|  |
| --- |
| UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI    https://lh7-us.googleusercontent.com/docsz/AD_4nXerJI51KeoYOPO4XMlHdhzPR1Gm4DKojp9aUMll7v0v5YpaGW0qNiy_UnV5K9u8mPLmXO9EvyTqjh7_ZMvdBgV2ghUgObW3z0RGnCZ5vaKB0hKkPJT4Uw6meMaaft4GA6T2G6pjr6SRLA82Kr78baf-m1UV?key=qovFjOnLIGOJmpEFlFVcdA  Research and Development  **BACHELOR THESIS**  By  **DANG LE PHUC AN - BI11-005**  Information and Communication Technology  (ICT)  Title:  **“Lung Nodules Classification Using Convolutional Neural Network (CNN)”**  **Supervisors: Dr. Nghiem Thi Phuong**  Lab name: USTH ICTLab  **Hanoi, 2024** |

# DECLARATION

Dang Le Phuc An, hereby declare that my thesis does not contain plagiarism from any other source. I certify that the work and the results are entirely of my own.

In case of plagiarism, I acknowledge the consequences and I accept that my thesis will not be evaluated. In this case, my bachelor’s thesis will be marked as "Failed".

July 1st, 2024

Dang Le Phuc An

# ACKNOWLEDGMENTS

First and foremost, I want to extend my gratitude to Dr. Nghiem Thi Phuong for helping me on this project, for being my supervisor, and for the invaluable guidance she has provided throughout my internship.

I would also like to thank Dr. Tran Giang Son and Mr. Kieu Quoc Viet for allowing me to use the school's server and for their tremendous support throughout my research. Their expertise and resources were instrumental in the successful completion of this thesis.

I also wish to convey my deepest thanks to my family for their encouragement from the very beginning of my journey at USTH. Their belief in my capabilities has been my source of strength and motivation throughout this endeavor.

July 1st, 2024

Dang Le Phuc An

# TABLE OF CONTENTS

[DECLARATION 1](#_Toc170828964)

[ACKNOWLEDGMENTS 2](#_Toc170828965)

[TABLE OF CONTENTS 3](#_Toc170828966)

[LIST OF ACRONYMS 5](#_Toc170828967)

[LIST OF TABLES 6](#_Toc170828968)

[LIST OF FIGURES 7](#_Toc170828969)

[ABSTRACT 8](#_Toc170828970)

[I. INTRODUCTION 9](#_Toc170828971)

[1.1. Motivation 9](#_Toc170828972)

[1.2. Background 9](#_Toc170828973)

[1.3. Technology Frameworks 10](#_Toc170828974)

[II. OBJECTIVE 11](#_Toc170828975)

[III. MATERIAL AND METHODS 12](#_Toc170828976)

[3.1. Data Acquisition 12](#_Toc170828977)

[3.2. Data Preprocessing 13](#_Toc170828978)

[3.2.1. Loading the 3D CT Scan Images: 13](#_Toc170828979)

[3.2.2. Normalization: 13](#_Toc170828980)

[3.2.3. Resizing: 13](#_Toc170828981)

[3.2.4. Data Augmentation: 14](#_Toc170828982)

[3.2.5. Data Split 14](#_Toc170828983)

[3.3. Model 14](#_Toc170828984)

[3.4. Training 15](#_Toc170828985)

[3.5. Validation 15](#_Toc170828986)

[IV. RESULT AND DISCUSSIONS 16](#_Toc170828987)

[4.1. Result 16](#_Toc170828988)

[4.2. Discussion 16](#_Toc170828989)

[V. CONCLUSION 17](#_Toc170828990)

[5.1. Summary 17](#_Toc170828991)

[5.2. Future works 17](#_Toc170828992)

[REFERENCE 18](#_Toc170828993)

# LIST OF ACRONYMS

CNNs: Convolution Neural Network

GPU: Graphics Processing Unit

CT Scan: Computed Tomography Scan

DICOM: Digital Imaging and Communications in Medicine

# LIST OF TABLES

# LIST OF FIGURES

[Figure 1: Nodule 12](#_Toc170829081)

[Figure 2: Non-nodule 12](#_Toc170829082)

[Figure 3: Resizing 14](#_Toc170829083)

# ABSTRACT

Lung cancer is one of the leading causes of cancer-related deaths worldwide. Early detection of lung nodules can significantly improve the survival rate of patients. This thesis presents a deep learning-based approach for the classification of lung nodules using Convolutional Neural Networks (CNNs). The model is trained and validated on a dataset of CT scan images, with the goal of distinguishing between nodules and non-nodules.

The project involves several key stages: data acquisition from the LUNA (LUng Nodule Analysis) dataset on Kaggle, preprocessing of the CT scan images to ensure consistency, and applying data augmentation techniques to enhance the model's robustness. A CNN architecture is designed specifically for this task, incorporating multiple convolutional layers to effectively extract features from the images. The model is trained using the PyTorch framework and optimized to achieve high accuracy.

The proposed method demonstrates the potential of CNNs in aiding radiologists for early detection of lung cancer, potentially leading to better patient outcomes and reduced healthcare costs. The results of this project show that the CNN model can achieve high classification accuracy, making it a promising tool for automated lung nodule detection.

**Keywords:** Lung Nodules, Convolutional Neural Networks, Deep Learning, Classification, Medical Imaging

# I. INTRODUCTION

## Motivation

Lung cancer is one of the most prevalent and deadly forms of cancer worldwide, leading to significant mortality rates. Early detection and accurate classification of lung nodules, which are small masses of tissue in the lungs, are crucial for improving the survival rates of patients. Traditional methods of detecting lung nodules involve manual examination by radiologists, which is time-consuming and prone to human error. Therefore, there is a pressing need for automated systems that can assist in the early detection and classification of lung nodules.

The advent of deep learning and convolutional neural networks (CNNs) has revolutionized the field of image recognition and medical imaging. By leveraging the power of CNNs, it is possible to develop robust models that can analyze medical images and detect anomalies such as lung nodules with high accuracy. This thesis aims to explore and implement a deep learning-based approach for the classification of lung nodules using CNNs.

## 1.2. Background

Convolutional Neural Networks (CNNs) are a class of deep learning models specifically designed for processing structured grid data, such as images. CNNs have demonstrated remarkable performance in various image recognition tasks, including object detection, facial recognition, and medical imaging. In the context of lung nodule classification, CNNs can be trained to differentiate between nodules and non-nodules by learning spatial hierarchies of features from the input images.

This thesis utilizes the LUNA (LUng Nodule Analysis) dataset, a publicly available repository of CT scan images, to develop and evaluate the CNN model. The LUNA dataset provides detailed annotations of lung nodules, making it an invaluable resource for this project.

## 1.3. Technology Frameworks

This project employs several technological frameworks and libraries to build and train the CNN model. The key technologies used include:

**Python:** A versatile programming language used for data manipulation, model building, and training.

**PyTorch:** An open-source deep learning framework that provides flexibility and ease of use for building and training neural networks.

**SimpleITK:** A library for reading and processing medical images in DICOM format.

**PIL:** The Python Imaging Library, used for image processing tasks such as resizing and converting images.

**Matplotlib:** A plotting library used for visualizing images and model performance metrics.

# II. OBJECTIVE

The primary objective of this thesis is to design, train, and evaluate a CNN-based model for the classification of lung nodules. The goal is to Lung classification nodules or non-nodules in CT scan images, thereby aiding in the early detection of lung cancer.

To achieve this objective, the following specific aims are defined:

1. **Data Acquisition:** Collect and preprocess a dataset of CT scan images containing lung nodules and non-nodules from the LUNA dataset.
2. **Model Development:** Design a CNN architecture tailored for lung nodule classification.
3. **Training and Optimization:** Train the CNN model using the collected dataset and optimize its performance through hyperparameter tuning.
4. **Evaluation:** Evaluate the model's performance on a validation set using various metrics, including accuracy, precision, recall, and F1-score.

By achieving these aims, the thesis seeks to contribute to the advancement of automated lung cancer detection systems and provide a foundation for future research in this area.

# III. MATERIAL AND METHODS

## 3.1. Data Acquisition

The dataset used in this project consists of CT scan images categorized into two classes: nodules (label nodule) and non-nodules (label non-nodule). The images were sourced from the publicly available LUNA (LUng Nodule Analysis) dataset on Kaggle.

The LUNA dataset is part of the LUNA16 Challenge, which is aimed at evaluating automated lung nodule detection systems in CT scans. The dataset comprises a comprehensive collection of CT scan images with detailed annotations provided by expert radiologists. These annotations include information about the presence, location, and characteristics of lung nodules, making the dataset an invaluable resource for developing and evaluating automated detection systems.

For this project, the CT scan images were downloaded and preprocessed to extract 2D slices centered on the lung nodules. The following code snippet demonstrates the process of loading the 3D CT scan images and converting them to 2D slices:



Figure : Nodule

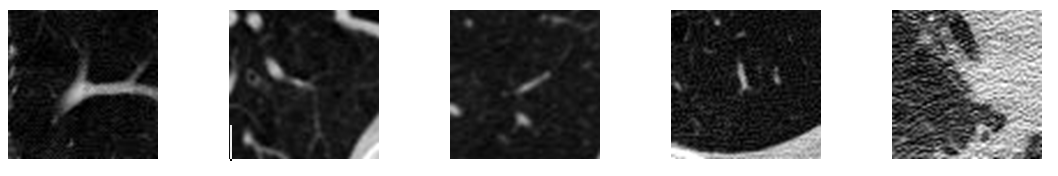


Figure : Non-nodule

## 3.2. Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for training the CNN model. The preprocessing pipeline for this project included the following steps:

### 3.2.1. Loading the 3D CT Scan Images:

The CT scan images were loaded using the load\_itk\_image function, which reads the DICOM files and extracts the image data along with metadata such as the origin and spacing.

### 3.2.2. Normalization:

The pixel values of the images were normalized to a range of [0, 1] by scaling that means converting floating-point feature values from their natural range [-1000, 400] into a standard range [0, 1]. This step helps improve the convergence of the training process.



Where:

x: The original value of the data point.

xmin: The minimum value in the datatset.

xmax: The maximum value in the datatset.

x′: The normalized value of the data point

### 3.2.3. Resizing:

All images were resized to a fixed dimension of 64x64 pixels to ensure consistency and compatibility with the CNN architecture.

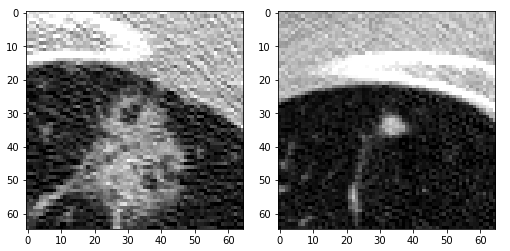


Figure 3: Resizing

### 3.2.4. Data Augmentation:

Data augmentation techniques, such as random rotations, horizontal flips, and scaling, were applied to the training images. Augmentation helps increase the diversity of the training data, thereby improving the model's robustness and generalization ability.

### 3.2.5. Data Split

Due to the large size of the dataset, the non-nodules class (non-nodule) was split into a smaller subset to balance the dataset.

## 3.3. Model

The CNN model used in this project is based on the MobileNetV3 architecture, which is known for its efficiency and effectiveness in image classification tasks.

MobileNetV3 is designed to be lightweight and suitable for deployment on mobile and edge devices, making it an ideal choice for this application.

## 3.4. Training

The training process involves the following steps:

Data Preparation: Split the dataset into training and validation sets and apply the necessary transformations.

Model Initialization: Initialize the MobileNetV3 model with the customized final layer.

Loss Function and Optimizer: Define the loss function and optimizer to be used during training.

Training Loop: Train the model for a specified number of epochs, including forward and backward passes, and optimize the model parameters.

## 3.5. Validation

# IV. RESULT AND DISCUSSIONS

## 4.1. Result

## 4.2. Discussion

# V. CONCLUSION

## 5.1. Summary

## 5.2. Future works

# REFERENCE

1. <https://www.kaggle.com/datasets/fanbyprinciple/luna-lung-cancer-dataset>
2. <https://luna16.grand-challenge.org/Tutorial/>
3. <https://www.researchgate.net/figure/MobileNetV3-Architecture-Based-On-21_fig5_357569687>