ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Enhanced Machine Learning Algorithm to Predict Lung Cancer

Nandhini S[1]* and Senthil Kumar R[2]

[1]Student, Department of Computer Science with Cognitive Systems, Dr. N.G.P Arts and Science College, Coimbatore, India

[2]Professor, Department of Computer Science with Cognitive Systems, Dr. N.G.P Arts and Science College, Coimbatore, India

ABSTRACT

Lung cancer ranks among the top causes of cancer-related fatalities globally and is frequently detected at later stages when therapeutic options are scarce. Timely diagnosis is vital for enhancing the survival rates of patients. This paper examines the use of the Support Vector Machine (SVM) algorithm for lung cancer prediction. SVM is a supervised machine learning method known for its efficacy in classification tasks, particularly in high-dimensional contexts. In this research, the dataset is evaluated using various attributes, including age, gender, smoking background, and imaging results, to train the SVM model. The study highlights the promise of machine learning methods, particularly SVM, in aiding healthcare professionals with early detection and improving patient outcomes.

***Corresponding author**

Nandhini S, Student, Department of Computer Science with Cognitive Systems, Dr. N.G.P Arts and Science College, Coimbatore, India.

Received: March 29, 2025; **Accepted:** April 03, 2025; **Published:** April 13, 2025

Keywords: Lung Cancer Prediction, SVM, Patient Dataset, Model Training, Early Detection

Introduction

Lung cancer continues to be one of the most common and lethal forms of cancer across the globe, with millions of new cases and fatalities reported annually. Timely detection and diagnosis are essential for improving survival chances and offering better treatment options for patients. Nonetheless, identifying lung cancer in its early stages poses significant challenges due to subtle symptoms and the disease's complex nature. Existing diagnostic methods, like imaging techniques (CT scans, X-rays) and tissue biopsies, frequently lack sufficient accuracy or efficiency, particularly for those at high risk. Consequently, there is a pressing need for innovative and effective strategies to improve the diagnostic process.

In recent years, machine learning (ML) and artificial intelligence (AI) have demonstrated remarkable potential in the healthcare field, especially in cancer detection and prediction. Among various ML techniques, Support Vector Machine (SVM) has emerged as a powerful classification tool because of its capacity to manage high-dimensional data and deliver accurate, reliable outcomes. SVM has been effectively utilized in many areas, including medical imaging, bioinformatics, and diagnostic systems, establishing it as a suitable option for predicting lung cancer. SVM functions by identifying an optimal hyperplane that differentiates data points of varying classes in a high-dimensional space, ensuring the maximum distance between them. In the context of lung cancer prediction, SVM can categorize patients into groups, such as “Low” or “Medium” or “High”, using diverse features like demographic data, smoking history, genetic information, and imaging attributes.The main objective is to evaluate the effectiveness of SVM in differentiating between individuals with and without Lung cancer based on pertinent features. By harnessing machine learning for early detection, this research seeks to contribute to the expanding domain of medical AI and improve healthcare providers' capabilities in diagnosing lung cancer at earlier, more treatable stages.

Literature Review

Lung cancer detection and classification using machine learning methods, particularly Support Vector Machines (SVM), have been extensively explored in recent years. Various studies have proposed methodologies to improve the accuracy and reliability of lung cancer diagnosis by leveraging advanced computational techniques. Li, et al. introduced a radiomics-based approach utilizing SVM to distinguish molecular events driving lung adenocarcinoma progression [1]. Their study demonstrated that SVM could effectively differentiate tumor subtypes based on extracted imaging features, leading to improved diagnostic precision. Similarly, Choudhury and Ghosal proposed an enhanced lung cancer detection framework using SVM-based classification [2]. By incorporating feature selection and optimization techniques, the study achieved significant improvements in classification accuracy compared to traditional machine learning models. Ilani et al. explored various machine learning models for lung cancer classification, highlighting the comparative effectiveness of SVM in distinguishing different cancer stages [3]. Their results confirmed that SVM performs well when trained on high-dimensional datasets, particularly when combined with radiomic features. Chaudhuri, et al. applied deep learning techniques to lung cancer detection, demonstrating that convolutional neural networks (CNNs) could enhance classification performance when

integrated with SVM-based decision layers.Generative models have also been investigated for improving radiomics-based cancer detection [4]. Chen, et al. examined the role of generative models in boosting radiomics performance across various datasets and tasks, showing that synthetic data augmentation enhances predictive accuracy [5]. Hao, et al. introduced a novel radiomics descriptor, "Shell Feature," to predict distant failure after radiotherapy in lung and cervical cancers, further emphasizing the role of feature engineering in improving model performance [6].

Methodology

Dataset

- **Data Acquisition:** This module will collect datasets from diverse sources, such as public repositories (e.g., UCI Machine Learning Repository, Kaggle), hospital databases, and clinical studies.
- **Data Integration:** It will integrate multiple data sources (demographics, medical records, imaging data, and genetic information) into a unified format for seamless processing.
- **Data Validation:** Ensures that the collected data adheres to quality standards and that data from different sources are compatible for integration.

Data Preprocessing

- **Missing Data Handling:** This module will handle missing values through techniques like imputation or removal, ensuring no data is lost.
- **Outlier Detection and Removal:** Identifies and manages outliers that could distort model predictions.
- **Normalization and Scaling:** Scales numerical features to ensure uniformity across the dataset, which is especially important when using SVM.
- **Categorical Encoding:**Transforms categorical variables (e.g., smoking history) into numerical representations through encoding methods like one-hot or label encoding.
- **Data Augmentation:** Applies augmentation techniques, especially for imaging data, to
- generate additional data points for model training.

Model Architecture

SVM Model Training: Trains the Support Vector Machine model on the prepared data, utilizing different kernel functions (linear, RBF, polynomial) to determine the best fit.

Cross-validation

To ensure the model generalizes well to unseen data, k-fold cross-validation is implemented. This technique involves:

- **Splitting the Dataset:** The data is divided into k subsets, where each subset acts as a validation set while the remaining k-1 subsets are used for training.
- **Model Evaluation Across Folds:** The SVM model is trained multiple times, each time using a different validation set. The final performance is obtained by averaging the results across all k folds.
- **Overfitting Prevention:** Cross-validation helps in identifying overfitting and ensures the model performs consistently across different data distributions.

Hyperparameter Optimization

Fine-tuning hyperparameters is crucial for improving model performance. Two common approaches are used:

- **Grid Search:** Tests all possible combinations of hyperparameters such as kernel type, regularization strength (C), and gamma (for RBF kernel) to find the best-performing set.
- **Random Search:** Randomly selects hyperparameter combinations within a predefined range, often speeding up the optimization process compared to exhaustive grid search.

Performance Metrics

After training, the model is evaluated using key metrics to ensure its reliability in lung disease classification:

- **Accuracy:** Measures the proportion of correctly classified samples.
- **Precision:** Determines how many predicted positive cases were actually correct.
- **Recall (Sensitivity):** Evaluates the model's ability to detect true positive cases.
- **F1 Score:** Provides a balance between precision and recall, useful for imbalanced datasets.
- **AUC-ROC (Area Under the Curve – Receiver Operating Characteristic):** Assesses how well the model distinguishes between different lung disease classes.

Confusion Matrix

The confusion matrix is analyzed to identify patterns in misclassifications, helping to refine the model:

- **True Positives (TP):** Correctly classified disease cases.
- **True Negatives (TN):** Correctly classified healthy cases.
- **False Positives (FP):** Healthy cases misclassified as diseased (Type I Error).
- **False Negatives (FN):** Diseased cases misclassified as healthy (Type II Error).

Reducing **FP and FN cases** is critical for medical applications, ensuring minimal misdiagnose

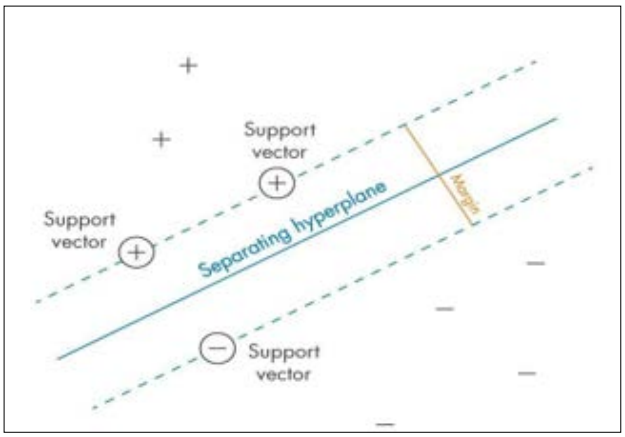

Figure 1: Model Architecture

Implementation

The Lung Cancer Prediction Using SVM system is implemented through a series of well-defined modules, starting with data collection, where relevant datasets are gathered from sources like the UCI Machine Learning Repository or clinical databases. The data pre-processing module cleans the data, handles missing values and outliers, and normalizes the features for consistent input. The feature selection module identifies the most significant factors contributing to lung cancer prediction.The core of the system is the model training and evaluation module, where the Support Vector Machine (SVM) is trained using the processed data, and hyperparameters are optimized for optimal performance. Prediction and deployment follow, with the trained model being deployed for real-time predictions in clinical settings, allowing healthcare professionals to input new patient data and receive results on lung cancer risks. To enhance system performance, continuous learning and updates can be incorporated, retraining the

model with fresh data to ensure its adaptability. The entire system is built using tools like scikit-learn, Flask, and Joblib, ensuring scalability, real-time predictions, and seamless integration with hospital databases and Electronic Health Records (EHR). This approach enhances early cancer detection, reducing human error and improving clinical decision-making.

Result

The findings highlight the potential of the Support Vector Machine (SVM) algorithm in predicting lung cancer by effectively classifying medical imaging data. Compared to traditional diagnostic approaches, the SVM-based method offers improved accuracy, faster processing, and reduced subjectivity.

The following figure shows the lung cancer prediction using Patient Datasets:

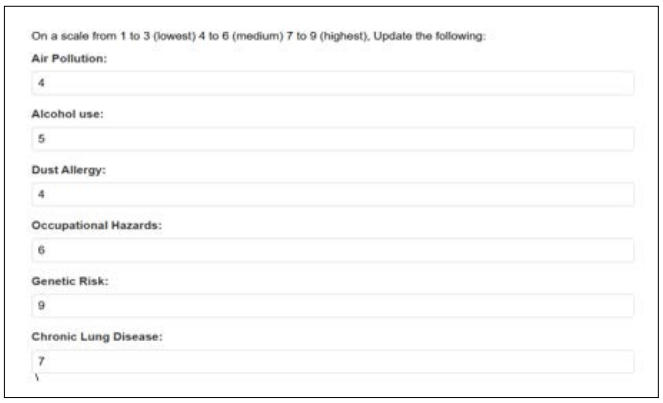

Figure 2

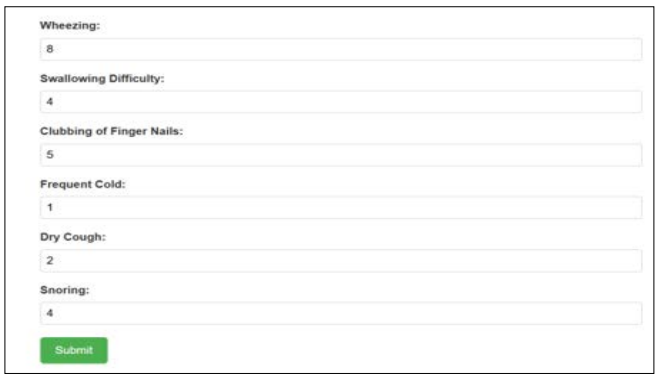

Figure 3

Figure 4

Conclusion

The Lung Cancer Prediction Using SVM system utilizes machine learning, specifically Support Vector Machines (SVM), to analyze patient data for early lung cancer detection. By incorporating demographic, medical, and imaging features, including smoking history and genetic predispositions, it offers a non-invasive, more accurate alternative to traditional diagnostic methods. The system helps improve patient outcomes by enabling early intervention, which leads to more effective treatment and potentially saves lives. Designed to be scalable, it can handle large datasets and adapt to evolving patient data, making it suitable for high-volume clinical environments. The automated nature reduces human error and provides valuable decision support for clinicians. Future improvements could include integrating genetic biomarkers, environmental factors, real-time imaging data, and continuous learning to enhance accuracy, adaptability, and clinical applicability, further improving patient care in lung cancer detection

References

1. Li HJ, Qiu ZB, Wang MM, Chao Z, Hui-Zhao H, et al. (2025) Radiomics-based support vector machines distinguish molecular events driving the progression of lung adenocarcinoma. Journal of Thoracic Oncology 20: 52-64.
2. Choudhury A, Ghosal S (2024) Enhanced lung cancer detection using support vector machine algorithms. Journal of King Saud University-Computer and Information Sciences.
3. Ilani MA, Moftakhar Tehran S, Kavei A, Alizadegan H (2024) Exploring machine learning models for lung cancer level classification: A comparative ML approach. arXiv preprint arXiv:2408.12838.
4. Chaudhuri A, Singh A, Gajbhiye S, Agrawal P (2025) Lung cancer detection using deep learning. arXiv https://arxiv.org/abs/2501.07197.
5. Chen J, Bermejo I, Dekker A, Wee L (2021) Generative models improve radiomics performance in different tasks and different datasets: An experimental study. arXiv preprint arXiv:2109.02252.
6. Hao H, Zhou Z, Li S, Genevieve M, Michael RF, et al. (2017). Shell feature: A new radiomics descriptor for predicting distant failure after radiotherapy in non-small cell lung cancer and cervix cancer. arXiv preprint arXiv:1709.09600.