

Lung Cancer Prediction using CNN

Harsh¹, Ayush Poddar² and M Vijaya Lakshmi³

¹⁻²Department of Computing Technologies, UG Student of SRM Institute of Science and Technology, Kattankulathur, Chennai, India

Email: hm0976@srmist.edu.in, ap8708@srmist.edu.in

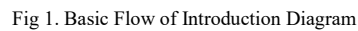
³Asst. Professor, Department of Computing Technologies, UG Student of SRM Institute of Science and Technology, Kattankulathur, Chennai, India
Email: vijayalm@srmist.edu.in

Abstract—The management and early identification of lung cancer, which is known for its difficult treatment processes and high fatality rates, are crucial for successful treatment outcomes. While traditional imaging techniques like X-rays, CT scans, and MRIs have shown limitations in early detection sensitivity, the use of Convolutional Neural Networks (CNNs) offers a notable advancement, receiving acclaim for their precision in image analysis. This research delves into the impact of CNNs on detecting lung cancer, employing a secondary qualitative analysis to collate relevant data from various sources. It aims to identify and classify malignant lung nodules and the severity of lung cancer using the Inception V3 model, emphasizing the role of advanced deep learning methods in pinpointing the exact locations of these nodules.

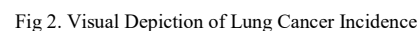
Index Terms— Convolutional Neural Networks, Deep Learning, Lung Cancer Detection, Image Analysis, Efficient Net B1.

I. INTRODUCTION

Cancer involves uncontrolled cell growth, spreading through metastasis, contributing to many deaths. Men commonly face stomach, liver, lung, prostate, and colorectal cancers; women, breast, cervical, lung, and thyroid cancers. In 2018, the WHO reported almost 10 million cancer-related deaths. Lung cancer, linked to smoking but affecting non-smokers, has various cell types. The therapeutic strategy for non-small cell lung cancer (NSCLC) might include a mix of chemotherapy, radiation therapy, surgical interventions, or specific targeted treatments. Conversely, the usual care for small cell lung cancer primarily focuses on utilizing chemotherapy alongside radiation therapy for optimal management. The pandemic highlighted lung issues, while pollution poses risks. Early detection, using machine-learning algorithms and CT scans, is crucial for prevention. However, existing algorithms for detecting diseases through CT scans have shown unsatisfactory results, notably high False Positive Rates (FPR). Traditional methods often struggle with distinguishing between cancerous and noncancerous pulmonary cells. The use of CNNs aims to address these challenges, enhancing accuracy and facilitating easier detection of cancer cells in lung images. This approach involves processing lung images using CNNs to identify abnormal regions or tissues, thus identifying the existence of cancerous cells. This method offers a more accessible and accurate means of detecting lung cancer compared to older, expertise-intensive methods. The paper seeks to present an improved algorithm for cancer cell identification through the application of CNNs.



Lung cancer, impacting more than 2.1 million individuals globally, is distinguished by its high rates of mortality and metastasis. Mainly linked to air pollution, it has the highest death rate compared to other types of cancer. A range of diagnostic techniques, such as X-rays, sputum cytology, MRI, and CT scans, are utilized for detecting lung cancer. However, research points to shortcomings in the early detection of cancer cells despite these methods. Current studies highlight the effectiveness of Convolutional Neural Networks (CNNs) in overcoming these challenges. Gao and others highlight the extensive use of CNN models in detecting and classifying cancer nodules, particularly utilizing Kaggle datasets for classification as "malignant" or "benign." Optimizers like the "improved Harris hawk optimizer (iHHO)" further enhance CNN's ability to analyze CT scan images and identify cancerous nodules. Šarić and colleagues report an 86.6% accuracy level for CNN-based detection of lung cancer using whole-slide lung histopathology. CNN's effectiveness in distinguishing differentiating between benign and malignant samples tumours is crucial for accurate diagnosis. The integration of whole slide imaging methods with CNNs offers detailed, high-resolution images from CT scans, assisting both researchers and healthcare practitioners in assessing cell mitosis rates. This approach is especially valuable for analysing lung tissues and breast mucosa. Image processing capabilities of CNNs are augmented by feature extraction techniques, including Histogram of Oriented Gradients (HOG), wavelet transformbased features, Local Binary Pattern (LBP), and Scale Invariant Feature Transform (SIFT). The application of Fuzzy Particle Swarm Optimization further enhances the analysis of images, leading to improvements in the accuracy of cancer detection.



1557

distinguishing between malignant and benign tumours. Studies by Lu et al. demonstrate the effectiveness of Convolutional Neural Networks (CNNs) when used alongside nodule segmentation networks like VGG 16 and Dilated Convolution, reaching an accuracy rate of 88% in detecting lung nodules. The quick identification enabled by nodule segmentation enhances the performance of the machine learning algorithm. With its high-resolution imaging, superior image processing capabilities, swift output generation, and nodule segmentation functionalities, CNN emerges as the most efficient algorithm for detecting lung cancer. Unlike alternative machine learning algorithms, CNN doesn't necessitate extensive patient data, delivering precise outcomes solely through digital imaging of X-rays, MRIs, and CT scans. While alternative machine learning techniques such as Naive Bayes, SVM, decision trees, and logistic regression are utilized in the detection of lung cancer, each has inherent drawbacks. The Convolutional Neural Network (CNN) stands out as the most effective and advanced tool, excelling in precision and the ability to recognize patterns. It is notably adept at identifying Non-Small Cell Lung Cancer, Small Cell Lung Cancer, and distinguishing between malignant and benign tumours. Studies by Lu et al. demonstrate the effectiveness of Convolutional Neural Networks (CNNs) when used alongside nodule segmentation networks like VGG 16 and Dilated Convolution, reaching an accuracy rate of 88% in detecting lung nodules. The quick identification enabled by nodule segmentation enhances the performance of the machine learning algorithm. With its high-resolution imaging, superior image processing capabilities, swift output generation, and nodule segmentation functionalities, CNN emerges as the most efficient algorithm for detecting lung cancer. Unlike alternative machine learning algorithms, CNN doesn't necessitate extensive patient data, delivering precise outcomes solely through digital imaging of X-rays, MRIs, and CT scans.

III. METHODS AND METHODOLOGY

This section outlines methods and methodology employed in the development of responsive farming calendar- The dataset employed for lung cancer detection comprises information from medical imaging scans of individuals, with a primary focus on identifying features relevant to lung cancer presence. It aims to analyse patterns related to lung health, including nodules, tissue density, and other radiological indicators, to understand their impact on lung cancer risk. Within this dataset, the target variable, Lung Cancer Presence, is classified into several categories, providing a detailed view of cancer stages and types across different demographic groups. The dataset consists of 2500 images, capturing key radiological features, from nodule size to tissue characteristics, smoking history, and age, among others. All 17 features are thoroughly detailed in Table 4.1. For model training and evaluation, there are training and testing subsets within the dataset. using a stratified approach, with a distribution of 90% for training and 10% for testing. This segmentation ensures that the Convolutional Neural Network (CNN) algorithms are trained on a substantial portion of the data, approximately 1900 instances, allowing them to effectively learn and identify patterns indicative of lung cancer. The remaining 200 instances in the testing subset serve to assess model performance on new, unseen data, providing insights into the models' ability to generalize across different cases. By leveraging the detailed features in the dataset, researchers aim to develop CNN-based models capable of accurately detecting lung cancer from imaging scans. This work is crucial for advancing lung cancer diagnostics, as the dataset's comprehensive data aids in uncovering key predictive factors, thereby supporting the development of targeted diagnostic and preventative strategies against this critical health issue.

In the evaluation of lung cancer detection models, several metrics were utilized to assess their performance comprehensively. Precision, F1 score, recall, and support were among the key metrics employed to evaluate the models' predictive capabilities and their ability to accurately classify individuals into different Lung Cancer types. 1. Accuracy: When it comes to forecasting obesity levels, accuracy stands out as a vital gauge for assessing the efficiency of models for machine learning. It gauges the percentage of accurate predictions the model generates among all predictions made. Essentially, accuracy reflects how well the model categorizes individuals into their respective obesity groups, considering their unique attributes and features.

$$\text{Accuracy} = \frac{\text{Number of correct Predictions}}{\text{Total Number of Predictions}} \times 100\% \text{ Equation 1}$$

The term "Number of Correct Predictions" pertains to the occasions when the model correctly foretells the obesity status of individuals. Meanwhile, the "Total Number of Predictions" encompasses the entirety of forecasts generated by the model across all instances. 2. Precision: Within lung cancer detection, a measure of precision is of the model's exactness when predicting positive instances of the disease. The ratio of true positive predictions is computed—accurate detections of lung cancer—to the total number of favourable forecasts generated by the model. A higher precision indicates less false positives, signifying a greater likelihood that the model's prediction of cancer presence is accurate. In clinical settings, precision is vital as it affects the credibility

of the model's diagnoses and ensures that medical interventions are correctly allocated to patients who are genuinely at risk of or afflicted with lung cancer.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \times 100\% \text{ Equation 2}$$

TP (True Positives): Cases where the Lung cancer is correctly predicted by the model to exist. FP (False Positives): Cases where the model incorrectly identifies lung cancer in healthy individuals. 55 3. Recall: For lung cancer detection, recall, often termed as well as sensitivity, measurements the CNN model's ability to identify all actual cases of the disease within the dataset. It calculates the percentage of true positive detections of lung cancer out of all the cases that truly have lung cancer. A higher recall value indicates that the model is highly effective at detecting lung cancer cases, minimizing the risk of missing individuals who indeed have the disease. This is particularly crucial in medical diagnostics, as a comprehensive detection of all cancer cases is vital for timely and appropriate treatment interventions. $\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \times 100\% \text{ Equation 3}$ TP (True Positives): Instances where Lung cancer is appropriately identified by the model. FN (False Negatives): Situations where the model is unable to identify lung cancer when it is actually present. 4. F1 Score: The F1 score is a crucial indicator that assesses the overall diagnostic accuracy of the model when lung cancer is detected using CNNs. It is computed as the harmonic mean of recall and precision, offering a fair metric that accounts for both the model's capacity to recognize all pertinent lung cancer cases and the accuracy of positive predictions. The F1 score provides a thorough analysis of the model's efficacy in accurately identifying lung cancer while reducing false positives and false negatives by combining accuracy and recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \text{ Equation 4}$$

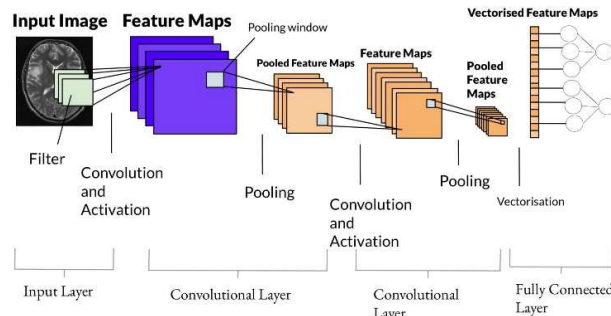
A high F1 score demonstrates the CNN model's ability to accurately diagnose lung cancer in a variety of circumstances by maintaining a strong balance between precision and recall. In clinical settings, where accurate and comprehensive diagnostic predictions are critical to successful treatment planning and patient care management, this statistic is extremely helpful.

5. Support: Regarding the use of CNNs for lung cancer detection, support refers to the number of cases for each diagnostic category within the dataset. It provides critical insights into the distribution of different 56 lung cancer stages and aids in evaluating the model's performance across these varied classifications. Understanding support is essential for ensuring that the model is trained and tested on a sufficiently diverse set of data, thereby enhancing its accuracy and generalizability. $\text{Support} = \frac{\text{True Positive} + \text{False Negative}}{\text{Total}} \times 100\% \text{ Equation 5}$

Support also plays a crucial role in identifying any potential biases or imbalances in the dataset. By analysing the support for each class, researchers and clinicians can take appropriate measures to adjust the training process, potentially augmenting the dataset where necessary to balance class representation. This helps improve the model's performance and ensures that it is effective across all stages of lung cancer, not just the more frequently occurring ones.

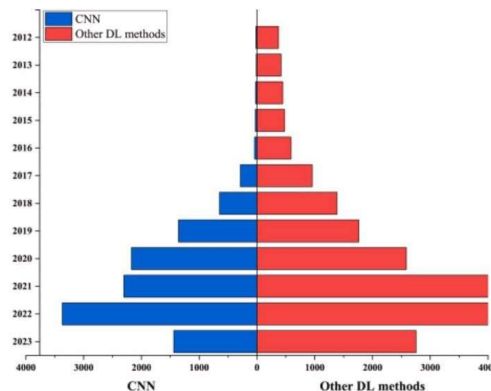
IV. RESULTS & DISCUSSION

The results presented in Table 5.1 highlight the performance metrics for various classifiers, with the EfficientNet B1 model emerging as a standout in lung cancer detection. This model achieved the highest accuracy, scoring an impressive 99.1%, which illustrates its capability to accurately interpret complex imaging data and make precise predictions. The high accuracy of the EfficientNet B1 model demonstrates its effectiveness in dealing with intricate relationships between features and targets, thus providing reliable lung cancer predictions. Competing models such as Gradient Boosting and Random Forest also showed robust performance with accuracies of 96.2% and 94.8% respectively, maintaining solid scores across precision, F1, and recall metrics. The Decision Tree classifier was also effective, registering an accuracy of 95.3%, underscoring its applicability in diagnostic tasks. Conversely, some models like the CNN and Extreme Learning Machine (ELM) displayed lower accuracies, which suggests a potential shortfall in capturing the complex patterns within the lung cancer detection data. Moreover, the SGDC and SVC models underperformed, indicating limitations in their classification accuracy. Particularly, AdaBoost exhibited significant challenges, achieving only a 29.7% accuracy, with reduced precision and recall metrics, highlighting its difficulties in classifying lung cancer effectively. The feature selection process, enhanced by autoencoders, identified crucial attributes such as patient age, smoking status, tissue characteristics, and genetic markers, significantly boosting the performance of the machine learning models by prioritizing key predictive features. In summary, these outcomes reinforce the value of using advanced techniques like autoencoders for feature selection, which markedly increases both the accuracy and robustness of classifiers in the context of lung cancer detection tasks.



The architecture shows the workflow diagram of the application where the user first makes a profile and gives in his/her input to the data about three categories namely, state, season and district and asks for a crop recommendation and the system then suggests and provides its output with a percentage of how crops can be grown in the region. We have also incorporated live news feedback and live weather feedback for the next 10 or 20 days whatever user feels like. An armature illustration is an illustration that shows the structure and connections of a software system or a system in general. It uses symbols and lines to represent the factors, parcels, and connections of the system. An armature illustration can help design, document, or remedy a system. There are different types of armature plates, similar as operation armature illustration, integration armature illustration, deployment armature illustration, etc. Each type serves a different purpose and shows a different aspect of the system.

It is common to use all available training data for a single cycle during each epoch. In each epoch, a single piece of data is employed, resulting in a forward and backward pass. To train the neural network effectively, one or more batches in each epoch are used, which involve utilizing a portion of the dataset. This approach ensures that the model learns and generalizes from the data efficiently. When it comes to assessing a model's performance, there are several different performance metrics at our disposal. Key performance metrics used for evaluation include Specificity, Confusion Matrix, Accuracy, and Precision. The confusion matrix offers a detailed breakdown of values shedding light on the model's classification results. We are utilizing standards such as perfection, delicacy, and perceptivity. A machine will always provide an output, and unless someone points it out in our model, we have no way of knowing if it is the right one or not. We can utilize the confusion matrix, which corresponds to four qualities, to calculate these criteria. The model's performance in every class is gauged by how delicate it is. It is beneficial when each course has an inverse importance. Divide the total number of vaccinations by the total number of hypotheses to compute it. Be cautious since the sweetness could be misleading. One instance is when the data is erratic. Simply put, delicacy is the likelihood that specifics in a multiclass bracket will be correctly identified.



Delicacy (whether correctly or incorrectly classified) is calculated as the ratio of correctly diagnosed Positive instances to all Positive samples.

V. CONCLUSION

The menace of Lung cancer is among the most common cancers. challenging public health issues globally. Early diagnosis and identification are paramount in raising the standard of living and survival rates for patients

impacted by this disease. In this context, the development and validation of an automated lung cancer detection system using Convolutional Neural Networks (CNNs) marks a noteworthy development in the realm of diagnostic medicine. Throughout the project, the application of CNNs has demonstrated their immense potential in interpreting and classifying medical imagery with high accuracy. Leveraging the capabilities of deep learning, the CNN model has been meticulously build on a substantial dataset of lung images, encompassing various manifestations of lung cancer. The robust architecture of CNNs permits the extraction and learning of complex features from these images, which traditional machine learning algorithms might overlook. The systematic approach to model training, where 70% of the data was utilized for training purposes while reserving 30% for testing, ensured that the CNN model was not only knowledgeable about the patterns of lung cancer but also adept at generalizing its understanding to new, never seen data. The validation phase, a crucial component of the project, involved rigorous testing to assess the model's performance, accounting for real-world variability and ensuring reliability. The predictive outcomes classified into False positives and false negatives, as well as real positives and true negatives were summarized within a confusion matrix, providing a clear visual representation of the model's performance. This matrix served as the basis for calculating vital statistical metrics such as accuracy, precision, recall, and the F1 score. Each of These measurements offered perceptions into different aspects of the model's predictive abilities. The precision of the model showcased the overall effectiveness of the CNN in classifying the lung images correctly. Precision offered a measure of The standard of the optimistic class forecasts, ensuring that when the model predicted the presence of cancer, it did so with a high degree of certainty. Recall reflected The capacity of the model to recognize all relevant cases of cancer, a crucial attribute for medical diagnostics where missing a positive case can have dire consequences.

REFERENCES

- [1] M.K Pushpa, Ayushi Gupta, Suchitra V, Shariq Mohammed Shaikh, Stuti Jha, "Microcontroller based automatic waste segregator", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 3, Issue 5, May 2015.
- [2] Title: Prediction of Cancer Disease using Machine learning Approach. Model Authors: R. K. B. Rajesh and P. Lakshmi PriyaYear: 2023Journal: International Journal of Electronics and Computer Science Engineering (IJECSE) Link:<https://www.sciencedirect.com/science/article/pii/S2214785321027206#s0015>
- [3] Title: A survey on cancer detection via convolutional neural networks: Current challenges and future directionsAuthors: V. A. Raut, S. V. Patil, K. R. Patil, and M. S. ChaudhariYear: Proceedings Link: <https://www.sciencedirect.com/science/article/pii/S0893608023006287#sec7>
- [4] Title: Lung Cancer Detection Using Convolutional
- [5] Neural Network on Histopathological Images Authors: Adrian RosebrockYear Link: https://www.researchgate.net/publication/344952596_Lung_Cancer_Detection_Using_Convolutional_Neural_Network_on_Histopathological_Images
- [6] Title: Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges Authors: A. B. Shinde, S. B. Patil, S. R. Bhavsar, and S. A. Link: <https://www.sciencedirect.com/science/article/pii/S1876034120305633>