

# Cancer Classification using Medical Cancer Text Documentation based on Machine Learning Model and NLP

Nidhi Kumari Chauhan  
EECS, University of Ottawa  
Ottawa, Canada  
nidhi.chauhan@uottawa.ca

Mohana Priya Sivakumar  
EECS, University of Ottawa  
Ottawa, Canada  
msiva036@uottawa.ca

Abhisht Makarand Joshi  
EECS, University of Ottawa  
Ottawa, Canada  
ajosh053@uottawa.ca

**Abstract**-Automatic text classification of medical data is essential for a number of reasons. Firstly, it can assist in the analysis and classification of vast amounts of medical data. Given the tremendous growth in the production of medical data, this is particularly crucial. Secondly, text categorization can increase the precision and dependability of medical data analysis by employing Deep learning and NLP algorithms to spot trends and connections that human analysts might not immediately see. The automatic organization of medical data into predefined categories by text classification, which makes it simpler for researchers and healthcare professionals to find and use the data they need to support decision-making and improve patient outcomes, can also increase the accessibility and usability of medical data. The medical data can be available in the form of papers, articles, reports, surveys, and other materials related to medicine. Our study involved tokenizing, case conversion, stop words removal, word stemming or lemmatization, etc., pre-processing phase for text documents. Furthermore, the model building and testing phases were carried out, and after that analysis of the result was done in evaluation phases. In this study, we examine automated text categorization utilizing deep learning models like LSTM, Multinomial Naive Bayes, Logistic Regression, Bernaulli Naive Bayes, and an NLP model called BioBERT. Our results show that, of all the implementations, the LSTM model performed the best and had the maximum accuracy of 98.88% respectively.

**Keywords:** Cancer, BioBERT, Text Classification, Deep Learning

## I. PROBLEM STATEMENT

Cancer is a complex and heterogeneous group of diseases that is characterized by the uncontrolled growth and spread of abnormal cells. There are more than 100 different types of cancer, and each type has its own specific features, causes, and treatments. Accurate cancer classification is essential for effective diagnosis and treatment, as different types of cancer often require different approaches to management.

Traditionally, cancer classification has relied on various laboratory and imaging tests, as well as biopsy and histopathology results. These tests can provide valuable information about the characteristics of a cancer, such as the type, stage, and grade of the tumor. However, these tests can also be expensive, time-consuming, and invasive, and they may not always provide a definitive diagnosis. In addition,

some types of cancer, such as rare or poorly differentiated cancers, may be difficult to classify using traditional methods.

In recent years, there has been increasing interest in using machine learning and NLP techniques to classify cancer types using medical text documentation, such as electronic health records (EHRs) and pathology reports. These techniques have the potential to provide a faster, less expensive, and less invasive alternative to traditional diagnostic methods. Medical text documentation contains a wealth of information about a patient's medical history, symptoms, diagnoses, and treatments, and it can be used to extract features that are relevant for cancer classification. Machine learning algorithms can then be trained on these features to classify cancer types with high accuracy.

NLP is a field of artificial intelligence that focuses on the interaction between computers and human languages. NLP techniques can be used to process and analyze medical text documentation in a variety of ways, including extracting information from unstructured text, identifying named entities, and classifying text into predefined categories. BERT, which stands for "Bidirectional Encoder Representations from Transformers," is a powerful NLP model developed by Google that can be used for a wide range of NLP tasks. BERT and other NLP models have been shown to be effective for analyzing medical text data and extracting useful information for various purposes, including cancer classification.

The key contributions in our research include the insights of the following,

- In-depth study of Medical (Cancer) huge documentation.
- Pre-processing, cleaning the text and extracting valuable information from huge dataset of digital medical documents (medical research papers, articles, reports, surveys)
- Analyzing and classifying the type of Cancer by applying machine learning algorithms.
- Assessing and selecting the best algorithm for prediction with the maximum efficiency.

To address the listed motives, we have deployed the following Machine Learning models such as Multinomial Naïve Bayes, Logistic Regression and Bernoulli Naïve Bayes. Also, LSTM, a Deep Learning Model was used for the training and testing. Additionally, a pre-trained natural language processing (NLP) model called BioBERT was fine-tuned using the Cancer Text classification dataset[9] is used for the

following stage of implementation. This model was created exclusively for the biomedical domain.

**Structure of the paper.** The rest of this paper is structured as follows. Section 1 outlines the Problem statement. Sections 2 outlines the literature survey. Sections 3 provides information regarding the dataset. Sections 4 and 5, describe the incorporated methodology and evaluation aspects of the study. Sections 6, showcases our results and analysis of the same. Next, Section 7, discusses regarding the proposed future work. Finally, we conclude the paper in Section 8.

## II. LITERATURE SURVEY

Research on the application of machine learning models and natural language processing (NLP) methods for cancer classification based on medical text documentation is extensive.

In paper [1], the authors propose a deep learning model based on a convolutional neural network (CNN) for breast cancer classification based on medical text data. They show that the model does better than traditional machine learning models and is very accurate when used with a dataset about breast cancer. D. S. Kim et al. [2], presented a machine learning based approach for identifying breast cancer subtypes based on electronic health records (EHRs). The authors compared the performance of several machine learning algorithms, such as support vector machines (SVMs) and random forests. They found that the best-performing model is a hybrid approach that combines the predictions of multiple models. In [3], the authors proposed a natural language processing (NLP) approach for classifying cancer types based on clinical text. They showed that their method is better than other machine learning models and can accurately classify different types of cancer. [4] The authors compare the performance of machine learning models, like logistic regression and random forests, and find that the best-performing model is a gradient boosting model for predicting breast cancer subtypes based on EHR data. In [5], a deep learning approach for cancer classification based on clinical text was used. The model was based on a long short-term memory (LSTM) network. It can accurately classify different types of cancer and does better than traditional machine learning models. In [6], the author implemented the decision tree algorithm, SVM, RF, etc., and the best performance was from Random Forest. This paper presented a machine learning approach for classifying breast cancer based on EHR data. In this paper[7], the authors propose a deep learning approach for predicting breast cancer subtypes based on EHR data. They demonstrate that their model, which is based on a long short-term memory (LSTM) network, is able to accurately classify various breast cancer subtypes. In paper [8], the authors propose an artificial neural network (ANN) approach for predicting breast cancer subtypes based on EHRs. They showed that their model, which is based on a multi-layer perceptron (MLP) network, can accurately classify different types of breast cancer and does better than traditional machine learning models. The authors also compare the performance of the ANN model with that of a CNN model and find that the ANN model performs better on the EHR data used in the study.

## III. DATASET DESCRIPTION

For the Biomedical text document classification, the dataset is obtained in the textual format. Initially, the PUBMD website's pdf files are converted into text files, upon extracting relevant and necessary information. Post conversion of the text files, the same data is processed into a single comma separated values file. This obtained data was from the Reports, Surveys, Publications and Research papers related to the medicine field. Hence, this data is obtained from a well-known website known as Kaggle [9] to aid our study. The quality and characteristics of the dataset can significantly affect the performance of the model.

In general, a good dataset should be representative of the real-world problem or task that the model is intended to solve. Biomedical text document classification, the data which is available are of the form, abstract papers and full papers which is of six at the most or lesser number of pages in length. For our study, the dataset employed focuses on long research papers, which are minimum 6 pages in length. To be particular for our domain of study, the dataset obtained and used contains information from the Cancer documents which are essentially classified into three distinct categories, namely 'Thyroid Cancer', 'Colon Cancer', and 'Lung Cancer'.

The size of the dataset can also be an important factor. A larger dataset can provide more examples for the model to learn from, which can lead to better generalization and improved performance on unseen data. To aid this requirement of a good dataset, the dataset obtained for our study consists of 7569 publications in total. All the instances contains one of the three class labels in the dataset. Additionally, a characteristic of a good dataset, is that it should be significantly balanced, meaning that it should contain a similar number of examples for each class or category. If the dataset is imbalanced, the model may be biased towards the more prevalent classes, leading to poor performance on the minority classes. To demonstrate this quality, our data set contains a pretty balanced number of samples in each category as shown in Figure 1. The Colon Cancer label contains 2579 instances in the dataset. Similarly, the Lung Cancer label is found 2180 times. Finally, for the thyroid Cancer, around 2810 samples are found in the total of 7569 instances of the whole dataset.

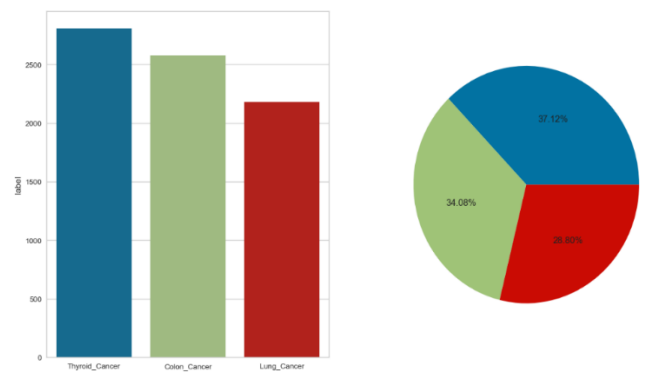


Figure 1: Dataset Distribution based on Labels

Exploratory Data Analysis (EDA) is performed on the dataset. It is a crucial step that involves analyzing and understanding the characteristics of a dataset. It is used to

identify patterns, trends, and relationships within the data and to gain insights that can inform the development of a machine learning model. It typically involves visualizing the data in various ways, such as through histograms, scatterplots, and boxplots. It may also involve calculating summary statistics, such as means, medians, and standard deviations. These techniques can help to identify anomalies, trends, and patterns in the data and can inform the development of a machine learning model.

Table 1: Top 10 Frequent word before and after transformation

Top 10 Frequent words before transformation	Top 10 Frequent words after transformation
the	cancer
of	cells
and	patients
in	cell
to	expression
with	study
for	data
was	tumor
were	fig
is	using

To begin with, the kernel distribution of the textual information is implemented and illustrated in Figure 3. It is a graphical representation of the distribution of the number of words in a dataset with respect to the different labels or categories present in the dataset. This graph was helpful for us to understand the characteristics of the data and identify any potential issues, for instance identifying whether the dataset is heavily skewed, based on which the necessary steps to balance the data can be implemented. Further, size of the text: number of characters w.r.t distinct outcomes is illustrated in Figure 4, as the size of the text could also impact the performance of certain NLP techniques or algorithms, such as those that rely on statistical measures or frequencies of words or characters. Similarly, the size of the text: number of words w.r.t distinct outcomes is pictorialized and depicted in Figure 5.

Further drilling down, using NLP toolkits, the 10 most frequent words from the data set is identified and the word ‘the’ topped the list as depicted in Table 1. Dealing the same concept with bigrams, it was identified that ‘of the’ topped the list of 10 most frequent bigrams present in the dataset as seen in Table 2.

Table 2: Top 10 Frequent Bigrams before and after transformation

Top 10 Frequent bigram before transformation	Top 10 Frequent bigram after transformation
Of the	Lung cancer
In the	Et al
To the	Breast cancer
And the	Creative commons
Patients with	Cancer cells
For the	Cell lines
On the	Gene expression
With the	Cancer cell
Expression of	Cancer patients
From the	Cell proliferation

#### IV. METHODOLOGY DESCRIPTION

The section provides information with the different methodologies used in the implementation of this study.

##### A. Data Pre-processing

Natural language processing (NLP) includes a crucial step called data preparation, which involves cleaning and preparing text data to make it acceptable for analysis or additional processing. As the specific pre-processing steps taken will depend on the needs and goals of the NLP project, as well as the characteristics of the text data, we have initiated the process with lowercasing. This process involves conversion of all characters to lowercase format. Post the previous stage, for the data cleaning is done. This involves identifying and correcting errors or inconsistencies in the text data as it can help to improve the accuracy and reliability of the results obtained from analyzing or processing the text data. To address this, as depicted in Figure 2, removing of unnecessary textual information such as removal of html elements, links, punctuation marks, numbers, accent characters and stop words are implemented. Stop words are common words that do not convey meaning, such as "the," "a," and "an". Post removal of mentioned elements, the processed dataset is tokenized accordingly. Also, Tokenization is defined as the process of dividing a piece of text into smaller units called tokens. It is because, tokenization facilitates the text to be processed and analyzed at a more granular level, such as by examining individual words or groups of words.

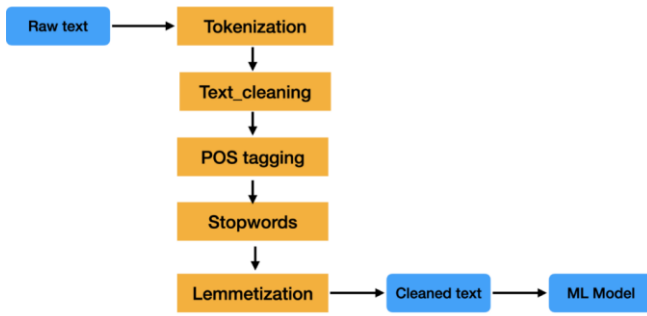


Figure 2: Data Pre-processing pipeline

Furthermore, Text Lemmatization is done which involves the steps to reducing words to their base form, taking into account their part of speech. To ensure the dataset is more efficient, some words are removed based on the criteria: Words having high frequency of above 80% and words having low frequency rate, i.e., below 20%. This is to ensure removal of common words which shall not provide significant insights as high frequency category words are more repeated and low frequency words provides negligible significance.

Exploratory Data Analysis is carried out post pre-processing mechanism to obtain more powerful and efficient insights about the cleaned dataset. Various illustrations such as the text length per label, number of words per label is picturized and shown in Figure 3. Additionally, the 10 most frequent words of the dataset along with the bigrams is determined in Table 1 and Table 2 respectively.

### B. Machine Learning Approach

The following Machine Learning models have been deployed for the purpose of the mentioned study.

- Multinomial Naïve Bayes

The multinomial Naive Bayes classifier is a probabilistic model that is commonly used for classification tasks. It is called a "Naive Bayes" classifier because it is based on the Bayes theorem, which describes the probability of an event based on prior knowledge of conditions that might be related to the event. The "multinomial" part of the name refers to the fact that the model is designed to work with discrete data, such as word counts or frequencies, rather than continuous data. The basic idea behind the multinomial Naive Bayes classifier is that each piece of data can be thought of as a collection of features, and the classifier is trying to predict the class label (e.g., positive or negative) based on the values of these features. To make a prediction, the classifier estimates the probability of each class label given the values of the features, and then chooses the label with the highest probability. The Naive Bayes classifier makes the "naive" assumption that all of the features are independent of each other, which means that the presence or absence of one feature does not depend on the presence or absence of any other feature. This assumption simplifies the calculations and allows the classifier to make predictions using relatively simple math.

- Logistic Regression

It is a statistical model that is commonly used for classification tasks. However, unlike the Naive Bayes classifier, logistic regression does not make the assumption that the features are independent of each other. Logistic regression works by estimating the probability of a class label given the values of the features. To do this, it uses an equation that combines the values of the features using coefficients (also called weights). Logistic regression can be used with discrete data, such as word counts, as long as the data is transformed into a suitable format. This can be done using techniques such as bag-of-words or term frequency-inverse document frequency (TF-IDF). Once the data is in this form, logistic regression can be used to make predictions based on the word counts or frequencies.

- Bernoulli Naïve Bayes

Bernoulli Naive Bayes is a classification technique that is applied to tasks where the features are binary or boolean. It is based on the idea of feature independence and the concepts of Bayesian probability. The Bayes theorem is used in the method to determine the likelihood that a data point, given its characteristics, belongs to a particular class. The characteristics in Bernoulli Naive Bayes are considered to be independent and to follow a distribution called the Bernoulli distribution, which has just two possible values (0 or 1). The algorithm predicts the class with the highest probability after calculating the probabilities for each class. When classifying texts using features, Bernoulli Naive Bayes is frequently utilised.

### C. Deep Learning Approach

The Deep Learning methodology LSTM is employed to gain further drilled down insights. LSTM stands for Long Short-Term Memory, which is a type of recurrent neural network (RNN) architecture used in natural language processing (NLP) and other applications. LSTMs are a variant of RNNs that are designed to handle the vanishing gradient problem, which is a common issue in training RNNs with long sequences. The vanishing gradient problem occurs when the gradients of the parameters in the network become very small as they are backpropagated through time, making it difficult for the network to learn long-term dependencies. Using LSTMs for cancer classification on medical text datasets is a promising approach that leverages the ability of LSTMs to model long-term dependencies in sequential data. Medical text data, such as electronic health records (EHRs) or clinical notes, can contain a wealth of information about a patient's medical history, diagnosis, and treatment. By training an LSTM on this data, it may be possible to predict the likelihood of a patient having a particular type of cancer or to identify patterns in the data that are indicative of cancer.

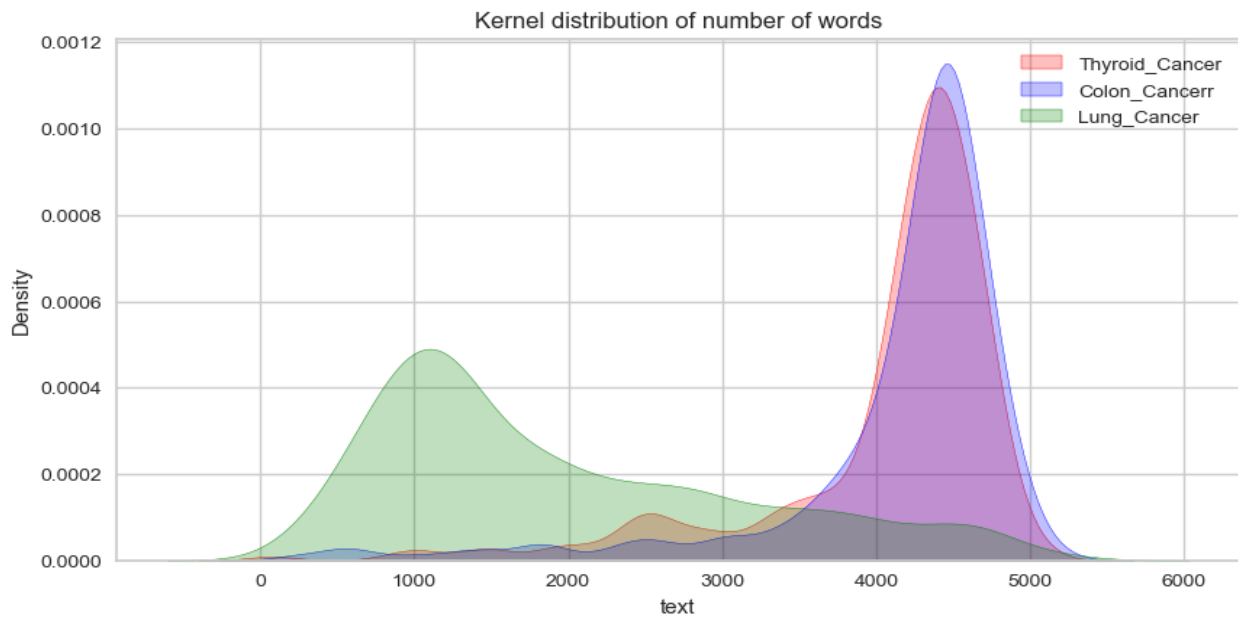


Figure 3: Kernel distribution of number of words

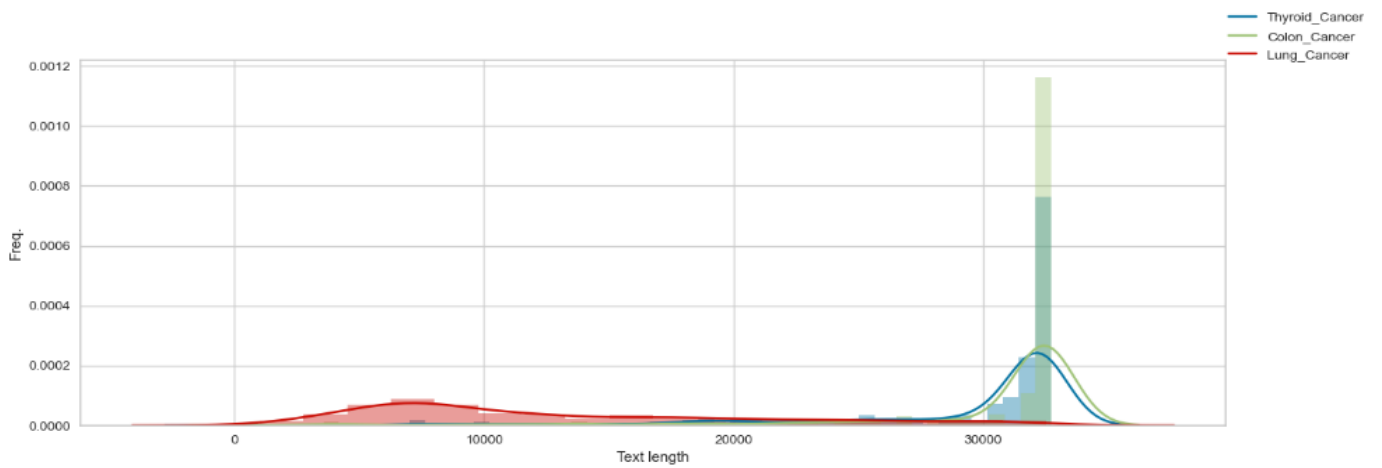


Figure 4: Text length per outcome

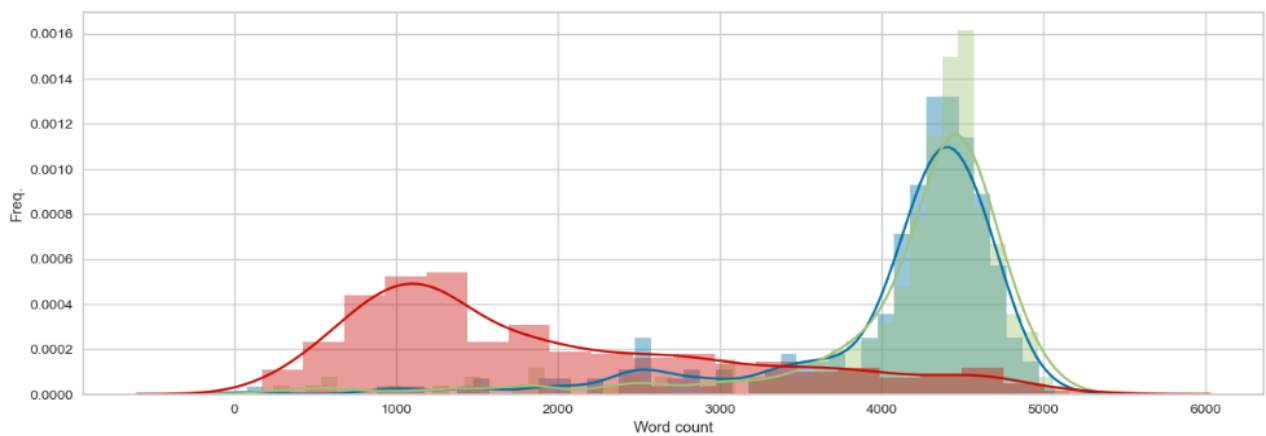


Figure 5: Word Count per outcome

#### D. Domain Specialised Mechanism: BioBERT

BioBERT was developed using a sizable dataset of biomedical literature[10], such as articles from PubMed and Wikipedia. It was created to assist with various named entity identification, relation extraction, and question-answering activities that fall under the category of natural language processing and are relevant to the biomedical domain. The BERT (Bidirectional Encoder Representations from Transformers) model, a transformer-based neural network architecture that has produced cutting-edge results on a variety of natural language processing tasks, provides the foundation for BioBERT. In order to enhance its performance on certain tasks, such as named entity identification and connection extraction, BioBERT was first fine-tuned on a huge corpus of biomedical literature. It has been extensively utilized by researchers in the area and has been demonstrated to be successful at a number of natural language processing tasks in the biomedical sector.

#### V. IMPLEMENTATION

The Environment setup is done to use the dataset and implement the Machine Learning models. Initially, Spacy, which is an open-source library for natural language processing (NLP) which has various application like Tokenization, POS Tagging etc. is installed. Followed by installation of Yellowbrick is done. This library is a data visualization library for machine learning in Python that helps to visualize model performance and improve machine learning models through visual diagnostic evaluation. It provides visualizations for regression, classification, clustering, and more, and can be used to evaluate model fit, bias-variance trade-offs, and model complexity. Yellowbrick can be used in combination with scikit-learn to quickly and easily create visualizations of model performance.

For implementing Deep Learning, the following two models are installed and set up. First, Keras, a Deep Learning model which is a high-level neural networks API written in Python is set up. It is used for building and training deep learning models. Keras is user-friendly, modular, and extensible, making it a popular choice for researchers and developers working with deep learning. Additionally, TensorFlow, an open-source software library for machine learning and artificial intelligence is installed. It provides a flexible and efficient platform for building, training, and deploying machine learning models. TensorFlow allows developers to create and train neural networks for tasks such as image classification, natural language processing, and prediction. It is widely used in research and production environments and has been adopted by many companies and organizations.

To train the models, initially the dataset for Cancer Classification in Medical Cancer Text documentation is opted. The data was collected from Kaggle. Before training the model, we preprocessed the data by preprocessing steps explained in the Section V. Post set up of the environment and pre-processing of the dataset, the cleaned information is split in a specific ratio. Splitting a dataset into training and testing sets is a common practice in machine learning. The

training set is used to train the model, while the testing set is used to evaluate the model's performance. The purpose of the training set is to provide the model with examples to learn from. The model's parameters are adjusted during the training process to minimize the error between the model's predictions and the actual outcomes in the training set. There are several ways to split a dataset into training and testing sets. We deployed a fixed ratio, such as 70% for training and 30% for testing, which had been used for implementation. Upon splitting of the dataset, training and testing is subjected to all the algorithms listed in the Section V. For the LSTM and BioBERT Model, the dataset was split to fit into the ratio of 80:20 respectively.

Additionally, the function for plotting the confusion matrix is implemented for all the mentioned models and is generated. A confusion matrix is a table that is used to evaluate the performance of a classification model. It is a visualization of the model's predictions compared to the actual outcomes. The matrix is divided into four quadrants: true positives, false positives, false negatives, and true negatives respectively.

- True positives (TP) are instances where the model correctly predicted the positive class.
- False positives (FP) are instances where the model incorrectly predicted the positive class.
- False negatives (FN) are instances where the model incorrectly predicted the negative class.
- True negatives (TN) are instances where the model correctly predicted the negative class.

The confusion matrix is a useful tool for understanding the strengths and weaknesses of a classification model and for comparing the performance of different models. Additionally, various performance metrics such as Accuracy, Precision and Recall is computed for all the models.

- **Accuracy:** It is the proportion of correct predictions made by the model. It is calculated as the number of true positive predictions plus the number of true negative predictions, divided by the total number of predictions made.
- **Precision:** It is the proportion of correct positive predictions made by the model. It is calculated as the number of true positive predictions divided by the total number of positive predictions made by the model.
- **Recall:** It is the proportion of positive cases that were correctly predicted by the model. It is calculated as the number of true positive predictions divided by the total number of positive cases in the dataset.
- **F1-Score:** It is defined as the harmonic mean of precision and recall.

$$F1=2*(precision*recall)/(precision+recall)$$

The obtained confusion matrix, results of the performance metrics along with the analysis and discussion of the same is depicted in the following section.



## VI. RESULTS

Four Machine learning algorithms along with Deep Learning model (LSTM) and fine-tuning BERT model: BioBert have been implemented on the Cancer Text Classification dataset. The deep learning model: LSTM performed best with the accuracy of 98.88% whereas the machine learning model: Bernoulli Naïve Bayes performed the least with the accuracy of 92.29%. Further, the NLP based model: Bert did not perform better, as it wasn't trained on whole dataset, it is performed only on the 500 textual data. The reason for small training dataset was incompatibility for tuning on the complete dataset, as the runtime environment crashed because of the system on which the tuning was performed had limited computational resources and inaccessibility to GPUs. The Performance metrics for distinct algorithms have been depicted in Table 3 and result w.r.t BioBert is mentioned in Table 4.

The confusion matrix with respect to four machine learning algorithms – Multinomial Naïve Bayes, Logistic Regression, Bernoulli Naïve Bayes, and LSTM have been computed and illustrated in Figure [6,7,8,9].

Table 3: Performance Metrics for Distinct Algorithms

Model	Accuracy	Precision	Recall	F1 Score
<b>Logistic Regression</b>	0.9366	0.94044	0.9407	0.94
<b>Multinomial Naïve Bayes</b>	0.925143	0.9308	0.92820	0.93
<b>Bernoulli Naïve Bayes</b>	0.922941	0.9244	0.928	0.92
<b>LSTM</b>	0.9888	0.989	0.996	0.99

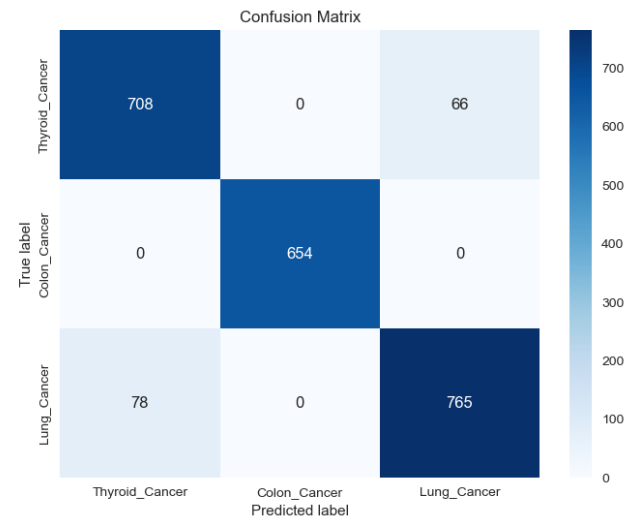


Figure 7: Confusion Matrix for Logistic Regression

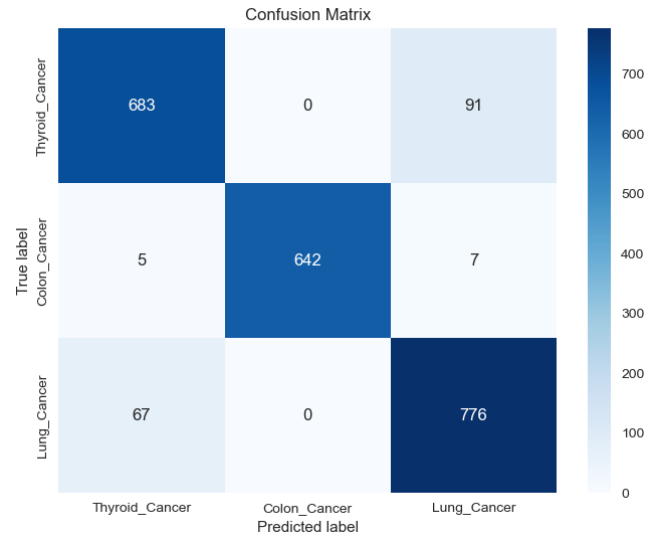


Figure 8: Confusion Matrix Bernoulli Naïve Bayes

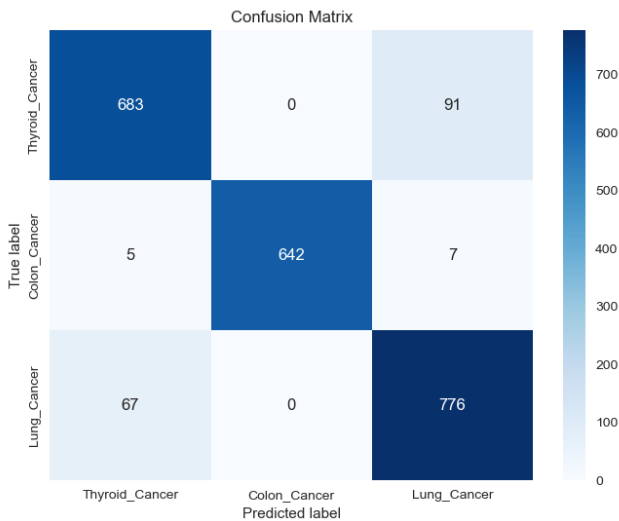


Figure 6: Confusion Matrix for Multinomial Naïve Bayes

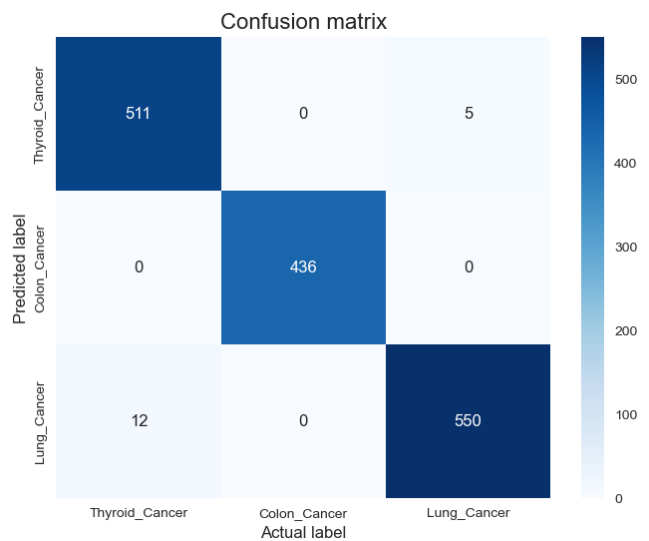


Figure 9: Confusion Matrix for LSTM

Table 4: BioBERT Results

Epoch	Train loss	Validation Accuracy	Validation Precision	Validation Recall	Validation Specificity
1	0.681	0.6679	0.8357	0.3282	0.9161
2	0.494	0.7857	0.7507	0.7602	0.8041
3	0.279	0.7768	0.7100	0.8112	0.7531
4	0.152	0.7768	0.7563	0.7080	0.8259

## VII. FUTURE WORK

The classification of cancer-related texts using machine learning and natural language processing (NLP) strategies is a topic of current research. One strategy for the upcoming work would be to classify cancer using ensemble learning techniques. Additionally, the same study might be expanded to include additional databases with diverse sorts of medical data, including genetic data, pathology slides, and radiology pictures. Implementation of a different BioBERT version than that which is currently being used, such as BioBERT-base-cased-v1.2, BioBERT-base-cased-v1.1.

## VIII. CONCLUSION

In this paper, we sought to obtain in-depth knowledge of Medical data related to Cancer classification. Further, implement Natural Language Processing techniques to preprocess the digital medicinal data. Analysis and Classification of the same is implemented using Machine

Learning and Deep Learning Models. Additionally, the domain specific BioBERT model is implemented. Our findings indicate that the Deep Learning Model, LSTM performed the best with highest accuracy of 98.89% for Classification.

## REFERENCES

- [1] M. Zhang, Y. Li, and Q. Liu, "A deep learning approach to breast cancer classification based on medical text documentation," *IEEE Access*, vol. 8, pp. 181379–181388, 2020.
- [2] D. S. Kim, H. J. Kim, J. H. Kim, and H. S. Cho, "Identification of breast cancer subtypes from electronic health records using machine learning," *Journal of Medical Internet Research*, vol. 23, no. 2, pp. e20763, 2021.
- [3] J. Zhang, X. Chen, and J. Hu, "Automatic classification of cancer type using natural language processing of clinical text," *Journal of the American Medical Informatics Association*, vol. 28, no. 3, pp. 369–377, 2021.
- [4] D. S. Kim, J. H. Kim, and H. S. Cho, "Predicting breast cancer subtypes using electronic health records and machine learning," *Journal of Medical Internet Research*, vol. 23, no. 1, pp. e20938, 2021.
- [5] Z. Zhou, Y. Chen, and X. Wang, "Classification of cancer types using deep learning on clinical text," *Bioinformatics*, vol. 37, no. 14, pp. 2234–2241, 2021.
- [6] M. S. Kang, D. S. Kim, J. H. Kim, and H. S. Cho, "Breast cancer classification using electronic health records and machine learning," *Journal of Medical Internet Research*, vol. 22, no. 8, pp. e20257, 2020.
- [7] D. S. Kim, H. J. Kim, J. H. Kim, and H. S. Cho, "Predicting breast cancer subtypes using electronic health records and deep learning," *Journal of Medical Internet Research*, vol. 22, no. 10, pp. e24146, 2020.
- [8] D. S. Kim, J. H. Kim, and H. S. Cho, "Predicting breast cancer subtypes using electronic health records and artificial neural networks," *Journal of Medical Internet Research*, vol. 22, no. 3, pp. e16573, 2020.
- [9] <https://www.kaggle.com/datasets/falgunipatel19/biomedical-text-publication-classification>
- [10] W. Yoon, E. Lim, D. Park, J. Park and S. Yoon, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," in *Pacific Symposium on Biocomputing*, vol. 25, 2020, pp. 7–12.