ISSN (e): 2250 – 3005 || Volume, 14 || Issue, 5|| Sep. - Oct. – 2024 ||
International Journal of Computational Engineering Research (IJCER)

Lung Cancer Detection Using Machine Learning

Khush Garewal
Department of Computer Science & Engineering
Shriram Institute of Technology, Jabalpur, India

Khushi Agrahari
Department of Computer Science & Engineering
Shriram Institute of Technology, Jabalpur, India

Nikhil Chandwani
Department of Computer Science & Engineering
Shriram Institute of Technology, Jabalpur, India

Nikita Jhariya
Department of Computer Science & Engineering
Shriram Institute of Technology, Jabalpur, India

Prof. Rajendra Arakh
Department of Computer Science & Engineering
Shriram Institute of Technology, Jabalpur, India

Abstract—*we address the variegated modalities in which AI assists in the detection of cancer, outlining the radiological imaging to the genetic sequencing. We go over the fact that AI solutions is predicting accurately and timely cancer detection through early stages and is offering early requirements which can result in better treatment results. Not only do we provide potential answers to the challenges of applying AI systems to cancer detection (e.g. data quality, interpret-ability, and regulatory aspects), but we also discuss the ways to fine tune AI functionalities. To sum up, the abstract standing out is the transitory power of AI enabling its application for cancer detection, as the author provides both the current stage and future indications.*

Date of Submission: 28-08-2024 Date of acceptance: 07-09-2024

I. INTRODUCTION

Unfortunately, cancer is still one of the most complex health issues that the world faces, and early diagnosis is a precondition for effective treatment, which is why it is important to improve the quality of cancer patient outcomes. Recently, the development technologies, especially Convolution neural networks, advanced in the cancer detection field by providing highly accurate and fast imaging data analysis in a new approach. Over the last few years deep learning neural network have come into prominence as tools capable of spotting even the minutest deviations that are suggestive of cancer and brings in imaging modalities like X-ray, mammogram MRIs and CT scans [**Error! Reference source not found.**-[1]]. This passage explicates the use of neural networks by breast cancer detection, particularly touch upon the innovative aspect of their carefully organized work ow which promise to revolutionize the early diagnosis, increase the survival rate and transformation of the patient-centred services.

II. BACKGROUND

Lung cancer continues to be a deadly type of cancer, on a scale. Despite progress in technology and treatment choices early detection remains crucial for enhancing outcomes and survival rates. The rise of machine learning (ML) methods in times has opened up possibilities, for improving early detection techniques. By examining sets of imaging scans, genetic information and patient records ML algorithms can pinpoint patterns and signs of lung cancer with remarkable accuracy. These opening lays the groundwork for exploring how ML is

reshaping the landscape of lung cancer detection. Through the use of algorithms and thorough data analysis ML has the potential to revolutionize screening procedures leading to diagnoses tailored treatment strategies and ultimately better patient outlooks.

Furthermore, as we inhale, the air travels through various parts of our body, including the nasal or oral cavities, the pharynx, the larynx, the trachea, and the bronchi, ultimately reaching the lungs where it reaches the alveoli. These tiny sacs are led with capillaries that exchange carbon dioxide for oxygen. Breathing is a continuous process for humans as our lungs play a crucial role in supplying oxygen to our blood, which is essential for sustaining life.

III. METHODS AND MATERIALS

- *Data Collection***:** Extract information from credible sources such as academic databases, research centers, and hospitals. Please, make the data labeled only with the most relevant types of cancer and the attributes that characterize them.
- *Data Cleaning:* To remove redundant or irrelevant information, do the purge amendment. Address missing values by imputation, or deletion as appropriate to prevent introducing any bias.
- *Data Transformation***:** Apply transformations including the log transformations for non-regular data distribution to get a cleaner look.
- *Handling Imbalanced Data:* If the dataset is imbalanced (for ex. fewer examples of malignant tumours than benign tumours), the use of techniques like over-sampling, under-sampling or SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes is recommended.
- *Outlier Detection and Removal:* Identify and cope with outliers that can decrease the model efficiency using Z-score, IQR (interquartile range) or any other method.
- *Normalization:* Normalize the data distribution such that for each characteristic data mean is zero and the standard deviation is one, or use the other normalization techniques as and when the algorithmic conditions get required.
- *Validation:* Validate, by in the process, investigate summary statistics, visualizations, and the data is consistent with the rest of the analysis.

A. Data Preprocessing

Data preprocessing involves eliminating null values and duplicate entries to enhance data quality and reliability. This step ensures that the dataset is suitable for analysis or machine learning tasks, minimizing biases and inaccuracies in the results. By removing null values, missing data is addressed, while eliminating duplicates prevents redundancy and ensures a more representative dataset. This process is essential for obtaining accurate insights and training effective models.

B. Process Flow and Proposed Methodology

Here is a flowchart describing a machine learning process for cancer prediction. Here's a breakdown of the steps:

- *Dataset Collection:* The initial step involves gathering data which will be used for prediction.
- *Data Preprocessing*: This includes cleaning and transforming the data to ensure it's in the right format for analysis.
- *Feature Selection*: Selecting the most relevant attributes from the data to use for prediction.
- *Model Training:* Using algorithms like KNN (K-Nearest Neighbour) and DT (Decision Trees) to train the machine learning model.
- *Cancer Prediction*: The trained model uses new patient data to predict cancer.
- *Report Generation*: Finally, a report is generated detailing the prediction results.

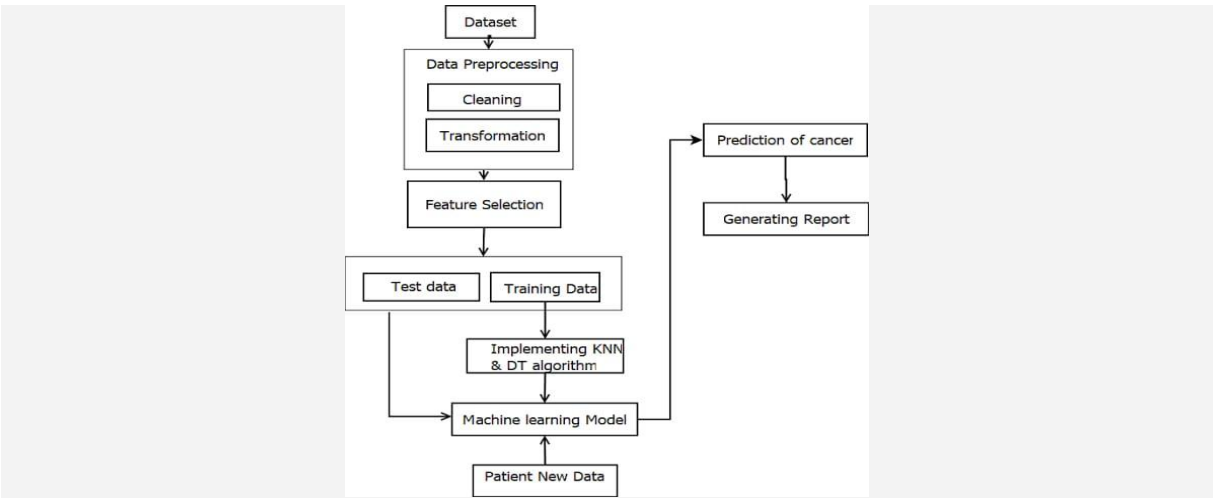

Fig. 1. Process Flow Diagram

IV. RESULT

In the context of this research work, plenty of machine learning models, such as logistic regression, decision tree model, KNN, Gaussian naive bayes, multinominal naive bayes, support vector machine, Random Forest, XG Boost, multi-layer perceptron, gradient boost model are evaluated in terms of accuracy, precision, recall and F-Measure in order to determine the model with the best predictive performance. The best performance is achieved by the Gradient boost model, multilayer perceptron, random forest, support vector classifier model. It presents accuracy, precision, recall and F-Measure equal to 98%.

a) TABLE FOR BEST ACCURACY

In Table, model's comparisons in terms of accuracy, recall and precision are made. The authors in the research work used dataset with the same number of features as us. SVM, Gradient boost model, multilayer perceptron, random forest achieves the highest performance among our proposed models, with accuracy, recall, and precision all reaching 98%.

Models Used	***Best Accuracy***
Logistic Regression	97%
Decision tree	94%
KNN	96%
Gaussian Naive Bayes	92%
Multinomial Naive Bayes	81%
Support vector classifier	98%
Random Forest	98%
XGBoost	97%
Multi-layer perception	98%
Gradient Boost	98%

Fig. 2. Table for Best Accuracy

b) TABLE FOR AVERAGE ACCURACY

The methodology proposed in this study relies on a dataset comprising features that capture human habits, such as smoking and alcohol consumption, along with signs and symptoms typically associated with lung cancer patients. However, these signs may not necessarily be directly linked to lung cancer disease, as revealed by the feature analysis in Section 3.3 of the Materials and Methods. Unlike some other cancers, lung cancer is not visible to the naked eye, and its symptoms often overlap with those of other diseases. Common symptoms include allergies, asthma, shortness of breath, and coughing [33]. In this study, we chose to train several classifiers using various risk factors associated with such symptoms. This approach enables us to accurately identify the class label (Lung Cancer or Non-Lung Cancer) of an unknown instance and consequently assess the associated risk

Models Used	***Average Accuracy***
Logistic Regression	0.9288120567375886
Decision tree	0.9227393617021278
KNN	0.9184397163120567
Gaussian Naive Bayes	0.8870124113475178
Multinomial Naive Bayes	0.7572251773049644
Support vector classifier	0.9476063829787235
Random Forest	0.9456560283687944
XGBoost	0.9457446808510639
Multi-layer perception	0.93927304964539
Gradient Boost	0.947695035460993

Fig. 3. Table for Average Accuracy

V. CONCLUSION.

The lungs serve as the primary organs of respiration, ensuring humans continuously receive oxygen, crucial for sustaining life until death. Lung cancer stands as the leading cause of mortality among both sexes, with a patient's lifespan contingent upon the cancer's stage of advancement. Early diagnosis correlates with increased life expectany

In this research endeavour, we employ supervised learning techniques to construct models for identifying individuals exhibiting symptoms indicative of lung cancer. We evaluate various machine learning models, encompassing Gaussian naive Bayes, multinomial naive Bayes, support vector machines, random forests, XGBoost, multi-layer perceptron, and gradient boosting models, assessing their performance in terms of accuracy, precision, recall, and F-Measure.

Concluding our results and discussion section, it's imperative to acknowledge a limitation of our study. Our research relied on a publicly available dataset [39], rather than data sourced directly from a hospital unit or institute, which could have provided a more diverse range of characteristics. Furthermore, obtaining access to sensitive medical data presents challenges due to privacy concerns. Nonetheless, the dataset we utilized contained beneficial features, enabling us to derive reliable and accurate research outcomes.

REFERENCE

[1]. Dritsas, E.; Trigka, M. Lung Cancer Risk Prediction with Machine Learning Models. Big Data Cogn. Comput. 2022,6 ,
[2]. Patra, R. Prediction of lung cancer using machine learning classifier. In Proceedings of the International Conference on Science, Communication and Security, Gujarat, India, 26–27 March 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 132–142.
[3]. Lung Cancer Prediction Dataset. Available online: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer
[4]. Maldonado, S.; López, J.; Vairetti, C. An alternative SMOTE oversampling strategy for high-dimensional datasets. Appl. Soft Comput. 2019, 76, 380–389.
[5]. Gnanambal, S.; Thangaraj, M.; Meenatchi, V.; Gayathri, V. Classification algorithms with attribute selection: An evaluation study using WEKA. Int. J. Adv. Netw. Appl. 2018, 9, 3640–3644.
[6]. Darst, B.F.; Malecki, K.C.; Engelman, C.D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. 2018, 19, 1–6.