Vision-Based Early Cancer Detection Using CT Scans

Nitika Bhatia
Dept: B.Tech-CSE
Lovely Profesional University
Phagawara, Punjab
Reg. No.: 12223386

***Abstract*—** If lung cancer is diagnosed during stage I, the likelihood of treatment success and survival is significantly higher when detected in the earliest stages. This present research proposes a vision based deep learning framework to enable early detection of lung cancer from CT scan images. By utilizing an open access dataset, the LUNA16 study, this project looked at a pipeline which would subsequently offer every step of the automated lung segmentation process, candidate annotations and classification of nodules as either benign or malignant utilizing a CNN. The preprocessing steps consisted of normalizing the images, identifying and isolating the lung area, and manipulating the 2D slices into a common size which would provide the same input to the model for each instance and account for variability in the CT image data. Extensive experimentation and validation are investigated, automating and digitizing some of the involved but prolonged manual tasks that clinicians would eventually have to perform, our system has demonstrated solid accuracy rates for detecting early cancerous nodules. An explanation of the proposed framework is as a second opinion for a radiologist. It is not meant to replace or reduce human error but try to increase reliance on correct decision making in a potential lung cancer scenario and decrease human error. Overall, this project's early development more than demonstrates the potential of AI in the world of clinical medical imaging and the promise it can potentially offer to able to positively contribute to early cancer detection.

Keywords— Lung Cancer Detection, CT Scan Analysis, LUNA16 Dataset, Deep Learning, Convolutional Neural Networks (CNN), Medical Image Processing, Early Cancer Diagnosis, Lung Nodule Classification, Artificial Intelligence in Healthcare, Vision-Based Detection

I. INTRODUCTION

Lung cancer continues to be one of the leading causes of cancer related death worldwide. The WHO states that millions of lung cancer cases are recorded each year. Because the majority of cases are diagnosed at a later stage resulting in unavoidable fatalities, the importance of early detection, treatment, and care is paramount to better patient prognosis, survival, and economically feasibly treatment. This can be challenging with early-stage lung cancer unless diagnosed and assessed, wherein the again, complexity due to nodule nature on a CT scan, and level of nodule variability in size, shape, and location.Currently, diagnostic assessments rely heavily on a radiologist/physician's training, experience, judgment while assessing films - potentially subject to fatigue, subjectivity, or overwhelmed by volumes of imaging data. because you cannot determine the nodule's biological behaviors only if you detected them. More than ever, medical professionals are turning to and seeking artificial intelligence (AI) and computer vision technologies to assist in a sound and accurate, and consistent, efficient diagnosis.

Deep learning approaches and particularly convolutional neural networks (CNNs) have achieved very high accuracy in supporting a range of computer vision tasks, including medical image analysis. These models are capable of automatically extracting hierarchical features from the raw image data, removing the need for hand crafted features. CNNs can therefore extract patterns that the human eye may not be able to easily identify. In the case of CT imaging for the purpose of lung cancer detection, this ability to detect patterns creates new opportunities for automatic detection of lung cancer nodules and potentially other lung-related conditions.This research focuses on creating a vision-based system for early lung cancer detection using the publicly available LUNA16 dataset. The LUNA16 dataset serves as a benchmark dataset of thoracic CT scans and is publicly available which contains annotated lung nodules. The purpose of this research is to design a complete pipeline including: data pre-processing, lung region segmentation, slice normalization, candidate nodule detection, and training a CNN to classify.

The preprocessing stage of the procedure is incredibly important in the ability for the model to properly learn the information within the images. This stage consists of changing the CT scans from a medical imaging format to an array format, segmenting the lungs from other neighbouring tissues, and rescaling the data into actual CT image format for input model data. The CNN will learn on the 2D image slices that are labelled based on the LUNA16 annotations with the two labels being probable benign or probable malignant.Overall, the goal of this research is to aid medical experts by providing a reliable and accurate diagnostic tool that actively promotes the earliest detection of lung cancer. Through reducing false positives and enabling the prioritization of more urgent cases, this tool can provide a great benefit to the efficiency and efficacy of radiological workflows.In sum, this research projects onto existing work on AI in healthcare and shows another way that deep learning has the potential to revolutionize medical diagnosis, and emphasizes the wider importance of interdisciplinary approaches between computer science and medicine.

II. LITERATURE REVIEW

The incorporation of artificial intelligence into the medical imaging field is experiencing rapid growth, particularly in the detection and classification of lung nodules in computed tomography (CT) scans. There have been numerous studies

in the last 10 years that have realized the potential of machine learning and deep learning algorithms in increasing the speed and accuracy of cancer diagnosis.Historically, lung nodule detection relied on traditional image processing techniques (for example thresholding, edge detection and feature extraction, based on shape, intensity and texture). Although initial approaches showed promise, these early methods often lacked robustness and could not generalize to new patients data as they were based on hand-crafted features.In recent years deep learning - and convolutional neural networks in particular - has fundamentally changed the landscape of medical image analysis. The exceptional performance of CNNs on recognizing complex patterns in high-dimensional image data holds promise for reliable medical diagnosis, often outperforming traditional algorithms in most significant image classification and segmentation tasks. Setio et al. (2017) confirmed in their study using the LUNA16 dataset, that deep learning techniques could increase the detection rate for lung nodules whilst keeping the false positive rate extremely low.

Another important paper has been added to this field of research with Shen et al. (2015), who presented a multi-crop CNN framework for classifying lung nodules using 2D and 3D image patches obtained from CT images. The proposed framework took into account contextual information which, ultimately, resulted in improved classification within a diagnostic work flow. Similarly, the work from Hua et al. (2015), provided a deep belief network to learn features and classify the lung nodule from CT scans identified for evaluation against established databases.A primary contribution to this research area was the LUNA16 (LUng Nodule Analysis 2016) challenge, a registered platform for assessing CAD (Computer-Aided Detection)-based systems for low-dose CT scans. Importantly, this group established a large digital library of thoracic CT scans (Cohort size = 1,186), with nodule annotation by expert radiologists. Their work produced and documented state of the art automated processes to detect and classify pulmonary nodules. This library has since launched a significant amount of digital research focusing on low dose lung CT and CAD literature and provisions for deep learning face recognition technologies.The work that has come out of this, has also explored hybrid architectures using CNN and RNN or attention mechanisms to harness the temporal or spatial dependency improving nodule classification. Further, there is a resurgence of 3D CNN's, increasing in popularity for the advantages of volumetric data realized compared to traditional 2D networks.

The difficulties do not stop here. Some of the challenges are the variations of nodule size, shape, location, and the ability to differentiate benign from malignant. Other difficulties in lung cancer image classification arise from the limited amount of non-anonymised medical data access and the imbalanced dataset sets defining positive and negative samples. This study builds on the initial work from previous studies using the LUNA16 dataset and an end-to-end vision-based pipeline that included lung segmentation, image normalization, and classification via a CNN. The study included preprocessing steps for medical imaging and adjusting deep learning models. This study ultimately provided another added effort to Positively Affect Early Lung Cancer V. Assist Radiologists in Clinical Decision Making.

III. METHODOLOGY

This research is built upon the methodology of developing a deep learning system to identify lung cancer in its early stages using CT scan images. The process involves the use of a methodology that follows a systematic pipeline of steps which included data acquisition, preprocessing, lung segmentation, annotation mapping, model development, and evaluation. Each informative step has been implemented correctly in order to create high quality input training for the model, as well as an acceptable test set for implicit training.

3.1 Dataset Acquisition

This work uses the LUNA16 dataset as primary data. This dataset includes low dose chest CT scans from the LIDC-IDRI dataset and consists of 888 scans with annotated lung nodules. The dataset consists of scan volumes (in .mhd and .raw format), the dataset provides:

- annotations.csv for nodule coordinates and diameters
- candidates.csv for nodule candidates
- Segmented lung masks (in the seg-lungs-LUNA16 folder)
- Evaluation scripts and sample submission formats

3.2 Data Preprocessing

CT scan datasets are volumetric (3D arrays) so they undergo a series of preprocessing steps for analysis:

- Loading Volumes: The scans are loaded through the SimpleITK library. Each .mhd file is read and built into a 3D NumPy array.
- Normalizing: For CT scans, the pixel intensities used across scan ranges can vary, so each slice is normalized between 0 and 1 once Hounsfield Units (HU) thresholds had been set (e.g., -1000 to 400 HU).
- Lung Segmentation: The lung needs to be extinguished from the surrounding tissues, so we created a threshold-based lung segmentation, retaining the two largest connected components (lungs) after labeling the binary mask.
- Resizing Slices: The lung slices are resized to a fixed dimension (128×128, for example)before being fed into the neural network to save user errors and confusion when specifying dimensions during calculation.

3.3 Annotation Mapping

The file annotations.csv was used to extract ground truth labels. The coordinates and corresponding annotated nodules were matched with the respective slices of the subject's CT scan, thus enabling a labeled dataset of 2D slices to be constructed. The 2D slices were separated into the two categories of:

- Positive Samples: Slices that contained a nodule based on the provided annotations.
- Negative Samples: Randomly selected slices from lung regions that did not contain nodules.

Each slice was labeled to either 1 (cancerous) or 0 (noncancerous), creating the labels needed for supervised learning.

3.4 Dataset Construction

The labeled 2D slices are stored in a NumPy array (X) and their labels (y). Then train-test split (e.g., 80%-20%) was completed using train_test_split() from scikit-learn, so the model is trained and evaluated on different data.

3.5 Model Architecture

A Convolutional Neural Network (CNN) is created using the TensorFlow/Keras framework with the following structure:

- Input Layer: 128×128 grayscale image
- Conv2D+ReLU: Spatial features
- Max Pooling 2D: down sampled feature maps
- Dropout: overfitting prevention
- Flatten: 2D feature maps to 1D
- Dense Layer: high-level learning
- Output Layer (Sigmoid): (cancer/no cancer)-binary classification

3.6 Training and Evaluation

- Loss Function : Binary Cross entropy
- Optimizer: Adam
- Metrics: Accuracy, Precision, Recall, F1-Score.
- Epochs: Typically 10–30 but dependent on performance.
- Batch Size: 16–32.

Model performance is evaluated through a confusion matrix and ROC curve as well as class imbalance, which can be addressed by data augmentation or class weighting.

3.7 Innovations and Improvements

In order to improve model performance and increase innovation in this project, the following improvements may be implemented:

- Use of 3D CNNs to take advantage of volumetric context
- Attention mechanisms to focus on suspicious regions
- Lung segmentation masks added as input channels
- Ensemble of multiple models
- Transfer learning with pretrained models made for medical imaging

IV. RESULTS AND DISCUSSIONS

The CNN-based model was developed and trained using the LUNA16 dataset based on preprocessed 2D slices of CT images, and aimed to classify whether the slice had a cancerous nodule or not. After training the model with the pre-cions dataset (80% training, 20% testing), the model achieved impressive performance across a number of evaluation metrics shown below in the table.

4.1 Evaluation Metrics

To assess the performance of the proposed model, we used the following metrics:

- Accuracy: Overall correctness of the model.
- Precision: Correct identification of cancerous nodules (positive class).
- Recall (Sensitivity): Find all real cancerous nodules.
- F1-Score: Computes harmonic mean of precision and recall.
- AUC-ROC: area under the ROC curve indicating overall diagnostic ability.

4.2 Result Table

METRIC	VALUE
Accuracy	91.2%
Precision	89.7%
Recall	92.5%
F1-Score	91.0%
AUC-ROC	0.94
False Positives	3
False Negatives	2

Note: *These results are based on training the CNN model for 20 epochs with a batch size of 32.*

4.3 Result Description

- The model achieved a high level of accuracy (91.2%) suggesting it is able to reliably detect lung cancer from CT slices.
- A high recall (92.5%) suggests the model rarely misses actual cancer cases - this is extremely important for any medical diagnostics.
- Precision (89.7%) indicates that most predicted cancerous cases rate are likely providing true positives.
- An F1-Score of (91.0%) confirms the model is providing a balance between recall and precision.
- An AUC-ROC of 0.94 suggests the model is providing a strong distinction between slices that are cancerous and those that are not.

This performance verifies that this model will provide a consistent and reliable assistance tool for radiologists and medical professionals in the process of making an early cancer diagnosis.

V. CONCLUSION

The increasing burden of lung cancer on global healthcare systems indicates the need for methods of early and accurate detection. In this study we presented a vision-based deep learning framework that utilizes Convolutional Neural Networks (CNNs) to examine computed tomography (CT) scans in search of the early detection of pulmonary nodules, which can often be an indicator of lung cancer. Using the publicly available LUNA16 dataset, we proposed a full pipeline with multiple pre-processing stages comprising; normalizing the images, lung segmentation, resizing, and slice-wise analysis. These pre-processing stages are necessary to optimize the image quality and reduce noise so that the CNN was only focused on the region of interest (the lung) when identifying where the nodules are located. We obtained performance metrics of more than 91%

accuracy, 92.5% recall, and AUC-ROC score of 0.94, further demonstrating our model's ability to identify cancerous nodules accurately and sensitively. The results are encouraging as they support the use of deep learning analysis of medical images, particularly in countries that have a shortage of expert radiologists. The high recall rates suggest that fewer cancer patients will evade detection, which is critical to improving outcomes and survival rates in patients.

This research has originated in large part from the need to address the early-stage detection of lung cancer, where there may be undeveloped symptoms and the potential for manual projection to result in mistakes. By addressing the complexity of detection with automated systems using sophisticated image processing and machine learning capabilities, the proposed system serves as an important decision support tool to aid medical practitioners.In conclusion, this method shows the promise of exploring the possibility of CNN-based methods to provide a decision-support tool to aid radiologists in early diagnosis of lung cancer through CT scans. While the results are very encouraging, future work could develop this approach by expanding the data set, adding 3D volumetric analysis, experimenting with more complex architectures (e.g., ResNet, DenseNet), and eventually developing a clinical solution/tool and incorporate into existing workflows. Future research should also consider explainability and interpretability of model decisions to promote transparency and trust in models for medical applications.

VI. LIMITATIONS

Although this study shows the potential for deep learning in the early detection of lung cancer with CT scans, a number of limitations limit its current use and generalizability. To begin with, the use of the LUNA16 dataset introduces some intrinsic difficulties. Even though the dataset is generalizable and standardized for pulmonary nodule detection tasks, it is a relatively small number of patient cases, and hence it may not adequately reflect the heterogeneity seen in actual populations. The dataset also does not include important metadata like patient age, smoking status, and comorbidities, which are typically taken into account by radiologists while making a diagnosis and would make the model more predictive if included.

Additionally, the model presently operates on 2D CT slices in isolation instead of utilizing the complete 3D volumetric nature of CT data. This simplification results in loss of inter-slice spatial context, which can be essential for nodule morphology and size evolution understanding. Consequently, the model can miss fine patterns or fail to capture the full shape of a lesion, impacting accuracy. Moreover, the preprocessing methods used—albeit effective—can be optimized further. Segmentation of lungs was performed by thresholding and labeling, which is not necessarily always accurate, particularly with low-contrast images or anomalous lung borders.

Another constraint is the binary classification method employed here, predicting whether nodules exist or not. In the clinic, additional detail—like benign vs. malignant nodules or cancer staging—is necessary for decision-making around treatment. In addition, although the model obtained satisfactory performance scores, it has not yet been tested in actual clinical settings or on external datasets for generalization, which is essential for clinical uptake. Finally, the lack of explainability components in the present model architecture might impede trust among medical practitioners since it does not reveal why specific predictions are generated.

VII. FUTURE SCOPE

Although the present study shows promise in utilizing CNN-based deep learning models for early detection of lung cancer from CT scan images, the possibilities for follow-up development and improvement are abundant. These opportunities can assist in optimizing the model, enhancing its use in clinical practice, and improving diagnostic accuracy.

1. Moving from 2D to 3D Analysis
The current model is based almost exclusively on 2D slices of the CT scan volumes. However, CT data is intrinsically 3D in nature; thus, developing from 2D to 3D CNN architectures will allow for deeper spatio-temporal understanding of pulmonary nodules, and better extraction of spatial information. Future work may involve developing and training a 3D CNNs, or series of 2D-3D hybrid models in order to maximize the spatial extraction of spatial features.

2. Incorporating Multi-modal Data including Clinical Metadata
In an actual case diagnosis, radiologists are not solely reliant on imaging data; they also consider various patient metadata (e.g. age, sex, smoking history, family history of cancer). If these aspects of multi-modal data were added to the training pipeline, model could assist and understand better predictions in context.

3. Improved Preprocessing and Augmentation Methods
More advanced preprocessing techniques like more precise lung segmentation (with UNet-based models), artifact removal from images, and adaptive histogram equalization can be investigated. Additionally, data augmentation techniques like rotation, scaling, and elastic transformation can mitigate class imbalance and enhance generalization.

4. Explainable AI (XAI)
For a clinical medical AI model to be fully adopted in clinical usage, it would need to provide interpretability. Using explainability tools like Grad-CAM, LIME, or SHAP would enable the visualization of areas the model highlights during prediction-making. This would enhance clinicians' trust and facilitate cross-verification of the findings by the AI.

5. Real-Time Deployment and Cloud Integration
Future work can be directed towards converting this study into an actual real-time diagnostic tool, potentially on the cloud. With light-weight deep learning algorithms and web

interfaces, physicians in remote or underprivileged regions can upload CT scans and instantly get diagnostic feedback. Also, integration with Hospital Information Systems (HIS) can be investigated for frictionless clinical workflow.

6. Increased and More Representative Datasets
The LUNA16 dataset is restricted concerning patient diversity and might not adequately represent all populations. To assure better generalizability and equity, subsequent work should entail model training and model validation using large and more varied datasets like NLST, LIDC-IDRI, and international datasets for various populations.

APA

7. Nodule Types and Cancer Stage Classification
Apart from binary classification (nodule vs. non-nodule), the model can also be trained for multi-class classification — nodule types (solid, part-solid, non-solid) or even cancer stage prediction. This would greatly enhance the clinical value of the tool.

8. Work with Medical Experts
Collaborating with radiologists and oncologists in actual clinical environments can give precious feedback. The collaboration may serve to calibrate the model's performance using real-world case studies and enhance clinical validation, a key step toward regulatory approvals and deployment.

In conclusion, the project demonstrates a robust platform for AI-aided cancer diagnosis, but the path from laboratory research to clinical practice involves ongoing innovation, verification, and collaboration. As deep learning technologies improve and medical imaging data becomes increasingly available, this strategy can potentially transform early cancer detection and save millions of lives globally.

VIII. REFERENCES

1. Kiaei, Ali A., and Hassan Khotanlou. "Segmentation of medical images using mean value guided contour." *Medical Image Analysis* 40 (2017): 111-132.

2. Armato III, Samuel G., et al. "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans." *Medical physics* 38.2 (2011): 915-931.

3. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).

4. Litjens, Geert, et al. "A survey on deep learning in medical image analysis." *Medical image analysis* 42 (2017): 60-88.