VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

MINI PROJECT-2 REPORT

On

Lung Cancer Prediction Using Machine Learning

Submitted by

Advithi D(1BM21CS009)
Anagha K S(1BM21CS021)
Dhanush H V(1BM21CS052)
Gagandeep Kattennanavar(1BM21CS064)

Under the Guidance of
Dr. Seema Patil
Assistant Professor, BMSCE

in partial fulfillment for the award of the degree of
BACHELOR OF ENGINEERING

in
COMPUTER SCIENCE AND ENGINEERING

B. M. S. COLLEGE OF ENGINEERING
(Autonomous Institution under VTU)
BENGALURU-560019

March 2024 to June 2024

B. M. S. College of Engineering,

Bull Temple Road, Bangalore 560019

(Affiliated To Visvesvaraya Technological University, Belgaum)

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project work entitled "Lung Cancer Prediction using Machine Learning" carried out by Advithi D(1BM21CS009), Anagha K S(1BM21CS021), Dhanush HV(1BM21CS052) and Gagandeep Kattennanavar(1BM21CS064) who are bonafide students of **B. M. S. College of Engineering.** It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visveswaraiah Technological University, Belgaum during the year 2023-2024. The project report has been approved as it satisfies the academic requirements in respect of **Mini Project-2 (22CS6PWMP2)** work prescribed for the said degree.

Signature of the Guide
Dr. Seema Patil
Assistant Professor
BMSCE, Bengaluru

Signature of the HOD
Dr. Jyothi S Nayak
Professor & Head, Dept. of CSE
BMSCE, Bengaluru

External Viva

Name of the Examiner Signature with date

1._____

2. _____

B. M. S. COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECALARATION

We, Advithi D (1BM21CS009), Anagha K S (1BM21CS021), Dhanush H V (1BM21CS052) and Gagandeep Kattennanavar (1BM21CS064) students of 5th Semester, B.E, Department of Computer Science and Engineering, B. M. S. College of Engineering, Bangalore, here by declare that, this Project Work-2 entitled " Lung Cancer Prediction using Machine Learning " has been carried out by us under the guidance of Dr. Seema Patil, Assistant Professor, Department of CSE, B. M. S. College of Engineering, Bangalore during the academic semester March-June 2024.

We also declare that to the best of our knowledge and belief, the development reported here is not from part of any other report by any other students.

Signature

Advithi D(1BM21CS009)

Anagha K S(1BM21CS021)

Dhanush H V(1BM21CS052)

Gagandeep Kattennanavar (1BM21CS064)

Abstract

The fact that lung cancer is still the primary cause of cancer-related death worldwide highlights the critical need for efficient diagnostic instruments for early identification and intervention. This study presents a unique method of lung cancer detection based on support vector machines (SVM) and k-Nearest Neighbors (k-NN) machine learning algorithms. Preprocessing lung cancer datasets to manage missing values, standardize data, and identify pertinent features are all part of the methodology.

Because it is easy to use and gives a baseline performance for initial classification, the k-NN algorithm is used. Then, to improve classification accuracy, the SVM algorithm—which is well-known for its resilience in handling high-dimensional data and modelling intricate decision boundaries—is applied. Both models' performances are assessed using metrics including F1-score, recall, accuracy, and precision.

The advantages and disadvantages of each algorithm in the context of lung cancer detection are highlighted by a comparative analysis. The findings show that combining k-NN and SVM greatly improves early detection skills, which may result in better patient outcomes by enabling prompt medical intervention. This effort lays the foundation for future study and advancement in automated cancer detection systems while highlighting the revolutionary potential of machine learning in improving medical diagnostics.

Chapter 1

1 Introduction

This study focuses on applying machine learning approaches to predict lung cancer early. One of the most common and deadly types of cancer in the world, lung cancer is frequently detected at an advanced stage, when there are few treatment choices and the chance of survival is greatly decreased. Improving patient outcomes and lowering associated death rates are directly correlated with early identification.

The potential for this issue to save lives by enabling prompt interventions and therapies makes it important to solve. Precisely recognizing those who are most likely to acquire lung cancer allows medical professionals to put preventative measures into place, do routine exams, and provide the right treatment, all of which improve survival and prognostic rates.

Our strategy is based on using machine learning algorithms to examine relevant clinical data and identify trends suggestive of the emergence of lung cancer. This covers a wide range of information, including genetic predisposition, medical records, smoking history, patient demographics, and results from diagnostic testing such as biomarker assays and imaging examinations. Our objective is to build a prediction model that can correctly classify people into high-risk and low-risk categories by training the algorithm on an extensive dataset that includes both healthy persons and lung cancer patients.

Our study's anticipated results and conclusions include an assessment of our prediction model's performance parameters, such as accuracy, precision, recall, and F1-score. We also intend to talk about possible research limits, directions for further investigation, and the therapeutic implications of our results. Overall, we anticipate that our method will be a useful resource for medical practitioners in identifying people who have a higher risk of developing lung cancer and in efficiently directing customized screening and intervention plans.

1.1 Motivation

Lung cancer is a major worldwide health concern that caused over 9.6 million deaths in 2018 alone. 85% of cases had smoking as a significant contributing factor, highlighting the vital need for early identification to increase survival chances. Late-stage diagnoses still have an adverse effect on patient outcomes even with improvements in therapy. Our goal is to create a prediction model for lung cancer risk assessment that is powered by machine learning.We aim to identify patients who are at heightened risk as soon as possible by looking at parameters including blood pressure, cholesterol levels, age, gender, diabetes status, and pulse rates. This strategy makes it easier to implement prompt treatments and individualized care plans, which might result in significant gains in early diagnosis, treatment effectiveness, patient mortality, and overall quality of life. This program is in line with international efforts to fight lung cancer from all angles.

1.2 Scope of the Project

The goal of this project is to utilize machine learning algorithms to evaluate data from Kaggle and forecast an individual's risk of developing lung cancer. Comprehensive activities including feature selection, data collection, preprocessing, model training, assessment, and validation are all included in the scope. The prediction model will be created to categorize people into high- or low-risk groups according to a variety of criteria, such as age, signs of obesity, lifestyle variables (including past smoking history), and results of diagnostic tests. Delivering a dependable and expandable early lung cancer prediction system that works for a range of patient demographics is the aim.

1.3 Problem statement

Using a Kaggle dataset, create a machine learning model to forecast the likelihood that a person would acquire lung cancer. A mix of genetic traits, medical records, smoking history, and demographic data will be used in this model. Improving early detection skills is the aim, which might lead to better patient outcomes.

Chapter 2

2 Literature Survey

A research evaluating many algorithms utilizing datasets like OASIS and BRATS was published in [1]. Methods such as logistic regression, random forest, SVM, ANN, naïve Bayes, and decision trees were evaluated. The results showed that ensemble approaches performed better than single classifiers, suggesting that mixing several algorithms can improve detection accuracy.

Research in [2] examined random forest, SVM, ANN, and naïve Bayes using data from Kaggle. SVM demonstrated its promise for early lung cancer prediction and its prospective uses in clinical diagnosis and therapy by achieving the best accuracy of 99.6%.

The work described in [3] analyzed SVM, k-nearest neighbor, random forest, ANN, and a voting classifier using the LIDC-IDRI and data.world datasets. It came to the conclusion that SVM might not be appropriate for big, noisy datasets, highlighting how crucial it is to choose algorithms depending on the features of the dataset.

Earlier studies, including [4], combined machine learning and image processing with the LIDC-IDRI dataset. It created a strong detection system by merging autoencoder systems, the OTSU algorithm, decision trees, and CNNs, demonstrating the value of multidisciplinary methods.

In order to forecast the risk of lung cancer, researchers in [5] used data from prefekt@nvs.ki.se and used the SGB model. The efficacy of the SGB model was shown across different smoking histories, highlighting the potential of machine learning to use symptom questionnaires for early identification.

In order to assess SVR, LSTM, and backpropagation, data from ten European nations spanning 42 years were used in the study covered in [6]. With its precise prediction of lung cancer incidence rates

and development of predictive models for public health policies, SVR emerged as the top performer.

Using data from Kaggle, the study in [7] examined SVM, KNN, MLP, and RBF. With an accuracy of 90%, MLP demonstrated the potential of neural networks in the prediction of medical data and emphasized the significance of feature engineering in improving predictive performance.

Research used RF, KNN, Decision Tree, LR, and Naïve Bayes to evaluate Kaggle data, as stated in [8]. With an accuracy rate of 88.5%, Random Forest showed the highest efficacy and proved the value of ensemble approaches in the early identification of lung cancer.

SVM, ANN, Naïve Bayes, and LR were examined in a survey conducted in [9] on extensive picture datasets. The predictive model that was created marked a substantial breakthrough in the field of medical image analysis and helped identify and diagnose lung cancer early on.

The NLST dataset and the LCP-CNN approach were used in a study by [10] to obtain excellent results in the identification of benign nodules. This demonstrates how deep learning can be used to use massive datasets for precise lung cancer detection.

Chapter 3

3 Design

3.1 High Level Design

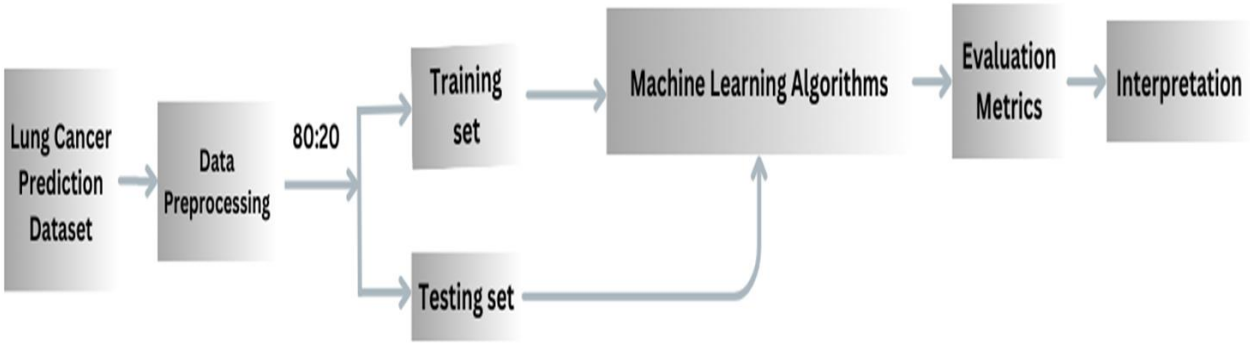


Fig 3.1. High-level design of the lung cancer prediction model

The flowchart illustrates a high-level design for predicting lung cancer. It begins with acquiring a lung cancer prediction dataset, followed by data preprocessing to clean and prepare the data for analysis. The dataset is then split into two subsets: 80% for training and 20% for testing. The training set is used to train various machine learning algorithms, enabling the models to learn patterns and make predictions. The trained models are evaluated using the testing set, with evaluation metrics applied to assess their performance and accuracy. Finally, the results are interpreted to provide insights and support decision-making in lung cancer prediction. This systematic approach ensures the development of robust and reliable predictive models for lung cancer.

3.2 Detailed Design

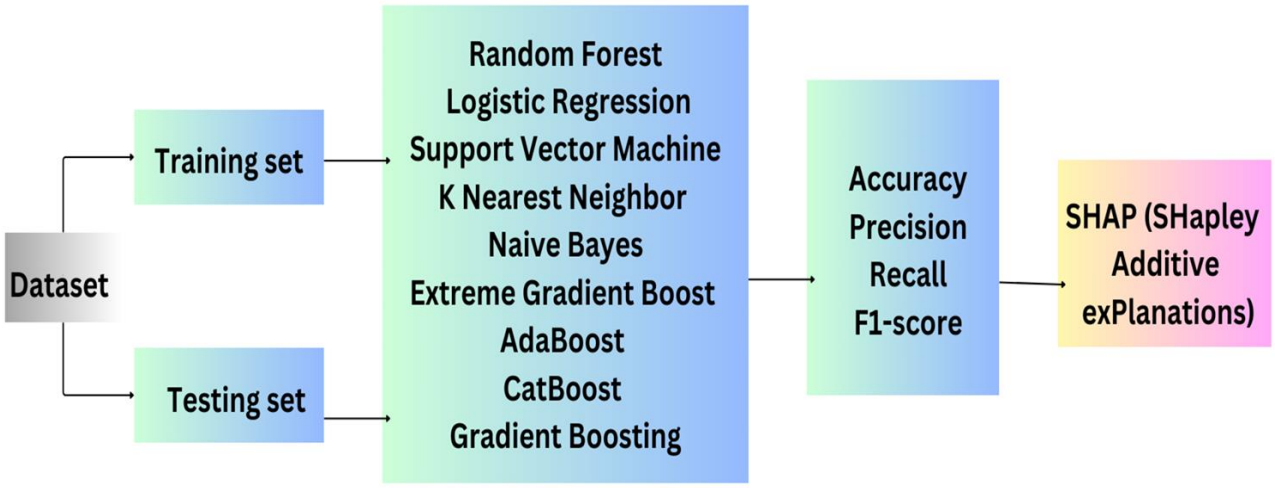


Fig 3.2 Detailed design of the lung cancer prediction model

This flowchart details a machine learning workflow for evaluating various models on a given dataset. The dataset is first divided into a training set and a testing set. The training set is used to train multiple machine learning algorithms, including Random Forest, Logistic Regression, Support Vector Machine, K Nearest Neighbor, Naive Bayes, Extreme Gradient Boost, AdaBoost, CatBoost, and Gradient Boosting. The performance of these models is then evaluated using metrics such as accuracy, precision, recall, and F1-score. Finally, SHAP (Shapley Additive Explanations) is employed to interpret the results, providing insights into the contribution of each feature to the predictions made by the models. This comprehensive approach ensures a thorough evaluation and interpretation of the models' performance.

Chapter 4

4 Implementation

4.1 Proposed methodology

The Kaggle Lung Cancer Prediction Dataset has 26 columns and 1000 rows; the 'index' column is removed during import. Preprocessing verified there were no duplicates or missing values. Target attribute imbalance was corrected by SMOTE after Label Encoder was used to convert categorical columns to numerical values. Using thresholds, outliers were located and replaced. With 'Level' as the goal, the dataset was divided into 80% training and 20% test sets.

Two algorithms were selected to pick the models:

1. K-Nearest Neighbors (KNN): Data points are classified using this supervised learning technique by being given the class label that appears the most often among their k nearest neighbors in the featurespace.
2. Support Vector Machine (SVM): SVM searches the feature space for the ideal hyperplane that best divides classes with the largest margin. By applying kernel techniques, it can handle non-linear data.

Evaluation Metrics:

F1 score, Accuracy, Precision, and Recall were used to gauge performance. Shapley Additive Explanations, or SHAP for short, was used to help in decision-making by helping to comprehend the contributions of features to model predictions.

Streamlit:

To improve the user interface and make input entering and stroke level prediction easier, Streamlit was used.

4.2 Algorithm used for implementation

• K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) are the two primary machine learning techniques we used to predict the risk of lung cancer. KNN is a simple yet powerful supervised learning algorithm that is well suited for our classification assignment. It classifies data points by determining the majority class among their k nearest neighbors in the feature space.

• SVM, on the other hand, is a strong algorithm that chooses the best hyperplane to divide classes with the biggest margin. Additionally, by employing clever kernel techniques, it efficiently handles non-linear data, improving its prediction power. To make sure both algorithms were successful in predicting the risk of lung cancer, they were assessed using measures including accuracy, precision, recall, and F1 score.

4.3 Tools and technologies used

Hardware:

- **RAM**: At least 16 GB of RAM was required to manage large datasets and complex computations.
- **CPU**: A multi-core processor with a minimum of 4 cores and 8 threads, such as the Intel Core i7 or AMD Ryzen 7, was used to enhance processing speed and performance.
- **Storage**: An SSD with at least 256 GB capacity was essential for quick data retrieval and storage capabilities.
- **Software:**

  • **Jupyter Notebook**: This interactive environment, which smoothly integrates code, visuals, and documentation, made it easier to design and test machine learning models.
  • **Streamlit Framework**: Streamlit was used to develop an intuitive online application that lets users enter information and get instantaneous risk estimates for lung cancer.
  • **Additional Python Libraries**: Pandas for data handling, NumPy for numerical

computations, and Scikit-Learn for machine learning method implementation were crucial Python libraries for data manipulation, analysis, and model construction.

4.4 Testing

The testing process included:

- **Data Splitting**: The data was divided into 80% for training and 20% for testing.
- **Model Evaluation**: Predictions on the test set were evaluated using metrics such as accuracy, precision, recall, and F1 score.
- **Performance Metrics**: Recall and precision measured true positive rates and decreased false negatives, accuracy evaluated overall correctness, and the F1 score offered a balance between the two.
- **Comparison with Training Performance**: In order to detect overfitting and confirm the model's ability to forecast fresh data, the test results were compared with the training performance. The model's potential for useful lung cancer risk prediction was confirmed by this technique.

Chapter 5

5 Results and Discussion

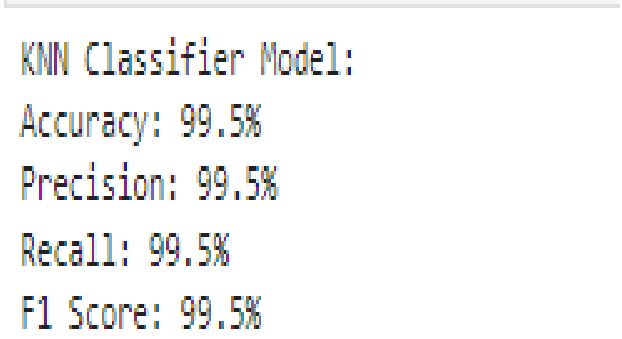


Fig 5.1. Result for KNN model

The above figure shows the output for the K-Nearest Neighbors model which includes the model's accuracy, precision, recall and F1-score.

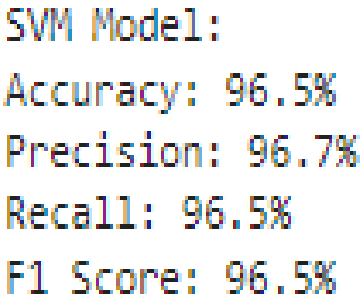


Fig 5.2. Result for SVM model

The above figure shows the output for the Support vector machine model which includes the model's accuracy, precision, recall and F1-score. These evaluation metrics help us differentiate between the best performing model.

SHAP:

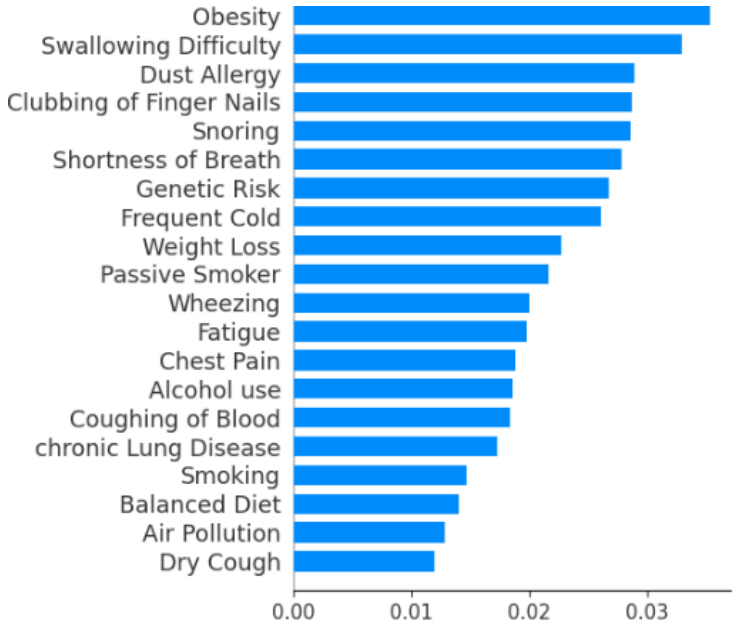


Fig 5.3. SHAP for KNN model

The above figure shows Shapley Additive Explanations which illustrates the attributes playing major role in predicting lung cancer using K-Nearest Neighbors.

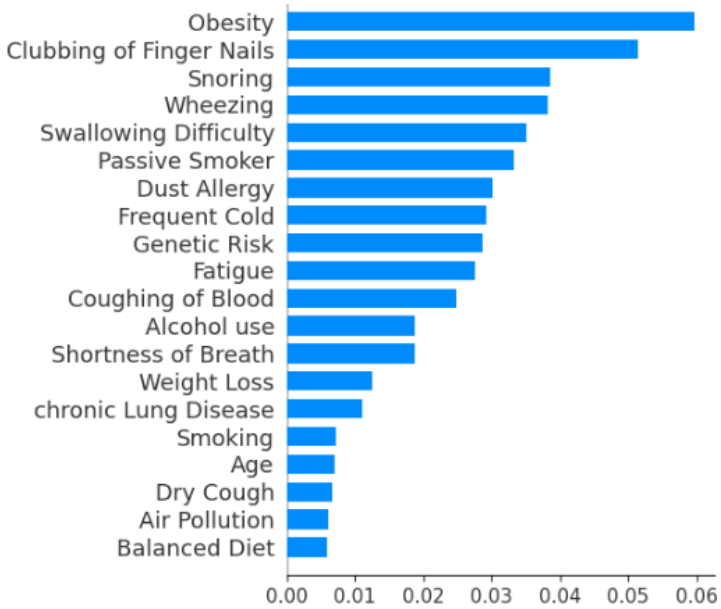


Fig 5.4 SHAP for SVM model

The above figure shows Shapley Additive Explanations which helps us improve the interpretability of the model by illustrating which attributes playing major role in predicting lung cancer using Support Vector Machine.

Chapter 6

6 Conclusion and Future Work

The significance of addressing imbalances and preparing data are among the study's main conclusions. Class imbalance was successfully handled by the use of SMOTE, and preprocessing procedures guaranteed a reliable dataset for model training. Support vector machines (SVM) and K-Nearest Neighbors (k-NN) shown to be effective in predicting the risk of lung cancer with respectable recall, accuracy, and F1 scores.

The application of SHAP values greatly improved the interpretability of the model and gave distinct insights into the significance of the features. In order to foster clinical trust and understanding, this honesty is crucial. To improve prediction performance, more machine learning algorithms like Random Forest, XGBoost, and neural networks should be investigated in future studies.

It is planned to enhance the Streamlit interface with more user-friendly input techniques and visuals. Using real-time data processing and other feature engineering approaches might provide more pertinent features and give up-to-date risk evaluations. As a result, the instrument would be more useful in clinical settings.

The very limited dataset, the inability of existing models to properly capture complicated data patterns, and the need for improvements to the user interface are the main drawbacks. It is advised to increase the dataset, experiment with deep learning architectures and ensemble techniques, and enhance the user interface. The purpose of these modifications is to increase the model's usability and resilience in clinical situations.

References:

[1] Stephen, Eali & Joshua, Eali & Bhattacharyya, Debnath & Janarthanan, Midhunchakkaravarthy. (2020). AN EXTENSIVE REVIEW ON LUNG CANCER DETECTION USING MACHINE LEARNING TECHNIQUES. Journal of Critical Reviews. 7. 2020. 10.31838/jcr.07.14.68.

[2] Jenipher, V. & Radhika, S.. (2020). A Study on Early Prediction of Lung Cancer Using Machine Learning Techniques. 911-916. 10.1109/ICISS49785.2020.9316064.

[3] C. Thallam, A. Peruboyina, S. S. T. Raju and N. Sampath, "Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1285-1292, doi: 10.1109/ICECA49313.2020.9297576.

[4] Pawar, Vikul. (2020). Lung Cancer Detection System Using Image Processing and Machine Learning Techniques. International Journal of Advanced Trends in Computer Science and Engineering. 9. 5956-5963. 10.30534/ijatcse/2020/260942020.

[5] Nemlander E, Rosenblad A, Abedi E, Ekman S, Hasselström J, Eriksson LE, Carlsson AC. Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers. PLoS One. 2022 Oct 21;17(10):e0276703.

[6] Nageswaran S, Arunkumar G, Bisht AK, Mewada S, Kumar JNVRS, Jawarneh M, Asenso E. Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. Biomed Res Int. 2022 Aug 22;2022:1755460. doi: 10.1155/2022/1755460. Retraction in: Biomed Res Int. 2024 Jan 9;2024:9851527. doi: 10.1155/2024/9851527. PMID: 36046454; PMCID: PMC9424001.

[7] Omar, Aya & Nassif, Ali. (2023). Lung Cancer Prediction using Machine Learning based Feature Selection: A comparative Study. 1-6. 10.1109/ASET56582.2023.10180436.

[8] S, Bharathy & R, Pavithra & B, Akshaya. (2022). Lung Cancer Detection using Machine Learning. 539-543. 10.1109/ICAAIC53929.2022.9793061.

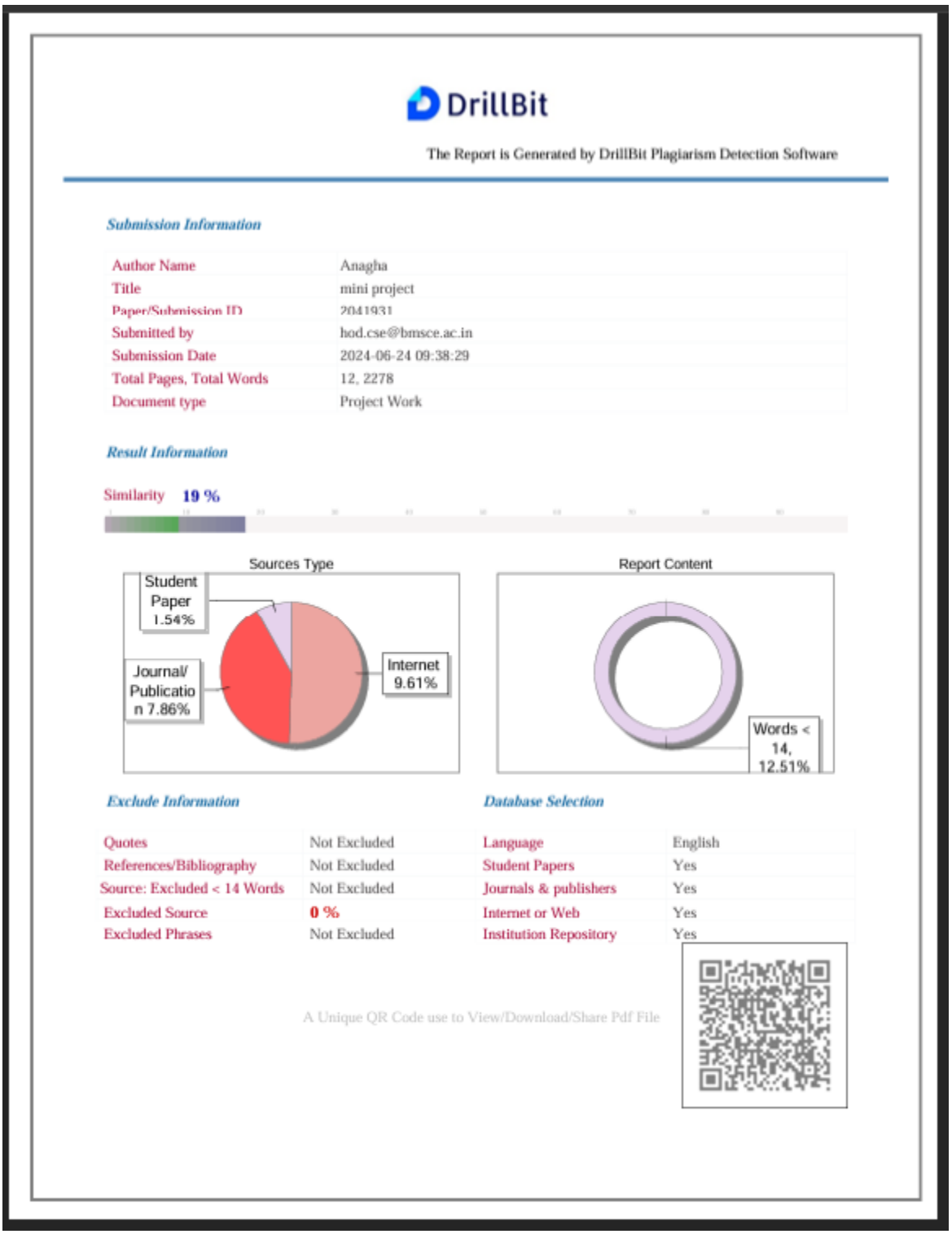
[9] Nageswaran S, Arunkumar G, Bisht AK, Mewada S, Kumar JNVRS, Jawarneh M, Asenso E.

Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. Biomed Res Int. 2022 Aug 22;2022:1755460. doi: 10.1155/2022/1755460. Retraction in: Biomed Res Int. 2024 Jan 9;2024:9851527. doi: 10.1155/2024/9851527. PMID: 36046454; PMCID: PMC9424001.

[10] N. Banerjee and S. Das, "Prediction Lung Cancer– In Machine Learning Perspective," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2020, pp. 1-5, doi: 10.1109/ICCSEA49143.2020.9132913.

APPENDIX:

Screenshot of the Plagiarism Check of the Report using Drillbit:



DrillBit

The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

Author Name	Anagha
Title	mini project
Paper/Submission ID	2041931
Submitted by	hod.cse@bmsce.ac.in
Submission Date	2024-06-24 09:38:29
Total Pages, Total Words	12, 2278
Document type	Project Work

Result Information

Similarity 19 %

Sources Type

Report Content

Exclude Information

Quotes	Not Excluded
References/Bibliography	Not Excluded
Source: Excluded < 14 Words	Not Excluded
Excluded Source	0 %
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

19	31	B
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	Medical Internet of things using machine learning algorithms for lung cancer det by Pradhan-2020	2	Publication
2	REPOSITORY - Submitted to Kalinga University, Raipur on 2024-03-19 15-50	2	Student Paper
3	link.springer.com	1	Internet Data
4	Assessing the lung cancer risk reduction potential of candidate modif, by Hoeng, Julia Maede- 2019	1	Publication
5	www.biorxiv.org	1	Internet Data
6	naac.mictech.edu.in	1	Publication
7	citeseerx.ist.psu.edu	1	Internet Data
8	assets.ctfassets.net	1	Publication
9	eprints.umm.ac.id	1	Internet Data
10	qdoc.tips	1	Internet Data
11	socj.telkomuniversity.ac.id	1	Internet Data
12	www.almabetter.com	1	Internet Data
13	www.nature.com	1	Publication

14	biomedcentral.com	<1	Internet Data
15	eurekaselect.com	<1	Internet Data
16	dochero.tips	<1	Internet Data
17	Innovation for and from emerging countries a closer look at the antecedents of by Giannetti-2019	<1	Publication
18	Multiparameter cell-tracking intrinsic cytometry for single-cell characterizatio by Apichitsopa-2018	<1	Publication
19	www.dx.doi.org	<1	Publication
20	academicjournals.org	<1	Publication
21	Application of Machine Learning Techniques for the Diagnosis of L- www.ijcaonline.org	<1	Publication
22	A biological model for lung cancer risk from 222Rn exposure by Naom-1996	<1	Publication
23	dev.to	<1	Internet Data
24	dokumen.pub	<1	Internet Data
25	mdpi.com	<1	Internet Data
26	RNAAgeCalc A multi-tissue transcriptional age calculator by Ren-2020	<1	Publication
27	scholar.sun.ac.za	<1	Publication
28	The COVID-19 pandemic and global environmental change Emerging research needs by Barouki-2021	<1	Publication
29	www.dx.doi.org	<1	Publication
30	American Institute of Aeronautics and Astronautics AIAA InfotechAer	<1	Publication

31	IEEE 2014 14th UK Workshop on Computational Intelligence (UKCI) - Br by	<1	Publication

Certificate of appreciation:

B.M.S. COLLEGE OF ENGINEERING

(Autonomous College Under VTU)

CERTIFICATE OF APPRECIATION

THIS CERTIFICATE IS PROUDLY PRESENTED TO

Advithi D (1BM21CS009), Anagha K.S (1BM21CS021), Dhanush H.V (1BM21CS052), Gagandeep Kattennana (1BM21CS064)

in recognition of their exceptional performance in their project "**Lung Cancer Prediction Using Machine Learning**" under the guidance of **Dr.Seema Patil** during the academic year 2023-2024.

Namratha M

Prof. Namratha M
MINI PROJECT CO-ORDINATOR

Dr. Jyothi S Nayak
HOD(CSE), BMSCE