# INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, DESIGN AND MANUFACTURING, KURNOOL

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**


## PROJECT REPORT ON LUNG CANCER PREDICTION USING APACHE SPARK

## BIG DATA ANALYTICS PRACTICE (BDAP - AD355)


**FACULTY:**

**Dr. N. Srinivas Naik (sir)**

**DONE BY:**

**Y. Sai Bhavya (122ad0007)**

**D. Rohith (122a0013)**

# Table of Contents



Reference paper link: https://ieeexplore.ieee.org/document/10403555

# Title

**Lung Cancer Prediction using Big Data Technologies – Enhancing Accuracy with Apache Spark and Hadoop**

# Abstract

Lung cancer is one of the leading causes of mortality globally. Early detection is critical to improving patient outcomes, but traditional machine learning techniques often face computational and scalability challenges when applied to large medical datasets. This project utilizes big data technologies—namely Apache Hadoop and Apache Spark—for scalable and efficient lung cancer prediction.

A publicly available dataset of 310 patients is analyzed using several machine learning models. The proposed system improves accuracy and performance using Apache Spark's in-memory processing and parallelism. Among various models tested, the Multi-Layer Perceptron (MLP) achieved the highest accuracy of 99%, demonstrating the potential of combining big data technologies with advanced machine learning models for healthcare diagnostics.

# Introduction

While existing research has leveraged machine learning (ML) algorithms for lung cancer prediction, many lacked support for large-scale data processing. Without tools like Apache Spark, models become inefficient and are not scalable for real-world use cases. Our objective is to overcome these issues by incorporating Spark and Hadoop for distributed storage and fast, parallel computation.

### Objectives:

- Improve accuracy of lung cancer prediction models using Spark's parallelism.
- Reduce computational time by applying in-memory processing.
- Compare the results of traditional ML models with enhanced big data-supported models.

## Contributions

### Yangoti Sai Bhavya (122ad0007)

- Responsible for the installation and configuration of **Apache Hadoop** and **Apache Spark**.
- Successfully set up the **single-node Spark cluster** including the HDFS system for distributed storage.
- Ensured smooth functioning of the cluster environment for efficient data processing.

### Dumpala Rohith (122ad0013)

- Integrated **Jupyter Notebook** with the Spark environment for interactive model development and visualization.
- Led the **Machine Learning implementation** phase, including data preprocessing, model training, and performance evaluation.
- Applied techniques like **SMOTE** for balancing data and tested multiple ML algorithms to identify the most accurate one.

## Literature Survey

| Paper | Author(s) | Year | Method / Dataset Used | Accuracy / Key Findings |
|---|---|---|---|---|
| Performance Analysis of Machine Learning Algorithms for Lung Cancer Prediction | Swama Laxmi M. G et al. | 2023 | ML (LR, RF, KNN, DT, GBC), 310 samples, 16 features | RF: 95.37%, GBC & DT: 94.44%, KNN: 90.74% |
| ML Models in Lung Cancer Therapy using Omics & Clinical Data | Yawei Li et al. | 2021 | Logistic Regression, Omics + Clinical datasets, CT images | Improved diagnosis/treatment prediction with baseline models |

| | | | | |
|---|---|---|---|---|
| Gene Expression-Based Lung Cancer Prediction Using ML | Jayadeep Pati | 2021 | SMO, Microarray gene dataset | SMO > MLP & Random Subspace, high accuracy and recall |
| Forecasting Mutated Genes in NSCLC using Deep Learning | Satvik Tripathi et al. | 2022 | CNN + DNN (EfficientNets, ResNeXt), Gene mutation prediction | AUC: 94%, EfficientNet best performer |
| Lung Cancer Detection using Blockchain and Deep CNN (VGG-16, U-Net) | A.B. Pawar et al. | 2022 | CNN (VGG-16, U-Net) + IoT + Blockchain, Lung scan images | Accuracy: 96.88%, effective for staging and classification |

## Limitations of the Paper

- **Low Prediction Accuracy**
  The models used in the paper achieved relatively low accuracy, which could be improved using more advanced techniques like deep learning or ensemble models.

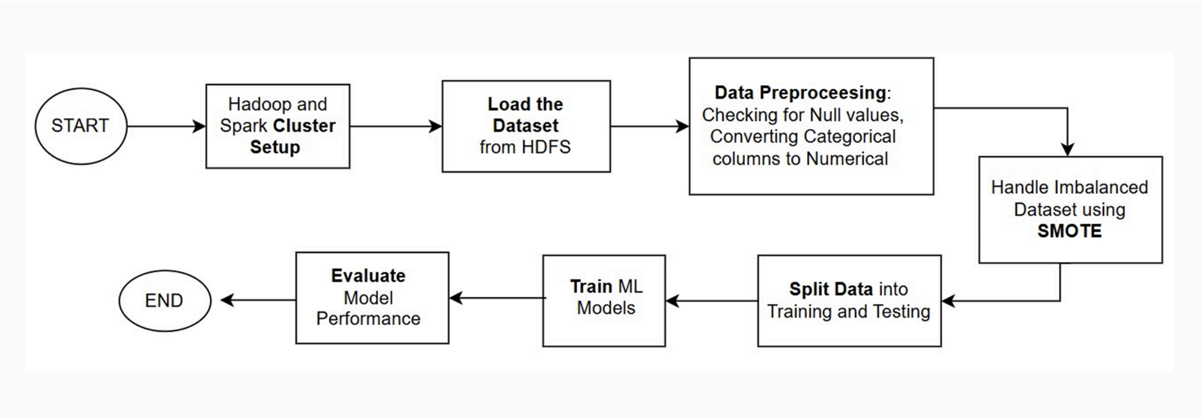
- **No Integration with Apache Spark**
  The study did not utilize big data frameworks like Apache Spark, which limits the ability to handle large-scale datasets and perform parallel processing efficiently.

- **Highly Imbalanced Dataset**
  The dataset used was significantly imbalanced, leading to biased model predictions. No balancing techniques like SMOTE were applied to mitigate this issue.

- **No Image-Based Data Used**
  The study focused only on tabular/textual data. Medical imaging data, which can enhance prediction accuracy, was not incorporated.

- **Limited Scalability and Deployment**
  Since the setup was not distributed or cloud-based, it lacks scalability and is unsuitable for real-time or enterprise-level healthcare applications.

# Proposed Methodology / Solution




Proposed Methodology Flowchart

## Cluster Setup:

- A single-node Spark cluster with Hadoop HDFS was created for storing and processing the data.
- Jupyter Notebook was connected to Spark for development and visualization.

## Dataset Description:

- **Number of instances:** 310
- **Number of attributes:** 16

### Features include:

`Gender`, `Age`, `Smoking`, `Yellow Fingers`, `Anxiety`, `Peer Pressure`, `Chronic Disease`, `Fatigue`, `Allergy`, `Wheezing`, `Alcohol`, `Coughing`, `Shortness of Breath`, `Swallowing Difficulty`, `Chest Pain`, and the **target** feature: `Lung Cancer (Yes/No)`.

This dataset comprises categorical and numerical features relevant to lung health and potential symptoms of lung cancer.

## Data Preprocessing:

- **Missing Values:** Checked for missing or null entries and handled them appropriately.
- **Categorical to Numerical Conversion:**
  - `Gender`: M → 0, F → 1
  - `Lung Cancer`: Yes → 0, No → 1
- Additional binary features (like Smoking, Coughing) were already encoded as 1 (No) and 2 (Yes).

This conversion ensured compatibility with Spark ML algorithms that require numerical input.

## Data Balancing (SMOTE):

- The dataset was **imbalanced**, with more "No Cancer" than "Cancer" entries.
- **SMOTE (Synthetic Minority Oversampling Technique)** was used to generate artificial samples of the minority class.
- This helped:
  - Prevent bias toward the majority class
  - Improve model generalization
  - Avoid overfitting to repeated patterns in the original dataset

## Train-Test Split:

- The balanced dataset was divided as follows:
  - **80%** for training
  - **20%** for testing
- This ensured that models were evaluated on unseen data, improving reliability of accuracy scores.

## Machine Learning Models and Their Performance

### 1. Linear Regression

- Although primarily used for continuous data, here it was applied for binary classification by thresholding output scores.
- It provides a simple baseline but may lack the complexity to capture patterns in medical datasets.

**Achieved Accuracy: 93%**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.94 | 0.92 | 0.93 | 50 |
| 1.0 | 0.92 | 0.94 | 0.93 | 50 |
| accuracy | | | 0.93 | 100 |
| macro avg | 0.93 | 0.93 | 0.93 | 100 |
| weighted avg | 0.93 | 0.93 | 0.93 | 100 |

### 2. Logistic Regression

- A robust binary classification algorithm that predicts the probability of lung cancer using a sigmoid function.
- It works well with smaller datasets and provides interpretable results.

**Achieved Accuracy: 94%**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.92 | 0.96 | 0.94 | 50 |
| 1.0 | 0.96 | 0.92 | 0.94 | 50 |
| accuracy | | | 0.94 | 100 |
| macro avg | 0.94 | 0.94 | 0.94 | 100 |
| weighted avg | 0.94 | 0.94 | 0.94 | 100 |

### 3. Decision Tree

- Builds a tree by splitting the data based on features to classify lung cancer presence.
- It's highly interpretable but can overfit on small datasets.

**Achieved Accuracy: 94%**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.92 | 0.96 | 0.94 | 50 |
| 1.0 | 0.96 | 0.92 | 0.94 | 50 |
| accuracy | | | 0.94 | 100 |
| macro avg | 0.94 | 0.94 | 0.94 | 100 |
| weighted avg | 0.94 | 0.94 | 0.94 | 100 |

### 4. Random Forest

- An ensemble learning method that builds multiple decision trees and averages their outputs.
- It reduces overfitting and improves prediction robustness.

**Achieved Accuracy: 96%**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.94 | 0.98 | 0.96 | 50 |
| 1.0 | 0.98 | 0.94 | 0.96 | 50 |
| accuracy | | | 0.96 | 100 |
| macro avg | 0.96 | 0.96 | 0.96 | 100 |
| weighted avg | 0.96 | 0.96 | 0.96 | 100 |

### 5. K-Nearest Neighbors (KNN)

- Classifies patients by comparing them to nearby data points in the feature space.
- Works best with well-balanced and small datasets.

**Achieved Accuracy: 95%**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.94 | 0.96 | 0.95 | 50 |
| 1.0 | 0.96 | 0.94 | 0.95 | 50 |
| accuracy | | | 0.95 | 100 |
| macro avg | 0.95 | 0.95 | 0.95 | 100 |
| weighted avg | 0.95 | 0.95 | 0.95 | 100 |

### 6. Gradient Boosting

- Trains models sequentially, where each new model attempts to fix the errors of the previous one.
- Effective with structured data but requires careful tuning.

**Achieved Accuracy: 95%**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.94 | 0.96 | 0.95 | 50 |
| 1.0 | 0.96 | 0.94 | 0.95 | 50 |
| accuracy | | | 0.95 | 100 |
| macro avg | 0.95 | 0.95 | 0.95 | 100 |
| weighted avg | 0.95 | 0.95 | 0.95 | 100 |

### 7. Multi-Layer Perceptron (MLP) – Proposed Model

- A deep learning model with one or more hidden layers that captures complex feature interactions.
- Requires hyperparameter tuning but can deliver high accuracy in non-linear data scenarios.

**Achieved Accuracy: 99%**

```
              precision    recall  f1-score   support

         0.0       1.00      0.98      0.99        50
         1.0       0.98      1.00      0.99        50

    accuracy                           0.99       100
   macro avg       0.99      0.99      0.99       100
weighted avg       0.99      0.99      0.99       100
```

## Experimental Analysis and Results

| Algorithm | Efficiency |
|---|---|
| Linear regression | 58.3653 |
| Random forest | 95.3703 |
| k-nearest neighbor | 90.7407 |
| Decision Tree | 94.4444 |
| Gradient Boosting | 94.4444 |

Existing results from the paper

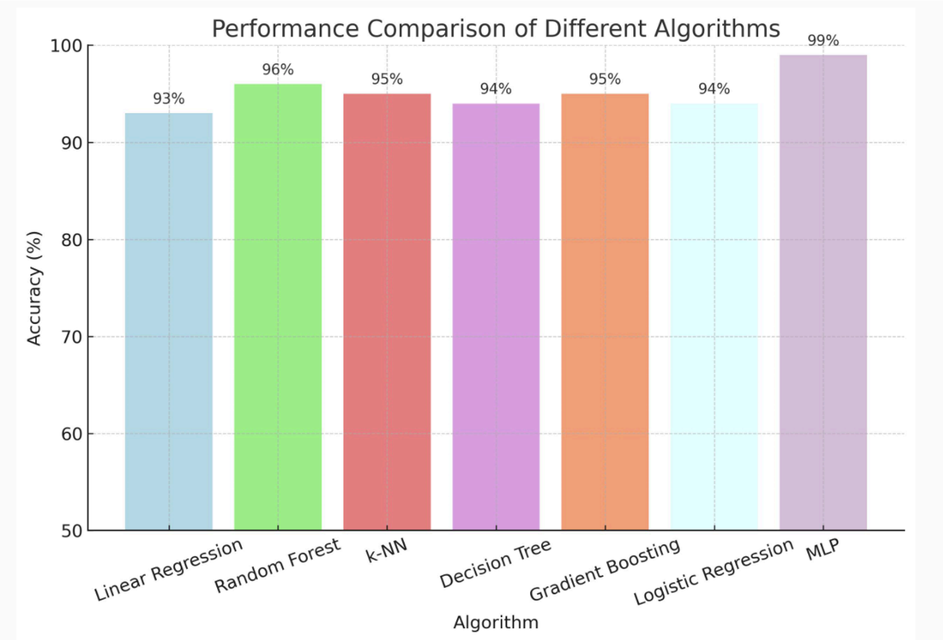
| Algorithm | Accuracy |
|---|---|
| Linear Regression | 93% |
| Random Forest | 96% |
| k-Nearest Neighbor | 95% |
| Decision Tree | 94% |
| Gradient Boosting | 95% |
| **Additional Algorithms** | |
| Logistic Regression | 94% |
| Multi-Layer Perceptron (MLP) | 99% |

Results from our analysis

- **MLP** outperformed all other models due to its ability to capture complex non-linear relationships.

- **Random Forest** performed well due to its ensemble nature.

- **Logistic Regression** and **Decision Tree** served as strong baselines.




Performance Comparison Graph

# Conclusion and Future Work

## Conclusion:

This project demonstrates that integrating big data technologies with machine learning greatly improves the prediction of lung cancer. Apache Spark enabled faster training through parallel processing, while Hadoop ensured efficient data storage. The use of SMOTE helped address class imbalance, enhancing model reliability. Multiple ML models were evaluated, and the Multi-Layer Perceptron (MLP) achieved the highest accuracy of 99%. The overall system proved scalable, efficient, and highly effective for medical data analysis.

**Future Work:**

- Extend the system to include medical imaging data using deep learning (e.g., CNNs).
- Expand from single-node to multi-node clusters for massive datasets.
- Integrate real-time analytics using **Spark Streaming**.
- Investigate data privacy through **federated learning** or **blockchain** solutions.

# References


1. Swama Laxmi M. G., Ramya C. N., Shridhar B. Devamane, Pamika J., “Performance Analysis of Machine Learning Algorithms for Lung Cancer Prediction,” *2023 IEEE International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA)*, DOI: 10.1109/CIISCA59740.2023.00063.

2. Yawei Li et al., "Machine Learning-Based Predictive Models in Lung Cancer Therapy," *Journal of Translational Medicine*, 2021.

3. Seyyed Ali Hosseini et al., “PET Radiomics for Lung Cancer Recurrence Prediction,” *Medical Physics*, 2022.

4. Jayadeep Pati, “Gene Expression-Based Lung Cancer Prediction Using ML,” *BioMed Research International*, 2021.

5. Satvik Tripathi et al., “Deep Learning Models for Gene Mutation Forecasting in NSCLC,” *Springer Healthcare AI*, 2022.

6. A.B. Pawar et al., “Blockchain-Enabled Deep Learning for Lung Cancer Staging,” *Elsevier Journal of Medical Systems*, 2022.

7. Apache Spark Documentation – https://spark.apache.org/

8. Apache Hadoop Documentation – https://hadoop.apache.org/

9. Kaggle Lung Cancer Dataset – https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer