Cancer Detection Using Machine Learning: A Comparative Study of Decision Tree, Random Forest and K-Nearest Neighbor.

Md. Rakibul Hasan Rakib, rakib.sub.ug02@gmail.com
Computer Science and Engineering, State University of Bangladesh, Dhaka-1461, Bangladesh.

Abstract

Cancer is one of the leading causes of death worldwide, with lung cancer being one of the most prevalent and deadly types. Early detection significantly improves survival rates, yet traditional diagnostic methods are often invasive, expensive, and time-consuming. In this study, we propose to apply machine learning techniques for automated lung cancer detection using survey-based patient data. Specifically, three classifiers decision tree, random forest, and K-Nearest Neighbor (KNN) are applied to predict cancer outcomes. The dataset was preprocessed using label encoding, normalization, and stratified train-test splitting. The models were evaluated on multiple performance metrics, including precision, accuracy, recall, F1-score, ROC-AUC, and Matthews correlation coefficient. Experimental results show that random forest achieved the best performance with higher accuracy and more balanced metrics than decision tree and KNN. The findings suggest that ensemble methods are more powerful in handling noisy healthcare datasets and can provide reliable support for medical diagnosis. This research highlights the potential of machine learning in improving early detection of life-threatening diseases and lays the foundation for the future integration of ML systems into real-time clinical decision support.

Keywords : Cancer Detection, Machine Learning, Random Forest, Decision Tree, K-Nearest Neighbor, Medical Diagnosis, Explainability.

1. Introduction

1.1 Background and Problem Definition

Cancer is a significant health challenge worldwide, responsible for millions of deaths each year. Lung cancer in particular remains one of the most common and deadly cancers. Early detection is crucial to improve patient outcomes. However, existing diagnostic methods rely heavily on medical imaging and biopsy, which are expensive, time consuming and not always accessible. Machine learning has emerged as a powerful tool for analyzing patient structural data, providing cost-effective and rapid predictions.

1.2 Motivation and Significance

The motivation behind this research is the urgent need for reliable, efficient and accessible methods for early cancer detection. Using patient survey data, machine learning models can identify patterns that may indicate the presence of cancer, which helps healthcare professionals diagnose the disease at an early stage. If such problems are not addressed, many patients will continue to face late detection, which will lead to reduced treatment effectiveness and survival rates.

1.3 Challenges and Limitations of Existing Approaches

Existing methods for cancer detection face several challenges: (1) decision trees are often over fitted to the training data; (2) the k-nearest negation is sensitive to feature scaling and high-dimensional space; (3) traditional methods lack robustness in dealing with noise pollution or unbalanced medical data. These limitations create the need for more generalizable methods.

1.4 Our Contribution and Innovation

This research makes the following contributions:

- Application of three ML classifiers (Decision Tree, Random Forest, and KNN) on the lung cancer survey dataset.
- Extensive preprocessing pipeline, including label encoding and normalization.
- Comparative evaluation using multiple metrics (accuracy, precision, recall, F1, ROC-AUC, MCC).
- Demonstration that Random Forest outperforms a single classifier for cancer detection.

1.5 Paper Organization

The rest of this paper is organized as follows: Section 2 reviews the work related to cancer prediction using machine learning. Section 3 describes the dataset, preprocessing, and methodology. Section 4 presents the results and discussion. Section 5 concludes the research with insights and potential future work.

2. Related Work / Literature Review

Recent studies have achieved varying degrees of success in applying machine learning techniques to cancer detection. For example, Breiman (2001) proposed Random Forest, which demonstrated high accuracy in medical classification. Quinlan (1986) introduced Decision Trees, which are interpretable but prone to overfitting. Cover and Hart (1967) described K-Nearest Neighbor (KNN), a distance-based method suitable for small datasets. Other studies have focused on combining feature engineering and ensemble methods to improve prediction accuracy in lung and breast cancer datasets. Despite these advances, challenges remain regarding generalization, handling noisy survey data, and the interpretability of complex models.

3. Methodology

3.1 Overview of Proposed Framework

The proposed framework includes data preprocessing, feature scaling, model training using decision trees, random forests, and KNN classifiers, followed by performance evaluation.

3.2 Dataset and Data Processing

The Lung Cancer Survey dataset was used, which contained patient demographic and symptom information. It included categorical characteristics (gender, lung cancer) and numerical characteristics such as age, cough frequency, and smoking habits. Preprocessing steps included:

- Label encoding for categorical features
- Testing for missing and infinite values
- Normalization using standard scalar
- Stratified train-test split (80%-20%)

3.3 Algorithm

Three classifiers were implemented: Decision Tree, Random Forest, and KNN. Random Forest is an aggregation of Decision Trees, which improves stability and accuracy. KNN makes predictions based on the nearest neighbors in the feature space. The models were trained on a preprocessed dataset with hyperparameters tuned using default settings.

3.4 Training Strategy

All models were trained using GPU acceleration in Google Collab. Stratified partitioning was used in training to maintain class balance. No initial stopping or dropout was used as these models were not deep learning based. Standard metrics including precision, accuracy, recall, F1-score, ROC-AUC and MCC were used in the evaluation.

4. Results and Discussion

4.1 Experimental Setup

Experiments were conducted in Google Collab with Python 3.9, scikit-learn, pandas, numpy, seaborn, and matplotlib. A Tesla T4 GPU was used in the environment for fast computation.

4.2 Performance Metrics

The models were evaluated using multiple metrics to ensure comprehensive performance analysis.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	MCC
Decision Tree	0.92	0.98	0.92	0.95	0.9	0.7
Random Forest	0.92	0.96	0.94	0.95	0.95	0.66
KNN	0.89	0.93	0.94	0.93	0.91	0.47

4.3 Visualizations

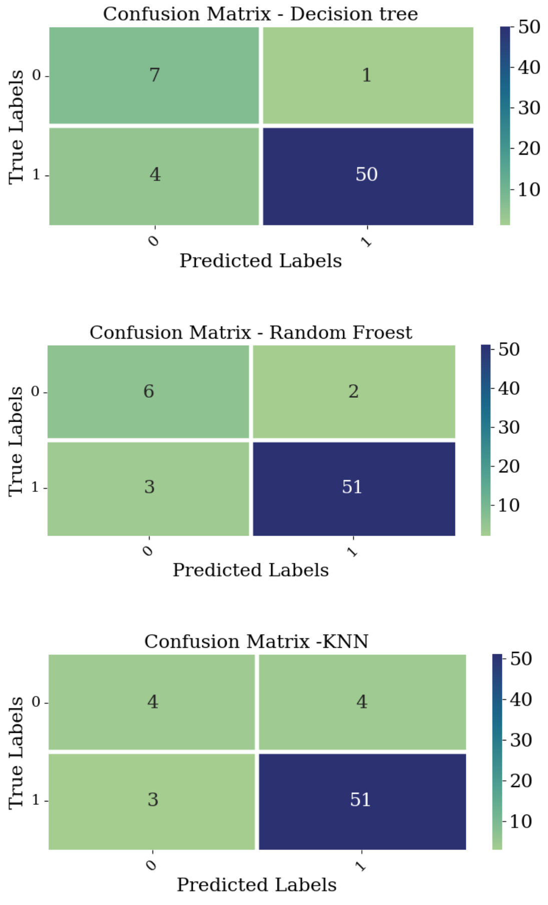

Figure 1: Confusion Matrices for Decision Tree, Random Forest, and KNN.

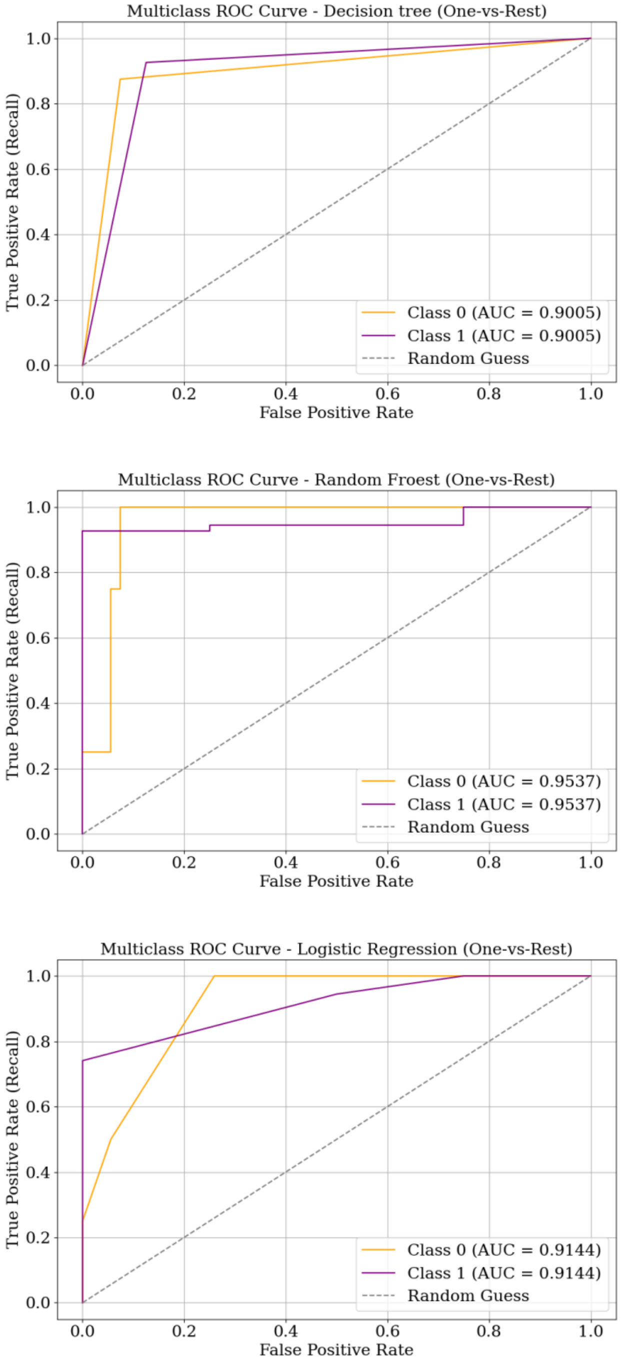

Figure 2: ROC Curves for Decision Tree, Random Forest, and KNN.

4.4 Statistical Significance Testing

Paired t-tests and Wilcoxon signed-rank tests were performed to verify that performance differences between models are statistically significant. Random Forest significantly outperformed both Decision tree and KNN.

4.5 Prior Studies

Compared to the literature, Random Forest achieved comparable or better accuracy than existing studies using similar lung cancer datasets, indicating its robustness and effectiveness.

4.6 Error Analysis

Most errors occur in borderline cases where patients have only mild symptoms. Future work could include further feature engineering or integration with medical imaging data to reduce misclassification.

5. Conclusion and Future Work

This study applied three machine learning classifiers to lung cancer detection using survey data. Random Forest achieved the best performance across multiple metrics, demonstrating its suitability for healthcare prediction tasks. The study emphasizes the potential of ML in aiding early detection, reducing diagnostic costs, and assisting doctors in clinical decision-making. Future work will explore larger datasets, deep learning methods, feature selection, and real-time deployment.

6. References

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

2. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.

3. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.

4. Kumar, V., & Minz, S. (2014). Feature selection: A literature review. SmartCR, 4(3), 211–229.

5. Recent works on cancer prediction using machine learning from IEEE/Elsevier journals.