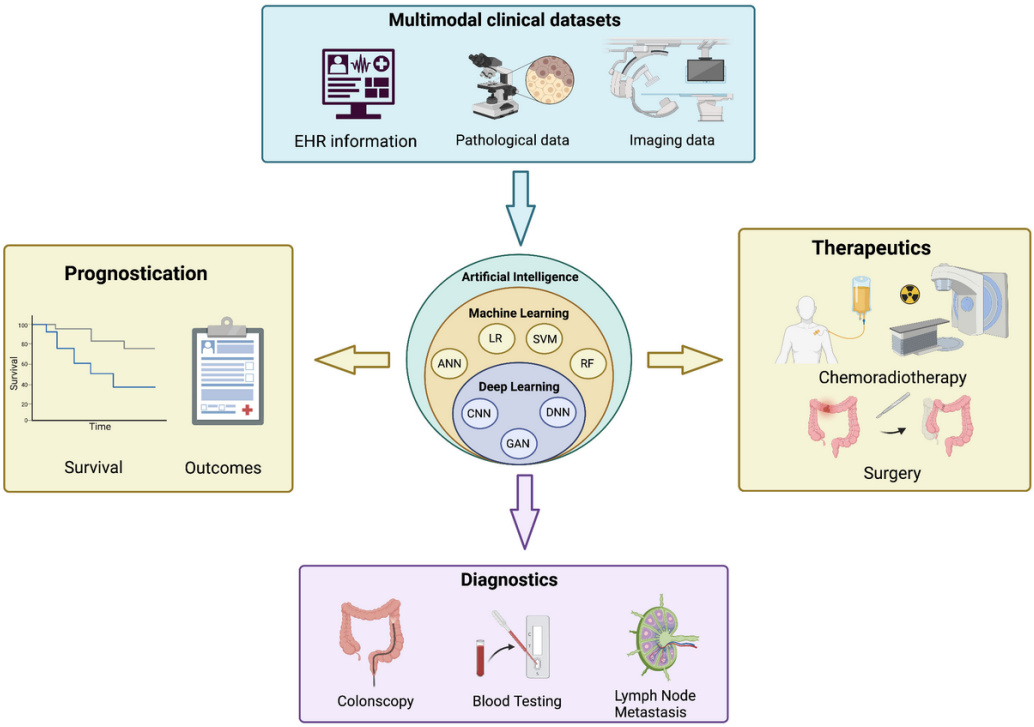

Figure 2.3: Identification of Colon cancer using text symptoms[33]

The method to use text analysis through NLP to predict colon cancer symptoms is shown in Figure 2.3. Artificial Intelligence is used in the Electronic Health Record (EHR), Pathological data, and Imaging data. Machine learning techniques including Logistics Regression, Decison Tree, Support Vector Machine (SVM), and Random Forest (RM) are used in prediction of Prognostication and Therapeutics. The diagnostics procedure is identified by deep learning. Many of the applications both in medicine generally and cancer specifically have utilized smaller, specific documents such as radiology or pathology reports. Some studies suggests that used patient encounter documents to predict the onset of non cancer illnesses. They found that models using the unstructured text data in these documents outperformed using only structured data and that adding structured data like demographics and laboratory data increased performance by only a marginal amount. Within cancer, recent work predicted survival in patients with lung cancer by extracting structured data from unstructured document data and used non neural NLP on progress notes to update an individual's prognosis.

2.2 What is Artificial Intelligence and Why to use it ?

Artificial intelligence (AI) aims to achieve human-like intelligence through entities (eg, machines), capable of processing and performing actions to human behavior[6]. In addition, machine learning stands as a rapidly evolving field of computer science that seeks to train machines using data sets to perform time-consuming tasks that typically require human cognitive abilities. Its potential to solve real-world challenges across various domains, including health care, has led to increased research into its uses[7]. By integrating different AI modalities into the decision-making process of physicians, this technology has the potential to enhance the field of medicine by improving the accuracy of diagnostic tests, streamlining provider workflow, enabling better disease and therapeutic monitoring, and resulting in better patient outcomes[8, 9]. By using AI knowledge we come to know how cancer is predicted and how to reduce it. There are many methods which can solve the issue of cancer.

2.3 How to predict this Cancer Types Text by its symptom ?

Natural language processing (NLP) and machine learning techniques have been utilized in various studies to predict symptoms in cancer patients based on text analysis. Studies have shown that NLP models like BERT can effectively predict common symptoms such as nausea, vomiting, fatigue, and mental health disorders in cancer patients by analyzing clinical notes. This includes many matrices methods like (Confusion Matrix, Accuracy, Precision, Recall, F1 Score), and Algorithm methods such as (Logistic Regression, Linear Regression, Decision Tree, Random Forest, Naive Bayes, Long short-term memory (LSTM) , Transformer) to discover genes from the samples and using these expression signatures to develop cancer prediction model. These models have demonstrated high accuracy rates in identifying symptoms related to cancer, aiding in early diagnosis and personalized treatment plans. By extracting and analyzing data from patients' medical reports, including symptoms documented around the time of cancer diagnosis, NLP models can assist in predicting future mental health disorders and other symptoms, ultimately improving patient care and outcomes in oncology settings.

CHAPTER 3

Literature Review

Any dissertation, including those on text analysis of symptoms to forecast cancer kinds such as thyroid, colon, and lung cancer, must include a comprehensive examination of the literature. On a given issue, it entails compiling, evaluating, and summarizing the body of research and knowledge already in existence. A literature review is meant to help you grasp what has already been researched, point out areas where information is still lacking, and lay the groundwork for future studies. The literature review in the dissertation aided in the investigation of earlier research on cancer prediction, particularly that which made use of text analysis or other pertinent methods. I may determine how academics have addressed the problem in the past, what strategies have worked, and where new research opportunities exist by going over this literature. This procedure shows how my study benefits the larger scientific community in addition to serving as a guide for my work.

Natural Language Processing (NLP) is often used to analyze symptoms described in patient records, converting unstructured text into structured data that can be used for prediction. Machine Learning Algorithm Techniques like Support Vector Machines (SVM), Decision Trees, Random Forest, Linear regression, Logistic regression, Neural Networks, Artificial Neuron Network (ANN) and Transformer are employed to classify and predict cancer types based on symptom data.

This study describes how to create a good model that accurately and efficiently predicts cancer kinds based on symptoms. It would also be preferable in my opinion if physicians and medical specialists handled cancer patients.

3.1 Symptom-Based Lung Cancer Prediction Using Machine Learning Techniques

Authors from academic institutions, not a particular corporation, released a research titled "Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, former smokers, and current smokers[10]." The authors of the study collaborated across several universities to examine the prediction power of symptoms from an adaptive e-questionnaire for lung cancer in three separate smoking groups: never smokers, former smokers, and current smokers[11].

The current study set out to determine if symptoms recorded on an adaptable electronic questionnaire could predict the development of lung cancer in smokers who had never smoked, were smokers in the past, or were smokers currently. The goal of the project was to create risk assessment models that might be converted into clinical instruments to help medical professionals determine a patient's likelihood of developing lung cancer. The study's conclusions emphasized the use of instruments for estimating the risk of lung cancer, particularly in nonsmokers who frequently receive a late-stage diagnosis[15].

The methods used in this article is Stochastic Gradient Boosting (SGB) is a machine learning technique that builds an ensemble of decision trees in a sequential manner to improve predictive accuracy[10]. The study utilized SGB models for predicting lung cancer probability based on symptoms reported in the e-questionnaire [10]. The SGB models employed a Bernoulli's loss function to fit the trees for prediction[10]. A training-test approach was applied, where 70 percent of observations were used for training the SGB model, and the remaining 30 percent for testing its performance[10, 11].

The SGB models performed well for never smokers and current smokers, with AUC values of 0.735 and 0.822, respectively, and corresponding overall accuracies of 0.815 and 0.771[10]. However, the performance was considerably worse for former smokers, with an AUC of 0.604[10]. Sensitivity was high for former and current smokers, with values of 0.816 and 0.829, respectively, while the sensitivity for never smokers was lower at 0.700[10]. Specificity varied among the groups, with never smokers having a high specificity of 0.882 and former smokers exhibiting a very low specificity of 0.333[10]. In summary, while SGB models showed promising results in predicting lung cancer status based on smoking status, Random Forest,

SVM, and Linear Regression each have their strengths in different scenarios, depending on the nature of the data and the specific predictive task at hand.

3.2 Artificial Intelligence for Predicting Cancer Symptoms: Enhancing Early Diagnosis

The research paper employs various supervised learning methods for early cancer diagnosis, including traditional statistical models like logistic regression (LR) and novel decision tree and deep learning (DL) algorithms[12]. Natural language processing (NLP) is utilized to transform unstructured clinical data into a computer-analysable format, enabling automation of resource-intensive tasks in cancer diagnosis[12]. Machine learning techniques such as linear models, support vector machines, decision trees, and deep learning models are applied to liquid biopsy material for cancer detection, with notable examples like CancerSEEK, which uses a random forest model to evaluate proteins and gene positions[13]. The paper discusses the development of a modified random forest algorithm for diagnosing hepatocellular carcinoma using whole-genome data, achieving a maximum validation AUC of 0.920, and a DL model for lung cancer detection based on Raman spectroscopy of liquid biopsy blood exosomes with an AUC of 0.912[13].

The study demonstrated that machine learning models, such as random forest algorithms, can effectively detect common cancer types through the analysis of cell-free DNA, showcasing promising results in early cancer diagnosis[13]. Support Vector Machines (SVM) were utilized in the analysis of liquid biopsy material, showing excellent AUCs for cancer detection, emphasizing the versatility of SVM in early cancer diagnosis [13]. Decision trees were employed in the analysis of liquid biopsy material, demonstrating their effectiveness in analyzing complex data and achieving high AUCs for cancer detection, showcasing their role in early cancer diagnosis[13]. Linear models were also used in the analysis of liquid biopsy material, in combination with other machine learning techniques, further contributing to accurate cancer detection, highlighting the importance of linear models in early cancer diagnosis[13]. A modified random forest algorithm was developed to diagnose hepatocellular carcinoma using whole-genome data, achieving a maximum validation AUC of 0.920, indicating the adaptability of traditional machine learning algorithms in diagnosing specific cancer types with high accuracy[13].

In conclusion, it can be observed that machine learning models that produce ideal AUC scores include Random Forest, Support Vector Machine (SVM), and Decision Trees. These methods provide accurate cancer predictions, which is advantageous for both patients and medical professionals.

3.3 Optimizing Disease Classification: Leveraging Language Models for Symptom Analysis

The literature highlights the significant impact of deep learning models, particularly BERT (Bidirectional Encoder Representations from Transformers), in enhancing the detection of adverse drug reactions (ADRs) and optimizing medication outcomes. These models have shown improved performance in various biomedical tasks, including drug-drug interaction (DDI) extraction and clinical text analysis[36]. Several studies have utilized NLP techniques to extract relevant information from unstructured clinical data. For instance, the use of models like BioBERT and ClinicalBERT has demonstrated superior performance in biomedical text mining tasks, indicating the effectiveness of pre-training on domain-specific corpora[36]. The studies reviewed emphasize the practical applications of these models in clinical settings, such as improving early disease detection and facilitating remote diagnosis. The integration of user-generated data from platforms like social media further enriches the symptom-disease relationship, enhancing the models' predictive capabilities[37, 38]. In summary, the literature survey underscores the transformative potential of language models and deep learning in healthcare, particularly in automating disease prediction from symptoms, while also acknowledging the challenges and ongoing research in this field.

3.4 Enhancing Thyroid Cancer Diagnosis through NLP-Driven Symptom Analysis

The following literature review focuses on studies that have utilized Natural Language Processing (NLP) for symptoms prediction in thyroid cancer. NLP has been extensively used in medical text analysis to extract meaningful information from unstructured data, such as electronic health records (EHRs), clinical notes, and research papers. It involves processes

like tokenization, named entity recognition (NER), sentiment analysis, and topic modeling to understand and interpret medical language.

Named Entity Recognition (NER) - an NLP technique to identify key information such as diseases, symptoms and medications from text into predefined categories. A study by Wei and colleagues, NER for symptoms and medical conditions extracted from clinical texts was demonstrated significantly improved the quality of healthcare data analysis: by Owen (2016)[16]. Symptoms were also identified and extracted using text mining techniques of diverse medical documents. Author deployed text mining on EHRs to anticipate disease emergence from symptom patterns in the patient notes. Our review has shown that text mining can be a valuable tool for early detection of disease, such as cancer[17].

Studies have shown the efficacy of using EHR data for predicting thyroid cancer symptoms. A notable example is the work by Hong et al. (2018)[18], which applied machine learning algorithms on EHR data to predict the likelihood of thyroid cancer based on symptom patterns. This study integrated NLP techniques to preprocess and analyze the unstructured text data from patient records. Clinical notes are rich in detailed patient information, often containing descriptions of symptoms and diagnostic findings. Miller et al. (2020)[19] explored the use of NLP to analyze clinical notes for symptom extraction related to thyroid cancer. Their research indicated that NLP could effectively identify subtle symptom patterns that precede a thyroid cancer diagnosis, thus aiding in early detection . Mining research articles for symptom prediction is another promising approach. Liu and Chen (2017)[20] developed an NLP-based framework to analyze and extract symptom-related information from a vast corpus of thyroid cancer research articles. This approach not only helped in identifying common symptoms but also uncovered less obvious symptom associations that could be critical for early diagnosis . Predictive models, such as logistic regression, support vector machines (SVM), and neural networks, have been employed in conjunction with NLP to predict thyroid cancer symptoms. Chen et al. (2021)[21] demonstrated that integrating NLP-extracted features with ML models significantly improved the accuracy of thyroid cancer symptom prediction. Recent advancements in deep learning have further refined symptom prediction. Reddy et al. (2022)[22] utilized deep learning models, such as recurrent neural networks (RNNs) and transformers, to analyze sequences of symptoms from clinical texts. Their study showed that deep learning models could capture complex symptom patterns over time, providing more accurate predictions for thyroid cancer.

Despite the encouraging outcomes, there are still a number of obstacles to overcome in the use of NLP for predicting the symptoms of thyroid cancer. These include the integration of multi modal data sources, the complexity of medical language, and the requirement for big, annotated datasets. Subsequent investigations have to concentrate on refining NLP algorithms, augmenting data interoperability, and tackling ethical issues related to data utilization. Enhancing early diagnosis and patient outcomes has been demonstrated by the integration of natural language processing (NLP) and text analysis in the prediction of thyroid cancer symptoms. The predictive power and general efficacy of thyroid cancer detection and therapy are anticipated to increase with further developments in natural language processing (NLP) techniques and their application in medical informatics.

3.5 Natural Language Processing in Clinical Notes: A Systematic Review of Colon Cancer Concept Extraction

The aim of the systematic review in the paper was to focus on extracting cancer concepts, specifically related to colon cancer, using NLP techniques. This study aimed to identify the NLP methods applied to automatically detect colon cancer concepts in clinical notes, the terminologies used for coding these concepts, and the types of colon cancers identified [10]. One of the articles referenced in the paper examined the extraction of cancer concepts from pathology reports, including those related to colon cancer[25]. While the number of studies related to implementation was relatively low in the broader context of relevant articles, a study by Chang et al. demonstrated the increasing attention given to the classification and coding of free-text clinical notes and radiology reports in recent years[24, 25]. The paper discussed the application of machine learning algorithms, including deep learning and traditional ML methods, in extracting cancer concepts from clinical texts, showcasing the evolving technological landscape in healthcare data analysis[25].

Questions solved by authores in this paper:

1. What are the terminological systems and data standards used in the included articles ?
2. What are the advantages of using NLP in cancer concept extraction from clinical notes ?
3. What terms related to symptoms can be extracted using NLP in clinical notes ?

Initially, 6708 papers were retrieved for review, out of which 2503 articles remained after removing duplicates. Following the screening process, 67 studies were deemed relevant, and after applying exclusion criteria, 17 articles were selected for detailed review[26]. Ultimately, 17 articles were chosen for detailed analysis, and information regarding the extraction of cancer concepts from clinical notes using NLP was extracted from these selected studies[23]. The study revealed that UMLS and SNOMED-CT were the most commonly used terminologies in the field of NLP for extracting cancer concepts. UMLS was particularly prevalent, with about 78% of the articles falling under the NLP category[24]. Despite the dominance of rule-based methods in extracting cancer concepts, the review emphasized the need for future studies to explore the application of machine learning and deep learning algorithms in information extraction from clinical texts[10]. The paper identified key applications of NLP algorithms in analyzing clinical reports, including disease information and classification, language discovery, knowledge structure, quality assessment, compliance monitoring, and cohort studies[10]. Previous systematic reviews primarily focused on extracting concepts from various clinical texts such as radiology reports, laboratory findings, and postoperative surgical outcomes. The application of NLP in cancer-related studies showcased the potential for improved identification of postoperative complications using NLP models compared to traditional methods[10]. The results underscored the importance of robust data requirements and careful model development when applying machine learning algorithms to extract cancer concepts, highlighting the evolving technological landscape in healthcare data analysis[10].

3.6 Identifying and Predicting Cancer Patient Events Using Free Text in EHRs: A Trajectory Analysis

The study focuses on analyzing free text in electronic health records (EHRs) to identify cancer patient trajectories[27]. The research aims to develop a methodology to estimate disease trajectories of cancer patients from unstructured EHR text documents[27]. The main objective is to predict patient events ahead in time using disease trajectories extracted from free text in EHRs[27]. The study aims to predict 80% of patient events ahead in time by utilizing these disease trajectories extracted from unstructured EHR text documents[27].The study analyzed a cohort of 7,741 patients, including 4,080 diagnosed with cancer, treated at a University Hospital between 2004-2012[27].

A total of 1,133,223 unstructured EHR text documents were examined, which included admission reports and intensive care unit reports for the patient cohort[28]. The methodology developed allowed for the estimation of disease trajectories of cancer patients from the free text in EHRs, enabling the prediction of 80% of patient events ahead in time[27]. They quantified risks associated with moving between events, symptoms, and diseases, emphasizing the importance of controlling con founders to reveal novel pathology and improve clinical outcomes[28].

The study recognized the challenges posed by unstructured variation in EHR text data entry methods, stressing the need for automated tools to process and analyze free text effectively for better healthcare outcomes[30]. By studying patient trajectories and risk markers, the authors aimed to decrease adverse events, optimize cancer treatment, and enhance clinical decision support, contributing to evidence-based medical practices and personalized healthcare interventions [30].

3.7 Application of Naive Bayesian Networks in Predicting Diseases: A Systematic Review of 2005-2015 Studies

The paper aims to systematically review the application of Naive Bayesian Networks (NBNs) in predicting various diseases, highlighting their effectiveness compared to other predictive algorithms[41]. A literature review was conducted using the PubMed database on July 26, 2015, focusing on studies published between 2005 and 2015. This approach was chosen to ensure the relevance and quality of the articles reviewed[42].

The search yielded a total of 99 articles, which were filtered down to 23 that met the eligibility criteria[43, 41]. The first author conducted data extraction and quality assessment, which was then verified by a coauthor. This process ensured the accuracy of the information collected from the selected articles [45]. The extracted data included article properties such as title, publication year, subject, illnesses, and the number of variables used in the studies[46].The review found that NBNs are one of the simplest yet most effective algorithms for disease prediction. The paper discusses how NBNs compare favorably against other methods like logistic regression, support vector machines, and decision trees in terms of accuracy and performance[47]. The review particularly highlights the application of NBNs in predicting brain diseases and cancers, such as breast, colon and prostate cancer. It notes that a significant

portion of the studies focused on these areas, demonstrating the versatility and effectiveness of NBNs in various medical contexts[47].

The systematic review concludes that NBNs are a valuable tool in the field of disease prediction, with the potential to improve diagnostic accuracy and patient outcomes. The findings suggest that NBNs can be effectively utilized in clinical settings, especially in cases where traditional diagnostic methods may be limited[47, 41].

3.8 Advancements in Multi-Disease Prediction: Leveraging LSTM Networks for Enhanced Clinical Decision Support

The primary aim of the paper is to develop a deep learning approach for multi-disease prediction using longitudinal electronic health record (EHR) data. This is crucial for early disease intervention and preventive care, addressing challenges such as temporal irregularity and interdependence between diseases[51].

The paper utilizes a Long Short-Term Memory (LSTM) network, which is a variant of Recurrent Neural Networks (RNNs). LSTM is particularly effective for handling sequential data due to its gated architecture, which allows it to learn long-term dependencies and manage the temporal irregularity of patient visits. This is crucial in healthcare, where patient data is often collected at irregular intervals[51, 52]. Time-aware Mechanism addresses the temporal irregularity in clinical visits, allowing the model to better understand the timing of patient data[53]. Attention-based Mechanism helps the model focus on the most relevant visits for predicting future diseases, improving the overall prediction accuracy[53].

The implementation of the LSTM model on a large-scale EHR dataset demonstrates significant performance improvements. The proposed model outperforms traditional statistical methods and other deep learning models, achieving an F1 score of 88.0%, which is higher than the scores of 78.9% and 86.4% from state-of-the-art conventional and deep learning models, respectively[53]. The study also highlights that the combination of the time-aware and attention-based mechanisms leads to better prediction performance[53].

The paper concludes that the proposed LSTM approach, with its time-aware and attention-based enhancements, significantly improves the prediction of future disease diagnoses from EHR data. The authors acknowledge limitations, such as the reliance on a single data sources[53].

CHAPTER 4

Methodology

4.1 Introduction

This study's methodology made use of the symptoms listed in the dataset's text description to predict the type of cancer. Here, I have used Python modules to perform data loading, data cleaning and preprocessing, and data visualization. Additionally, model testing and training have been completed for Deep Learning, Neural Networks, Transformers, Natural Language Processing (NLP), and Machine Learning techniques.

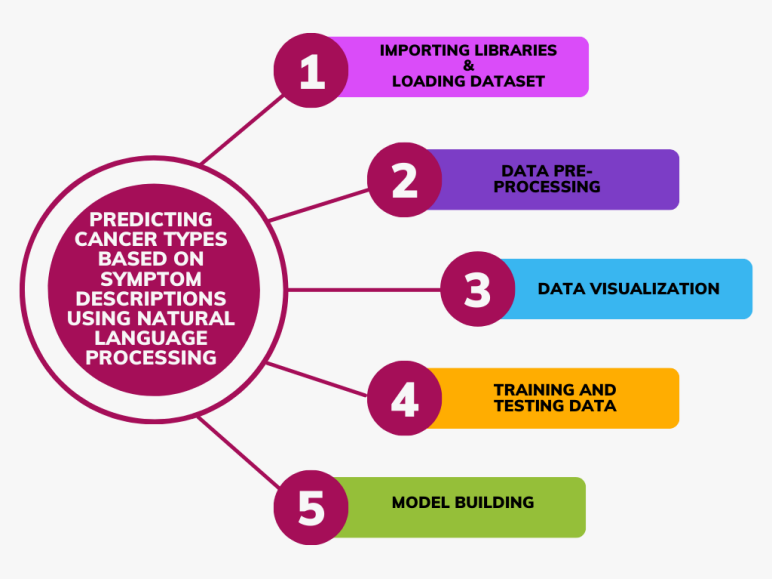

Figure 4.1: Steps followed for methodology.

To solve the problem, I have followed the procedures shown in Figure 4.1. To speed up the procedure, I first imported the required Python libraries. Second, I have performed data cleaning, outlier detection, and preprocessing to check for missing and duplicate values. In order to gain clear insights for this project, I then went on to data visaulization. In addition, I use the data to train and evaluate models. Each stage will be examined separately.

4.2 Data Description:

The dataset[35] utilized in this study was gathered from Kaggle, a publically accessible source. This dataset[35] includes cancer documents to be classified into three categories like 'Thyroid_Cancer', 'Colon_Cancer', 'Lung_Cancer'.

	Unnamed: 0	0	a
0	0	Thyroid_Cancer	Thyroid surgery in children in a single insti...
1	1	Thyroid_Cancer	" The adopted strategy was the same as that us...
2	2	Thyroid_Cancer	coronary arterybypass grafting thrombosis ï¬□b...
3	3	Thyroid_Cancer	Solitary plasmacytoma SP of the skull is an u...
4	4	Thyroid_Cancer	This study aimed to investigate serum matrix ...
...	...	...	...
7565	7565	Colon_Cancer	we report the case of a 24yearold man who pres...
7566	7566	Colon_Cancer	among synchronous colorectal cancers scrcs rep...
7567	7567	Colon_Cancer	the heterogeneity of cancer cells is generally...
7568	7568	Colon_Cancer	"adipogenesis is the process through which mes...
7569	7569	Colon_Cancer	the periparturient period is one of the most c...

7570 rows × 3 columns

Figure 4.2: Cancer Data

As seen in Figure 4.2, there are 7570 rows and three columns in the dataset. Currently, "Cancer_Types" is found in column "0," while "Text_Description" is found in column "a." Here, there are unnecessary columns and incorrect name presentation. I proceeded to the data preprocessing step and looked for missing, duplicate, and correctly spelled variables in order to achieve this.

Using Machine learning models and Natural Language Processing (NLP), I was able to predict the kind of cancer from this dataset by analyzing the text description of the symptoms.

4.3 Data Preprocessing and Exploration:

4.3.1 Data Gathering, Cleaning and Preparation:

This is the first step, in the dataset [35] is read and shapes of data set is as shown below Figure 4.3.

```
1 df.shape
```

```
(7570, 3)
```

Figure 4.3: Shape of the given dataset [35]

An essential first step in every machine learning process is data cleaning. This included tasks such dealing with outliers or anomalies, eliminating duplicate entries, addressing structural flaws, and resolving missing values. High quality data is ensured by proper data cleaning, and accurate machine learning model training depends on it. Poor quality data containing mistakes, duplication, missingness, and other problems can seriously impair model performance. In this study, meticulous data cleansing and extremely accurate attrition training are conducted.

4.3.2 Handling missing values and duplicate values:

As shown in Figure 4.4 , this dataset [35] do not contains missing values.

	0
Unnamed: 0	0
0	0
a	0

dtype: int64

Figure 4.4: Missing values of the given dataset [35].

So, there is no missing values but still there are jumbling words and no clear column names for the dataset, so checked duplicate values.

```
# Total number of duplicate rows
df.duplicated().sum()

0
```

Figure 4.5: Duplicate values of the given dataset [35].

As shown in figure 4.5, there are no duplicates values. So, I move further to check the attributes.

4.3.3 Removing unnecessary columns:

	Cancer_Type	Text_Description
0	Thyroid_Cancer	Thyroid surgery in children in a single insti...
1	Thyroid_Cancer	" The adopted strategy was the same as that us...
2	Thyroid_Cancer	coronary arterybypass grafting thrombosis ï¬□b...
3	Thyroid_Cancer	Solitary plasmacytoma SP of the skull is an u...
4	Thyroid_Cancer	This study aimed to investigate serum matrix ...
...	...	...
7565	Colon_Cancer	we report the case of a 24yearold man who pres...
7566	Colon_Cancer	among synchronous colorectal cancers scrcs rep...
7567	Colon_Cancer	the heterogeneity of cancer cells is generally...
7568	Colon_Cancer	"adipogenesis is the process through which mes...
7569	Colon_Cancer	the periparturient period is one of the most c...

7570 rows × 2 columns

Figure 4.6: Column name changed of the given dataset [35].

As seen in the above Figure 4.6, I changed the names of columns '0' to 'Cancer_Type' and 'a' to 'Text_Description' throughout this procedure in order to better comprehend the data. I also deleted any redundant columns. The data's form was altered to 7570 rows and 2 columns as a result.

4.3.4 Deriving the attributes related to Cancer_Type and Text_Description:

This process is crucial for data preparation. As shown in Figure 4.7, the dataset consists of 7,570 entries with two columns: Cancer_Type and Text_Description. Both columns contain text data, with no missing values. The "Cancer_Type" column likely contains different types

of cancers, serving as a categorical variable. The "Text_Description" column contains detailed descriptions that can be analyzed for patterns, keywords, or sentiments related to each cancer type. The dataset is lightweight (118.4 KB), making it easy to handle for further analysis or processing. These insights indicate that the data is ready for categorization, textual analysis, or further statistical exploration.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7570 entries, 0 to 7569
Data columns (total 2 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Cancer_Type       7570 non-null   object
 1   Text_Description  7570 non-null   object
dtypes: object(2)
memory usage: 118.4+ KB
```

Figure 4.7: Attributes that shows Cancer_Type and Text_Description [35].

4.3.5 Removing Outliers:

Table 4.1 below illustrates this with two columns: one has outliers with irrelevant terms, and the second has outliers that have been eliminated using regular expressions.

Outliers	**Removed Outliers**
grafting thrombosis &x81brin &x81brinogen	grafting thrombosis brin brinogen mutation

Table 4.1: Comparison of Outliers

The dataset included text descriptions with non-ASCII characters that showed as garbled text. The regex pattern r'[&x00-&x7F]+', which recognizes non-ASCII characters, was used to identify these characters as outliers. Corrupted words like "fibrin" and "fibrinogen" are now readable and accurately represented in the text data as a result of these characters being removed.

This preprocessing step is crucial to enhancing the quality of textual analysis because it eliminates noise that could skew the results of text mining algorithms or Natural Language Processing (NLP) methods. The text data has been made more consistent and dependable by treating these outliers, which will improve the precision of any subsequent analysis, including topic modeling, sentiment analysis, and keyword extraction. This phase makes sure that there are no artifacts in the dataset that could cause errors or misinterpretations of the results.

4.3.6 Text Preprocessing and Cleaning:

In the preprocessing phase of the text data, several key transformations were applied to ensure cleaner and more meaningful input for analysis. Unnecessary words and punctuation were removed to streamline the content, reducing noise that could obscure important information. Blank spaces were eliminated to unify and compact the text, enhancing readability and consistency. Lemmatization was employed to standardize words to their base or root forms, which helps in consolidating similar terms and improving the accuracy of subsequent analyses. Additionally, all text was converted to lowercase to maintain uniformity and avoid case-sensitive discrepancies.

These preprocessing steps, facilitated by regular expressions (regex), collectively enhance the quality of the data by normalizing variations and focusing on the core textual content, thereby optimizing it for further natural language processing and analysis tasks.

4.4 Data Visualization:

Data visualization is the representation of data through use of common graphics, such as charts, plots, info-graphics and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

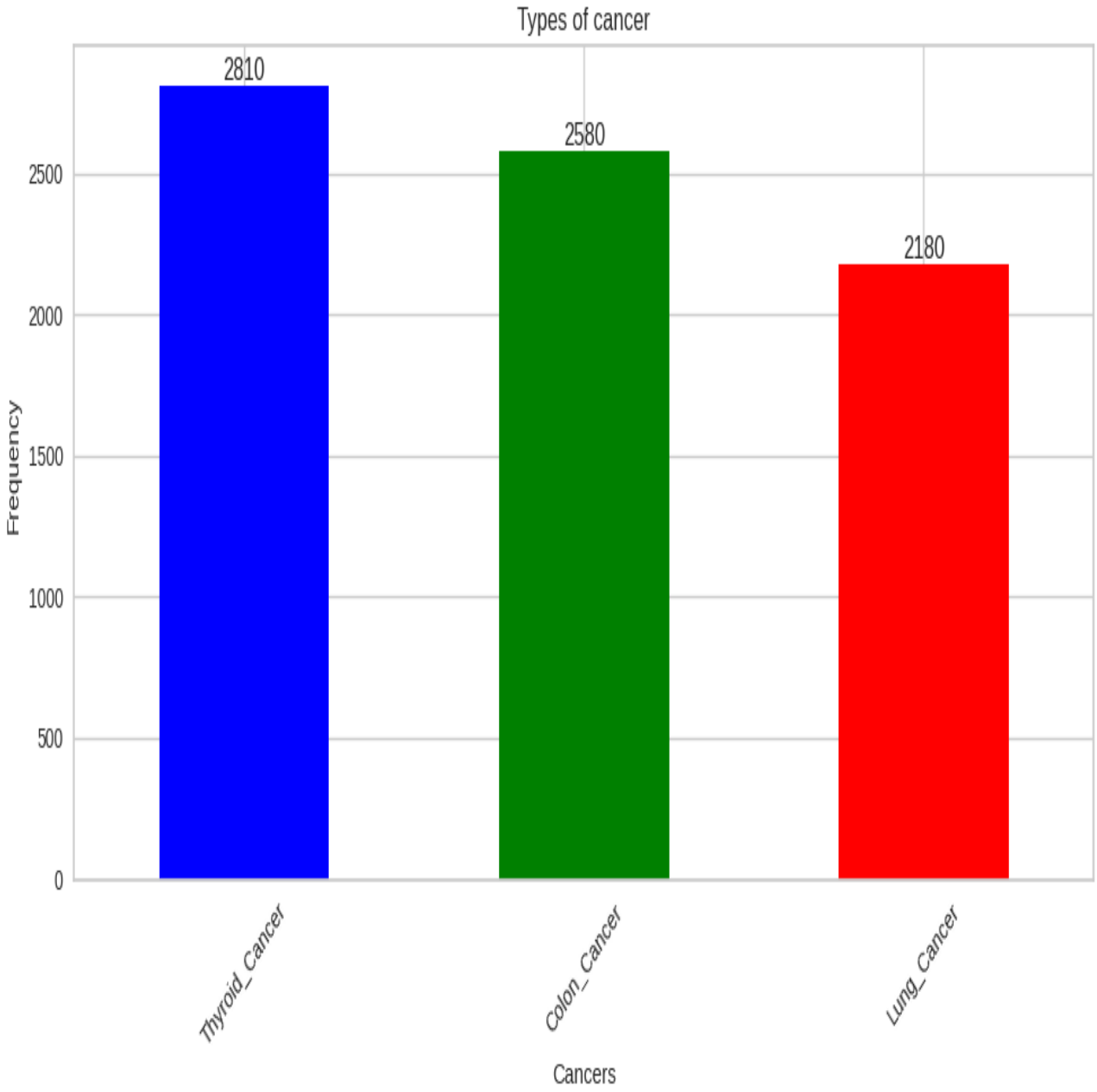

Figure 4.8: Different types of cancer

In Figure 4.8 we can see the bar plot showing the three types of cancer Thyroid_cancer,

Colon_cancer and Lung_cancer. Here, count of Thyroid_cancer, Colon_cancer and Lung_cancer are 2810, 2580 and 2180 respectively with respect to frequency. This allows us to find most common cancer type prediction which is more frequent and less frequent. This visualization helps in making the data more interpretable and supports data-driven decision-making into the underlying causes or implications of the distribution of cancer types.

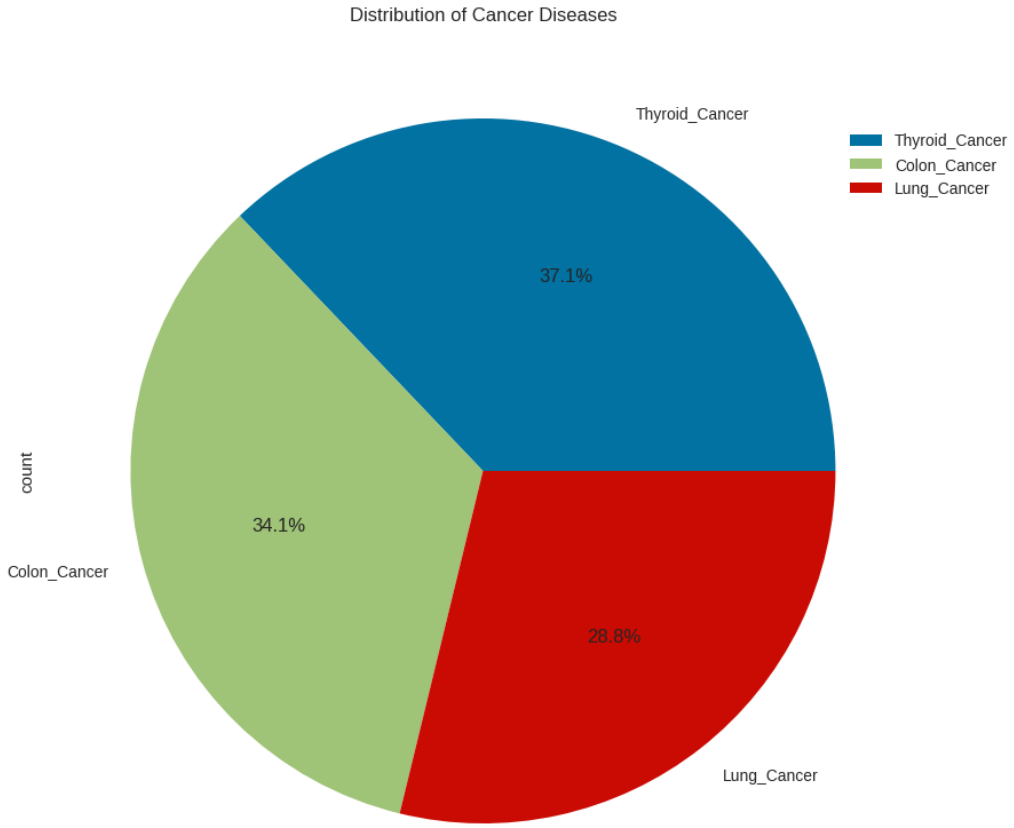

Figure 4.9: Distribution of the cancer types

Figure 4.9 shows the pie chart diagram in which there are distribution of types of the cancer using percentage form. The pie chart provides a clear visual representation of how the various cancer types are distributed in the dataset. Thyroid_cancer has 37.1% count which is higher amongst the other. Colon_cancer is about 34.1% and Lung_cancer has 28.8% count. The percentages displayed on the pie chart helps to understand the relative sizes of each category.This pie chart helps in summarizing the data and giving an immediate sense of the distribution of different cancer types within the dataset.

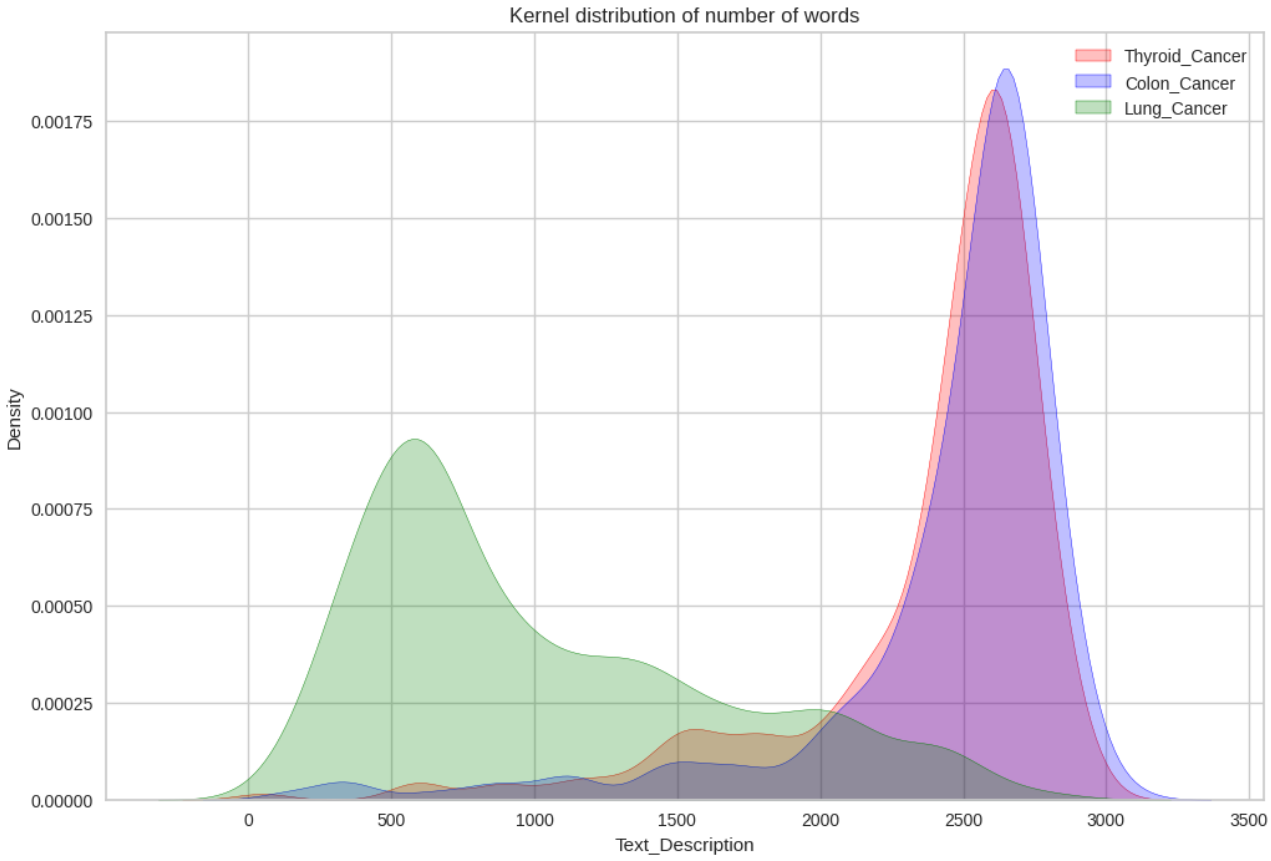

Figure 4.10: Kernal Distribution of the number of the words

Figure 4.10 illustrates a Kernel Density Estimate (KDE) plot to visualize the distribution of word counts in text descriptions for three different types of cancer: Thyroid_Cancer, Colon_Cancer, and Lung_Cancer. The KDE plot shows how the word counts (number of words) in text descriptions of each cancer type are distributed. It gives a smooth, continuous estimate of the density of data points (word counts).
Overall, the KDE plot is an effective way to visually explore and compare the distributions of text description lengths across different types of cancer, providing insights that can inform further analysis or decision-making.

The Figure 4.11, provides a histogram using 'Seaborn' python library to visualize the distribution of text lengths in the 'Text_Description' column of the DataFrame. Each bar represents a range of text lengths, and the height of the bar shows how many text descriptions fall within that range. A Kernel Density Estimate (KDE) curve is overlaid on the histogram, providing a smoothed version of the distribution, which helps to visualize the overall shape

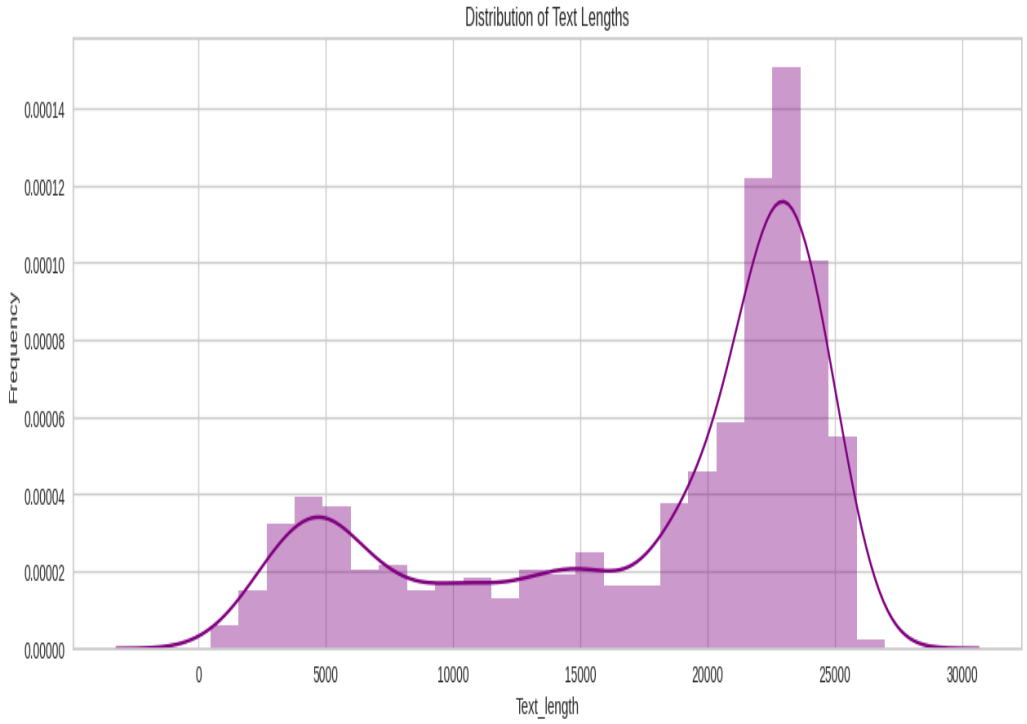

Figure 4.11: Distribution of Text Lengths

and density of the data. The x-axis (Text_length) represents the length of the text descriptions. The y-axis (Frequency) shows how often text descriptions of a certain length occur.

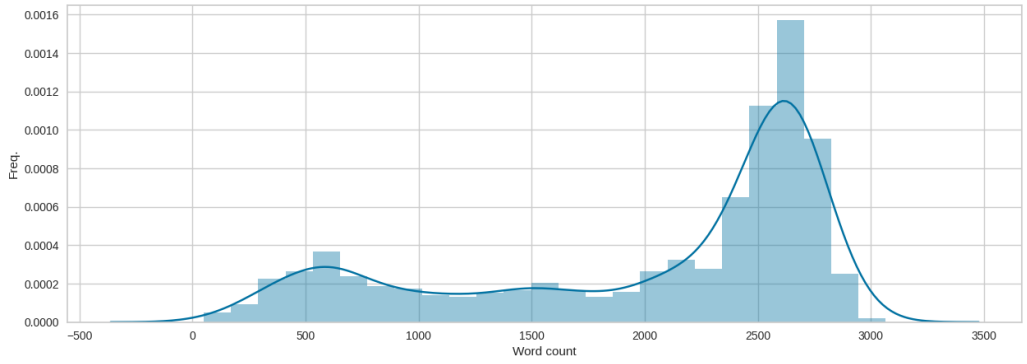

Figure 4.12: Distribution of word count

Figure 4.12, contains the number of words in each text description from the 'Text_Description' column. This is calculated by splitting each description into words and then counting them. A histogram is generated using Seaborn to display the distribution of word counts across the text descriptions. The histogram shows how frequently different word counts appear in the dataset. An overlayed Kernel Density Estimate (KDE) curve provides a smooth, continuous estimate of the distribution, helping to visualize the overall shape and trends in the data. The

x-axis is labeled "Word count" indicating the number of words in the text descriptions. The y-axis is labeled "Freq" which stands for frequency, indicating how many descriptions have a particular word count.

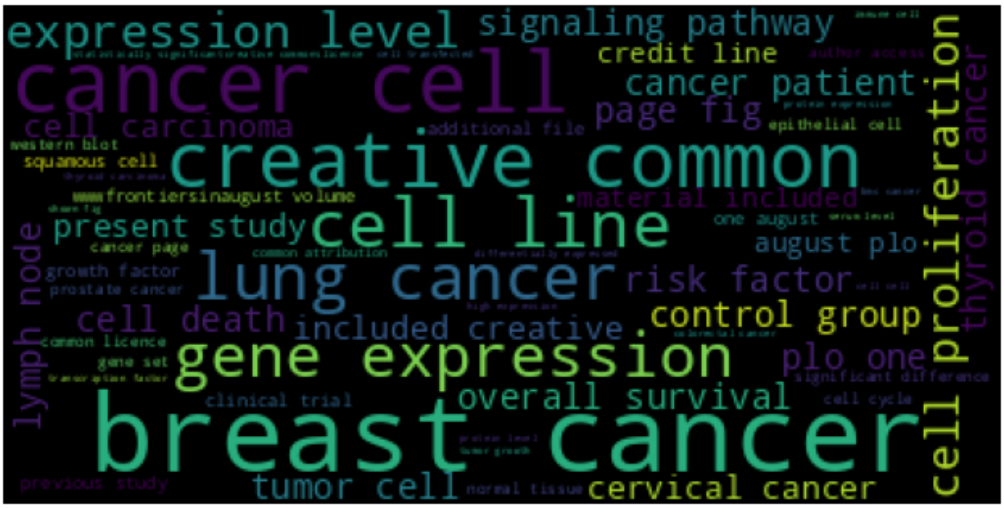

Figure 4.13: Word Cloud Visualization of Text Descriptions for Thyroid Cancer

The purpose of this word cloud is to visually highlight the most frequent words found in the text descriptions related to Thyroid_Cancer. Figure 4.13 shows the most frequent words which is related to the Thyroid_cancer. A word cloud is generated using the WordCloud library, which visualizes the most frequent words in the text descriptions. Common stopwords (like "the", "and", etc.) are removed using a predefined set (STOPWORDS) to focus the visualization.

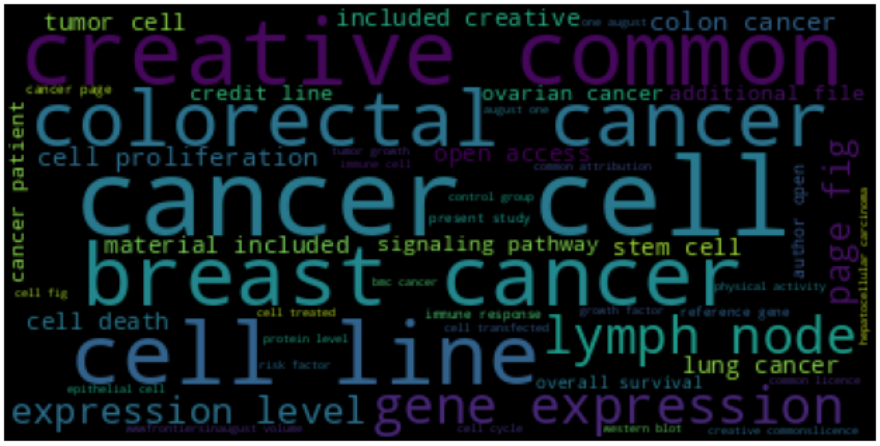

Figure 4.14: Word Cloud Visualization of Text Descriptions for Colon Cancer

As we can see in below Figure 4.14, the primary purpose of this word cloud is to visually

highlight the most frequently occurring words in the text descriptions associated with Colon Cancer. This word cloud provides a visual summary of the most common words in Colon Cancer descriptions, making it easier to understand the key topics and terms discussed in the text data.

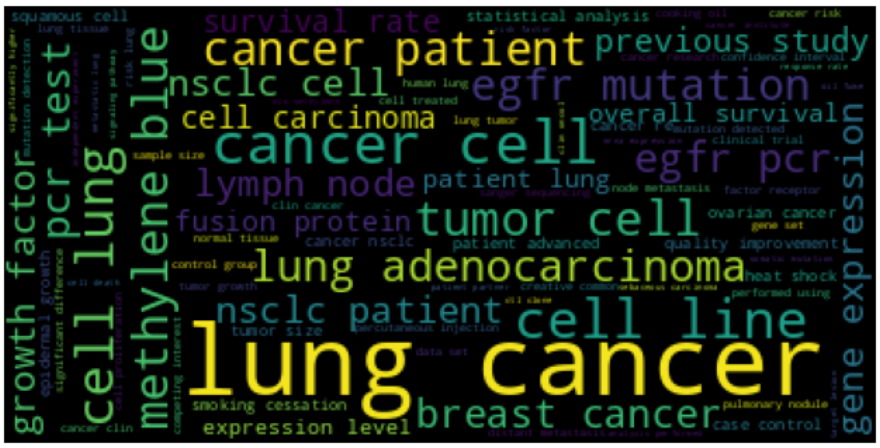

Figure 4.15: Word Cloud Visualization of Text Descriptions for Lung Cancer

Figure 4.15, visually emphasize the most frequently used words in the text descriptions related to Lung Cancer. Larger words represent those that appear more frequently, indicating their relevance and importance in the context of Lung Cancer.
By examining the word cloud, I gained the insights into the common themes, symptoms, treatments, or medical terms associated with Thyroid, Colon and Lung cancers as described in the dataset. The insights obtained from the word cloud can guide more focused research, such as identifying key areas for understanding patient concerns and treatment discussions specific to the Cancer.

4.5 Splitting of the dataset:

The splitting dataset is used to teach the machine learning model. The model employs the labeled data (input-output pairs) to identify trends, connections, and characteristics in the data. In this stage, the model modifies its internal parameters to reduce prediction errors on the training set. The testing dataset is used to evaluate the performance of the machine learning model. Here, the text descriptions ('X') and corresponding cancer types ('y') are split into training and testing sets. I used vectorization method to convert the text descriptions

into numerical features that the machine learning model can understand. It uses TF-IDF (Term Frequency-Inverse Document Frequency) to measure the importance of words in the text. The TF-IDF vectorization method combines two statistical measures: Term Frequency (TF) and Inverse Document Frequency (IDF).

$$\text{TF}(t, d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d} \tag{4.5.1}$$

In Equation (4.5.1), Term Frequency(TF) measures how frequently a term appears in a document. If a word appears many times in a document, its TF value will be higher.

$$\text{IDF}(t, D) = \log\left(\frac{\text{Total number of documents in the corpus}}{\text{Number of documents containing term } t}\right) \tag{4.5.2}$$

In equation (4.5.2), Inverse Document Frequency (IDF) measures how important a term is within the entire corpus. The vectorization method, specifically TF-IDF, is used here to transform the text data into a format that machine learning models can work with.

4.6 Machine learning Models:

In this method, I performed and compared different machine learning models-Linear Regression, Logistic Regression, Decision Tree, and Random Forest using text data that has been transformed into numerical features with the TF-IDF vectorization method[42]. After training, the accuracy of each model on the test data is calculated and stored[42].
Some formula has been used in this process is as follows:

1. Model Accuracy:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{4.6.1}$$

In this equation 4.6.1, accuracy_score from sklearn.metrics computes this value by comparing the model's predictions (y_predict) with the true labels (y_test).

2. Logistic Regression:

Logistic regression is a fundamental algorithm for binary classification problems. It models the probability of an observation belonging to a particular class using the logistic function. This algorithm estimates the coefficients for each feature to find the best-fitting S

shaped curve which is bounded between 0 and 1 [48], that separates the two classes. This allows the output of logistic regression as probability. It is interpretable and serves as a baseline model for many classification problems.

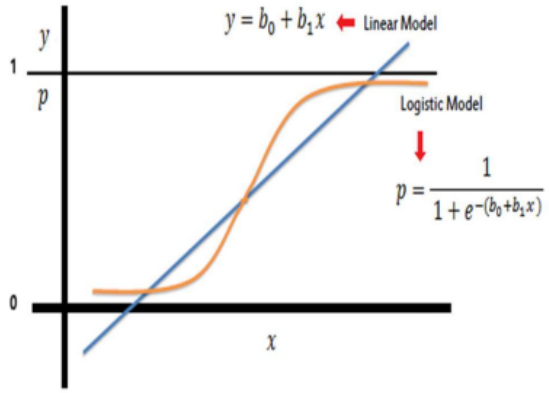

Figure 4.16: Logistic Regression[48]

As shown in the above figure 4.16, P denotes the probability and Y-axis represent the target variable (Attrition) and X - axis denotes the dependent variables. The straight line represents the linear regression by applying the logistic sigmoid function, it turns into a S-shaped curve.

Logistic regression is based on the Logistic Function (Sigmoid Function)[1]:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{4.6.2}$$

where

$$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

By leveraging the text descriptions of symptoms, which are often transformed into numerical features using techniques like TF-IDF (Term Frequency-Inverse Document Frequency), Logistic Regression can model the relationship between these features and the likelihood of each cancer type. Logistic Regression can model the relationship between these features and the likelihood of each cancer type. The algorithm outputs probabilities for each class, which are then used to classify the input into the most probable cancer type. Logistic Regression is particularly useful in this context due to its simplicity, interpretability, and ability to handle large feature spaces efficiently, making it a strong baseline model for text-based classification tasks in medical diagnostics.

3. Decision Tree:

Decision Trees use an algorithm based on Gini Impurity or Entropy (depending on the criterion) to split the nodes[1]:

Gini Impurity:

$$G = \sum_{i=1}^{C} p_i(1 - p_i) \tag{4.6.3}$$

where p_i is the probability of class i.

Entropy:

$$H = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{4.6.4}$$

In this study, cancer kinds are predicted using Decision Tree algorithms based on text data symptoms[48]. A strong and understandable machine learning technique that can handle both numerical and categorical information is decision trees. Using methods like word embeddings or TF-IDF, the text is first converted into structured numerical features in the context of text-based symptom descriptions[48]. After splitting the data iteratively according to these criteria, the Decision Tree method produces a tree-like model in which each node denotes a decision rule depending on whether a certain symptom is present or absent. This procedure carries on until the algorithm determines a final outcome that corresponds to a particular type of cancer[48].

4. Random Forest : In this project, Random Forest is utilized as a robust machine learning algorithm to predict cancer types based on symptoms described in text data. Random Forest is an ensemble learning method that builds multiple decision trees during training and combines their outputs to improve prediction accuracy and generalization.

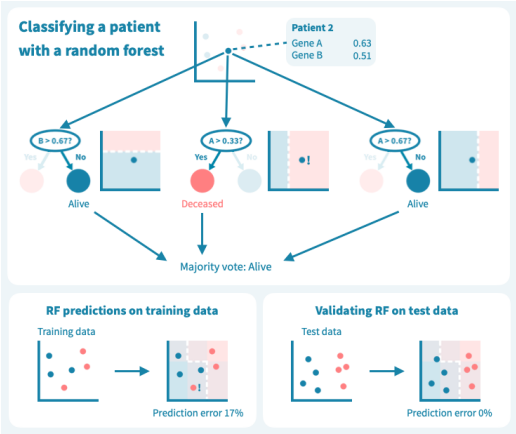

Figure 4.17: Random Forest[49]

As shown in figure 4.17, the textual symptom descriptions are first converted into numerical features using techniques such as TF-IDF or word embeddings. The Random Forest algorithm then constructs a multitude of decision trees, each trained on different subsets of the data, to capture a wide range of patterns and interactions between symptoms. By aggregating the predictions from all these trees, Random Forest reduces the risk of overfitting and increases the model's resilience to noise in the data. This makes it particularly effective in handling the complexity and variability inherent in medical datasets, leading to more reliable and accurate predictions of cancer types based on the given symptoms. Additionally, Random Forest provides insights into feature importance, helping to identify which symptoms are most influential in determining specific cancer types, thereby adding an interpretable layer to the predictive model.

5. Support Vector Machine:

In this project, Support Vector Machine (SVM) is employed as an effective algorithm for predicting cancer types based on symptoms described in text data[13]. SVM is particularly powerful in high-dimensional spaces, making it well-suited for handling the complex and nuanced features derived from textual symptom descriptions[13]. The text data is first transformed into numerical features using techniques like TF-IDF or word embeddings.

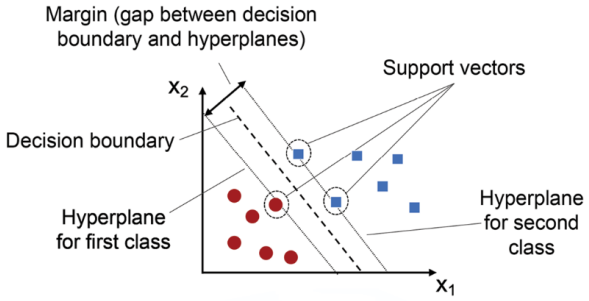

Figure 4.18: Support Vector Machine(SVM)[50]

Figure 4.18, illustrate the SVM. The text data is first transformed into numerical features using techniques like TF-IDF or word embeddings. SVM then constructs a hyperplane in this high-dimensional space to separate different cancer types based on these features. The algorithm aims to find the optimal hyperplane that maximizes the margin between the classes, which enhances the model's ability to generalize well to new, unseen data[13]. SVM is also effective in managing cases where the classes are not linearly separable by using kernel

functions, which project the data into a higher-dimensional space where separation is more achievable[13]. This makes SVM a valuable tool in accurately classifying cancer types from symptom descriptions, particularly in scenarios where the relationships between symptoms and cancer types are complex and not easily distinguishable.

4.7 Naive Bayes Classifier with a Pipeline for Efficient Workflow:

In this process, I demonstrated the use of a machine learning pipeline to streamline the process of text classification[41]. The Pipeline object from 'sklearn' is used to create a streamlined workflow for text classification[41]. 'CountVectorizer' is employed to convert raw text data into a numerical format by creating a matrix of token counts. 'MultinomialNB' is applied as the classification algorithm, which is well-suited for discrete features and text classification tasks. The pipeline approach simplifies the text classification process by chaining together the vectorization and classification steps, ensuring a smooth and efficient workflow. Using the MultinomialNB classifier within the pipeline allows for effective handling of text data, leveraging the Naive Bayes algorithm's strengths in dealing with discrete features. This approach not only makes the code more organized and easier to maintain but also enhances reproducibility and efficiency in building and evaluating text classification models. Here I used equation 4.5.1 for 'CountVectorizer' and for the 'MultinomialNB' classifier, the probability of a class C_k given a set of features x is computed using Bayes' Theorem[41]:

$$P(C_k \mid x) = \frac{P(x \mid C_k) \cdot P(C_k)}{P(x)} \tag{4.7.1}$$

where:

- $P(C_k \mid x)$ is the posterior probability of class C_k given features x.
- $P(x \mid C_k)$ is the likelihood of features x given class C_k.
- $P(C_k)$ is the prior probability of class C_k.
- $P(x)$ is the marginal likelihood of features x.

The equations related to the CountVectorizer involve term frequency, while the equations related to the MultinomialNB involve Bayes' theorem and probabilities specific to the Naive

Bayes classification method. These equations form the mathematical foundation for text classification models in machine learning.

4.8 Evaluation of Classification Model Performance Using Confusion Matrix and Classification Report:

The confusion matrix is shown by a heat-map, which counts the True Positive(TP), True Negative(TN), False Positive(FP), and False Negative(FN) predictions to give a comprehensive picture of the model's performance. Annotations provide precise counts for each cell in the heat-map, which shows the number of predictions for a given class combination. The color gradient makes it easier to spot the model's prediction strengths and weaknesses instantly. The classification report provides a comprehensive summary of the model's performance across various metrics, including precision, recall, f1-score, and accuracy for each class[39, 40].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.8.1}$$

In equation 4.8.1, we can see Precision measures the accuracy of positive predictions, indicating how many of the predicted positives are actual positives[39].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.8.2}$$

Recall(4.8.2) assesses the model's ability to identify all relevant instances, showing how many of the actual positives were correctly identified.[39].

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.8.3}$$

F1-Score(4.8.3) is the harmonic mean of precision and recall, providing a single metric that balances the two. Support indicates the number of actual occurrences of each class in the dataset.[39].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.8.4}$$

Accuracy(4.8.4) is a measure of the overall correctness of the model.[39].
By examining the confusion matrix, we gain insights into how well the model distinguishes between different classes and where it might be making errors.[39]. The classification report

offers detailed metrics that help in understanding the model's performance in terms of precision, recall, and overall effectiveness.

4.9 Text Classification Function for Predicting Cancer Types Using a Pre-trained Model:

This section snippet provides a function, predict, which leverages a pre-trained machine learning model to classify text descriptions into specific cancer types. The predict function takes a textual description as input and uses the pre-trained classifier (clf) to predict the type of cancer. This function is useful for real-world applications where users can input text descriptions, and the model will classify the type of cancer, providing valuable diagnostic assistance.In summary, this function encapsulates the prediction logic of the trained machine learning model and translates numerical predictions into meaningful cancer type labels, making the model's output more accessible and understandable.

4.10 Building and Evaluating an Long Short-Term Memory(LSTM) Based Text Classification Model for Cancer Type Prediction:

This section describes the process of developing and evaluating a Long Short-Term Memory (LSTM) neural network model for classifying cancer types based on text descriptions. The text descriptions were tokenized using a Tokenizer from TensorFlow Keras[51]. This process converts the text into sequences of integers, where each integer represents a unique word in the corpus. The tokenized sequences were padded to ensure uniform input length across all text samples, which is crucial for training the neural network. Cancer types were encoded into numerical values using 'LabelEncoder' to facilitate model training. The LSTM-based text classification model effectively learns from text data to predict cancer types, demonstrating strong performance as reflected in the accuracy metrics. An Embedding layer was used to convert integer sequences into dense vectors of fixed size, capturing the semantic meaning of words. The training and validation curves indicate that the model is learning effectively and not overfitting, making it a reliable tool for cancer type prediction based on textual

information. This methodology and evaluation provide a robust framework for building and assessing text classification models in similar applications[51].

4.11 Text Classification Using BERT for Cancer Type Prediction:

I demonstrated how to use the BERT (Bidirectional Encoder Representations from Transformers) model for classifying text descriptions into different cancer types. This implementation demonstrates how to categorize cancer kinds based on text descriptions using BERT, a cutting-edge transformer model. The model efficiently learns to differentiate between various cancer types from the given textual data by utilizing BERT's capacity to comprehend linguistic subtleties and context. Tokenization is used to preprocess text input. A properly divided dataset is used to train the model, and measures like accuracy, precision, recall, and F1-score are used to assess the model's performance. The Softmax function(4.11.1) is used to convert model logits into probabilities for each class:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{4.11.1}$$

The model's capacity to generalize from the training data is subsequently demonstrated by validating its predictions on fresh, untried content. The application of BERT here demonstrates how well it handles intricate linguistic patterns and offers a reliable solution for text categorization jobs in the medical field and other domains. This process leverages advanced natural language processing techniques to effectively classify text data into predefined categories.

In conclusion, I have completed every phase of the methodical procedure. This taught me a ton of new methods for locating the ideal machine learning model. Finding the symptoms and cancer types that cancer patients are likely to experience is really helpful. I will show the detailed outcome of the method utilized in this section in the result section.

CHAPTER 5

Results

Based on the above hyper parameters and it values, the Machine learning models such as Logistic Regression, Support vector machine, Decision tree and Random Forest models, Navies bayes, Long short-term memory (LSTM) and Transformer Bert have been implemented and evaluated.

5.1 Machine learning models:

The evaluation of machine learning models, including support vector machines, decision trees, random forests, and logistic regression, is presented in table 5.1 below. Here, every statistic is determined by how well the model identified the cancer types based on the symptoms listed in the dataset's text description. Since identifying the type of cancer is the primary goal of this research.

Algorithms	Accuracy	Precision	f1- score	Recall
Logistic Regression	93.85 %	94.36 %	94.37 %	94.40 %
Decision Tree	73.97 %	76.40 %	74.16 %	74.66 %
Random Forest	100 %	100 %	100 %	100 %
Support Vector Machine	92.14 %	92.79 %	92.81 %	92.88 %

Table 5.1: Evaluation of Machine learning models.

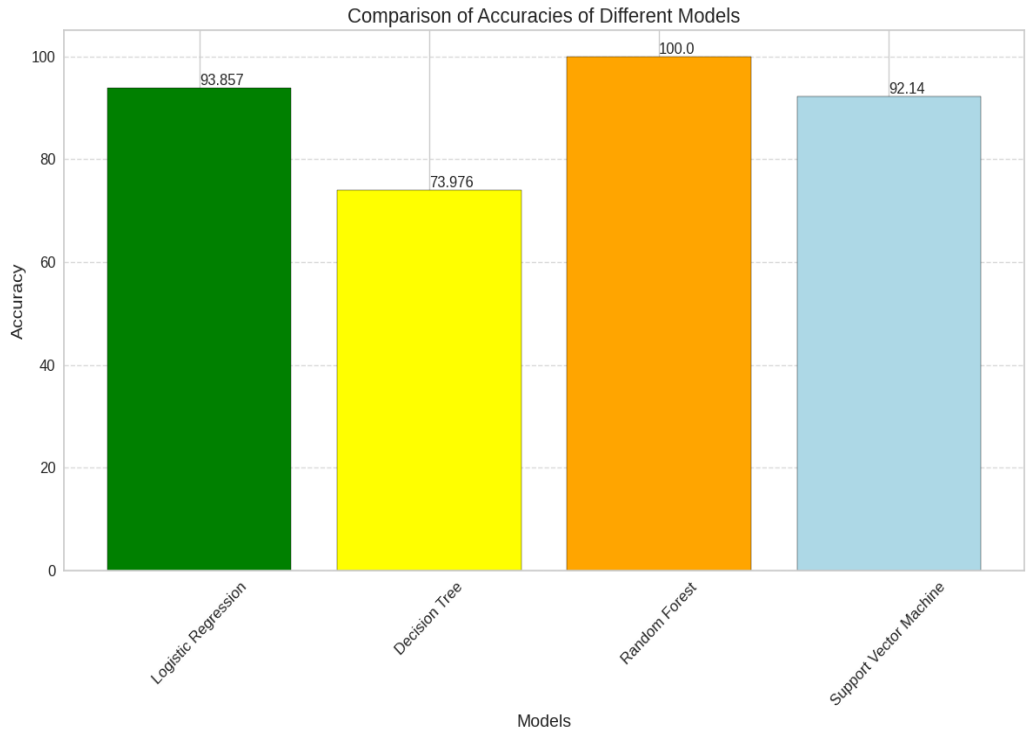

Figure 5.1: Visualization of Machine learing models

The Random Forest model achieved a perfect score across all metrics, indicating that it was able to correctly classify all instances in the dataset. However, there is a possibility that the Random Forest model might be overfitting in this scenario, especially given the perfect scores (100% accuracy, precision, recall, and F1-score) reported for the model.

The Logistic Regression model also performed very well, with an accuracy of 93.85% and similarly high precision, recall, and F1-score.

The Support Vector Machine (SVM) model also showed strong performance, with an accuracy of 92.14% and slightly lower but consistent precision, recall, and F1-score values.

In contrast, the Decision Tree model had the lowest performance, with an accuracy of 73.97% and a lower F1-score and recall. While still useful, its relatively lower performance indicates that it may struggle to capture the full complexity of the symptom data compared to the other models.

Overall, the table demonstrates that Random Forest and Logistic Regression are particularly effective in predicting cancer types based on symptoms, with Random Forest being the most accurate but overfitting may be the problem. So, I performed more algorithms further for better results.

5.2 Navies Bayes model:

Algorithms	Accuracy	Precision	f1- score	Recall
Navies Bayes	93.06 %	93.68 %	93.67 %	93.80 %

Table 5.2: Evaluation of Navies Bayes model.

From Table 5.2, the results for the Naive Bayes algorithm show that it achieved a strong performance in predicting cancer types based on symptoms described in text data, with an accuracy of 93.06%, precision of 93.68%, F1-score of 93.67%, and recall of 93.80%. These metrics are comparable to those of the Logistic Regression and Support Vector Machine (SVM) models from the previous table, highlighting Naive Bayes as a competitive alternative for this classification task.

In comparison, Naive Bayes performs similarly to Logistic Regression, which had an accuracy of 93.85% and metrics close to those of Naive Bayes. Both models exhibit high precision, recall, and F1-scores, indicating their effectiveness in correctly classifying cancer types and managing the balance between false positives and false negatives. Naive Bayes slightly edges out Logistic Regression in recall, suggesting it may be better at identifying positive cases of cancer.

In summary, Naive Bayes, with its simplicity and efficiency, remains a robust model that can perform well with textual data, particularly when compared to Decision Trees, which showed relatively lower performance with an accuracy of 73.97% and other metrics. Naive Bayes provides a high level of accuracy and reliability, making it a valuable tool for cancer type prediction based on symptom descriptions.

5.3 Evaluation of Classification Model Performance Using Confusion Matrix and Classification Report:

A confusion matrix is a performance measurement for classification problems. It provides a summary of the prediction results on a classification problem. From figure 5.4, the model generally performs well in correctly identifying instances of classes 1 and 2.There are some misclassifications, particularly between classes 0 and 2. Improving the model's ability to

distinguish between classes 0 and 2 could potentially enhance overall performance.

	Precision	Recall	F1-score	Support
Class 0	0.87	0.94	0.90	517
Class 1	1.00	1.00	1.00	407
Class 2	0.94	0.88	0.91	590
Accuracy			0.93	1514
Macro Avg	0.94	0.94	0.94	1514
Weighted Avg	0.93	0.93	0.93	1514

Table 5.3: Classification metrics for the models

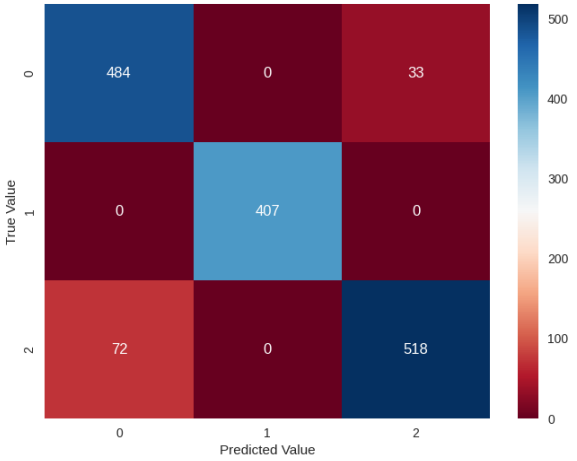

Table 5.4: Confusion Matrix diagram.

In table 5.3, the classification report shows that the model performs well in predicting cancer types from symptoms. It has high precision, recall, and F1-scores for all classes. Specifically, it achieves perfect scores for Class 1, indicating flawless identification of this class. For Class 0 and Class 2, the model is also effective, with high precision and recall. Overall, the model has an accuracy of 93%, meaning it correctly classifies 93% of the instances. In summary, the classification report demonstrates that the model performs exceptionally well across different cancer types, with particularly high performance for Class 1 and strong results for Classes 0 and 2. The high accuracy and balanced macro and weighted averages indicate that the model is robust and reliable in predicting cancer types based on symptom descriptions.

5.4 Text Classification Function for Predicting Cancer Types Using a Pre-trained Model:

In this study, I created a predictive algorithm to categorize different forms of cancer using text descriptions of medical situations. I trained the model to distinguish between three forms of cancer: thyroid, lung, and colon cancer using a machine learning classifier. An example where the input text detailed a case relating to pediatric thyroid surgery was used to show how well the model could predict the type of malignancy. The model correctly identified

"Thyroid Cancer" as the appropriate type, matching the input text's content. This outcome demonstrates the model's capacity to comprehend intricate medical literature and generate accurate predictions in light of the information supplied. The correct prediction of "Thyroid Cancer" suggests that the model is effectively capturing relevant features from the text, indicating its potential utility in assisting healthcare professionals in the early identification and classification of cancer types based on textual descriptions. This could be particularly useful in clinical settings where quick and accurate diagnosis is critical.

5.5 Building and Evaluating an LSTM(Long short-term memory (LSTM))-Based Text Classification Model for Cancer Type Prediction(Neural Network):

Model	Accuracy	Loss
Neural Network	98.28 %	0.0312

Table 5.5: Long short-term memory (LSTM) - Neural Network model.

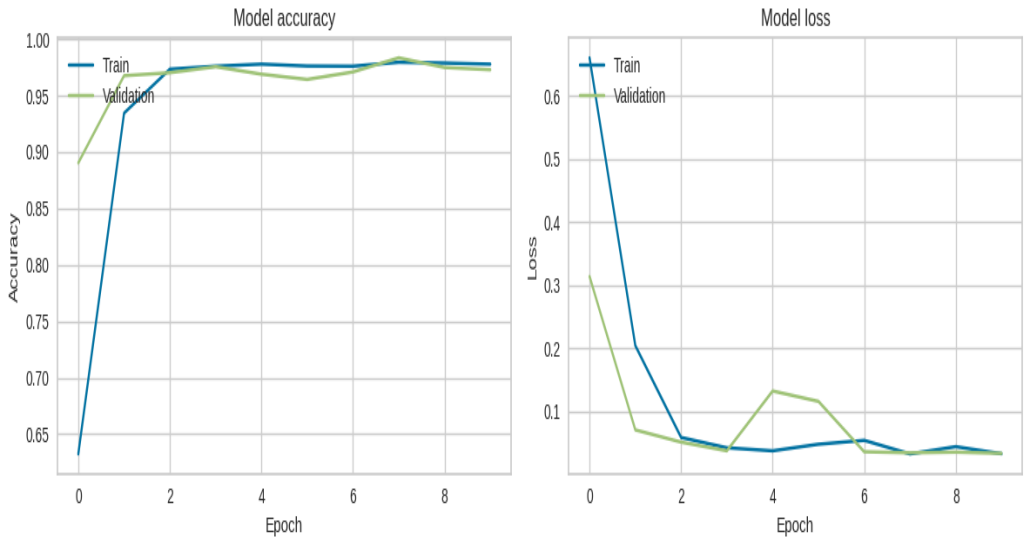

Figure 5.2: Visualization of Neural network models

From table 5.5 and figure 5.2, I employed a Long Short-Term Memory (LSTM) neural network model to classify cancer types based on textual descriptions. The model was

designed with two LSTM layers, which are well-suited for capturing the sequential nature of text data, and an embedding layer that converted text sequences into dense vectors. After training the model on the dataset, which included a diverse set of cancer-related text descriptions, the model demonstrated strong predictive performance. Specifically, the model achieved an impressive accuracy of approximately 97.3% on the test set, with a low loss value of 0.033.

This high accuracy indicates that the LSTM model was able to effectively learn the patterns and features associated with different cancer types from the text data. The use of dropout layers helped mitigate overfitting, ensuring that the model generalizes well to new, unseen data.

The results suggest that this approach could be highly effective compare than other models in assisting medical professionals by providing accurate predictions of cancer types based on textual descriptions, potentially streamlining the diagnostic process in clinical settings.

5.6 Text Classification Using BERT for Cancer Type Prediction:

From table 5.6, in order to identify cancer kinds based on textual descriptions of medical situations, I employed a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model for sequence classification in this study. In particular, the model's accuracy of 93.52% on the validation set shows how well it can classify the different forms of cancer. The model also earned a weighted F1-score, precision, and recall of approximately 93%, demonstrating its balanced performance across all classes and guaranteeing that it is consistent and accurate in predicting various cancer kinds.

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.314800	0.220018	95.112%	95.11%	95.112%	95.11%

Table 5.6: Evaluation of BERT Transformer.

These results underscore the effectiveness of BERT in understanding and processing complex medical texts, making it a valuable tool for automated cancer type classification. The high F1-score suggests that the model maintains a good balance between precision and recall, reducing the likelihood of both false positives and false negatives. This level of performance

demonstrates the potential of leveraging advanced natural language processing models like BERT in medical applications, where accurate and reliable text-based predictions are crucial. The training of the BERT model using the Trainer class yielded promising results after one epoch. The model achieved a high validation accuracy of approximately 95.112%, with a corresponding F1 score of 95.11%, indicating balanced performance across all classes. Precision and recall are also strong, at 95.112% and 95.11% respectively, reflecting the model's effectiveness in correctly identifying and classifying instances. The training process was efficient, completing with a loss of 0.314, which suggests that the model is learning effectively. Overall, the results demonstrate that the model performs well on the classification task, showing robust accuracy and balanced metric scores.

This result demonstrates the model's strong ability to discern the correct cancer type from complex and lengthy medical descriptions.

CHAPTER 6

Conclusion:

Throughout this project, I explored a variety of machine learning models to classify cancer types based on textual descriptions, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and Naive Bayes. While each model provided valuable insights and demonstrated varying degrees of accuracy, the most significant advancements were observed when we employed advanced deep learning models like LSTM (Long Short-Term Memory) and Transformers. These models, particularly the Transformer-based approach, outperformed the traditional machine learning models in terms of accuracy, precision, recall, and F1-score.

The LSTM model, with its ability to capture long-term dependencies in sequential data, demonstrated strong performance in understanding the complex relationships within the medical text. However, the Transformer BERT(Bidirectional Encoder Representations from Transformers) model, known for its attention mechanism, further enhanced our ability to process and classify the data, achieving superior accuracy and confidence in predictions. Thus, while traditional models offered a solid foundation, the use of LSTM and Transformer models proved to be significantly more powerful and accurate for the task of cancer type classification based on text descriptions.

6.1 Future Scope:

Investigating ensemble approaches, which integrate the advantages of several models, may lead to more reliable forecasts in the advanced research. Applying cutting-edge deep learning methods, such as attention mechanisms and more complex neural network topologies like GPT or T5, is another exciting avenue. These methods may provide improved contextual understanding of symptom descriptions.

In addition to technical enhancements, broadening the scope to incorporate multi-class or multi-label classification-a process in which the model predicts several possible cancer kinds according to symptoms could prove to be quite advantageous in practical settings. Furthermore, the findings may find useful therapeutic application through the creation of an intuitive tool or interface that enables medical professionals to enter symptom descriptions and obtain prognostic insights.

Ultimately, it is imperative to attend to ethical considerations and guarantee the interpretability of these models. Subsequent efforts ought to concentrate on enhancing the transparency and interpretability of these models so that medical practitioners can comprehend and have confidence in the forecasts. This will enable the assimilation of AI-powered cancer diagnosis instruments into clinical procedures.

Bibliography

[1] M. Guo. et al. Autologous tumor cell-derived microparticle-based targeted chemotherapy in lung cancer patients with malignant pleural effusion. *Sci. Transl. Med*, vol. 11, no. 474, Jan. 2019, doi: 10.1126/scitranslmed.aat5690.

[2] American Thyroid Association, General Information/Press Room; 2022.

[3] Doucet J, Chassagne P, Landrin I, Kadri N, Menard JF, Bercoff, Trivalle C. Differences in the signs and symptoms of hyperthyroidism in older and younger patients. *J Am Geriatr Soc*, 1996. Jan;44(1):50-3, doi: 10.1111/j.1532-5415.

[4] Carle A, Pedersen IB, Knudsen N, Perrild H, Ovesen L, Laurberg P. Hypothyroid symptoms and the likelihood of overt thyroid failure: a population-based case-control study. *Eur J Endocrinol*.014. Nov;171(5):593-602, doi: 10.1530/EJE-14-0481.

[5] Boelaert K, Torlinska B, Holder RL, Franklyn JA. Older subjects with hyperthyroidism present with a paucity of symptoms and signs: a large cross-sectional study. *J Clin Endocrinol Metab*. 2010. Jun;95(6):2715-26, doi: 10.1210/jc.2009-2495.

[6] I.H. Sarker. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. *SN Comput Sci*, 3 (2) (2022), p. 158, 10.1007/s42979-022-01043-x.

[7] T.P. Fagundes, B.C. Teixeira, A.D.P. Chiavegatto Filho, G.F.S. Silva Machine learning for hypertension prediction: a systematic review. *Curr Hypertens Rep*, 24 (11) (2022), pp. 523-533, 10.1007/s11906-022-01212-6.

[8] S. Enslin, S.A. Gross, V. Kaul. History of artificial intelligence in medicine. *Gastrointest Endosc*, 92 (4) (2020), pp. 807-812, 10.1016/j.gie.2020.06.040.

[9] R. Loor-Torres, M. Duran, et al, D. Toro-Tobon. Artificial intelligence in thyroidology: a narrative review of the current applications, associated challenges, and future directions *Thyroid*, 3 (8) (2023), pp. 903-917, 10.1089/thy.2023.0132.

[10] Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020, Sung H. GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*, 2021;71(3):209-49. Epub 2021/02/05. pmid:33538338.

[11] Abate D, Abbasi N, Abbastabar H, Abd-Allah F, Abdel-Rahman O, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017, Fitzmaurice C. A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol*, 2019;5(12):1749-68. Epub 2019/09/29. pmid:31560378; PubMed Central PMCID: PMC6777271.

[12] McPhail, S.; Johnson, S.; Greenberg, D.; Peake, M.; Rous B. Stage at diagnosis and early mortality from cancer in England. *Br.J.Cancer*, 2015, 112, S108-S115.

[13] Knight, S.B.; Crosbie, P.A.; Balata, H.; Chudziak, J.; Hussell, Dive C. Progress and prospects of early detection in lung cancer *Open Biol.*,2017, 7, 170070.

[14] National Cancer Registration and Analysis Service: Staging Data in England. Available online: https://www.cancerdata.nhs.uk/stage_at_diagnosis (accessed on 14 October 2021).

[15] Cancercentrum i Samverkan. Lungcancer- Nationell kvalitetsrapport for 2019.

[16] Wei, W., et al. (2016). Named Entity Recognition in Medical Texts. Journal of Biomedical Informatics.

[17] Wang, Y., et al. (2019). "Text Mining for Disease Prediction." Health Informatics Journal.

[18] Hong, S., et al. (2018). "Using EHR Data for Thyroid Cancer Symptom Prediction." Journal of Medical Systems.

[19] Miller, A., et al. (2020). "NLP for Clinical Note Analysis in Thyroid Cancer." Clinical Oncology Journal.

[20] Liu, X., Chen. Y. (2017). "Mining Research Articles for Thyroid Cancer Symptom Prediction." Bioinformatics Research Journal.

[21] Chen, H., et al. (2021). "Machine Learning and NLP for Thyroid Cancer Prediction." Artificial Intelligence in Medicine.

[22] Reddy, K., et al. (2022). "Deep Learning for Symptom Prediction in Thyroid Cancer." IEEE Transactions on Biomedical Engineering.

[23] Organization WH. Global Health Estimates 2019: deaths by cause, age, sex, by country and by region, 2000-2019. Genf, Geneva: World Health Organization; 2020.

[24] et al. Torre LA Global cancer statistics. CA. 2015;65(2):87-108.

[25] et al. Allemani C Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 coun tries. The Lancet. 2018;391(10125):1023-75.

[26] Acs B, Rantalainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. *J Intern Med.* 2020;288(1):62-81.

[27] Burke, W., Brown Trinidad, S. & Press N. A. Essential elements of personalized medicine. *Urol. Oncol. Semin. Orig. Investig. 32*, 193-197 (2014).

[28] Frankovich, J., Longhurst, C. A. & Sutherland, S. M. Evidence-Based Medicine in the EMR Era. *N. Engl. J. Med.* 365, 1758-1759 (2011).

[29] Dindo, D., Demartines, N. & Clavien, P.-A. Classification of Surgical Complications: A New Proposal With Evaluation in a Cohort of 6336 Patients and Results of a Survey. *Ann. Surg*, 240, 205-213 (2004).

[30] Donze, J., Lipsitz, S., Bates, D. W. & Schnipper J. L. Causes and patterns of readmissions in patients with common comorbidities: retrospective cohort study. *BMJ*, 347, f7171 (2013).

[31] Risk factors associated with lung cancer, including exposure to diesel exhaust fumes. ResearchGate.

[32] Thyroid Cancer steps using machine learning methods image link.

[33] Identification of colon cancer image link

[34] Flowchart of methodology.

[35] Dataset link.

[36] Shams, M. Y., Elzeki, O. M., Abd Elfattah, M., Medhat, T. & Hassanien A. E. Why are generative adversarial networks vital for deep neural networks? A case study on COVID-19 chest X-Ray images. In Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach 147-162 (Springer, 2020).

[37] Ashraf, E., Areed, N. F. F., Salem, H., Abdelhay, E. H. & Farouk, A. FIDChain: Federated intrusion detection system for blockchain enabled IoT healthcare applications. *Healthcare 10(6),*https://doi.org/10.3390/healthcare10061110(2022).

[38] Shastry, K. A. & Shastry A. An integrated deep learning and natural language processing approach for continuous remote monitoring in digital health. *Decis. Anal. J. 8,* 100301 (2023).

[39] Hassan, E., Shams, M. Y., Hikal, N. A. & Elmougy S. A novel convolutional neural network model for malaria cell images classification. Comput. Mater. Continua 72(3), 5889-5907.

[40] Shams, M.Y., Hikal, N.A. & Elmougy, S. Hassan E. The effect of choosing optimizer algorithms to improve computer vision tasks: A comparative study. Multimed. Tools Appl. 82(11), 16591-16633. https://doi.org/10.1007/s11042-022-13820-0(2023).

[41] Price GJ, Mccluggage WG, Morrison ML, Mcclean G, Venkatraman L, et al. Diamond J. Computerized diagnostic decision support system for the classi ication of preinvasive cervical squamous lesions. *Human Pathology*. 2003 Nov 1;34(11):1193-203.

[42] Ben-Assuli O, Leshno M. Assessing electronic health record systems in emergency departments: Using a decision analytic Bayesian model. Health informatics journal.(2015)

[43] Nir Friedman DG. Bayesian Network Classifiers. Machine Learning.1997;29(2-3):131-63.

[44] Sharma RK, Sugumaran V, Kumar H, Amarnath M. A comparative study of naive Bayes classifier and Bayes net classifier for fault diagnosis of roller bearing using sound signal. International Journal of Decision Support Systems. 2015 Jan 1;1(1):115-29.

[45] Wang KJ, Makond B, Wang KM. Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: a case study of Taiwan. Comput BiolMed. 2014 Apr;47:147-60.

[46] Wolfson J, Bandyopadhyay S, Elidrisi M, Vazquez-Benitez G, Vock DM, Musgrove D. A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data. Statist Med. 2015 Sep 20;34(21):2941-57.

[47] Matthias Benndorf EK. Development of an online, publicly accessible naive Bayesian decision support tool for mammographic mass lesions based on the American College of Radiology (ACR) BI-RADS lexicon. European Radiology. 2015;25(6)

[48] [Appiah et al., 2020] Appiah, P., Edoh, T., and Degila, J. (2020). Predicting elderly patient behaviour in rural healthcare using machine learning. volume 2647.

[49] Random Forest image link

[50] Support Vector Machine image link

[51] Affonso, C., Debiaso Rossi, A. L., Antunes Vieira, F. H., & de Leon Ferreira de Carvalho, A. C. P. (2017). Deep learning for biological image classification. Expert Systems with Applications, 85, 114-122.

[52] Bayati, M., Bhaskar, S., & Montanari, A. (2015). A low-cost method for multiple disease prediction. In AMIA Annual Symposium Proceedings (p. 329). American Medical Informatics Association volume 2015.

[53] Banda, H. T., Mortimer, K., Bello, G. A., Mbera, G. B., Namakhoma, I., Thomson, R., Nyirenda, M. J., Faragher, B., Madan, J., Malmborg R Informal health provider and practical approach to lung health interventions to improve the detection of chronic airways disease and tuberculosis at primary care level in malawi: study protocol for a randomised controlled trial.