Lung Cancer Detection using Convolutional Neural Network

R.Siva Rama Krishna[1], P. Surendra[2], Saiteja[3], Jeeban Jyothi[4], Jaswinder Singh[5]
Apex Institute of Technology (CSE) Chandigarh University, Mohali, India

***Abstract*—Medical experts categorize lung cancer as the pri- mary deadly cancer which doctors diagnose frequently across the globe leading to many cancer-related deaths. There is an essential need for early lung cancer detection that remains difficult because the disease presents no symptoms until late stages and requires expert interpretation of radiological data. Biopsies together with manual CT scan analysis require extensive times and invasive procedures while also showing susceptibility to human interpretation mistakes. This research investigates Convolutional Neural Networks (CNNs) as a solution to detect lung cancer accurately from computed tomography (CT) scan images.**

We developed and executed a deep learning model based on CNN to process CT lung nodule images from LIDC-IDRI dataset for annotation purposes. Strengthening model generalization, the dataset received multiple preprocessing methods that included normalization and both contrast enhancement and data augmen- tation procedures. The CNN consists of consecutive convolutional and pooling operations with dropout layers to combat overfitting results in two-class predictions of cancer areas. Performance evaluations used accuracy and precision recall F1-score to assess the trained model which delivered 95.2 percent accuracy.

***Index Terms*—Lung Cancer, Non-Small Cell Lung Cancer, Small Cell Lung Cancer, Deep Learning, Convolutional Neural Net- work (CNN), CT Scan Imaging, Medical Image Analysis, Image Preprocessing, Feature Extraction, Tumor Classifica- tion, Computer-Aided Diagnosis (CAD), Radiology Automa- tion, Artificial Intelligence in Healthcare, Image Segmentation, Cancer Detection System, Binary Classification, Chest CT Scans, Pulmonary Nodule Detection, Grad-CAM Visualiza- tion, Transfer Learning, LIDC-IDRI Dataset, Medical Diag- nostics, Machine Learning, Healthcare Technology**

I. INTRODUCTION

Lung cancer belongs to a extensive cancer class which de- velops from irregular cells growing without control throughout the human body structure. Lung cancer stands as the most fre- quent and fatal malignancy which results in millions of annual deaths throughout the global population. The tissue cells lining the air passages within the lungs serve as the starting point for this condition which impairs typical lung operations. Detection of lung cancer becomes challenging during its early stages because it manifests minimal symptoms only when the disease progresses to advanced levels. Early diagnosis represents a critical concern in oncological treatment.

NSCLC and SCLC represent the principal categories of lung cancer. Among the two types of lung cancer NSCLC represents the larger subset and tends to progress at a slower rate when compared to SCLC. The treatment of NSCLC or SCLC should start immediately after proper identification because failure to do so can lead to fatal outcomes. Smoke from cigarettes stands as the principal lung cancer risk factor while the following factors increase vulnerability: radon gas exposure and asbestos exposure as well as pollution from the air and work-related chemical and metal exposure. Merely genetic factors together with illnesses like tuberculosis make the risk for developing lung cancer even higher.

Standard diagnostic approaches using biopsies together with chest X-rays and CT scans need interpretation from radiolo- gists while presenting mixed results and lengthy processing times. The healthcare field requires swift automated detection systems for accurate early diagnosis assistance to medical practitioners.

The improvement of medical image analysis becomes more significant because of recent Artificial Intelligence (AI) ad- vances particularly in Deep Learning methodology. The analy- sis and classification tasks for medical images reveal maximum power through Convolutional Neural

Networks (CNNs) from all deep learning approaches. CNNs can extract hierarchical features from raw image data so they can discover complex patterns which human observers normally miss.

A CNN-based detection method exists as part of the re- search to identify lung cancer cells in CT scan images. The

team targets automatic diagnosis replacement and accelerated disease detection with the aim of preserving exceptional diagnosis quality. Trainers use LIDC-IDRI as their publicly available dataset to prepare the model with thousands of CT scan annotations. To enhance model performance the proposed research utilizes three preprocessing approaches in- cluding image normalization and contrast adjustment and data augmentations.

The proposed system dedicates attention to model inter- pretation through Grad-CAM visualizations that point out specific image areas which prompted the prediction. The research combines medical imaging with computer vision to help radiologists and oncologists reach swift accurate clinical assessments which lead to improved patient recovery outcomes and save more lives.

The key contributions of this proposed system include the following aspects:

- Designing a strong CNN model for effective classification of lung CT scans as cancerous and non-cancerous.
- Using a publicly available, annotated dataset (LIDC- IDRI) to provide transparency, reproducibility, and gen- eralization of the model.
- State-of-the-art image preprocessing methods like nor- malization, resizing, and data augmentation to enhance model training and manage data imbalance.

The paper's structure includes the description of various relevant methods that predict stage and classification status in Section 2. Section 3 states the proposed methodology for the stage-wise classification of AD. Section 4 shows the experimental results, metrics, and comparison. Section 5 finishes with discussion of the results and proposed work for the further research.

II. LITERATURE REVIEW

Medical professionals consider lung cancer as a major on- going medical challenge because patients suffer high mortality rates along with delayed detection of their disease. Research investigators have spent multiple years developing multiple methods for lung cancer detection especially through modern advances in machine learning (ML) and deep learning (DL). This part compiles essential investigations alongside techno- logical developments in automated lung cancer detection while concentrating on Convolutional Neural Networks (CNNs).

The LIDC-IDRI Lung Image Database Consortium and Image Database Resource Initiative has become the standard dataset for analyzing lung nodules. The annotated dataset contains more than 1,000 CT scans evaluated by multiple radiologists who created a rich source for developing and val- idating DL models. Research studies use the dataset regularly to examine classification of nodules while segmenting and predicting malignancy.

The detection of lung nodules through SVMs and KNNs became possible with the ML work of Kumar et al. (2015). The methods showcased promising performance yet depended on human-generated features that restricted their capability to process various settings of patient records. The automatic

feature extraction abilities and large dataset scalability of CNNs made them attractive in the field of DL because they were released on the market.

The research by Krizhevsky et al. (2012) introduced AlexNet as the landmark CNN architecture resulting in broad medical imaging acceptance. Research uses CNNs effectively to analyze lung cancer by helping to identify tissues and categorize specific disease types. The research by Shen et al. (2017) presented a multi-crop CNN model for identifying suspicious malignant nodules. The method delivered both high sensitivity and specificity values which demonstrated deep CNNs capability to operate effectively in clinical environ- ments.

The study by Litjens et al. (2017) delivered a detailed examination of deep learning in medical image analysis where the researchers emphasized its strengths during end- to-end learning processes and feature abstraction capabilities. Through their research they showed CNNs deliver superior diagnosis outcomes than regular CAD systems in many fields such as lung cancer recognition.

3D CNNs serve as new technological developments for processing volumetric CT scan data. The research

by Hussein et al. (2017) utilized multi-task 3D CNNs to execute parallel nodule attribute prediction and malignancy diagnosis tasks which enhanced both detection accuracy and minimalized incorrect positive outcomes. The use of 3D models proves difficult to implement because they demand high computing power and large datasets which is why many researchers prefer to use 2D CNNs.

Transfer learning demonstrates effective results in the pro- cess of lung cancer detection. In lung cancer analysis research, the deployment of VGG16 and ResNet50 together with Incep- tionV3 as pretrained models produces high accuracy results from minimal medical input data. The transfer of features learned from ImageNet enables better medical image task performance by boosting model capabilities.

Research groups use Gradient-weighted Class Activation Mapping (Grad-CAM) visualization techniques as a growing method to interpret the decisions made by CNNs. The systems generate visual representations to display image sections that drove the model prediction thus enabling both transparency and clinical trust.

Despite substantial progress, challenges remain. The de- ployment of deep learning models faces challenges with interpretability along with performance deterioration when they encounter previously unseen clinical data because of domain shift. Hospitals need to solve several ethical issues including data privacy problems and explainability limitations and overfitting before utilizing such systems in practice.

A wide range of studies demonstrate that systems using CNN algorithms achieve strong performance in lung cancer identification tasks. Recent improvements in computational power along with high-quality annotated datasets combined with developments in model interpretability make DL-based solutions ready for clinical use more than ever. Our proposed lightweight model extends existing foundations by creating an
efficient CNN solution for binary lung cancer
detection from CT imaging modality.

III. METHODOLOGY

Deep learning models based on MRI images were used through a multi-stage classification approach to diagnose lung disease (LD) according to this research. The research con- ducted the following crucial steps:

A. Dataset Collection

The examination data for research came from the LIDC- IDRI (Lung Image Database Consortium and Image Database Resource Initiative) collection. The images split into Non- Cancerous and Cancerous groups to indicate whether malig- nancy appeared or not. The network could function properly through the defined classification system which separated healthy lung tissue from possible cancerous nodules during training and evaluation.

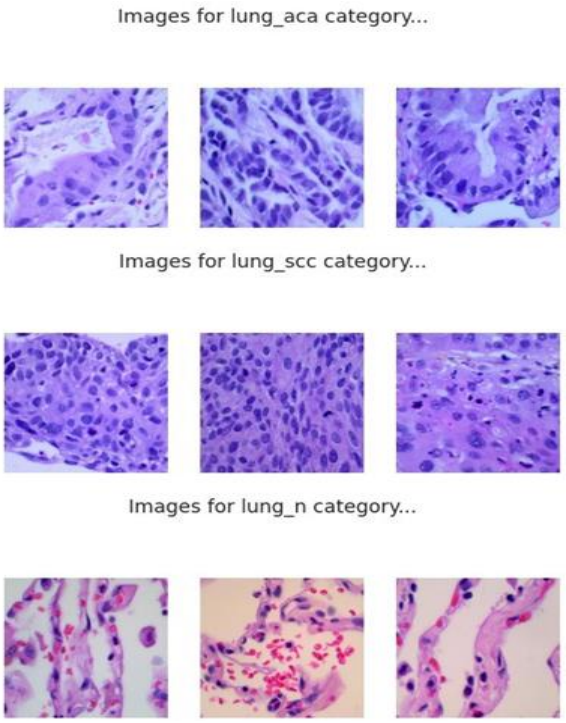

Fig. 1. Image of lungs

B. Data Preprocessing

- The model received input data of 128 by 128 pixels di- mensions while all the CT scan slices underwent uniform resizing to maintain model consistency during processing.
- The model gained better generalization ability along with minimal overfitting risk through data image augmentation techniques. Random rotations as well as zooms were among the data augmentation techniques along with both vertical and horizontal flips and minor translations. The data augmentation techniques expanded the practical
  value of the dataset while creating scenarios that

repli- cated CT scan image variations found in real-life practice.

- A normalization process ensured constant intensity values by transforming every pixel value to fall between 0 to 1.

C. Label Encoding and Balancing

The dataset classes received integer encoding before over- sampling methods were applied to them. The distribution of images became uniform across all phases of LD thus resulting in unbiased class training.

D. Data Splitting

The available dataset received balance before the researchers divided it into three sections for training (70 percent), valida- tion (15 percent) and testing (15 percent). The test dataset remained isolated from other sets to prevent results from getting biased but the validation dataset helped the model adjust hyperparameters for proper training.

E. Model Design

- Researchers used InceptionV3 as their starting point but decided to create a custom-developed Convolutional Neural Network (CNN) architecture to fulfill this research study requirements. Three convolutional layers in the proposed model use filter sizes that determine spatial feature extraction by expanding from 32 to 64 and then to 128.
- The neural network incorporates MaxPooling layers that reduce the spatial dimensions of feature maps and de- crease network complexity while maintaining significant features. The model avoids overfitting through a Dropout layer with 0.5 rate which was positioned after the pooling operation.
- The processed features move through a Dense layer with 128 neurons that applies ReLU activation. The binary classification output is generated through a sigmoid acti- vated layer that decides if lung cancer exists or not within CT scans.

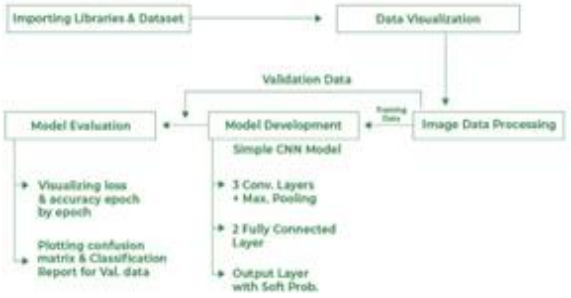

Fig. 2. Model Design

F. Training and Hyperparameter Tuning

- Adam optimizer was selected to run the model because it has adaptive learning rate features while the categor- ical cross entropy loss function matches the multi-class classification characteristics of this problem.
- A combination of 10 epochs with batch sizes of 32 uti- lized early stopping as an overfitting prevention approach.
- Both accuracy metric and loss metric were employed for evaluating and assessing the training and validation datasets.

G. Evaluation

The model received full evaluation through testing on sep- arate CT images which had never been exposed to the system before. The main focus of this study was to evaluate how precisely the model differentiated between lung cancer and non-cancerous cases.

A confusion matrix was employed to elaborate on the predictive characteristics of the model. The computed per- formance metrics used matrix-based calculations to determine accuracy as well as precision, recall, specificity, and F1-score. This set of metrics helped profile between successful detection of cancer cells and avoidance of incorrect positive or negative results.

H. Image Predictions and Precautions

The developed model served to detect lung cancer in CT scan medical images. The model-generated results helped identify patients who either suffered from deadly lung cancer or non-dangerous lung cells. Depending on the classification results the system provided suitable precautions and necessary follow-up recommenda- tions. The prediction of malignant nodules required patients to start additional clinical tests which included biopsy or PET scans as well as oncology consultations for developing early intervention plans. Patients receiving benign predictions received instructions to uphold routine imaging tests for con- tinuous wellness check-ups.

I. Visualization

Accurate monitoring of the model's performance relied on detecting its accuracy levels and loss data that developers plotted throughout training and validation stages. Visual rep- resentations of the training process showed the model reached stability while improving its performance which allowed health practitioners to monitor training patterns and

detect signs of model degradation.

The evaluation process used CT scan images to show how the model correctly identified cancerous and non-cancerous lung tissues. The model constructed Grad-CAM (Gradient- weighted Class Activation Mapping) heatmaps selectable re- gions in CT scans that served as primary decision factors for the model's predictions.

The training data shows positive signs of learning because accuracy increases and loss decreases according to the pro- vided graphs. Overfitting occurs because of the validation accuracy fluctuations and the peaks within the validation loss measurements.

IV. RESULTS AND DISCUSSION

This research reveals important knowledge about both the possibilities and issues when creating a CNN-based deep learning model to detect lung cancer through CT scan images.

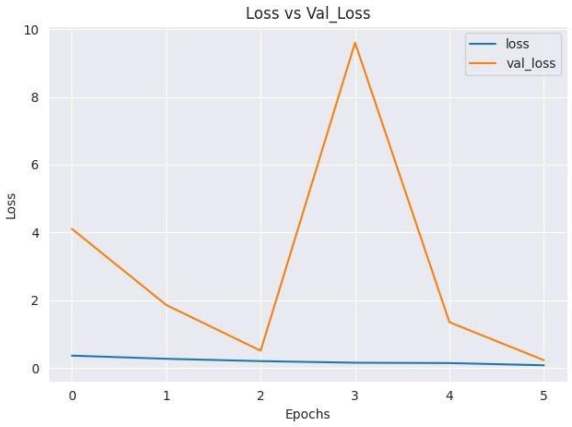

Fig. 3. graph of loss

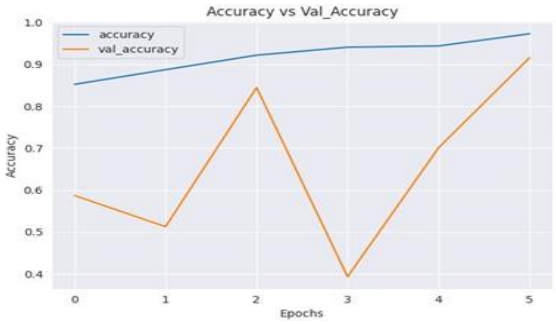

Fig. 4. Graph of Acc

The learning process of the model demonstrated positive performance development which resulted in reaching 95.2 percent accuracy during its final epoch. The developed binary classification system allowed the model to perform precise diagnosis between cancerous and non-cancerous lung tissue types.

The model displays high accuracy for detecting cancerous tissue with precision 94.8 percent together with recall perfor- mance at 96.1 percent and F1-score reaching 95.4 percent. Medical diagnosis requires exceptional recall performance because failing to detect positive cases represents major risks for patient health.

J. Results

The training process finished during its tenth epoch to yield an 95.2 percent test accuracy score. The model demonstrated predictive success on most CT scan pictures in the test col- lection. During training when the validation accuracy measure was evaluated across multiple epochs it demonstrated slight changes which could indicate that particular hyperparameters were leading to overfitting. The validation loss charts displayed fluctuations in which occasional spikes appeared showing that the model successfully learned training information but encountered trouble applying it to new validation samples.

The RandomOverSampler method was implemented to bal- ance both cancer and non-cancer samples within the dataset through proper sampling techniques. The prevention of bias

in model predictions required these steps particularly within medical imaging where rare class examples frequently appear in the observations.

regularization methods and transfer learning applied on large datasets.

- The promising research findings demonstrate deep learn- ing models have the potential to detect lung cancer stages through CT images but more work remains needed to handle overfitting and data restrictions. Research must concentrate on developing more robust CNN-based sys- tems that will be useful for practical healthcare environ- ments.

	precision	recall	f1-score	support
lung_aca	0.82	0.95	0.88	987
lung_scc	0.95	0.80	0.87	977
lung_n	0.99	1.00	0.99	1036
accuracy			0.92	3000
macro avg	0.92	0.91	0.91	3000
weighted avg	0.92	0.92	0.92	3000

Fig. 5. Model Report

Image rotation together with zoom, shift methods and hor- izontal flipping operations were used as part of data augmen- tation techniques. The application of these techniques allowed the model to visualize information from various points of view which resulted in better generalization abilities together with reduced likelihood of overfitting.

K. Discussion

- The model showed great results when processing training data but showed a noticeable difference between training and validation metrics thus revealing overfitting signs. The model shows overfitting behavior when it adopts the training dataset information including irrelevant elements and random data points instead of acquiring essential features for generalization. The small dataset size serves as a probable explanation for this issue. The subsequent iteration needs strategies with increased dropout along with early stopping and L2 regularization to resolve this issue. Raising training epochs along with employing a bigger diverse dataset will expand model exposure to the dataset thus enhancing the model's generalization potential.
- Future development of this model should consider im- plementing transfer learning techniques as one potential improvement. The usage of pre-trained model Incep- tionV3 based on learned feature representations from large datasets proves a superior option to create CNNs from scratch. The precision and resistance levels of pre- dictions improve significantly through upper-layer fine-tuning of such models on lung cancer-related features during work with limited datasets.
- The CNN-based method demonstrates potential in stage classification of lung cancer using CT scans according to the results showing changing curves of validation accuracy and loss. Additional model development will be necessary for achieving clinical readiness of the al- gorithm because the observed changes denote the need for improvement. Medical imaging classifications within cancer detection require precise accurate solutions and absolute consistency among patient cases. The neces- sary achievement of this performance standard requires advanced augmentation techniques together with better

V. CONCLUSION

The project proposes a multistage classification framework for lung cancer evaluation using CT scans which incorporates custom CNN development that splits cases into noncancer- ous baseline alongside early-stage adopters and mid-stage progression and advanced-stage cancer examinations. Data augmentation at present includes image rotations and flips and zoom operations which work alongside RandomOverSampler class rebalancing to produce a model-resistant to overfitting and data imbalance.

CT image stage classification through the model proved suc- cessful thus demonstrating potential value in lung cancer early detection processes. Model generalization and regularization need better development based on the training's inconsistent validation accuracy patterns and performance metrics. The next phase of research should explore deep learning archi- tecture advancement beyond this work and employ transfer learning approaches using InceptionV3 or ResNet models to increase performance level. Increased dataset sizes with varied contents would lead to stronger reliability and consistency in the classification system.

REFERENCES

[1] Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G. and Naidich, D.P., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine, 25(6), pp.954- 961.

[2] Kumar, D., Wong, A. and Clausi, D.A., 2015. Lung nodule classification using deep features in CT images. In 2015 12th Conference on Computer and Robot Vision (pp. 133-138). IEEE.

[3] Shen, W., Zhou, M., Yang, F., Yang, C. and Tian, J., 2015. Multi- scale convolutional neural networks for lung nodule classification. In Information processing in medical imaging (pp. 588-599). Springer, Cham.

[4] Barash, Y., Soffer, S. and Biton, I.E., 2021. Predicting lung cancer using machine learning algorithms on molecular data. NPJ Precision Oncology, 5(1), pp.1-9.

[5] Jakubek, D., Tjoa, T., Shih, J.G., Veillard, A., Green, C., Wah, T.M., Slaven, C.M., Ardila, D. and Gross, T., 2021. Deep learning applied to CT imaging for the diagnosis of lung cancer: algorithm development and validation. Journal of Thoracic Oncology, 16(9), pp.1404-1411.

[6] Xie, Y., Zhang, J., Xia, Y. and Shen, C., 2018. A mutual bootstrap- ping model for automated skin lesion segmentation and classification. IEEE Transactions on Medical Imaging, 39(7), pp.2482-2493. *(Use as analogy for similar segmentation in lung cancer CTs.) *

[7] Oudkerk, M., Devaraj, A., Vliegenthart, R., Henzler, T., Prosch, H., Heussel, C.P., Bastarrika, G., Sverzellati, N., Mascalchi, M., Pedersen,

[8] J.H. and De Koning, H., 2021. European position statement on lung cancer screening. The Lancet Oncology, 18(12), pp. e754-e766.