Automated Lung Cancer Detection Using CNN from CT Scan Images: A Deep Learning Approach

Kuldeep Prajapati [1], Suraj Yadav[2]

[1]M.Tech. CSE Scholar, Department of Computer Science, Jagan Nath University, Jaipur
[2]Assistant Professor-CSE, Department of Computer Science, Jagan Nath University, Jaipur

Abstract:
Lung cancer remains one of the leading causes of cancer-related mortality worldwide, largely because it is diagnosed at advanced stages when treatment options are limited. Conventional diagnosis relies on radiologists manually reviewing CT scans, a process that is time-taking and prone to error. This paper presents an lung cancer detection framework using a Convolutional Neural Network (CNN) to classify chest CT images as cancerous or normal without relying on hand-crafted features. The pipeline covers dataset acquisition from Kaggle, image preprocessing (resizing, normalization, noise reduction, and augmentation), automatic CNN-based feature extraction, and binary classification. The proposed CNN achieved 98.73% accuracy, 99% precision, 99% recall, an F-measure of 98.52%, and an error rate of 1.27%, outperforming four classical machine learning baselines — Decision Tree, Logistic Regression, Random Forest, and Naive Bayes — by a substantial margin.

Keywords: Lung Cancer Detection, CNN, Deep Learning, CT Scan Classification, Medical Image Analysis

1. Introduction

Lung cancer is among the most commonly diagnosed and lethal cancers globally, arising from the unregulated and abnormal growth in lung tissue. Cigarette smoking is the dominant risk factor, responsible for over 80% of cases, while second-hand smoke, radon exposure, asbestos, occupational carcinogens, air pollution contribute to disease incidence.

A central challenge is that lung cancer is frequently asymptomatic in its early stages; by the time symptoms such as consistent cough, hemoptysis, pain in chest, or unexplained weight loss appear, the disease has often progressed to a stage where treatment options are limited.

Because early detection strongly predicts survival, diagnostic tools that can flag suspicious CT findings quickly and consistently are urgently needed. Classical machine learning algorithms such as Decision Trees, Logistic Regression, Random Forest, and SVM have been widely applied to lung cancer prediction, but depend on manually engineered features that limit their ability to capture the full complexity of images. Convolutional Neural Networks (CNNs), by contrast, learn hierarchical spatial features directly from raw pixel data, making them well suited to CT-based lung cancer detection.

This paper reports a CNN-based lung cancer detection system trained and tested on a labelled CT image dataset.

2. Related Work

A substantial body of research has explored classical machine learning and deep learning approaches to lung cancer detection. Several studies compare Naive Bayes, Random Forest, Logistic Regression, and Decision

Tree classifiers on Kaggle-sourced datasets, generally finding tree-based methods such as Random Forest outperform simpler linear models [1], [2], [3]. Symptom- and lifestyle-based prediction using structured clinical records has also been explored with SVM, k-NN, Gradient Boosting, XGBoost, and CatBoost, with SVM and Logistic Regression frequently emerging as strong performers [4], [5], [3].

Within imaging-based detection, CNNs and CNN-hybrid architectures dominate recent literature. Several works combine CNN-based feature extraction with an SVM classifier to filter irrelevant information before classification, achieving accuracies close to 98% [6]. Others apply dedicated CT-scan benchmarks such as the IQ-OTH/NCCD dataset using handcrafted texture descriptors (Gabor filters, Gray-Level Co-occurrence Matrix) with SVM kernels, or CNN pipelines on Iraqi hospital CT datasets that later became reused community benchmarks [7], [8], [9]. Comparative studies of transfer learning architectures — MobileNet, Xception, VGG-16, ResNet-50, and Inception V3 — show these can match or exceed custom CNNs on nodule classification, though a well-tuned custom CNN can outperform larger pre-trained networks on smaller, task-specific datasets [10], [11].

Other studies pursue optimization-driven or multi-modal strategies, including DenseNet combined with AdaBoost [12], Genetic Algorithm-optimized ANN training [13], Gravitational Search Algorithm-guided deep networks [14], [15], multi-omics fusion via PCA-SMOTE-CNN [16], multi-modality CT/MRI/ultrasound systems [17], and hybrid handcrafted-plus-autoencoder features [18]. Segmentation-focused approaches using watershed algorithms and geometric nodule descriptors isolate overlapping nodules prior to classification [19], [20], while histopathological frameworks combining frequency- and wavelet-domain features with multi-channel CNNs target related lung and colon tissue datasets [21]. Chest X-ray screening using DenseNet-121 has also been explored as a lower-cost alternative to CT, albeit with lower accuracy [22]. Across this body of work, careful preprocessing consistently improves downstream classification reliability regardless of classifier choice [23].

Taken together, the literature indicates two consistent findings that motivate this study: CNN-based models that learn features automatically tend to outperform classical machine learning models relying on manual feature engineering, and classification accuracy remains sensitive to preprocessing quality, dataset size, and class balance — leaving room for improvement through better preprocessing pipelines, the gap this paper addresses.

3. Problem Statement

The literature review highlights several recurring limitations in existing lung cancer detection systems: prediction accuracy on many existing datasets remains inconsistent; traditional feature-extraction methods struggle to extract the full complexity of images of medical; several existing models exhibit relatively high misclassification rates; adaptive, generalizable approaches for lung disease identification are limited; and performance metrics such as precision, recall, and F-measure are often not jointly optimized.

4. Proposed Methodology

The overall approach follows systematic pipeline in which the CT scan dataset is first collected and preprocessed, after which the CNN architecture is configured with appropriate layers and hyperparameters. The model is then trained on the prepared dataset to learn the distinguishing patterns between benign and malignant cases, and subsequently tested on unseen data to assess its classification capability. Finally, the trained model's performance is evaluated using standard quantitative metrics — accuracy, precision, recall, and F1-score — to validate its effectiveness for early and accurate lung cancer diagnosis. Fig. 1 depicts the methodology.

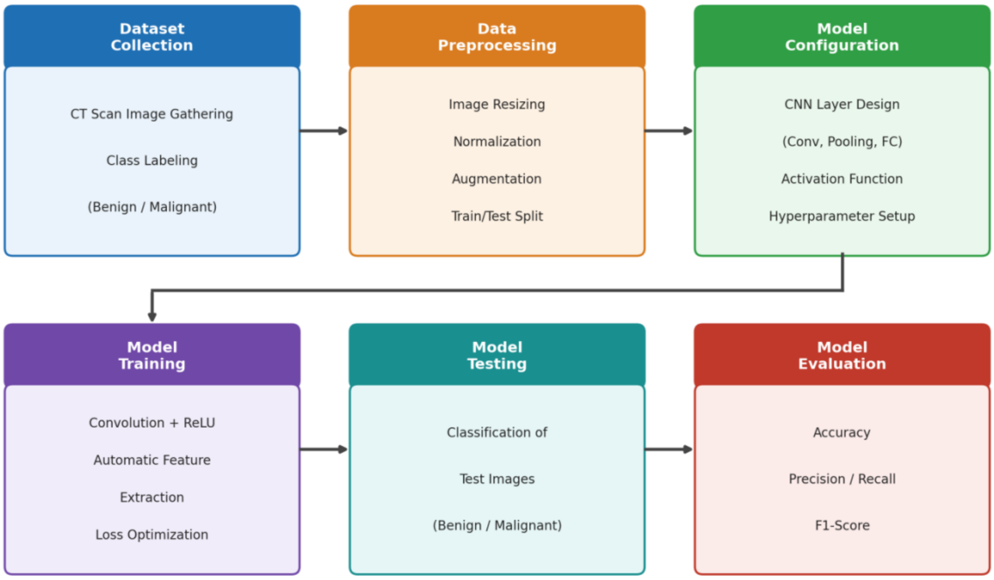


Figure 1: Proposed Research Methodology for Lung Cancer Detection

4.1 Dataset Collection

The Public dataset used in this study is lung disease CT image collection obtained from the Kaggle dataset repository. It consists of labelled CT images indicating whether a given scan corresponds to a cancerous or non-cancerous case. Standard data management practices were followed to keep the dataset clean and consistent, and the data was split into training and test subsets to allow the model to be evaluated on entirely unseen samples.

4.2 Dataset Image Preprocessing

The following preprocessing operations are applied to normalize the data and enhance features relevant to classification: images are resized to a standard resolution; pixel intensities are normalized to accelerate convergence; noise reduction filters (Gaussian or median) improve image clarity; and data augmentation — rotation, flipping, zooming, and brightness adjustment — increases robustness and reduces overfitting. CT images were converted to 2D NumPy array matrices of 512 x 512 pixels using an 8-bit unsigned integer (uint8) representation, and multiple OpenCV thresholding variants — BINARY, BINARY_INV, TRUNC, TOZERO, and TOZERO_INV — were evaluated to identify the representation that best isolates structural abnormalities for the CNN.

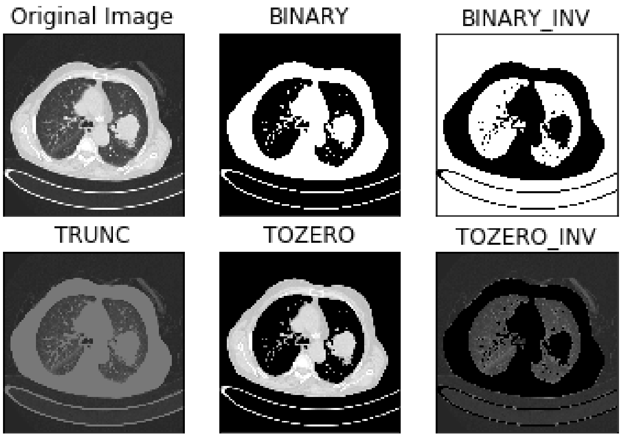
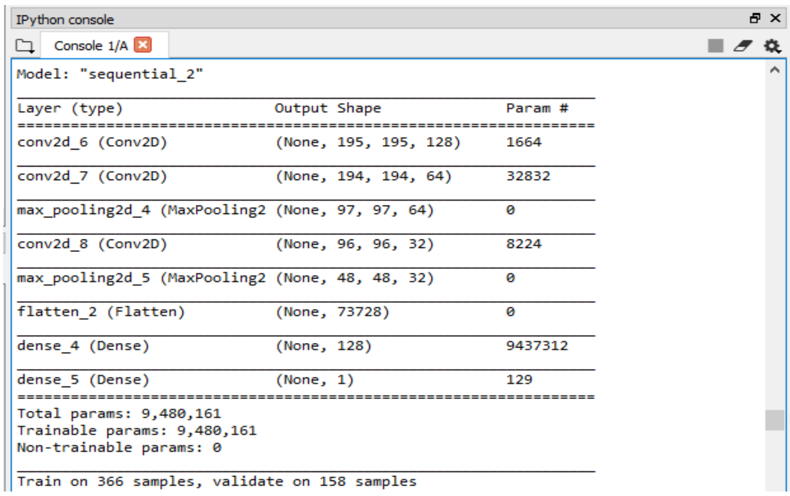


Figure 2: Image Preprocessing / Thresholding Variants

4.3 CNN Model Configuration

Rather than manually engineering features, the proposed system uses convolutional layers to automatically learn edges, textures, shapes, and spatial hierarchies from preprocessed images, with feature abstraction increasing at deeper layers. Each convo layer is followed by a max-pooling layer, reducing dimensionality, computational cost, and overfitting.

The CNN architecture follows a standard structure: an input layer; convo layers that extract spatial features; ReLU activation function for non-linearity; pooling layers to reduce feature-map dimensionality; fully connected layers that integrate learned features; and an output layer producing a probability score for cancer presence. The network is trained via backpropagation using the ADAM optimizer and binary cross-entropy loss, suited to this binary classification task.



```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_6 (Conv2D)            (None, 195, 195, 128)     1664
_________________________________________________________________
conv2d_7 (Conv2D)            (None, 194, 194, 64)      32832
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 97, 97, 64)        0
_________________________________________________________________
conv2d_8 (Conv2D)            (None, 96, 96, 32)        8224
_________________________________________________________________
max_pooling2d_5 (MaxPooling2 (None, 48, 48, 32)        0
_________________________________________________________________
flatten_2 (Flatten)          (None, 73728)             0
_________________________________________________________________
dense_4 (Dense)              (None, 128)               9437312
_________________________________________________________________
dense_5 (Dense)              (None, 1)                 129
=================================================================
Total params: 9,480,161
Trainable params: 9,480,161
Non-trainable params: 0
_________________________________________________________________
Train on 366 samples, validate on 158 samples
```

Figure 3: CNN Model Summary

4.4 Model Training

The system was implemented in Python using the Spyder IDE, with core functionality provided by scikit-learn, NumPy, pandas, Matplotlib, Seaborn, and OpenCV. The CNN was trained on the training set using binary cross-entropy loss and the ADAM optimizer; training loss decreased to 0.0015 by the final epoch, indicating rapid, stable convergence and strong generalization to unseen samples.

4.5 Model Testing

On the 158-sample test set (59 normal-class and 99 cancer-class images), the model achieved an F1-score of 0.98 for the normal class and 0.99 for the cancer class, with overall precision, recall, and F1-score of 99% across macro, micro, and weighted averages. The resulting confusion matrix is shown in Table 1.

Table 1: Results of Confusion metrics

	Predicted: Cancer	**Predicted: Normal**
Actual: Cancer	TP = 57	FN = 0
Actual: Normal	FP = 2	TN = 99

Among 158 test images, the model correctly identified all 57 true cancer cases with zero false negatives, misclassified only 2 normal cases as cancerous, and correctly identified all 99 normal cases. The complete absence of false negatives is significant in a clinical screening context, since missed cancer cases carry a higher cost than false alarms, which can be resolved through follow-up confirmation.

4.6 Overall Performance Summary

Table 2 summarizes the final performance metrics achieved by the CNN model of this study.

Table 2: Model performance results

Sr. No.	Parameter Name	Value
1	Precision	99%
2	Recall	99%
3	F_Measure	98.52%
4	Accuracy	98.73%
5	Error Rate	1.27%

5. Comparison with Baseline Methods

To contextualize these results, the proposed CNN was benchmarked against four classical machine learning baselines — Decision Tree, Logistic Regression, Random Forest, and Naive Bayes — evaluated on the same task. As given in Table 3, Random Forest was the strongest classical baseline at 94% accuracy, followed by Decision Tree and Naive Bayes at 93%, while Logistic Regression trailed at 82% accuracy and 83% F1-score. The proposed CNN outperformed every baseline, achieving 98.73% accuracy alongside 98.52% precision, recall, and F1-score, with an error rate of just 1.27%.

Table 3: Results Comparison with baseline models

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree [1]	93	94.3	92	92.6
Logistic Regression [1]	82	83	83.6	83
Random Forest [1]	94	94.3	89	90.3
Naive Bayes [1]	93	94.3	92	92.6
Proposed CNN [1]	98.73	98.52	98.52	98.52

This performance gap is consistent with the literature reviewed in Section 2: models relying on manually engineered features are limited by the completeness of those descriptors, whereas the CNN's convolutional layers learn hierarchical, task-specific representations directly from pixel data. The combination of targeted preprocessing and automatic CNN feature learning used here meaningfully reduces both false positives and false negatives relative to classical approaches, supporting CNN-based tools as a second-opinion decision-support aid in clinical lung cancer screening.

6. Conclusion and Future Scope

This paper presented a CNN-based deep learning pipeline for automated lung cancer detection from CT scan images, spanning dataset acquisition, preprocessing, automatic CNN feature extraction, classification, and rigorous performance evaluation. On a CT-image Kaggle-sourced dataset, the proposed model achieved 98.73% accuracy, 99% precision, 99% recall, and a 98.52% F-measure, with an error rate of just 1.27%, substantially outperforming Decision Tree, Logistic Regression, Random Forest, and Naive Bayes baselines. These results confirm that CNN-driven automatic feature learning offers a more robust and scalable alternative to manual feature engineering for CT-based lung cancer screening, and could serve as a fast, consistent second opinion that helps prioritize cases for radiologist review.

Future work could integrate high-resolution MRI data alongside CT scans to improve soft-tissue contrast and tumor localization; extend the system with real-time screening, 3D modelling, and more interactive diagnostic interfaces; broaden the dataset to cover other respiratory conditions such as tuberculosis, pneumonia, and COVID-19-related complications for multi-class diagnosis; and systematically compare the CNN against alternative architectures — SVM, Random Forest ensembles, and LSTM-based models — to identify configurations offering better accuracy, generalizability, or computational efficiency.

REFERENCES:

[1] P. S, V. B, and L. Krishnasamy et al., "Lung cancer prediction using classical machine learning classifiers on a Kaggle dataset," in Proc. 2025 3rd Int. Conf. on Communication, Security, and Artificial Intelligence (ICCSAI), 2025.

[2] T. R. Baitharu and S. K. Pani, "Comparative analysis of classical data mining classification techniques on a UCI lung cancer dataset," Int. J. Computer Trends and Technology, 2015.

[3] T. R. Ojha, "Comparative evaluation of six machine learning algorithms for lung cancer risk classification using patient history," J. Artificial Intelligence and Capsule Networks, 2023.

[4] M. A. Hamim, M. N. Haq, F. M. Tanmoy et al., "Machine learning-based lung cancer detection using symptom-based clinical record data," in Proc. 2023 IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications (TrustCom), 2023.

[5] N. N. Y. Dinni, A. B. Hassan, A. B. A. Mahfooz, and M. R. Ali, "Ensemble machine learning-based system for early lung cancer detection using clinical and lifestyle data," Int. J. Engineering and Science Research, 2026.

[6] A. Y. Saleh, C. K. Chin, V. Penshie, and H. R. H. Al-Absi, "A hybrid CNN-SVM approach for lung CT image classification using a computer-aided detection framework," Int. J. Advances in Intelligent Informatics, 2021.

[7] H. F. Kareem, M. S. AL-Husieny et al., "The IQ-OTH/NCCD lung cancer CT-scan dataset and SVM-based classification approach," Indonesian J. Electrical Engineering and Computer Science, 2021.

[8] H. F. Al-Yasriy and M. S. AL-Husieny, "A computer-aided system for detecting lung cancer using a CNN trained on an Iraqi hospital CT-scan dataset."

[9] H. F. Al-Yasriy et al., "A CNN-based computer-aided system for lung cancer detection using an Iraqi hospital CT-scan dataset."
[10] S. P. Nisha, S. Srividhya, and V. S. Thangarasu, "Deep learning-based lung cancer detection using CT images from the LIDC-IDRI dataset," Int. J. Innovative Research in Technology (IJIRT), 2023.
[11] M. Mamun, M. I. Mahmud et al., "Lung cancer diagnosis using CT scan images: a custom CNN model compared against ResNet-50, Inception V3, and Xception."
[12] N. Kalaivani et al., "Deep neural network approach for detecting lung cancer from CT images using DenseNet-AdaBoost," IOP Conf. Series: Materials Science and Engineering, 2020.
[13] M. K. Abdullah and S. Kurnaz, "A Genetic Algorithm-optimized Artificial Neural Network (GA-ANN-BP) for lung cancer diagnosis," Int. J. Computer Science and Mobile Computing, 2019.
[14] S. K. Lakshmanaprabu and S. N. Mohanty, "Automated diagnosis using an Optimal Deep Neural Network guided by a Modified Gravitational Search Algorithm."
[15] S. K. Lakshmanaprabu et al., "An automated diagnosis approach for lung cancer using an Optimal Deep Neural Network trained via a Modified Gravitational Search Algorithm."
[16] T. I. A. Mohamed and A. E. Ezugwu, "Deep learning-based classification and prediction of Lung Adenocarcinoma using integrated multi-omics data," IEEE Access, 2024.
[17] A. Asuntha, A. Brindha, S. Indirani, and A. Srinivasan, "Multi-modality lung cancer detection system for CT, MRI, and ultrasound images using PSO, genetic optimization, and SVM," J. Chemical and Pharmaceutical Sciences, 2016.
[18] L. Li and J. Yang, "A hybrid feature extraction approach combining GLCM texture features, Haralick descriptors, and deep autoencoder features for lung cancer detection."
[19] S. Makaju, P. W. C. Prasad, A. Alsadoon et al., "Computer-aided diagnosis system for lung cancer detection using watershed segmentation and SVM classification," Procedia Computer Science, 2018.
[20] P. Joshi and R. Sharma, "Automatic lung cancer detection from CT images using image processing-based nodule feature analysis."
[21] M. Masud, N. Sikder et al., "A classification framework for lung and colon tissue types using 2D-DFT, 2D-DWT, and a multi-channel CNN on the LC25000 dataset."
[22] W. Ausawalaithong and A. Thirach, "DenseNet-121 with transfer learning for lung cancer classification from chest X-ray images."
[23] H. Ahmed, A. Hussain, and M. Ali, "Deep learning-based lung cancer classification using convolutional and fully connected layers on the Kaggle Lung Cancer dataset."