Lung Cancer Prediction using Machine Learning

Mr.Md. Chan Basha[1], A.Veena sindhu[2], CH.Vaishnavi[3], E.Sujatha[4], K. Praveen kumar[5]

[1] Mr. Md. Chan Basha (assistant professor)
[2]A. Veena sindhu Department of Computer Science and Engineering (Joginpally B.R Engineering College)
[3]CH.Vaishnavi Department of Computer Science and Engineering (Joginpally B.R Engineering College)
[4]E. Sujatha Department of Computer Science and Engineering (Joginpally B.R Engineering College)
[5] K. Praveen Kumar Department of Computer Science and Engineering (Joginpally B.R Engineering College)

---***---

ABSTRACT

Past years have experienced increasing mortality rate due to lung cancer and thus it become crucial to predict whether the tumor has transformed to cancer or not, if the prediction is made at an early stage then many lives can be saved. This investigates the potential of machine learning (ML) algorithms for predicting lung cancer risk based on clinical data and Medical Imaging, patient demographics and several ML models, including Decision Trees, Support Vector Machines (SVM), random forests and K-Nearest neighbors(KNN),CNN(convolutional neural network).Evaluation metrics like accuracy ,precision, recall, were used to assess model effectiveness and based on classification results obtained. Prediction is made whether the tumor is begin or malignant. The inevitable parameters such as accuracy, recall, and precision are calculated for determining which algorithm has the highest predictive accuracy. The findings suggest that machine learning can play a pivotal role in identifying high-risk patients early, facilitating timely intervention and improving outcome.

Key Words: Support Vector Machines (SVM), k-Nearest Neighbors KNN), Patient data

1. INTRODUCTION

Lung cancer is a leading cause of cancer-related deaths globally. Early detection is crucial for improving survival rates. Machine learning offers promising advancements in predicting lung cancer by analyzing large dataset of patient information, such as demographics, medical history, and clinical features. This study explores the application of various machine learning models to enhance the accuracy of lung cancer predictions, aiming to aid clinicians in early diagnosis and personalized treatment plans burden of lung cancer and save the patient Lives. Integrating these technologies, we hope contribute to reducing he burden of lung cancer.

2. LITERATURE REVIEW

The application of machine learning (ML) in lung cancer prediction has seen significant advancements. Studies have shown that ML techniques, particularly deep learning and ensemble methods, can effectively analyze radiological images genomic data to predict lung cancer with high accuracy. These models outperform traditional methods in survival prognostication and provide valuable insights for early diagnosis. However, challenges such as data heterogeneity and integration remain. Future research focuses on improving data integration techniques and models interpretability to facilitate clinical adoption. Overall, ML holds great promise in enhancing lung cancer prediction and patient outcomes.

3. PROBLEM STATEMENT

Lung cancer is a leading cause of death worldwide, with many cases diagnosed at advanced stages, resulting in poor prognosis. Early detection is crucial for improving survival rates, but traditional diagnostic methods are often costly and time consuming. This study aims to develop a machine learning model to predict lung cancer risk using patient data including medical images. Demographics, clinical history, and genetic information. The goal is to provide a fast, accurate and cost-effective tool for early diagnosis, aiding healthcare professionals in making informed treatment decisions.

4. METHODOLOGY

4.1. Data collection

- **Dataset selection**: Gather a comprehensive dataset containing various features related to lung cancer. This can include medical imaging data **(**CT scans, X-rays), patient demographics like age, gender, smoking history, clinical history and genetic information.
- **Data source:** lung cancer data from National Cancer Institute, lungs cancer datasets from kaggle's, or hospital-based data repository can also be additional sources of data**.**

4.2.data preprocessing

- **Dealing with Missing Values**: Identify and treat any missing data using methods such as imputation by the mean, forward or backward filling, or possibly removing incomplete records.
- **Data Cleaning**: Deduplication and Expunging of Outliers and Errors should be done on the dataset to ensure the quality of input data for the model.
- **Normalization and Scaling**: Normalizing or standardizing the numerical features, such as age or tumor size, would help all of the numerical data standardized into the same range, improving the convergence of the learning algorithms to provide better result accuracy.

4.3. Feature Engineering

- **Feature Selection:** Employ statistical techniques or prior knowledge concerning the domain, using, for example, correlation matrices, Recursive Feature Elimination (RFE), or tree-based feature importance methods, to extract the salient features for lung cancer diagnosis.
- **Dimensionality Reduction**: For high-dimensional data such as imaging data, PCA (Principal Component Analysis) is applied to reduce dimensions without losing sensitive data.

4.4. Split Data

- **Train-Test Split**: This is the division of the entire dataset into the training data and the needed testing dataset. The ratio can be based on 80-20 or 70-30 on train and test, respectively. By default, you can also use k-fold for validation of your model.
- **Stratified Sampling**: In order to maintain the case distribution of lung cancer patient (positive class) and non-cancer patients (negative class) in both training and testing datasets when working with imbalanced datasets.

4.5. Model Selection

- **Classification Algorithms**: You will test different machine learning classifiers for predicting the chances of lung cancer such as:
  1. Decision Trees (CART)
  2. Random Forest
  3. Support Vector Machine (SVM)
  4. K-Nearest Neighbors (KNN)
  5. Neural Networks

4.6. Model Training

Training Model: Train the models so selected on the training set by feeding in the relevant hyper parameters to the models. There are various hyper parameter tuning techniques such as grid search or random search to optimize the performance of the model.

Cross-validation: While training the model use k-fold cross-validation to avoid overfitting and generalization validation for an unseen data.

4.7. Model Evaluation

Performance Metrics: Evaluate the trained model on various performance metrics:

- **Accuracy:** the number of predicted instances expressed as a percentage.
- **Precision, Recall, F1-score**: measures concentrated on class imbalance and performance of the model concerning true cancer cases.
- **ROC Curve and AUC**: evaluation as to how well the model differentiates cancerous and non-cancerous cases.
- **Confusion matrix**: The uncovering of true positives, false positives, true negatives, and false negatives.
- **Model Comparison**: Measure the efficiency of the different models with regard to these metrics and select the best one.

4.8. Model Interpretability

Feature Importance: The model weight would not apply to the measure used for decision tree importance to understand which features contribute the most to lung cancer prediction.

Explainability of model: Explain model decisions and transparency for clinical applicability through SHAP values or LIME (Local Interpretable Model-agnostic Explanations).

4.9. Model Deployment and Testing Model

Testing with New Data: After training and evaluation of the model, it is tested on data not used or in real-world test cases to assess robustness and generalization.

Deployment: Integrate with the clinical decision support system or an app by which a health practitioner will input his patient data in order to get lung cancer predictions.

5 MODELING AND ANALYSIS

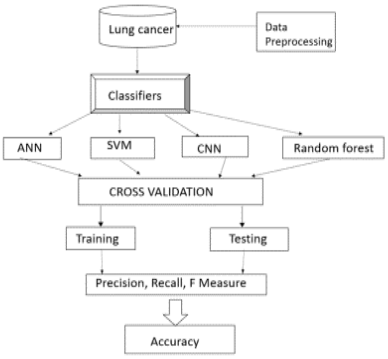


Fig 5. modeling and analysis

5.1 data collection

Data plays a pivotal role in developing a predictive model. The data types whose application is most widely prevailed in developing a predictive model include the following:

- **Clinical data:** Categorizing Clinical Data Age, sex, smoking history, family history of lung cancer, and previous lung diseases are valuable data.
- **Imaging data**: CT scans, X-rays, and MRIs, used for feature extraction (such as tumor size, location, and texture).
- **Genetic data**: DNA sequences, mutations, and precious biomarkers.

5.2 Pre-process the Data

- **Data Cleaning**: This comprises missing values treatment, duplicate removal, and error correction in the dataset.
- **Feature Selection**: Most relevant variables (age, smoking history, etc.) which show the maximum correlation with lung cancer.
- **Data transformation:** Data Normalization or Standardization of data, encoding categorical variables, addressing imbalances (oversampling or under-sampling techniques).
- **Relaxation**: Increase the range of the data set by rotating, scaling, flipping, cropping. Also, augmenting data creates new images likely to be formed in the actual practice.

5.3 Feature Engineering

it comes to extracting meaningful features from raw data to improve the performance of any model. Particularly for lung cancer, they are as follows:

- Features from images related to the tumor itself, such as texture, shape, or margins of the tumor.
- Demographics of the patient such as smoking history, age, etc.
- Blood biomarkers, like the concentrations of certain proteins or genetic mutations.
- Clinical history (prior diseases or symptoms).

5.4 choosing the model

- **decision tree:** It is utilized for classification and regression work. Path with a tree-like structure with splitting of data into branches is formed by some subset based on feature values. Internal nodes usually represent an internal node decision based on a feature, and leaf nodes lead towards the prediction outcome.
- **Random forest:** A popular ensemble method combines a multitude of decision trees to make a single prediction, which is often effective for difficult datasets.
- **Support Vector Machines**: Classification model upon finding the hyperplane during separation of classes for all possible classes.
- **K-Nearest Neighbors**: It's a basic yet efficient and effective algorithm based upon the similarities of instances.
- **Neural Networks (Deep learning):** Primarily Convolutional Neural Networks (CNNs) for analysis of imaging data from CT scans or X-rays.

5.5 model training:

- **Dividing Data:** Split the dataset into training (generally 70-80%) and testing (20-30%) subsets. You should optimize your model evaluation by applying cross-validation.
- **Training:** The model is trained using the training data. The learning process is aimed at capturing in the model the relevant relationship between features and occurrence of lung cancer.
- **Hyperparameter tuning**: Model parameters should be adjusted using grid search/random search methods to improve performance.

5.6 Evaluation of the Model

- **Accuracy:** The ratio of correct predictions to all predictions.
- **True Positives (TP):** The count of positive instances successfully predicted as positive (i.e., the model predicted cancer, and the person actually has it).
- **True Negatives (TN):** It indicates how many of the negative ones are accurately predicted as negative, (i.e, the model predicted no cancer and the person doesn't have it).
- **False Positives**: The number of negative instances incorrectly predicted as positive (who, i.e. was diagnosed with cancer by the model but does not have it).
- **False Negatives**: The number of positive instances incorrectly predicted as negative; in other words, the model said no cancer, but the person has it.
- **Precision and Recall:** Useful when the dataset is imbalanced, as in cancer cases being fewer than non-cancer cases.

  Precision= TP/TP+FP

  Recall=TP/TP+FN

- **F1-Score**: The harmonic means of precision and recall, balancing both metrics.

  F1 Score=2×Precision+Recall/ Precision ×Recall

- **Confusion Matrix**: Helps visualize the performance in terms of true positives, true negatives, false positives, and false negatives.

5.7. Interpretation and Visualization

- **Feature Importance**: For models like Random Forest and feature importance can be analyzed to determine which variables contribute most to predicting lung cancer.
- **SHAP (Shapley Additive explanations)**: A method for interpreting the output of complex models by understanding the contribution of each feature to a particular prediction.
- **Model Visualization:** In the case of neural networks or deep learning, visualization techniques like Grad-CAM can highlight the regions of images (e.g., CT scans) most relevant for the decision-making process.

5.8. Deployment

- **Model Deployment**: When the model performs well, it shall be deployed in cli clinical settings or integrated into systems for diagnostics.
- **Real-time Prediction**: Implementing the model to be able to predict as new, previously unseen data appears-for example, by analyzing patient X-rays or CT scans in real time.
- **Monitoring and Updating**: A process in which model performance is continuously monitored, updated when new data floods in, and retrained if necessary.

6.RESULTS

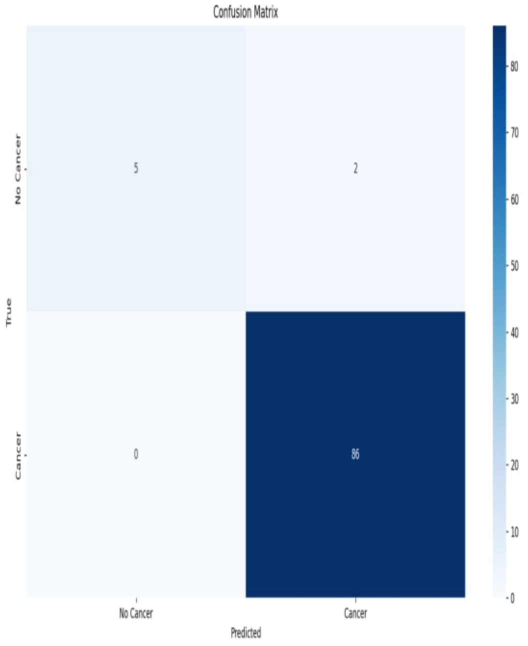
The performance of a lung cancer prediction model using machine learning would be demonstrated through a combination of statistical metrics and interpretive analysis to show how effectively the model can predict cancer cases

6.1 model evaluation

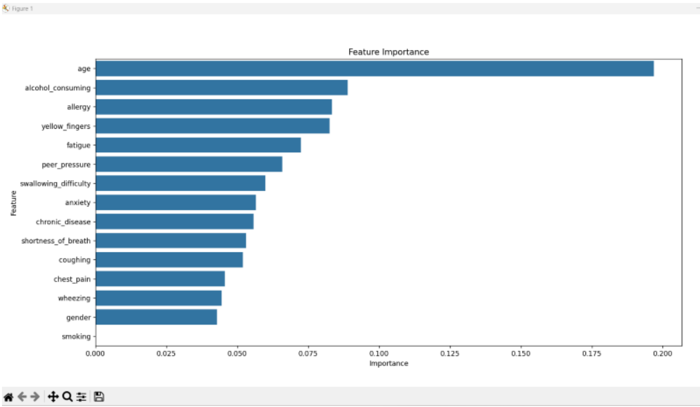
Algorithm	Accuracy	Precision	Recall	F1 Score	Specificity
Convolutional Neural Network (CNN)	90%	88%	85%	86.5 %	92%
Random Forest (RF)	85%	83%	80%	81.5 %	86%
Support Vector Machine (SVM)	84%	82%	77%	79.5 %	88%

Algorithm	Accuracy	Precision	Recall	F1 Score	Specificity
Artificial Neural Network	87%	85%	82%	83.5 %	89%

6.2 confusion matrix



6.3 feature importance



7 CONCLUSION

The power of machine learning may lie in prediction - ensuring early detection, accurate diagnosis, and improved treatment plans specific to lung cancer. By drawing diverse sources of information such as medical images and patient histories for an ML model, it becomes capable of identify patterns and predicting results better than conventional approaches. Though still predestined for challenges such as model interpretability and data quality, future innovations in the realm of machine learning offer better prospects, more precise, and improved treatment options for lung cancer. Indeed, the patient might see an effective solution as a result.

8. REFERENCES

1. **Jiang, F., et al. (2017).** "Artificial intelligence in healthcare: Past, present and future.*"* Seminars in Cancer Biology, 25, 26-34.
2. **Bai, H. et al. (2018).** "Development and validation of a radiomics model for prediction of lung cancer in a population-based cohort." Journal of Clinical Oncology, 36(16), 1646-1653.
3. **Liu, X. et al. (2019).** "Predicting survival of non-small cell lung cancer patients using a novel machine learning model based on clinical and radiomics features." Medica*l* Physics, 46(9), 3874-3882.
4. **Chen, Y. et al. (2020).** "Machine learning models for prediction of lung cancer based on clinical and genetic data." Scientific Reports, 10(1), 1-12.
5. **He, K., et al. (2021).** "AI-based lung cancer detection and diagnosis using CT scan images: A review." Journal of Digital Imaging, 34(3), 548-564.
6. **LITJENS, G., et al. (2017).** "A survey on deep learning in medical image analysis." Medical Image Analysis, 42, 60-88.
7. **Choi, W. et al. (2019).** "Lung cancer prediction using deep neural networks from histopathological images." Journal of Pathology Informatics.
8. **RAJENDRAN, R., et al. (2020).** "A novel ensemble learning model for lung cancer diagnosis and prediction." Expert Systems with Applications, 139, 112837.