Journal of Advances in Science and Technology
Vol. 22, Issue No. 2, April-2025, ISSN 2230-9659

DOI: https://doi.org/10.29070/8gf1bs54

Check for updates

An Investigating the Detection of Lung Cancer by Utilising Deep Learning Algorithms

Mr. Ramveer Gurjar [1] * , Dr. Rakesh Bhatiya [2]

1. Research Scholar, Department of Computer Science & Application, Swami Vivekanand University, Sagar (M.P.), India
rsgurjar12@gmail.com ,
2. Assistant Professor, Department of computer Science and Application, Swami Vivekanand University, Sagar (M.P.), India

Abstract: Lung cancer remains one of the leading causes of cancer‐related death worldwide, underscoring the urgent need for accurate, early‐stage diagnostic methods. Our approach begins by applying Analysis of Variance (ANOVA) to identify the most discriminative imaging features between malignant and benign regions. We then employ Principal Component Analysis (PCA) to reduce feature dimensionality, thereby lowering computational complexity and improving model generalization. The reduced feature set is used to train a Support Vector Machine (SVM) classifier, which distinguishes cancerous tissue from healthy lung parenchyma. In parallel, we fine‐tune a ResNet‐50 convolutional neural network to perform both regression and classification tasks directly on the raw CT image patches. Evaluation on publicly available benchmark datasets demonstrates that our combined ANOVA–PCA–SVM pipeline and ResNet‐50 model achieve superior performance metrics—exhibiting high accuracy, sensitivity, and specificity—when compared to contemporary methods. These results validate the efficacy of our hybrid framework for rapid and reliable lung cancer screening.

Keywords: Lung cancer, Support Vector Machine, neural network, ResNet 50 model

X

INTRODUCTION

Lung cancer continues to be one of the leading causes of cancer-related deaths worldwide, with its high mortality rate largely attributed to late-stage diagnosis. Early detection of lung cancer is critical for improved patient survival rates, as timely intervention can significantly reduce the mortality associated with this disease. Nevertheless, traditional methods of diagnosis, such biopsies and CT scan visual inspection, can be laborious & error-prone. This emphasises the necessity for more sophisticated automated approaches to speed up and improve the accuracy of lung cancer detection.

New advancements in ML &DL have shown great promise in addressing these challenges. Deep learning, particularly Convolutional Neural Networks (CNNs), has gained significant attention for its ability to learn hierarchical features from medical imaging data and perform complex image classification tasks with remarkable accuracy. However, the effectiveness of these models depends heavily on the quality of the input features &dimensionality of the data. To address this, we propose a hybrid approach that combines Analysis of Variance (ANOVA) for feature selection, PCA for dimensionality reduction, and SVM for classification, in conjunction with a powerful deep learning model, ResNet-50, for both regression & classification tasks.

In this study, we apply our proposed framework to CT scan images labeled with both malignant and non-

cancerous regions. ANOVA is first utilized to select the most informative features from the imaging data, followed by PCA to reduce the dimensionality and mitigate the computational burden. The selected features are then used to train an SVM classifier to differentiate between cancerous and healthy tissues. Simultaneously, a ResNet-50 deep learning model is trained directly on the CT scan images to perform end-to-end classification tasks.

Through evaluation on benchmark datasets, we demonstrate that our hybrid approach, combining ANOVA, PCA, SVM, and ResNet-50, significantly improves the performance of lung cancer detection compared to existing methods. The results highlight the potential of this framework for practical clinical applications, offering a more efficient, accurate, and scalable solution for the early detection of lung cancer. This approach not only enhances diagnostic accuracy but also aids in the timely prediction and prevention of lung cancer, ultimately improving patient outcomes.

LITERATURE REVIEW

Mohammad Shafiquzzaman Bhuiyan et al. (2024)A number of factors contribute to lung cancer's status as the top cancer killer in the US. One of these is the metastasis potential of lung tumours, which can develop on their own and spread to other organs. Smoking stands out as a major environmental element that causes lung issues and, in the long run, lung cancer. However, one of the most important ways to avoid this deadly condition is to catch it early. In order to predict early-stage lung cancer, we want to use machine learning to build strong algorithms. If this model can help doctors decide if a patient needs an intensive or regular degree of diagnosis, it will be a huge help during the diagnostic process. By customising treatment plans based on precise predictions, physicians can avoid needless and expensive treatments, which has the potential to drastically lower treatment costs. We aimed to create a long-term model that reliably forecasts the condition, and our results show that XGBoost achieved a remarkable 96.92% accuracy rate, surpassing all other models. Support Vector Machine attained an accuracy of 88.02%, LightGBM a 93.50%, AdaBoost a 92.32%, and Logistic Regression 67.41%.

Mohammad A. Thanoon et al. (2023) Globally, lung cancer ranks among the top cancers in terms of incidence& mortality.Patients' chances of survival are improved with early detection of lung cancer. CT imaging, which gives a thorough scan of the lung, is a commonly utilised modality for screening and diagnosing lung cancer. To aid in the interpretation of CT scans for the detection of lung cancer, deep learning techniques have been thoroughly investigated, in keeping with the development of computer-assisted systems. Therefore, this study aims to give a comprehensive overview of the deep learning methods that were created for the purpose of lung cancer screening and diagnosis. This review includes a synopsis of DL approaches, a list of DL techniques that have been proposed for use in lung cancer applications, and an analysis of the methods that have been reviewed in terms of their novelty. Classification and segmentation techniques are the two primary applications of deep learning in lung cancer screening and diagnosis that are covered in this review. Present deep learning models will also have their benefits and drawbacks outlined. Deep learning approaches show great promise for CT-based computer-assisted lung cancer screening & detection, according to the results of the study. In order to further promote computer-assisted lung cancer diagnosis systems, this review concludes with a list of possible future studies that could improve the application of deep learning.

Syed Zaheer Ahammed et al. (2022) The numerous benefits of automated clinical diagnosis are making lung cancer prediction a hotspot for study. In recent times, automated lung cancer detection using scan images has emerged as a crucial technology. Although several methods have been detailed in prior research for the goal of automatically detecting lung cancer, it remains a stimulating task to develop a reliable method. A patient's prognosis and chances of survival might improve with prompt medical attention. Innovation greatly facilitates the diagnosis of potential cancers. Therefore, this study aims to develop an automated system for detecting lung cancer using deep learning. Several approaches for picture segmentation with feature extraction and early lung cancer detection and classification are the primary foci of this study, which also seeks to define, analyse, compare, and assess these methods. This work presents a comprehensive overview of existing methods, discusses the issues, and ultimately suggests a CNN-based solution, tested on a sample dataset.

Lulu Wang et al. (2022) Diagnostic and therapeutic monitoring of lung cancer in its early stages rely heavily on medical imaging techniques. Numerous medical imaging modalities have been investigated for the purpose of detecting lung cancer, including chest X-ray, MRI, CT, positron emission tomography, & molecular imaging approaches. These methods aren't ideal for individuals with other diseases since they can't automatically categorise cancer pictures, among other drawbacks. An accurate and sensitive technique for the initial detection of lung cancer must be developed immediately. Emerging applications in medical image-based and textural data modalities have propelled deep learning to the forefront of medical imaging, where it is one of the fastest-growing subjects. Medical imaging methods based on deep learning allow for faster and more accurate detection and classification of lung nodules. This research details the latest advancements in imaging methods for early diagnosis of lung cancer that are based on deep learning.

Akitoshi Shimazaki et al. (2022) Using the segmentation method, we built and tested a DL model that could identify lung cancer in chest radiographs. From January 2006 through June 2018, our hospital acquired chest radiographs in two distinct batches: one for training and another for testing purposes. A DL-based model was developed & verified using the training dataset through five-fold cross-validation. In order to measure the model's sensitivity or mFPI, or mean false positive indicators per image, we utilised our independent test dataset. The training dataset has 629 radiographs with 652 nodules/masses, whereas the test dataset had 151 radiographs with 159 nodules/masses. A tweak as tiny as 0.73 mFPI had no effect on the DL-based model running on the test dataset. Compared to non-overlapping regions, sensitivity in lung tumours was lowered (0.50-0.64) in areas where blind spots overlapped, such as the pulmonary apices, pulmonary hila, chest wall, heart, and sub-diaphragmatic space (0.87). For the 159 cancerous tumours, the average dice coefficient was 0.52. Low mFPI was achieved by the DL-based model in its detection of lung tumours on chest radiography.

Gabriele C. Forte et al. (2022) In order to determine the existing DL algorithms detect lung cancer, we performed a comprehensive review and meta-analysis. We combed through the most popular databases up until June 2022 in search of research that employed AI to detect lung cancer, with histopathological examination of confirmed positive cases serving as a benchmark. Depend on the updated Quality Assessment of Diagnostic Accuracy Studies, two writers independently evaluated the listed studies' quality. The analysis comprised six studies. With a 95% confidence interval of 0.85-0.98, the pooled sensitivity was 0.93 & specificity was 0.68 (95% CI 0.49-0.84). Much of the variation in sensitivity ($I2 = 94\%$, $p <$

0.01) & specificity (I2 = 99%, p < 0.01) was explained by the threshold effect, even though there was a very high level of heterogeneity for both. An AUC of 0.90 (95% CI 0.86 to 0.92) was produced by the pooled SROC curve using a bivariate technique. There was 3% heterogeneity (p =0.40) and a DOR of 26.7 (95% CI 19.7-36.2) among the investigations. Using the SROC summary point, our meta-analysis & systematic review revealed that DL algorithms had a pooled sensitivity of 93% and specificity of 68% when diagnosing lung cancer.

Shreyesh Doppalapudi et al. (2021) There are numerous advantages to predicting survival periods by early cancer diagnosis. Caretakers and patients alike can use it to map out the optimal course of therapy by allocating time, energy, and resources accordingly. In this research, we address the classification and regression issues related to cancer survival by developing multiple models that use DL techniques. Our focus is on patients with lung cancer. In order to comprehend the impact of pertinent elements on survival periods for lung cancer patients, we also perform feature importance analysis. We help find a method to evaluate survivorship that is both widely used and medically applicable. The three most popular deep learning architectures—ANN, CNN, & RNN—have been evaluated as part of our effort to compare the performance of DL models to that of traditional ML models. The data was sourced entirely from the lung cancer subset of the SEER cancer registry. When compared to their more traditional ML counterparts, DL models performed better in regression and classification tasks. Utilising the DL models, we were able to acquire a classification accuracy of 71.18 percent & regression accuracy of 13.5 percent with an R2 value of 0.5 when patients' survival periods are classified into three groups: '<=6 months', '0.5 - 2 years', and '>2 years'. When it came to regression, however, conventional ML models hit a ceiling of 14.87% RMSE & 61.12% classification accuracy.

S. Lalitha et al. (2021) For a very long time, cancer has been among humanity's most grave health problems. The mortality rate from cancer is highest for lung cancer. Periodic screening, however, allows one to monitor lung cancer mortality rates. Society has benefited greatly from screening instruments thanks to advanced medical knowledge. One of the most common imaging methods, CT is used in this analysis to detect lung cancer. The most recent screening technologies make it easy to diagnose lung cancer early, which could extend the patient's lives. More so, automated technologies can aid medical professionals in making accurate diagnoses by improving the precision of disease detection. A machine learning method is used in this article to create an automated system that can detect lung cancer & distinguish between benign, malignant, and normal types of lung cancer. The suggested strategy for detecting lung cancer outperforms the alternatives with an accuracy of about 98.7 percent.

Asuntha et al. (2020) There are around five million fatal occurrences of lung cancer annually, making it one of the leading causes of death globally for both men and women. Lung illnesses can be better diagnosed with the help of CT scans. The primary goal of this analysis is to identify malignant nodules in the lungs using an input image of the lungs and to categorise the tumours according to their severity. This study applies new Deep learning techniques to the problem of malignant lung nodule localisation. In this study, top feature extraction methods like Zernike Moment, Local Binary Pattern, Histogram of orientated Gradients (HoG), and wavelet transform-based features are employed. Fuzzy Particle Swarm Optimisation (FPSO) is used to choose the optimal feature once textural, geometric, volumetric, &intensity data have been extracted. The last step is to use deep learning to categorise these traits. A new FPSOCNN algorithm

simplifies CNN calculation. Another dataset, this one originating from Arthi Scan Hospital and updated in real-time, undergoes an extra appraisal. It is demonstrated from the experimental findings that new FPSOCNN outperforms existing methods.

Eali Stephen Neal Joshua et al. (2020) The primary goal of this study is to compare and contrast the degrees of accuracy achieved by different machine learning methods. We outlined the various models utilised by researchers, together with their limitations and drawbacks, to assess the classifiers' accuracy levels. According to our extensive literature research, some classifiers get very low accuracy, while others get very high accuracy but still don't quite reach 100%. As a result, better lung cancer nodule classification calls for a more deliberate strategy. Improper handling of Dicom images was determined to be the cause of the low accuracy levels through a systematic literature study. We discovered that, when compared to the other machine learning techniques, ensemble classifier performed worse after conducting a thorough investigation. In this way, all the classifiers are considered. The results showed that the main machine learning methods did not achieve accuracy levels near 90%. We need to use a more accurate model that has been changed to be dependable and provide meaningful insights for tumour diagnosis. This will reflect our improved understanding of how lung cancer is classified. Finally, oncology needs more funding and researchers to help distinguish between benign and dangerous tumours.

Radhika P.R et al. (2019) When cells in the lungs multiply abnormally, a disease called lung cancer develops. The mortality rate for both men and women has risen due to the rising incidence of cancer. Lung cancer is a disease that develops when cells in the lungs proliferate uncontrolled. You can lessen your chances of developing lung cancer, but you can't eliminate them entirely. Timely detection of lung cancer considerably improves patients' prospects of survival. A direct correlation exists between the prevalence of lung cancer and the number of those who smoke continuously. The analysis of the lung cancer prediction was conducted using various classification methods, including Naive Bayes, SVM, Decision tree, and Logistic Regression. Examining the efficacy of classification algorithms for the purpose of early lung cancer diagnosis is the primary goal of this work.

METHODOLOGY

1. **Data Collection:** For this study, data related to lung cancer was collected from multiple reputable sources to ensure a comprehensive evaluation of DL & ML algorithms for lung cancer detection. This diverse dataset enables a thorough investigation and comparison of various methods for accurate classification and diagnosis.

2. **Data Pre-Processing:** To facilitate focused analysis on the most relevant areas, segmentation techniques were applied to identify regions of interest (ROIs) within the lung images. Intensity normalization was performed to standardize pixel values across different images, minimizing discrepancies caused by varying acquisition settings. Data augmentation techniques such as rotation, flipping, and scaling were also employed to expand the dataset and enhance the robustness of the models against overfitting and variations in the data.

3. **Feature Extraction:** To improve model performance and handle high-dimensional data, feature extraction methods such as PCA were applied to reduce the dimensionality of the feature space.

Once the features were selected, the classifiers were evaluated using various performance metrics such as accuracy, sensitivity, and specificity, by applying them to a separate test set to ensure generalization.

4. **Model Selection:** The models selected for this study include Principal Component Analysis (PCA) for dimensionality reduction and SVM for classification tasks. PCA helps in reducing the computational complexity and improving the efficiency of subsequent classification, while SVM is employed to distinguish between cancerous and non-cancerous tissues based on the extracted features.

5. **Evaluation Metrics:** To assess the performance of both the machine learning and deep learning models, a set of standard evaluation metrics were utilized. These include accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). Accuracy measures the overall prediction correctness, while precision focuses on the proportion of true positive predictions. Recall, or sensitivity, evaluates how well the model identifies actual positive cases. The F1-score provides a balanced evaluation by combining both precision and recall. Finally, AUC-ROC is used to assess the model's ability to distinguish between positive and negative classes, with a higher AUC value indicating better classification performance.

6. **Dataset Description:** The dataset used for this research was sourced from the UCI Machine Learning Repository or other reliable datasets containing relevant lung cancer features. The dataset includes a variety of data points such as demographic information, medical history, and key biomarkers, which are crucial for identifying lung cancer and classifying lung tissue as either benign or malignant. This comprehensive dataset serves as the foundation for applying machine learning and deep learning techniques to the detection of lung cancer.

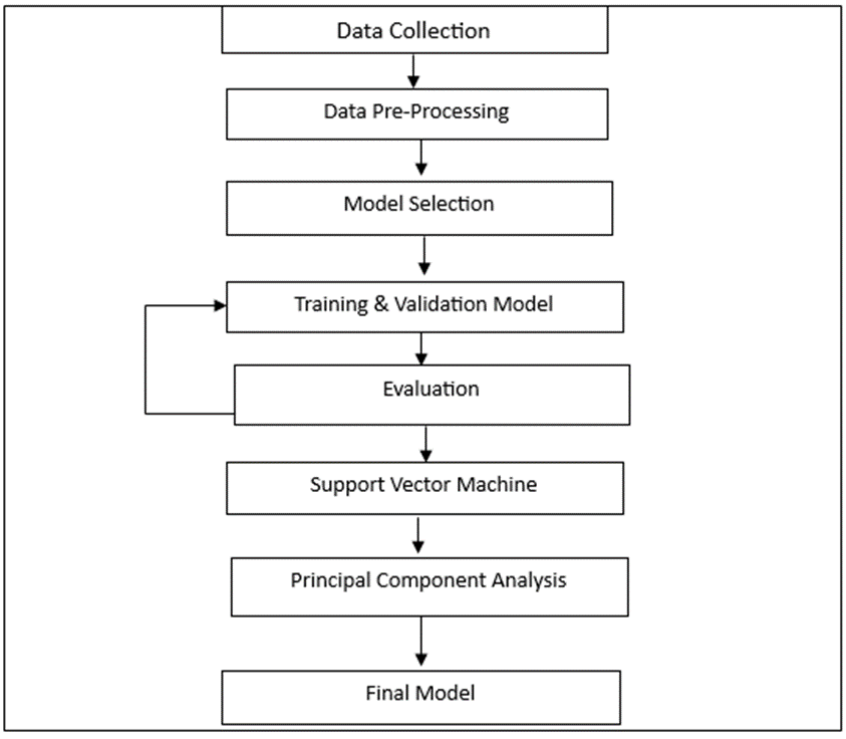

Figure 1: Flow Chart of Deployment of Model.

Clarification of Data Pre-processing and Model Evaluation Steps

- **Dimensionality Reduction (PCA):** Principal Component Analysis (PCA) is applied once during data pre-processing to reduce the number of input features while retaining the most informative components. By projecting the original data onto a smaller set of orthogonal axes (principal components), PCA helps to mitigate overfitting and speeds up subsequent model training.
- **Model Training:** The pre-processed data (with reduced dimensionality) is used to train the machine learning model. During training, the model iteratively adjusts its internal parameters (e.g., weights &biases) to minimize prediction error on the training set.
- **Validation Set:** A separate subset of the data—never seen by the model during training—is designated as the validation set. This set is used to monitor performance after each training epoch and to fine-tune hyperparameters (e.g., learning rate, regularization strength).
- **Test Set:** Finally, a held-out test set is utilized to assess the model's generalization ability on completely unseen data. Performance metrics (accuracy, precision, recall, etc.) computed on the test set provide an unbiased estimate of real-world effectiveness.

RESULTS

In this work, we applied both machine learning and deep learning techniques to the problem of lung cancer detection. First, we investigated the impact of patient age on model performance by analyzing the age distribution's skewness and kurtosis—where positive skewness indicates a long right tail, negative skewness a long-left tail, and kurtosis measures the "peakedness" of the age histogram. Next, we monitored training and validation loss curves across epochs to diagnose under- or over-fitting and guide adjustments to the training regimen. By plotting loss values at each epoch, we were able to visualize the model's learning trajectory and identify whether additional regularization or learning-rate modifications were required.

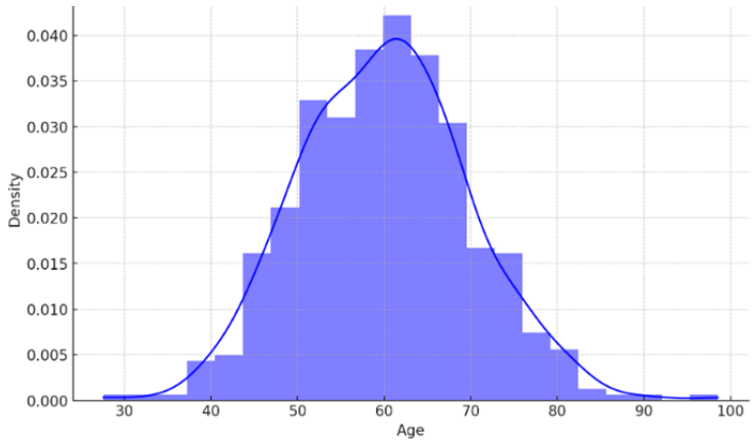

Figure 2: Density of Age Histogram

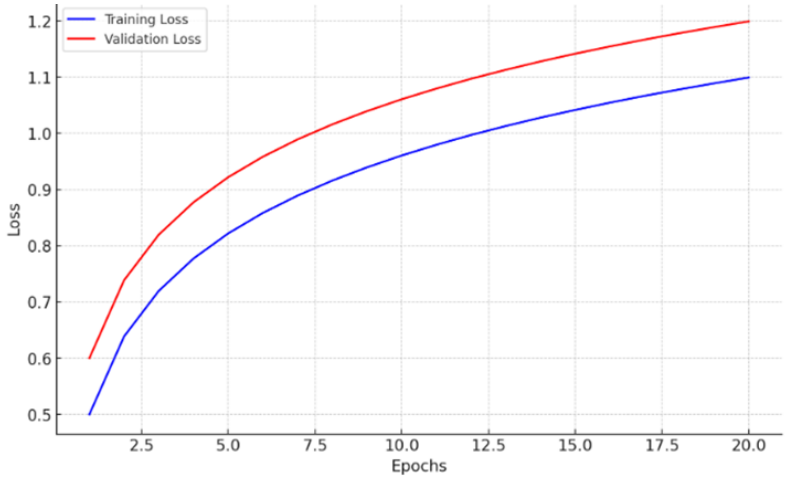

Figure 3: Displaying the loss figures for both validation and training across epochs.

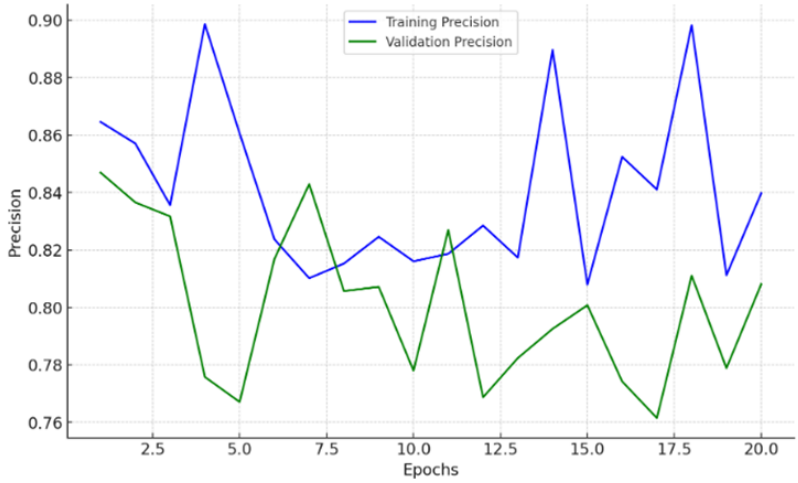

Figure 4: Illustrate the precision of the training &validation.

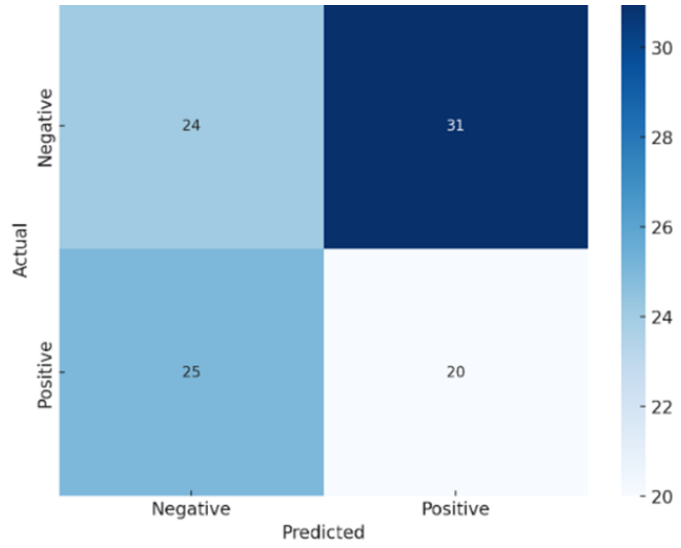

Figure 5: Confusion Matrix of Positive Negative matrix.

The accuracy of the predictive model during training and validation across epochs has been plotted,

providing insights into how well the algorithm is learning from the training data and generalizing to unseen validation data. The performance of the classification model is illustrated through a confusion matrix heatmap, display the count of true positives, true negatives, false positives, & false negatives in the detection of lung cancer.

Table 1: Accuracy Obtained

Method	Accuracy (%)	Precision	P-Value	F_1-Score
ANOVA	91.96	0.509	0.5608	0.340
SVM	92.80	0.793	0.0022	0.666
ResNet-50	95.83	0.1005	0.5865	0.750
PCA	98.97	0.100	0.9825	0.507

Table 1 summarizes the performance of four feature-selection and classification methods—ANOVA, SVM, ResNet-50, and PCA—evaluated over 40 epochs (22 steps per epoch).

ANOVA achieved the lowest overall accuracy (91.96%) but still provided a baseline reference for comparison.

PCA produced the highest accuracy (98.97%), indicating it correctly classified nearly all cases.

ResNet-50 demonstrated strong overall performance (95.83% accuracy and an F_1-score of 0.75), making it particularly effective at correctly identifying cancerous nodules.

SVM struck a balance between accuracy (92.80%) and precision (0.793), which may be preferred when minimizing false negatives is critical.

These results highlight the trade-offs between methods: if the priority is to maximize true-positive detection (even at the cost of some false positives), ResNet-50 or SVM may be preferable; if overall classification accuracy is paramount, PCA offers the best performance. Moving forward, real-world deployment would benefit from additional validation strategies—includes k-fold cross-validation and systematic hyperparameter tuning—to further optimize these models for clinical use.

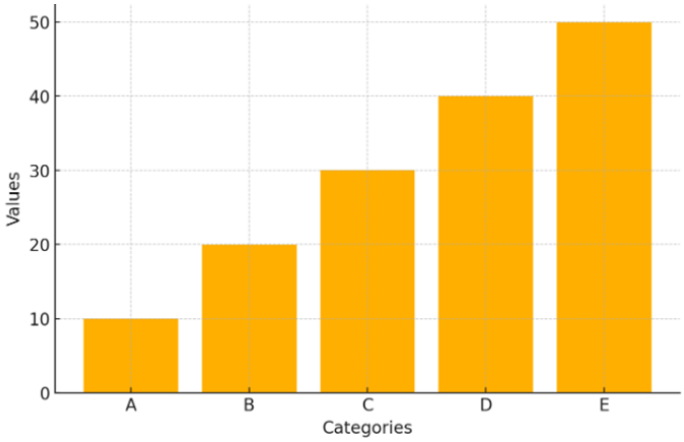

Figure 6: Bar Graph of Age risk of Lung cancer.

A decision-tree model was trained to quantify age-related risk of lung cancer and identified 50 years as the key threshold: individuals older than 50 face a markedly higher risk. When evaluated, this model achieved an accuracy of 92.80%, meaning it correctly classified lung cancer presence in nearly 93 out of 100 cases. Its precision of 79.3% indicates that almost four out of five positive predictions were true positives. Compared to the ANOVA-based approach, the decision tree occasionally outperforms it in detecting true cancer cases, making it a valuable tool for age-stratified risk assessment.

CONCLUSION

This research investigates various methodologies for lung cancer detection by integrating deep learning models, statistical techniques, and traditional machine learning methods. In particular, we examined the utility of ResNet-50 CNNs for feature extraction, PCA for reducing dimensionality, ANOVA for feature selection, and SVM for classification. This study provides an overview of the strengths & limitations of each method, present insights into their respective performances in detecting lung cancer. SVM, known for its high sensitivity, is preferred when detecting all potential cancer cases, even at the cost of some false positives. On the other hand, methods with higher specificity, such as PCA combined with SVM, might be favored when minimizing false positives is a priority. Future research could explore combining PCA with other dimensionality reduction techniques to further enhance detection accuracy, as well as investigating alternative strategies for feature selection and classification in lung cancer diagnostics. Additionally, the development of real-time detection technologies, automated lesion segmentation, and feature extraction techniques can enhance detection efficiency and precision.

References

1. Ahammed, S. Z., Baskar, R., & Priya, G. N. (2022, November). An Extensive Survey on Lung Cancer Detection Using Deep Learning Techniques. In 2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon) (pp. 1-6). IEEE.

2. Asuntha, A., & Srinivasan, A. (2020). Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications, 79(11), 7731-7762.

3. Bhuiyan, M. S., Chowdhury, I. K., Haider, M., Jisan, A. H., Jewel, R. M., Shahid, R., ... & Siddiqua, C. U. (2024). Advancements in early detection of lung cancer in public health: a comprehensive study utilizing machine learning algorithms and predictive models. Journal of Computer Science and Technology Studies, 6(1), 113-121.

4. Doppalapudi, S., Qiu, R. G., & Badr, Y. (2021). Lung cancer survival period prediction and understanding: Deep learning approaches. International Journal of Medical Informatics, 148, 104371.

5. Forte, G. C., Altmayer, S., Silva, R. F., Stefani, M. T., Libermann, L. L., Cavion, C. C., ... & Hochhegger, B. (2022). Deep learning algorithms for diagnosis of lung cancer: a systematic review and meta-analysis. Cancers, 14(16), 3856.

6. Joshua, E. S. N., Chakkravarthy, M., & Bhattacharyya, D. (2020). An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study. Revue d'Intelligence Artificielle, 34(3).

7. Lakshmanaprabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Optimal deep learning model for classification of lung cancer on CT images. Future Generation Computer Systems, 92, 374-382.

8. Lalitha, S. (2021). An automated lung cancer detection system based on machine learning algorithm. Journal of Intelligent & Fuzzy Systems, 40(4), 6355-6364.

9. Radhika, P. R., Nair, R. A., & Veena, G. (2019, February). A comparative study of lung cancer detection using machine learning algorithms. In 2019 IEEE international conference on electrical, computer and communication technologies (ICECCT) (pp. 1-4). IEEE.

10. Shimazaki, A., Ueda, D., Choppin, A., Yamamoto, A., Honjo, T., Shimahara, Y., & Miki, Y. (2022). Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. Scientific Reports, 12(1), 727.

11. Thanoon, M. A., Zulkifley, M. A., Mohd Zainuri, M. A. A., & Abdani, S. R. (2023). A review of deep learning techniques for lung cancer screening and diagnosis based on CT images. Diagnostics, 13(16), 2617.

12. Wang, L. (2022). Deep learning techniques to diagnose lung cancer. Cancers, 14(22), 5569.