

European Journal of Technology (EJT)









Big Data-Driven Approach for Lung Cancer Identification via Advanced Deep Transfer Learning Models

Rajiv Chalasani, Venkataswamy Naidu Gangineni, Sriram Pabbineedi, Mitra Penmetsa, Jayakeshav Reddy Bhumireddy, Mukund Sai Vikram Tyagadurgam



Big Data-Driven Approach for Lung Cancer Identification via Advanced Deep Transfer Learning Models

 Rajiv Chalasani^{1*},  Venkataswamy Naidu Gangineni²,  Sriram Pabbineedi³, 
Mitra Penmetsa⁴,  Jayakeshav Reddy Bhumireddy⁵,  Mukund Sai Vikram
Tyagadurgam⁶

¹Sacred Heart University, ²University of Madras, Chennai, ³University of Central Missouri,

⁴University of Illinois at Springfield, ⁵University of Houston, ⁶University of Illinois at
Springfield



Article history

Submitted 03.07.2025 Revised Version Received 01.07.2025 Accepted 04.07.2025

Abstract:

Purpose: To develop and evaluate a highly accurate computer-aided diagnosis (CAD) model based on ResNet-50 for the early identification of lung cancer-related pulmonary nodules using the publicly accessible LIDC-IDRI CT image dataset.

Materials and Methods: This study utilizes the LIDC-IDRI dataset, which comprises CT scans of pulmonary nodules for the detection of lung cancer. The preprocessing pipeline involves converting all CT images to grayscale, resizing them to a consistent dimension, and applying data augmentation techniques such as rotations and flips to enhance the model's robustness. A refined ResNet-50 convolutional neural network is employed for classification to extract deep characteristics and differentiate between benign and malignant nodules. Two baseline models, the Feed Forward Back Propagation Neural Network and the Support Vector Machine (SVM), are also used for comparison to assess the efficacy of this strategy.

Findings: The ResNet-50 model demonstrated superior performance across all evaluation metrics, achieving an accuracy of 99.38%, an F1-score of 99.37%, a precision

of 99.91%, and a recall of 98.76%. ResNet-50 showed a high capacity to reliably detect pulmonary nodules from CT images by consistently outperforming both the SVM and the Feed Forward Back Propagation Neural Network when compared to the baseline models.

Unique Contribution to Theory, Practice and Policy: Based on the findings, it is recommended that the ResNet-50-based CAD model be integrated into clinical radiology workflows to facilitate the early diagnosis of lung cancer. For broader applicability, further validation should be conducted using multi-center and prospective datasets to ensure the model's generalizability. Additionally, incorporating real-time preprocessing and inference mechanisms within existing PACS (Picture Archiving and Communication System) platforms could streamline diagnostic processes and improve radiologist efficiency.

Keywords: Lung Cancer, Deep Learning (DL) Techniques, ResNet-50, LIDC-IDRI Dataset

JEL Classification: C63, I10, I18, O33

INTRODUCTION

Prompt treatment of lung, breast, and other cancers contributes significantly to higher survival rates. Although cancer has consistently been a serious public health issue, its incidence, particularly in lung cancer, continues to grow [1][2]. The diagnosis of cancer began decades ago with physicians relying on a single test, such as mammograms, ultrasounds, MRIs, or PET scans. In the context of lung cancer, ultrasound helps assess surrounding structures. At the same time, PET scans are particularly valuable, using radioactive tracers like F-fluorodeoxyglucose to detect early cellular changes and locate tumours in the lungs [3][4]. Dynamic MRI aids in evaluating blood vessels and detecting potential metastases, especially in advanced cases of lung cancer. While mammography is primarily used for breast cancer, similar imaging techniques using X-rays are applied to visualize lung tissue and identify abnormalities such as nodules or masses, enabling earlier detection and handling.

Among lung cancers, cancer is the leading cause of mortality globally[5]. Lung cancer is the primary cause of cancer-related mortality, accounting for 18% of all cancer-related deaths. Lung cancer is mostly caused by smoking, and the problem has either become very high or is still increasing in various parts of the world. Thus, they should expect the number of lung cancer cases to continue rising for the foreseeable future. If lung cancer is found early and treated correctly, patients may recover much better [6][7]. Ten to twenty per cent of people with lung cancer will eventually make it to five years after diagnosis. Medical professionals often use MRI and computed tomography (CT) to detect illnesses early, which helps improve patients' chances of survival.

In past lung cancer detection methods, the top-detection features were manually selected through methods such as the Sequential Forward Floating Selection Algorithm (SFFSA) and Genetic Algorithms (GA)[8]. SFFSA is a greedy search-based algorithm that evaluates data combinations and generates optimal feature sets that achieve improvements in model performance (SFFSA) [9]. GA is a bio-inspired optimization algorithm that leverages natural selection to identify relevant features from the groups of potential empirical data features. These and analogous feature selection methods assembled and formed initial machine learning pipelines for use; nevertheless, they were highly dependent on hand-crafted feature engineering and manual fine-tuning, and were slow to make decisions while being susceptible to human bias.

Conversely, state-of-the-art deep learning (DL) approaches – specifically, convolutional neural networks (CNNs) and models like ResNet-50 – directly learn multi-level, high-dimensional features from raw image data in an automatic way. The capability to Learning characteristics in a hierarchical way improves scalable and objective diagnostic support or solutions over earlier methods.

The recognition of cancer becomes more difficult over time. To determine whether cancer is benign or malignant, doctors must now sort through a large number of tests and factors while attempting to avoid prejudice. It became evident that automation was necessary since the human mind is not always objective [10]. For example, in mammography, physicians often have to evaluate a large number of images, which can result in mistakes. The diagnosis of lung cancer presents a similar difficulty, since deciphering intricate imaging data may be laborious and prone to mistakes.

Recent studies stress the importance of ML in lung cancer diagnosis/prognosis as it offers accurate data-driven results. ML is required because of the convoluted complexity in medical imaging and clinical data [11][12][13]. Therefore, ML algorithms can quickly realize outputs and statistically recognize patterns in CT scans, pathology slides, and genomic profiles in a

manner that is often much faster than humans. ML models can also support clinical decisions, such as diagnosis (early detection), classification, or treatment planning. Thus, ML systems can help decrease the probability of human error as well as the time physicians take to decide on treatment modalities.

Problem Statement

The complexity, amount, and unpredictability of imaging data make early and accurate lung cancer diagnosis a major clinical issue even with advances in medical imaging and diagnostic techniques. Traditional methods, which rely heavily on hand-crafted features and manual interpretation, are often time-consuming, prone to human error, and lack scalability. While algorithms like SFFSA and Genetic Algorithms have improved early detection to some extent, their dependency on manual feature selection and limited adaptability reduces their effectiveness in diverse real-world scenarios. Moreover, the rising incidence and mortality of lung cancer globally underscore the urgent need for automated, objective, and high-performing diagnostic systems. Therefore, there is a critical need to develop intelligent, deep learning-driven approaches that can autonomously learn discriminative features from raw medical data, enhance diagnostic accuracy, reduce radiologist workload, and facilitate timely clinical decision-making.

Significance and Contribution

This study is important because it applies advanced deep learning (DL) methods to enhance lung cancer detection using computed tomography imaging, which is essential for improved treatment outcomes and earlier diagnosis. The primary cause of cancer-related fatalities is still lung cancer, and increasing survival rates require early identification. Automated detection systems based on DL have been shown to reduce radiologists' workload and minimize diagnostic errors, particularly in high-volume clinical settings [14][15]. By utilizing the comprehensive LIDC-IDRI dataset and a robust pre-processing pipeline, the study addresses challenges like image variability and limited annotated data. Using data augmentation and deep residual learning together helps the model become more reliable for clinical purposes, since it can apply its learning in various circumstances. In conclusion, the study promotes earlier and more precise identification of lung cancer, advancing medical imaging procedures. An overview of the study's contributions may be found below:

- Modified a processing cycle that included turning images into grayscale, altering their sizes, and augmenting the data to suit lung CT images.
- A ResNet-50 design was used to efficiently detect different features that are important for lung nodule classification.
- Examined the model by evaluating several performance metrics to test its robustness and generalizability.
- Compared the suggested DL approach to baseline and conventional models for lung cancer detection, demonstrating its superiority and clinical relevance.

Justification and Novelty of the Study

The primary benefit of the study is the utilization of the ResNet-50 network, which helps avoid the vanishing gradient effect and successfully obtains more detailed and complex features from medical images. Having this skill helps identify the subtle signs of lung cancer in CT scans. Completing pre-processing steps, such as converting to grayscale, adjusting image size, and applying more data, improves the training data and helps the model handle new situations better. The chosen strategy can be applied to various use cases and aligns well with clinical routines,

making radiologists' work more efficient. This combination of technical rigour and clinical focus justifies the approach as a valuable step toward reliable, automated detection of lung cancer.

Structure of the Paper

The erection of this research is to conduct surveys: Section II examines relevant research on the use of DL and medical imaging methods for the goal of finding lung cancer. Section III explains the process, including the DL model architecture, preparation procedures, and dataset specifics. Section IV displays the performance analysis and experimental findings. Finally, Section V presents an overview of the study's main conclusions and offers suggestions for future research.

LITERATURE REVIEW

In this segment, provide a literature review of recent studies using advanced DL methods for Lung cancer detection. The papers reviewed are summarized in Table I, which includes the methods used, datasets employed, main findings, and limitations or proposed future work.

Saric et al. (2019) propose a completely automated technique that uses convolutional neural networks (CNNs) for classification in order to detect lung cancer in entire slide photographs of lung tissue samples. After comparing two CNN designs, the study concludes that CNN-based methods can significantly expedite the procedure and assist pathologists [16].

Hussein et al. (2019) suggest both unsupervised and supervised ML techniques to enhance the characterisation of tumours using radiological images. The supervised technique integrates task-dependent feature representations into CAD systems using DL algorithms like Transfer Learning and 3D CNN. For unsupervised tumour classification, the unsupervised method investigates deep features and employs proportion-SVM to handle sparse labelled training data. The suggested algorithms demonstrate cutting-edge sensitivity and specificity when tested on lung and pancreatic tumours[17].

Gao et al. (2019) investigate the use of longitudinal data for lung nodule identification and cancer prediction. Heterogeneous, irregular acquisitions in clinical imaging are accommodated by generalising the Long Short-Term Memory (LSTM) model. Both regular and irregular data are subjected to the Distanced LSTM (DLSTM) model, which incorporates a Temporal Emphasis Model (TEM). Three datasets are used to assess the DLSTM framework, and it performs competitively on both simulated and routinely sampled datasets The proposed DLSTM achieves a 0.8905 AUC score[18].

Perumal and Velmurugan (2018) used the enhanced artificial bee colony optimisation (EABC) method to identify probable cancerous regions in CT (Computerised Tomography) scan pictures. Utilizing MATLAB software, the suggested EABC implementation section utilizes CT (Computerized Tomography) scanned lung images. Radiologists and medical professionals can use this technique to identify syndromes in their early stages and prevent cancer from progressing to more serious phases[19].

Zhang et al. (2018) analyzed 231 urine samples from various cancers, including lung cancer, to develop non-invasive biomarkers for lung cancer diagnosis. Random forest (RF) modeling was used by the researchers to identify urine proteins that might distinguish lung cancer from other types of cancer. After choosing five urine indicators, they developed a combinatorial model that accurately categorises instances of lung cancer. This set of indicators not only distinguishes lung cancer from other common tumors, but it also distinguishes it from control groups [20].

Table I presents a literature review related to Lung cancer detection, describing the methodology used, datasets employed, main findings, identified limitations, and suggested future directions.

Table 1: Summary of Literature Review based on Lung Cancer Detection using ML AND DL

Study	Methodology	Dataset	Key Findings	Limitations	Future Work
Saric et al. (2019)	Fully automatic lung cancer detection using CNNs on pathology whole slide images	Whole slide lung tissue images	CNN significantly improves classification speed and aids pathologists	Limited to pathological slides; lacks multimodal validation	Apply to multimodal datasets; integrate into real-time clinical workflows
Hussein et al. (2019)	Supervised: 3D CNN + Transfer Learning; Unsupervised: Proportion-SVM with deep features	1018 lung CT & 171 pancreas MRI scans	achieved cutting-edge tumour categorisation sensitivity and specificity.	Data limited to specific imaging centers; high training complexity	Expand datasets; integrate graph-based feature learning
Gao et al. (2019)	Distanced LSTM (DLSTM) with Temporal Emphasis Model (TEM) for irregular longitudinal imaging	Three longitudinal datasets (clinical imaging)	AUC of 0.8905 for lung cancer prediction; handles irregular acquisition well	May underperform with very sparse temporal data	Adapt for multi-center real-world datasets; temporal attention refinements
Perumal & Velmurugan (2018)	Enhanced Artificial Bee Colony (EABC) optimization + CT preprocessing in MATLAB	CT scan images of lungs (MATLAB environment)	Identifies suspicious lung cancer regions effectively for early-stage detection	Classical method; lacks validation with larger DL models	Combine EABC with CNN or hybrid DL for improved classification
Zhang et al. (2018)	Random Forest classification using urinary proteomics biomarkers	231 urine samples (lung & other cancers)	Five urinary biomarkers used for accurate noninvasive lung cancer detection (AUC 0.87–0.99)	Small sample size for each cancer subtype	Validate on larger clinical cohorts; integrate with DL-based biosignal analysis

Research Gap

Although DL and conventional ML methods for lung cancer diagnosis have made significant strides, several research gaps remain. Most existing studies focus either on imaging-based approaches (e.g., CNN, 3D-CNN, LSTM) or non-imaging bio signal methods (e.g., urine proteomics), but seldom integrate both to enhance diagnostic accuracy and robustness. Additionally, methods like Enhanced Artificial Bee Colony Optimization and Proportion-SVM have shown promise but lack validation across large-scale, diverse, and multi-institutional datasets. There's also limited exploration of how unsupervised deep learning models can be effectively combined with clinical temporal data for continuous monitoring and early prediction. Furthermore, current CAD systems often require extensive labeled data, which is not always feasible in clinical settings, highlighting a gap in developing generalized, semi-supervised, or label-efficient models for real-world deployment. Therefore, there is a strong need for hybrid, interpretable, and scalable lung cancer diagnostic models that can handle heterogeneous data sources while addressing data scarcity and clinical applicability.

MATERIALS AND METHODS

This methodology presents a systematic method for detecting lung cancer by evaluating a several-phase DL framework using the LIDC-IDRI dataset. Data methods for enhancing datasets are used to boost their variety and resilience after data preparation, which involves converting photos to greyscale and resizing them for standardization. After cleaning the data, divide it into two sets: one for training purposes and another for validation. This will ensure that the model is evaluated correctly. The core classification is performed using a ResNet-50 architecture, leveraging its deep residual learning capabilities for accurate feature extraction and pattern recognition in medical images. Finally, we thoroughly evaluate the model's ability to distinguish between lung cancer and non-cancerous tissue using widely used metrics like F1-score, recall, accuracy, and precision. All outcomes should be methodically documented and examined to confirm the clinical applicability of the suggested method. From using datasets to evaluating the outcomes, Figure 1 illustrates the workflow of the proposed technique.

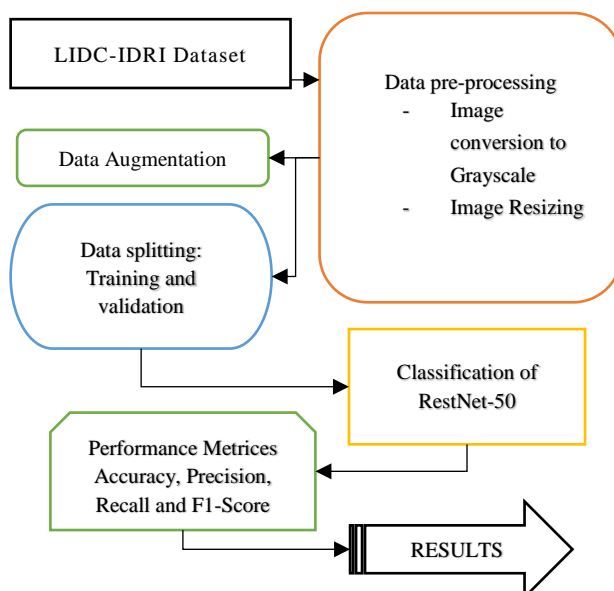


Figure 1: Flowchart Depicting Lung Cancer Detection Methodology

A brief discussion of the phases of the suggested methodology is provided below:

Data Collection

A diagnosis and the LIDC-IDRI include chest CT scans that have been evaluated for lung cancer. Possible applications include creating, training, and testing CAD algorithms for lung cancer detection and classification. In the context of DICOM, LIDC-IDRI, and low-dose CT scans, it contains all pertinent information about nodules, such as their location, size, diagnostic results, and other pertinent information. It has 244,527 photos and 1018 instances, with a total size of around 131 GB. Figure 2: dataset samples.

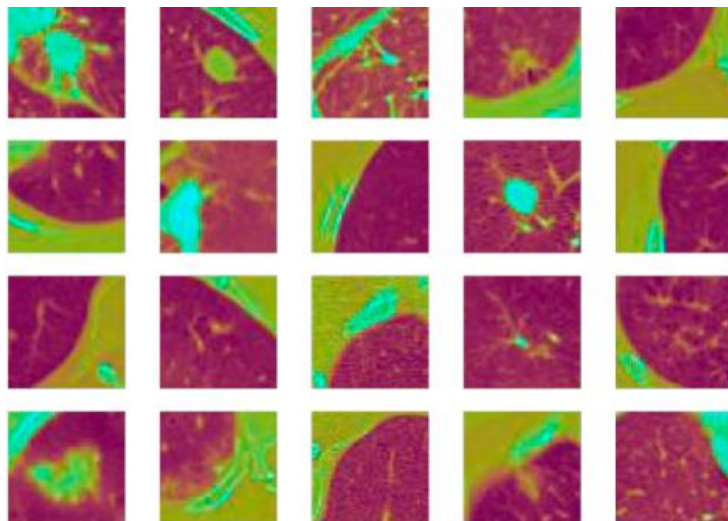


Figure 2: Sample Image of the Dataset

Figure 2 displays a 4×5 grid of 20 colorized microscopic images, each showing cellular or tissue structures with a false-colour scheme. The images feature predominantly magenta/purple backgrounds contrasted by bright cyan, green, and yellow regions that likely represent different biological components such as cell nuclei, cytoplasm, or specific markers. The varying patterns across the tiles suggest different samples or regions, with some showing scattered bright spots while others display larger coloured areas or linear structures, typical of fluorescence microscopy or histological imaging used in biomedical research.

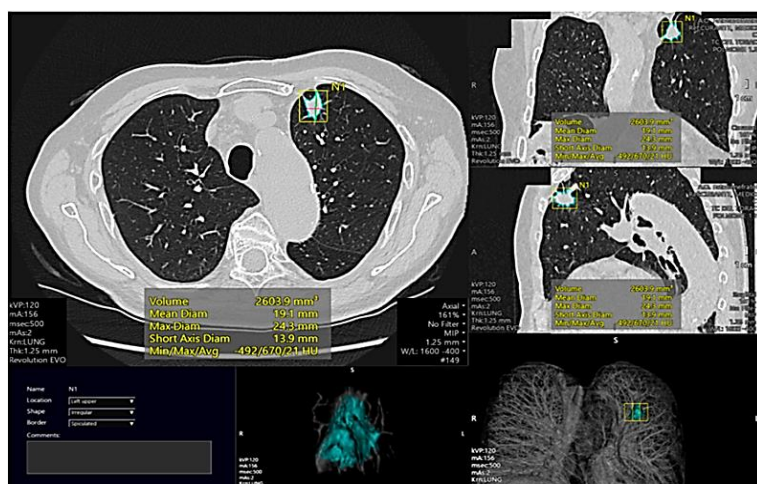


Figure 3: Multi-Panel Chest CT Scan Showing Axial and Coronal Lung Views

Figure 3 shows a multi-panel chest CT scan display with axial and coronal lung views, featuring measurement annotations for volume, diameter, and other parameters, along with 3D reconstructions used for pulmonary assessment.

Data Preprocessing

Pre-processing entails transforming unprocessed data into an appropriate format for examination. Concerning lung cancer data in the LIDC-IDRI dataset, pre-processing includes converting colour images to grayscale, resizing, and applying data expansion techniques to help enhance model accuracy and robustness. The following steps of pre-processing are given below:

- **Image conversion to Grayscale:** A digital image in grayscale mode has each pixel reflect the brightness or darkness captured by the camera. This kind of picture only includes very dark black and very bright white tones. Moreover, the image displays only three colours: black, white and Gray, in which Gray can have several levels.
- **Image Resizing:** Image resizing is the act of changing the image dimensions to a fixed size, usually to meet the input requirements for DL models. In this study, we resized all of the photos to 224×224 pixels, giving a consistent input structure and reducing computational complexity and yet allowing preservation of essential diagnosis features needed for accurate lung cancer classification.

Data Augmentation

Data augmentation involves altering pictures by methods such as flipping, cropping, rotating, scaling and so on. Additional samples produced by data augmentation do not alter the class under the main category.

Data Splitting

The data splitting method entails dividing the dataset into its component elements: a training set (80%) and a validation set (20%). Overfitting may be prevented by comparing the model's performance as it passes through its testing on the training dataset, followed by the validation dataset.

Classification of ResNet50 Model:

ResNet-50, an innovative CNN variant, combines the rest of the module the potential 50-layer deep structure ResNet-50 is composed of 48 convolutional layers, 1 FC layer, and 1 max pooling layer. A key benefit of ResNet-50 is that it utilizes leftover units [21]. The problem of vanishing angles that was present in earlier deep systems is well handled by these units. Skipping associations, the remaining units are shown in every section inside the ResNet-50 architecture.

Within ResNet-50's framework are fifty distinct instances of convolutional layer, The 164 channels have dimensions of 7 by 7 and a walk of 2. Next, with a walk of 2, the max pooling module lowers the convolution estimate. Three convolution layers with 64 estimated channels each are typically composed of 256 1×1 channels, 64 3×3 channels, and a 1×1 matrix. In this instance, convolutional layers are employed. The next four convolutional layers are configured using 512 estimated 1×1 channels, 128 estimated 3×3 channels, and 128 measured 1×1 channels. In the second layer, there are a total of 1024 estimation channels (1×1), 256 estimation channels (1×1) that are iteratively used six times, 256 measurement channels (3×3), and 256 estimation channels (1×1).

Among the last convolution layers of the ResNet-50 network are 2048 sensors, 512 estimated 3×3 channels, 512 estimated 1×1 channels, and 1×1 channels. At its highest level, this structure has a normal pooling layer or FC layer that communicates with the final highlight vector over a thousand tests. Classifying photos into distinct categories is achieved via the use of a "SoftMax" implementation. With a total of 44.5 million training parameters, RenNet101 uses the ImageNet dataset to train its 101 layers for distinction.

Evaluation Metrics

512 approximate 3*3 channels, 512 approximate 1*1 channels, and 1*1 channels.. The confusion matrix displays the model's capacity to classify from TP, TN, FP, and FN. If we want to know how effectively the models can differentiate between malignant and non-cancerous instances, we require these measures. A strong empirical validation of the model relies on the correct interpretation of these values. This part of the work is crucial for determining which model best fits a given situation that can be applied in healthcare.

Accuracy: Accuracy measures how well the model anticipates or predicts things in general by displaying the percentage of correct predictions. It presents how accurately the model can predict. The way to find it is given by formula Equation (1):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots (1)$$

Precision: This metric measures whether the model makes correct positive predictions without also labeling something as positive that is negative. It helps us judge how much we can depend on the model's predictions. Equation. (2) shows the way the formula is calculated:

$$Precision = \frac{TP}{TP+FP} \dots\dots\dots (2)$$

Recall: This technique, known as the TPR, demonstrates how well the model detects positive cases of lung cancer. It clarifies the model's capacity to diagnose instances of lung cancer. The equation for the formula is shown as Equation. (3):

$$Recall = \frac{TP}{TP+FN} \dots\dots\dots (3)$$

F1 Score: A reasonable evaluation of the F1 Score considers both recall and accuracy, which are indicators of the model's effectiveness. It is very useful when trying to reach an agreement between these two. This calculation is done in Equation. (4):

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \dots\dots\dots (4)$$

Additionally, several models for lung cancer diagnosis may have their net recovery compared using these criteria.

Result Analysis And Discussion

This section presents the findings of a machine learning (ML) system developed to utilize computed tomography (CT) images for the detection of lung cancer. The machine used for these experiments featured a 3.3 GHz Intel dual-core i6 processor, 1 TB of RAM, and Windows 11 Pro. Results were gauged using the metrics: Recall, accuracy, precision, and F1-score. Table II displays the results of the identification of lung cancer using the RestNet50 model.

Table 2: Model Performance of Proposed RestNet50.

Performance Metrics	RestNet50 (%)
Accuracy	99.38
Precision	99.91
Recall	98.76
F1-Score	99.37

The ResNet50 model's efficacy in lung cancer classification is outstanding, as evidenced by its high evaluation metrics. The accuracy rate of 99.38% indicates that the tool performs well in predicting both positive and negative cases. The fact that the model is almost perfectly correct at 99.91% means it rarely misinterprets regular test results as cancerous cells. Such a recall of 98.76% shows that the model rarely fails to detect actual cancer cases. In addition, the F1-score

of 99.37% indicates that the model performs accurately in both detecting positive records and avoiding false positives. This supports that ResNet50 is dependable and effective for spotting lung cancer on CT scans.

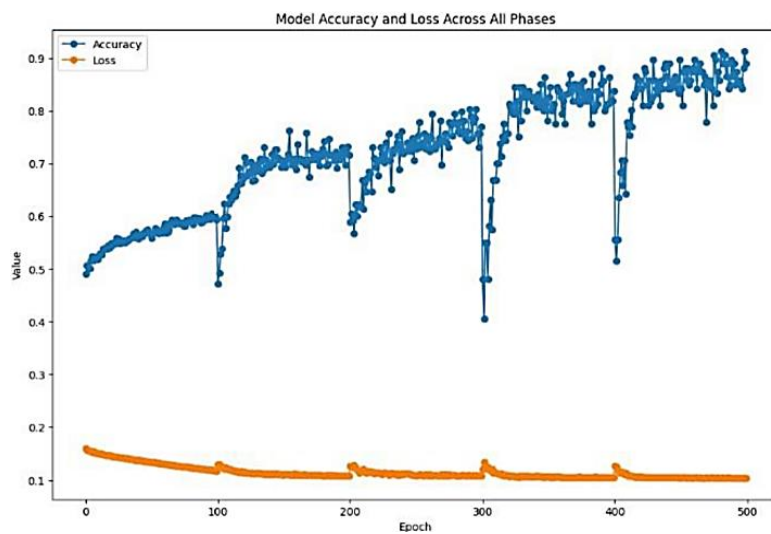


Figure 4: Training Loss and Accuracy

Figure 4 presents a chart of both accuracy and loss during every phase of training the model. Training accuracy is represented by the blue curve, which increases in value as the training process progresses. Spikes in the graph might occur because the learning rate was changed for different phases of training. With every epoch, the orange curve systematically goes down, proving the model can lower the loss. All of these curves indicate that the model is both stable and improving, achieving an accuracy rate of nearly 90% and a loss rate lower than 0.1 in the final stages of training.

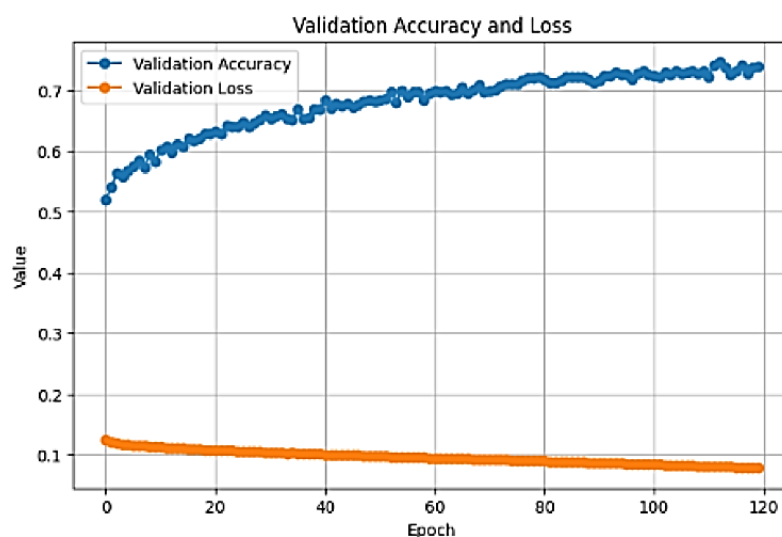


Figure 5: Validation Loss and Accuracy

The model's performance at validation was observed over 120 epochs and is presented in Figure 5, which shows trends in accuracy and loss. As the blue curve shows the validation accuracy, we can see that the model improves in handling unseen data, increasing from a score of around 0.5 to over 0.7. The orange line indicates validation loss, and you can notice it dropping from about 0.13 to a value below 0.1. Since accuracy decreases with loss, the model is fine-tuned well and is less likely to overfit, meaning it can generalize effectively to the validation set.

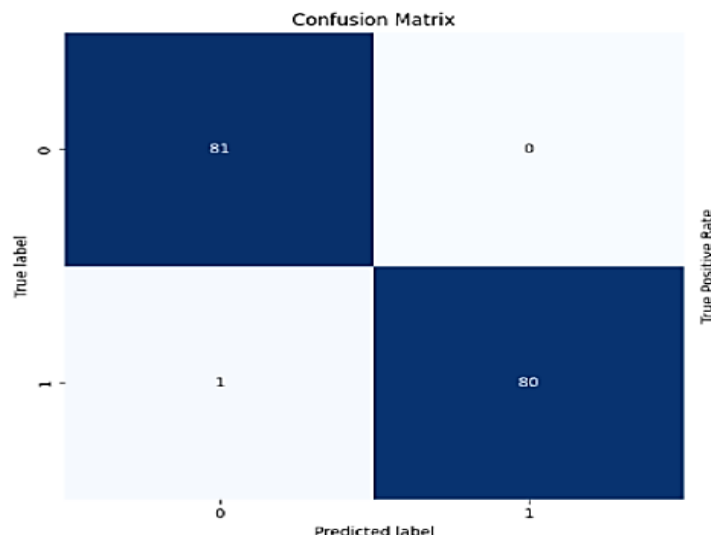


Figure 6: Confusion Matrix for RestNet50 Model

The confusion matrix used to gauge the model's classification performance is shown in Figure 6. With a total of 161 predictive classifications, the model achieved 81 correct classifications for class 0 (true negatives) and 80 correct classifications for class 1 (true positives). There were only two non-matching outcomes: one mislabeled instance (class 1 turned into class 0) and no false positive cases. As a result, the model exhibits good accuracy and a clear capability to differentiate between the two classes, with only minor mistakes.

Comparative Analysis and Discussion

The part consists of an analysis comparing several methods for detecting lung cancer with different DL methods. For example, models were created or implemented such as Feed Forward Back Propagation Neural Network (FF-BPNN), CNN, SVM and ResNet50. Assessments of these models were conducted using various evaluation methods, with a focus on their accuracy, as shown in Table III.

Table 3: Comparison Analysis of the Proposed and Baseline Models Based on Lung Cancer Detection

Model	Accuracy (%)
FF-BPNN[22]	97.7
CNN[23]	86.4
SVM[24]	97.6
RestNet50	99.37

The proposed ResNet50 model outperformed the competition in lung cancer classification, with a remarkable 99.37% accuracy rate, which surpasses all baseline models discussed in Table III. On the other hand, FF-BPNN and SVM were significantly more accurate, with results of 97.7% and 97.6%, respectively, while the common CNN achieved 86.4%. ResNet50 improves significantly thanks to its residual learning ability, which addresses the vanishing gradients problem and enables the model to discover more complex features. According to the findings, ResNet50 is considered both reliable and accurate for identifying lung cancer on CT scans.

The adopted ResNet50 design brings several benefits to achieving better results in lung cancer classification. The framework facilitates the creation of very deep networks, which enhances the extraction of important features from CT images. Therefore, BERT reports fairer accuracy and better scores than standard models. Moreover, the hierarchical learning function of ResNet50 enables it to detect subtle signs in medical images, which is crucial for identifying

lung cancer early. Due to its stability and ability to manage information, Hadoop is widely used in healthcare.

CONCLUSION AND RECOMMENDATIONS

Conclusion

Approximately 72% of cancer-related deaths are caused by lung cancer, which is defined by abnormal lung cell growth. This study employed a DL model to screen the LIDC-IDRI dataset for lung cancer. Thanks to its pre-processing system and the use of ResNet-50, the proposed method successfully finds and classifies pulmonary nodules from CT scans. The model was correct 99.38% of the time and also exhibited excellent precision, recall, and F1-score, demonstrating that it works well and is reliable. Following a comparison with baseline models, it was confirmed that ResNet-50 is more effective. These findings suggest that the proposed method of analysis can enhance the accuracy and timeliness of lung cancer screenings.

Recommendations

Future studies could attempt to expand the data to encompass more imaging options and various nodule appearances, thereby enhancing the model's reliability. If explainable AI techniques are applied, the predictions made by models can be more understandable, which could support clinical decisions. Including both patient information and genomic tests may enhance the detection of diseases. Running the model and applying it in clinical settings would enhance its practical use beyond the classroom.

REFERENCES

- [1] J. L. Causey *et al.*, “Highly accurate model for prediction of lung nodule malignancy with CT scans,” *Sci. Rep.*, 2018, doi: 10.1038/s41598-018-27569-w.
- [2] V. KOLLURI, “Machine Learning in Managing Healthcare Supply Chains: How Machine Learning Optimizes Supply Chains, Ensuring the Timely Availability of Medical Supplies,” *Int. J. Emerg. Technol. Innov. Res.*, pp. 2349–5162, 2016.
- [3] S. G. Armato *et al.*, “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans,” *Med. Phys.*, 2011, doi: 10.1118/1.3528204.
- [4] V. Kolluri, “An Innovative Study Exploring Revolutionizing Healthcare with AI: Personalized Medicine: Predictive Diagnostic Techniques and Individualized Treatment,” *Int. J. Emerg. Technol. Innov. Res.*, vol. 3, no. 11, pp. 2349–5162, 2016.
- [5] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA. Cancer J. Clin.*, 2018, doi: 10.3322/caac.21492.
- [6] J. Abraham, “Reduced lung cancer mortality with low-dose computed tomographic screening,” *Community Oncol.*, 2011, doi: 10.1016/S1548-5315(12)70136-5.
- [7] R. Wender *et al.*, “American Cancer Society lung cancer screening guidelines,” *CA. Cancer J. Clin.*, 2013, doi: 10.3322/caac.21172.
- [8] H. Hijazi and C. Chan, “A Classification Framework Applied to Cancer Gene Expression Profiles,” *J. Healthc. Eng.*, vol. 4, no. 2, pp. 255–283, Jan. 2013, doi: 10.1260/2040-2295.4.2.255.
- [9] K. Homsapaya and O. Sornil, “Modified Floating Search Feature Selection Based on Genetic Algorithm,” *MATEC Web Conf.*, vol. 164, p. 01023, Apr. 2018, doi: 10.1051/matecconf/201816401023.
- [10] T. Kadir and F. Gleeson, “Lung cancer prediction using machine learning and advanced imaging techniques,” 2018. doi: 10.21037/tlcr.2018.05.15.
- [11] S. Garg, “Predictive Analytics and Auto Remediation using Artificial Intelligence and Machine learning in Cloud Computing Operations,” *Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci.*, vol. 7, no. 2, 2019, doi: <http://dx.doi.org/10.5281/zenodo.15362327>.
- [12] R. Gruetzemacher, A. Gupta, and D. Paradise, “3D deep learning for detecting pulmonary nodules in CT scans,” *J. Am. Med. Informatics Assoc.*, 2018, doi: 10.1093/jamia/ocy098.
- [13] S. Garg, “AI/ML DRIVEN PROACTIVE PERFORMANCE MONITORING, RESOURCE ALLOCATION AND EFFECTIVE COST MANAGEMENT IN SAAS OPERATIONS,” *Int. J. Core Eng. Manag.*, vol. 6, no. 6, pp. 32–45, 2019.
- [14] D. Ardila *et al.*, “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography,” *Nat. Med.*, 2019, doi: 10.1038/s41591-019-0447-x.
- [15] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, “Artificial intelligence in radiology,” 2018. doi: 10.1038/s41568-018-0016-5.

- [16] M. Saric, M. Russo, M. Stella, and M. Sikora, "CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images," in *2019 4th International Conference on Smart and Sustainable Technologies, SpliTech 2019*, 2019. doi: 10.23919/SpliTech 2019.8783041.
- [17] S. Hussein, P. Kandel, C. W. Bolan, M. B. Wallace, and U. Bagci, "Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches," *IEEE Trans. Med. Imaging*, vol. 38, no. 8, pp. 1777–1787, 2019, doi: 10.1109/TMI.2019.2894349.
- [18] R. Gao *et al.*, "Distanced LSTM: Time-Distanced Gates in Long Short-Term Memory Models for Lung Cancer Detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019. doi: 10.1007/978-3-030-32692-0_36.
- [19] S. Perumal and T. Velmurugan, "Lung cancer detection and classification on CT scan images using Enhanced Artificial Bee Colony Optimization," *Int. J. Eng. Technol.*, 2018, doi: 10.14419/ijet.v7i2.26.12538.
- [20] C. Zhang *et al.*, "Urine Proteome Profiling Predicts Lung Cancer from Control Cases and Other Tumors," *EBioMedicine*, 2018, doi: 10.1016/j.ebiom.2018.03.009.
- [21] M. Talo, "Automated classification of histopathology images using transfer learning," *Artif. Intell. Med.*, 2019, doi: 10.1016/j.artmed.2019.101743.
- [22] J. Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for CT images," *Comput. Methods Programs Biomed.*, 2014, doi: 10.1016/j.cmpb.2013.10.011.
- [23] W. Li, P. Cao, D. Zhao, and J. Wang, "Pulmonary Nodule Classification with Deep Convolutional Neural Networks on Computed Tomography Images," *Comput. Math. Methods Med.*, vol. 2016, pp. 1–7, 2016, doi: 10.1155/2016/6215085.
- [24] W. J. Choi and T. S. Choi, "Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach," *Entropy*, 2013, doi: 10.3390/e15020507.
- [25] Kalla, D. (2022). AI-Powered Driver Behavior Analysis and Accident Prevention Systems for Advanced Driver Assistance. *International Journal of Scientific Research and Modern Technology (IJSRMT) Volume, 1*.
- [26] Kuraku, D. S., Kalla, D., & Samaah, F. (2022). Navigating the link between internet user attitudes and cybersecurity awareness in the era of phishing challenges. *International Advanced Research Journal in Science, Engineering and Technology*, 9(12).
- [27] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2022). Enhancing Early Diagnosis: Machine Learning Applications in Diabetes Prediction. *Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-205. DOI: doi. org/10.47363/JAICC/2022 (1), 191, 2-7*.
- [28] Kalla, D., Kuraku, D. S., & Samaah, F. (2021). Enhancing cyber security by predicting malwares using supervised machine learning models. *International Journal of Computing and Artificial Intelligence*, 2(2), 55-62.
- [29] Katari, A., & Kalla, D. (2021). Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies. *ESP Journal of Engineering & Technology Advancements (ESP-JETA)*, 1(1), 150-157.

- [30] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2021). Facial Emotion and Sentiment Detection Using Convolutional Neural Network. *Indian Journal of Artificial Intelligence Research (INDJAIR)*, 1(1), 1-13.
- [31] Chinta, P. C. R., Katnapally, N., Ja, K., Bodepudi, V., Babu, S., & Boppana, M. S. (2022). Exploring the role of neural networks in big data-driven ERP systems for proactive cybersecurity management. *Kurdish Studies*.
- [32] Routhu, K., Bodepudi, V., Jha, K. M., & Chinta, P. C. R. (2020). A Deep Learning Architectures for Enhancing Cyber Security Protocols in Big Data Integrated ERP Systems. *Available at SSRN 5102662*.
- [33] Chinta, P. C. R., & Katnapally, N. (2021). Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures. *Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures*.
- [34] Katnapally, N., Chinta, P. C. R., Routhu, K. K., Velaga, V., Bodepudi, V., & Karaka, L. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. *American Journal of Computing and Engineering*, 4(2), 35-51.
- [35] Chinta, P. C. R. (2022). Enhancing Supply Chain Efficiency and Performance Through ERP Optimisation Strategies. *Journal of Artificial Intelligence & Cloud Computing*, 1(4), 10-47363.
- [36] Sadaram, G., Sakuru, M., Karaka, L. M., Reddy, M. S., Bodepudi, V., Boppana, S. B., & Maka, S. R. (2022). Internet of Things (IoT) Cybersecurity Enhancement through Artificial Intelligence: A Study on Intrusion Detection Systems. *Universal Library of Engineering Technology*, (2022).
- [37] Karaka, L. M. (2021). Optimising Product Enhancements Strategic Approaches to Managing Complexity. *Available at SSRN 5147875*.
- [38] Chandrasekaran, A., & Kalla, D. (2023). Heart disease prediction using chi-square test and linear regression. *Computer Science & Information Technology*, 13, 135-146.
- [39] Kalla, D., & Kuraku, S. (2023). Phishing website url's detection using nlp and machine learning techniques. *Journal of Artificial Intelligence*, 5, 145.
- [40] Kuraku, D. S., & Kalla, D. (2023). Impact of phishing on users with different online browsing hours and spending habits. *International Journal of Advanced Research in Computer and Communication Engineering*, 12(10).
- [41] Kuraku, S., Kalla, D., Samaah, F., & Smith, N. (2023). Cultivating proactive cybersecurity culture among IT professional to combat evolving threats. *International Journal of Electrical, Electronics and Computers*, 8(6).
- [42] Kuraku, D. S., Kalla, D., Smith, N., & Samaah, F. (2023). Exploring How User Behavior Shapes Cybersecurity Awareness in the Face of Phishing Attacks. *International Journal of Computer Trends and Technology*.
- [43] Chinta, P. C. R. (2023). Leveraging Machine Learning Techniques for Predictive Analysis in Merger and Acquisition (M&A). *Journal of Artificial Intelligence and Big Data*, 3(1), 10-31586.
- [44] Kuraku, D. S., Kalla, D., Smith, N., & Samaah, F. (2023). Safeguarding FinTech: elevating employee cybersecurity awareness in financial sector. *International Journal of Applied Information Systems (IJAIS)*, 12(42).

- [45] Moore, C. (2023). AI-powered big data and ERP systems for autonomous detection of cybersecurity vulnerabilities. *Nanotechnology Perceptions*, 19, 46-64.
- [46] Chinta, P. C. R. (2023). The Art of Business Analysis in Information Management Projects: Best Practices and Insights. *DOI*, 10.
- [47] Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. *J Contemp Edu Theo Artific Intel: JCETAI-104*.
- [48] Maka, S. R. (2023). Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. *Available at SSRN 5116707*.
- [49] Routhu, KishanKumar & Katnapally, Niharika & Sakuru, Manikanth. (2023). Machine Learning for Cyber Defense: A Comparative Analysis of Supervised and Unsupervised Learning Approaches. *Journal for ReAttach Therapy and Developmental Diversities*. 6. 10.53555/jrtdd.v6i10s(2).3481.
- [50] Chinta, Purna Chandra Rao & Moore, Chethan Sriharsha. (2023). Cloud-Based AI and Big Data Analytics for Real-Time Business Decision-Making. 36. 96-123. 10.47363/JAICC/2023.
- [51] Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. *J Contemp Edu Theo Artific Intel: JCETAI-104*.
- [52] Bodepudi, V. (2023). Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. *Journal of Artificial Intelligence and Big Data*, 3(1), 10-31586.
- [53] Polu, A. R., Buddula, D. V. K. R., Narra, B., Gupta, A., Vattikonda, N., & Patchipulusu, H. (2021). Evolution of AI in Software Development and Cybersecurity: Unifying Automation, Innovation, and Protection in the Digital Age. *Available at SSRN 5266517*.
- [54] Polu, A. R., Vattikonda, N., Buddula, D. V. K. R., Narra, B., Patchipulusu, H., & Gupta, A. (2021). Integrating AI-Based Sentiment Analysis With Social Media Data For Enhanced Marketing Insights. *Available at SSRN 5266555*.
- [55] Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., Polu, A. R., Vattikonda, N., & Gupta, A. K. Advanced Edge Computing Frameworks for Optimizing Data Processing and Latency in IoT Networks.
- [56] Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., Polu, A. R., Narra, B., & Gupta, A. K. Predictive Analytics in E-Commerce: Effective Business Analysis through Machine Learning.
- [57] Jha, K. M., Bodepudi, V., Boppana, S. B., Katnapally, N., Maka, S. R., & Sakuru, M. Deep Learning-Enabled Big Data Analytics for Cybersecurity Threat Detection in ERP Ecosystems.

License

Copyright (c) 2025 Rajiv Chalasani, Venkataswamy Naidu Gangineni, Sriram Pabbineedi, Mitra Penmetsa, Jayakeshav Reddy Bhumireddy, Mukund Sai Vikram Tyagadurgam



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Authors retain copyright and grant the journal right of first publication with the work simultaneously licensed under a [Creative Commons Attribution \(CC-BY\) 4.0 License](https://creativecommons.org/licenses/by/4.0/) that allows others to share the work with an acknowledgment of the work's authorship and initial publication in this journal.