

**AUTOMATED LUNG CANCER CELL DETECTION THROUGH  
ADVANCED MACHINE LEARNING ALGORITHM BASED ON  
ANALYTICAL PREDICTION METHOD**

**BY**

**Sumayarof Mita  
ID: 182-15-11501**

This Report Presented in Partial Fulfillment of the Requirements for the  
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

**Nushrat Jahan Ria**  
Lecturer  
Department of CSE  
Daffodil International University

Co-Supervised By

**Lamia Rukhsara**  
Senior Lecturer  
Department of CSE  
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

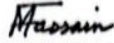
**DHAKA, BANGLADESH**

**JANUARY 20**

## APPROVAL

This Project titled "Automated lung cancer cell detection through advanced Machine learning Algorithm Based on analytical prediction method", submitted by Sumayarof Mita, ID: 182-15-11501 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 25 January 2024.

### BOARD OF EXAMINERS



**Dr. Md. Fokhray Hossain (MFH)**

**Professor**

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

**Chairman**



**Md. Sadekur Rahman (SR)**

**Assistant Professor**

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

**Internal Examiner**



**Most. Hasna Hena (HH)**

**Assistant Professor**

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

**Internal Examiner**



**Dr. S. M. Hasan Mahmud (SMH)**

**Assistant Professor**

Department of Computer Science

American International University-Bangladesh

**External Examiner**

## DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Nusrat Jahan Ria, Lecturer, Department of CSE Daffodil International University.** We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

**Supervised by:**

*On behalf*  
*Tazmin*

**Nusrat Jahan Ria**  
Lecturer  
Department of CSE  
Daffodil International University

**Co-Supervised by:**

*Lamia Rukhsara*

**Lamia Rukhsara**  
Senior Lecturer  
Department of CSE  
Daffodil International University

**Submitted by:**

*Sumayaro Mita*

**Sumayaro Mita**  
ID: 182-15-11501  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Nushrat Jahan Ria, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Sheak Rased Haider Noori**, Professor, and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

## ABSTRACT

Lung cancer is the leading cause of cancer-related deaths worldwide. According to the World Health Organization (WHO) 1.80 million people died in 2020 because of lung cancer [1]. Lung cancer remains a leading cause of cancer-related mortality worldwide, emphasizing the critical need for early and accurate detection methods. This research paper introduces a novel approach to automated lung cancer cell detection utilizing an advanced machine learning algorithm based on an analytical prediction method. We utilized a Kaggle dataset named lung-and-colon-cancer-histopathological-images, which encompasses three distinct classes, namely lung\_aca (Lung Adenocarcinoma), lung\_n (Lung normal Tissue), lung\_scc (Lung Squamous Cell Carcinoma) [2]. This dataset was categorized based on these class attributes. Technology plays a pivotal role in enhancing cancer detection methods, and numerous researchers have proposed diverse approaches in this regard. In our study, we employed five classification models, namely CNN, Xception, VGG16, ResNet-50, and Inception-v3, to identify early-stage lung cancer (LC) using the provided dataset of histopathological image. The research findings revealed that the VGG16 algorithm exhibited the highest classification accuracy, achieving 99.35% for LC detection. In comparison, ResNet-50 achieved 99.26%, CNN attained 97.72%, Xception reached 89.79% and Inception-v3 achieved 85.25. These results underscore the significance of proper system design, tuning, and the selection of machine learning methods in achieving accurate and efficient for detection lung cancer in its early stages using clinical data.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Approval	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
List of Figure	vii
List of Table	viii
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-6</b>
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	3
1.4 Research Questions	4
1.5 Expected Output	4
1.6 Report Layout	5-6
<b>CHAPTER 2: LITERATURE REVIEW</b>	<b>7-14</b>
2.1 Introduction	7
2.2 Related Works	7-11
2.3 Comparative Analysis	11-13
2.4 Scope of the Problem	13-14
2.5 Challenges	14

<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>15-32</b>
3.1 Introduction	15
3.2 Research Subject and Instrumentation	15
3.3 Data Collection Procedure	16
3.4 Statistical Analysis	16-19
3.5 Proposed Methodology	20-32
<b>CHAPTER 4: EXPERIMENTAL RESULTS &amp; DISCUSSION</b>	<b>33-39</b>
4.1 Introduction	33
4.2 Experimental Results	33-38
4.3 Descriptive Analysis	38-39
4.4 Summary	39
<b>CHAPTER 5: IMPACT ON SOCIETY &amp; SUSTAINABILITY</b>	<b>40-42</b>
5.1 Introduction	40
5.2 Impact on Society	40-41
5.3 Ethical considerations	39
5.4 Sustainability	42
<b>CHAPTER 6: FUTURE SCOPE &amp; CONCLUSION</b>	<b>43-45</b>
5.1 Summary of the Study	43-44
5.2 Implication for Further Study	44
5.3 Recommendations	44-45
5.4 Conclusions	45
<b>APPENDX</b>	<b>46</b>

**LIST OF FIGURES**

Figure 1: Dataset classes	17
Figure 2: Data visualization for lung_aca	18
Figure 3: Data visualization for lung_n	18
Figure 4: Data visualization for lung_n	18
Figure 5: Split dataset	20
Figure 6: Methodology flowchart	20
Figure 7: CNN algorithm architecture	21
Figure 8: Xception algorithm architecture	22
Figure 10: VGG16 algorithm architecture	23
Figure 11 : ResNet50 algorithm architecture	25
Figure 12: InceptionV3 algorithm architecture	26
Figure 13: Training and validation accuracy and loss for CNN	28
Figure 14: Training and validation accuracy and loss for ResNet50	28
Figure 15: Training and validation accuracy and loss for VGG16	29
Figure 16: Training and validation accuracy and loss for Xception	29
Figure 17: Training and validation accuracy and loss for InceptionV3	29
Figure 18: Models accuracy	38



## LIST OF TABLES

Table 1: Related Work table	8
Table 2: Required Equipment	15
Table 3: List of dataset	16
Table 4: Algorithm Comparison Table	27
Table 5: Confusion Matrix	30
Table 6: Confusion Matrix Score	34
Table 7: Classification Report	36

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

Lung cancer is a malignant condition characterized by uncontrolled cell growth in the tissues of the lungs. It is one of the most prevalent and lethal forms of cancer, responsible for a significant portion of cancer-related deaths globally. Lung cancer stands as the foremost cause of cancer-related fatalities globally, a sobering reality underscored by the World Health Organization's report indicating 2.21 million people are affected by lung cancer in 2020[1]. Causes for Lung cancer are smoking, secondhand smoke, air pollution, family history and so on. General symptoms for lung cancer persistent cough, shortness of breath, chest pain, unexplained weight loss, Fatigue, Coughing up blood. To detect lung cancer first suggest some test like - Chest X-rays, CT scans, and MRIs which help visualize abnormalities in the lungs. The symptoms and the results of certain tests may strongly suggest that the patient has lung cancer, but the actual diagnosis is based on examination of lung cells in the laboratory. A histopathological image is a visual representation of a thin section of tissue that has been stained to highlight specific structures or features. These images are typically obtained through biopsy or surgical resection procedures. The staining techniques help distinguish different types of cells, identify abnormalities, and aid pathologists in making diagnostic assessments. In the context of lung cancer, histopathological images of lung tissue are examined to identify cancerous cells or tumors. Pathologists analyze these images to determine the type of lung cancer, its stage, and other relevant information. Lung cancer is often classified into different subtypes (e.g., adenocarcinoma, squamous cell carcinoma) based on histopathological characteristics, which can influence treatment decisions. In our research we used histopathological images because in lung cancer detection histopathological image involve the microscopic examination of lung tissue samples to diagnose and characterize lung cancer. These images are fundamental in understanding the cellular morphology and pathology associated with the disease.

Recurrent respiratory infections Early and accurate detection methods are imperative to mitigate the impact of this pervasive health concern. This research endeavors to address this critical need through the introduction of an innovative approach—automated lung cancer cell detection. Our methodology harnesses the power of advanced machine learning algorithms grounded in an analytical prediction method. By amalgamating cutting-edge image processing techniques with a sophisticated machine learning model, we aim to significantly enhance the efficiency and accuracy of lung cancer cell identification.

## **1.2 Motivation**

The alarming statistics of lung cancer-related deaths prompt a compelling motivation for this research. Lung cancer is particularly deadly when diagnosed in later stages. Early detection significantly improves survival rates, and automated systems can analyze medical images much faster and more consistently than humans, potentially identifying suspicious cells at earlier stages. Though society's control and spreading awareness about LC may help in the early detection of cancer, the most common and efficient way to identify the presence of a cancerous tumor is by tissue biopsy [11]. Human analysis of medical histopathological images can be subjective, time consuming and prone to error. Cell size and shape alone can't accurately identify tumors when the cells come in many different shapes and sizes. Low-quality images make it hard to tell cancerous cells from healthy ones even if you look at their size and shape [14]. Due to this, employing diagnostic assistance tools in such analyses proves beneficial, as it minimizes the need for pathologists' involvement. This, in turn, shortens the time required for each patient's examination and releases pathologists from the task of scrutinizing numerous straightforward samples. Machine learning algorithms trained on large datasets can achieve higher accuracy and consistency in identifying cancerous cells, potentially reducing misdiagnoses. Analyzing medical images for lung cancer is a time-consuming and laborious task for radiologists. Automation can free up their time for more complex tasks and consultations, improving overall efficiency and patient care. Every year millions of people died because of misdiagnosis and mistreatment. If it is possible to make easy to diagnose and detect lung cancer it can help to reduce the

death rate. The urgent necessity to improve early detection methods is evident, and the potential of automated systems powered by advanced machine learning algorithms presents a promising avenue for significant advancements in this field. The motivation to contribute to the reduction of mortality rates associated with lung cancer through technological innovation propels this study.

### **1.3 Rationale of the study**

The rationale behind this research is rooted in the critical gap that exists in the current landscape of lung cancer detection. Smoking and air pollution are the primary reason of affected by lung cancer, family history and others also play a significant role to affected our lung. In this current era, because of use of different fuels, industrial process, waste burning, deforestations are common, so air pollution rate is become higher. Because of that's reason lung cancer rate in non-smoker range also increases. People affected by different type of lung disease; among them lung cancer is crucial. People are not aware that much about lung cancer; besides awareness we need to make a system which help to test their lung to find out is it affected by cancer or not. So, it's very essential to find a way to detect lung cancer to save peoples life. Traditional methods are often constrained by limitations in accuracy and speed. There is an urgent requirement for computer-assisted diagnostic (CAD) tools that can perform precise and effective quantitative analysis of histopathology images [13]. This study aims to fill this gap by introducing an automated system that leverages advanced machine learning algorithms, offering a more precise and timely approach to lung cancer cell identification. By implementing this work, Patients don't need to wait 2 or 3 days to get their pathological test, within 1-2 hours will be enough to detect the cancerous cell. The rationale extends to the potential impact on patient outcomes through early-stage detection and intervention. By overcoming the challenges and ensuring responsible development and implementation, this study has the potential to save lives and improve the lives of many patients facing this critical disease. This study yearns to make a substantial contribution to the field of medical research by bridging the gap between technological advancements and clinical diagnosis needs. The development of an automated lung cancer detection system has the potential not only to refine diagnostic

practices but also to serve as a benchmark for future research in the intersection of artificial intelligence and healthcare.

#### **1.4 Research Questions**

- What is the effectiveness of the proposed automated lung cancer cell detection system in comparison to traditional methods?
- How do different machine learning algorithms contribute to the accuracy and efficiency of early-stage lung cancer detection?
- What specific type of advanced machine learning algorithm and analytical prediction method are most effective for automated lung cancer cell detection?
- How does the performance of the proposed system compare to traditional human-based lung cancer diagnosis approaches?
- What are the clinical implications of using this system, such as potential improvements in patient outcomes and healthcare costs?
- What are the ethical considerations and challenges associated with implementing this system in clinical practice, and how can they be addressed?
- Can the system be trained and adapted to perform personalized risk assessments and treatment planning for lung cancer patients?
- What are the potential limitations and uncertainties associated with this technology, and how can they be mitigated?

#### **1.5 Expected output**

The paramount goal of this research is to yield a meticulously validated and optimized automated system for the detection of lung cancer cells. The expected output concentrate on a significant enhancement in the accuracy of lung cancer detection. Exploring sophisticated machine learning models, our research goals to outshine existing benchmarks, providing a more reliable and precise tool for Computer aided diagnosis (CAD). This heightened accuracy is anticipated to streamline the diagnostic process, reducing false negatives and false positives. This anticipated output holds

significant promise in contributing to advancements in early-stage lung cancer identification. Beyond the development of a highly accurate automated system for discerning lung cancer cells from histopathological images, our study aims to uncover invaluable insights into the nuanced performance variations exhibited by distinct machine learning models. This exploration is poised to play a pivotal role in refining and selecting optimal algorithms, thereby elevating the efficacy of early-stage lung cancer detection. The research output, thus, encompasses not only the tangible development of a sophisticated automated system but also the intangible yet crucial understanding of how different machine learning models can be harnessed to enhance detection performance. The output of this study may help the people detect their disease at early stage, reduce the workload of a pathologist, and cost will be lessening. Besides practical applications, our research exertions to contribute to the academic knowledge based on the fields of medical imaging, machine learning, and cancer diagnostics. The methodologies, insights, and comparative analyses presented in this study are expected to serve as valuable references for researchers, paving the way for further advancements in the intersection of technology and healthcare

## **1.6 Report layout**

This research provides 6 chapters to mark up its efficiency for readers and other researchers.

**In Chapter 1:** Discuss about the Lung cancer detection necessity. Motivation behind this research, Rationale of the Study. Added some Research Questions, explained about expected Output and details of report Layout.

**In Chapter 2:** Literature Review sessions discussed about the related work of other researchers and reviews. How others ideas to detect LC and about their used method and result analysis. This chapter also provide a comparative analysis which shows the difference of this research paper to others and how effective or efficient it is to detect lung cancer. Scope of the Problem, Challenges are also discussed.

**In Chapter 3:** Research Methodology parts discussed about the used methodology, work procedures and analysis. Which instruments needed for this research, data collection and process, statistical analysis and model implementation requirements are explained in this chapter. About the used algorithm are also discussed. A description analysis given in this chapter about all procedures and processes were follow in this research paper.

**In Chapter 4:** Provides results of the research. Result described in analytical and statistical way. Try to give enough idea about the effectiveness of this research. Table and diagram are used so that reader or other researchers can understand the outcome of this study. Accuracy rate of used algorithm, some others results which helps to utilize mode ls efficiency, analytical results description also discussed.

**In Chapter 5:** Discussed about the impact on the society. How this research help to detect lung cancer and what will be happen after implementation this thought. And sustainability of this research, Future work, knowledge gain, implementation thoughts are discussed. How its impact on our life, ethical consideration, and how sustainable the work will be talked.

**In Chapter 6:** provide the Summary parts which includes Conclusion, Recommendation and Implication for Future Research, talks about Summary of the Study, Conclusions, Recommendations Implication for Further Study.

## **CHAPTER 2**

### **LITERATURE REVIEW**

#### **2.1 Introduction**

Here we talk about the difficult topic of lung cancer, which has been a problem for a long time. Lung cancer has long been a crucial issue in the present decade. In human history, the development of the disease has been characterized by a significant degree of suffering, in addition to a significant number of deaths. Because lung cancer is considered to be a potential threat to human life. That's why a significant amount of research has been carried out in order to reduce the cruelty of the mischievous consequences of this disease. This chapter discuss about an overview of the previous work and background knowledge that are necessity for the evaluation. Here we are going to review some of the important pieces of literature that have been written on this same topic. We talked about their technique and successive rate so that we can compare our work with them. In summary, a comparative research study about lung cancer detection undertaken to assess the impact of our present work, ensuring due recognition is accorded to effort.

#### **2.2 Related Works**

As it's a clinical need to identify cancer cell in appropriate time and start treatment to reduce ta death rate, so many researchers and publishers work on it. Different type of methodology and data used to find best approach to detect LC. Some talk about the crucial part of lung cancer, risk factor, symptoms and necessity of early detection. There has been huge amount of deep learning methods that have been developed and proposed for histopathological image classification [26].

Here some related works are given in below:



Table 1: Related Work table

No	Author	Title	year	Methods are used	Finding
01	Sundaresh Ram, Wenfei Tang, Alexander J. Bell, Ravi Pal, Cara Spencer, Alexander Buschhaus, Charles R. Hatt, Marina Pasca diMagliano, Alnawaz Rehemtulla, Jeffrey J. Rodríguez, Stefanie Galban, Craig J. Galban. [3]	Lung cancer lesion detection in histopathology images using graph-based sparse PCA network	2023	GS-PCANet	Detection accuracy 0.908.
02	Jie Ji, Weifeng Zhang, Yuejiao Dong et al. [4]	Automated Lung Cancer Detection using Histopathological Images	2023	U-Net, U-net++, ResNet34, DenseNet-121	Used 5X magnification and 512×512 patches acquired 0.934 accuracy.

03	Faria, Nelson, Sofia Campelos, and Vítor Carvalho. [5]	A Novel Convolutional Neural Network Algorithm for Histopathological Lung Cancer Detection	2023	CancerDetectNN V5	Archived precision-0.972, an area under the curve (AUC) of 0.923, and F1-score 0.897.
04	Radical, Rakhman, Wahid., Chilyatun, Nisa., Rahayu, Prabawati, Amaliyah., Eva, Yulia, Puspaningrum [6]	Lung and colon cancer detection with convolutional neural networks on histopathological images	2023	ShuffleNetV2, GoogLeNet, ResNet18	ResNet18 obtained highest accuracy for lung cancer 98.82%
05	Basra Jehangir, Soumya Ranjan Nayak, Sourav Shandilya [7]	Lung Cancer Detection using Ensemble of Machine Learning Models	2022	Hybridized model of CNN, SVC, Random Forest, XG Boost	Obtained overall accuracy 99.13%
06	Aayush Rajput, Abdulhamit Subasi [8]	Lung cancer detection from histopathological lung tissue images using deep learning	2023	ResNet combined with SVM	Achieved accuracy 98.57%

07	Manop Phankokkruad [9]	Ensemble Transfer Learning for Lung Cancer Detection	2021	VGG16, ResNet50 V2, DenseNet 201	Validation accuracy for ensemble model is 91%
8	Nicolas Coudray, Paolo Ocampo, Theodore Sakellaropoulos, Andre Moreira, David Fenyő, Narges Razavian, Aristotelis Tsirigos 10]	Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning	2018	Inception V3	Obtained 0.97 AUC for LC classification
9	Javier, Civit-Masot., Alejandro, Bañuls-Beaterio., Manuel, Domínguez-Morales., M., Rivas-Perez., Luis, Muñoz-Saavedra., José, María,	Non-small cell lung cancer diagnosis aid with histopathological images using Explainable Deep Learning techniques	2022	Color CNN, Greyscale CNN	System accuracy between 97.11% and 99.96% based on classes

	Rodríguez, Corral [11]				
10	Prabira Kumar Sethy, A. Geetha Devi, Bikash Padhan, Santi Kumari Behera, Surampudi Sreedhar, Kalyan Das [12]	Lung cancer histopathological image classification using wavelets and AlexNet	2022	AlexNet, wavelet, SVM	10-fold cross validation method achieved 99.3% accuracy

### 2.3 Comparative Analysis

The field of tumor detection and classification within commonly available microscopy software has seen extensive development in feature extraction techniques. These techniques encompass analyses of size, shape, and morphological features [15], texture features such as local binary patterns (LBP) [16], local Fourier transforms [17], co-occurrence matrices, and fractal features [18], as well as energy minimization and optimization approaches [19]. Over-generalization makes these methods too clunky for real-world scenarios. We need approaches that can better adapt to diverse datasets and handle complex images. Convolutional Neural Network (CNN) was used to detect lung cancer in my study. I used a total of five models (CNN, Xception, VGG16, ResNet50, InceptionV3) in my study. Each of them showed a comprehensive level of accuracy. Convolutional Neural Networks (CNNs) have emerged as a transformative force in image-based tasks, particularly in the domain of medical imaging.

This study proposes a graph-based sparse PCA network for detecting lung cancer lesions in histopathological images. The emphasis is on leveraging graph structures and sparse principal component analysis for improved accuracy in lesion identification [3].

The ensemble model using 5X magnification and 512X512 patches and achieved an accuracy of 0.934, sensitivity of 0.877, specificity of 0.948, and dice similarity coefficient of 0.840 on the test dataset. The performances of using the 5X magnification outperformed those of using the 10X and 20X magnifications. The study explores advanced image analysis techniques to enhance the efficiency of detection, showcasing the potential for automated diagnostic systems [4].

Radical proposed a model where ShuffleNet, GoogleNet, ResNet18 used as a base model. These models trained on the LC25000 dataset contains 25,000 histopathological color image samples of colon and lung tissues. ResNet18 achieved the highest accuracy of 98.82% for classifying lung cancer. ResNet18 had the shortest training time of 1749.5 seconds for lung cancer classification [6]. Used hybrid model of CNN, Random Forest, XcBoost obtained accuracy 99.13% which may a good validation model for lung cancer detection [7]. Combined model of VGG16, ResNet, SVM scores 98.57% accuracy [9]. Sethy et al. employ a combination of wavelet transformations and the AlexNet architecture for histopathological image classification in lung cancer. This study achieved 99.3% accuracy. The study explores the fusion of wavelet features with deep learning for improved accuracy [12].

The research paper presents a computer-aided diagnosis (CAD) system for lung cancer classification in computed tomography (CT) scans. The system utilizes a 3D Convolutional Neural Network (CNN) for nodule detection and classification, achieving a test set accuracy of 86.6%. The authors discuss the significance of deep learning in medical imaging, particularly the application of CNNs in pattern recognition and machine learning. They also highlight the promising results of deep learning in medical imaging, citing previous research in the field. The paper provides a detailed description of the CAD system's architecture, training process, and evaluation metrics. Additionally, it discusses the potential for further improvement and the importance of addressing challenges in lung cancer detection and classification [27]

This comparative analysis seeks to unravel a compelling narrative where CNN algorithm demonstrates an unprecedented level of accuracy in lung cancer detection. My CNN model which is not pretrained an accuracy of 97% where 72% was the validation accuracy. The exception model was used, which is a pretrained model that showed an overall accuracy of more than 89%. VGG16 and ResNet50 were also used. These two are both pretrained models. They showed the accuracy of more than 99% which is quite astounding. Inception V3 is another pretrained model that was used and came up with accuracy of 85%. These results were commendable given the large dataset I used. Compared to other studies my dataset was bigger and the accuracy was overall better or similar.

## **2.4 Scope of the problem**

Lung cancer, a formidable global health challenge, demands innovative approaches to detection for timely intervention and improved patient outcomes. Traditional diagnostic methods often fall short in terms of efficiency and accuracy, necessitating a paradigm shift toward automated image analysis. In 2020, there were an estimated 12,999 new cases of lung cancer in Bangladesh, accounting for 8.3% of all new cancer cases [20]. A 2017 study reported a near 200% increase in lung cancer cases at the National Institute of Cancer Research and Hospital (NICRH) between 2014 and 2017 [21]. This rate increases day by day. Due to this reason its very important to make an evolution in our pathological medical sector by using computer-aided diagnostic tools.

ML algorithms, inspired by the human visual system, have proven to be adept at discerning patterns and features within medical images, holding great promise in the early detection of lung cancer. The problems that might arise are as follows; Lung nodules can have a diverse range of appearances, making accurate detection difficult. Factors like size, shape, texture, and surrounding tissue characteristics can lead to misidentification. Acquiring large, high-quality datasets of annotated lung cancer images is essential for effective training of model. However, privacy concerns and resource limitations often hinder data collection. While CNNs excel at prediction, understanding their internal decision-making process is crucial for building trust and refining models. Lack of interpretability makes it difficult to pinpoint potential biases

or flaws in the model. Seamless integration of AI-powered detection tools into existing clinical workflows is necessary for smooth adoption and practical use by healthcare professionals. Training complex CNNs requires significant computational resources, which can be prohibitive for smaller institutions or resource-constrained settings.

## **2.5 Challenges**

Lung cancer detection poses a unique set of challenges primarily due to the intricate nature of medical imaging associated with the respiratory system. The diverse range of lesions, subtle anomalies, and overlapping structures within the lungs demand a level of sophistication in image analysis that often stretches the limits of conventional diagnostic methods. One of the primary challenges lies on dataset itself. The Kaggle dataset, while diverse, may have inherent limitations, including size, representativeness, and potential biases. The effectiveness of the developed system is contingent on the dataset's ability to encapsulate the diversity of lung cancer histopathological images encountered in real-world clinical settings.

The adoption of computer aided diagnosis in clinical settings introduces a critical dimension related to the interpretability of AI-generated results. As these algorithms operate as complex black boxes, elucidating the decision-making process becomes imperative for gaining the trust of healthcare professionals and ensuring seamless integration into existing diagnostic workflows. Bridging the gap between the powerful predictive capabilities of ML algorithms and the need for transparent, interpretable results remains an ongoing challenge that requires innovative solutions. While these models exhibit significant promise, there exists a notable variation in the performance of different models across diverse datasets and patient populations. Factors such as algorithmic biases, dataset imbalances, and variations in imaging techniques contribute to disparities in performance. Ensuring the robustness and reliability of research across varied scenarios demands a comprehensive understanding of these factors, paving the way for the development of more universally applicable models.

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Introduction

This chapter outlines the methodology employed to conduct the research on automated lung cancer cell detection through advanced machine learning algorithms based on an analytical prediction method. The research methodology encompasses data collection, preprocessing, model development and evaluation. This chapter emphasize the summary of data collection, processing, model applying on in this research.

#### 3.2 Research subject and Instrumentation

At the heart of our investigation lies the research subject—automated lung cancer cell detection. With lung cancer being a leading cause of global mortality, our focus is to harness the potential of cutting-edge technology to revolutionize early detection. The research subject encapsulates the utilization of state-of-the-art machine learning algorithms, emphasizing precision and efficiency in identifying malignant cells within histopathological images of the lung. The chosen instrumentation forms the base of our empirical exploration. In this research, we employ a robust set of instruments to navigate the complexities of automated lung cancer cell detection, data collection and data processing,

Here the list of equipment needed for the Model:

Table 2: Required Equipment

Hardware and software	Development Tools
11th Gen Intel(R) Core (TM) i3-1115G4 @ 3.00GHz 3.00 GHz	Windows 11
1 TB HDD	Python 3.10.11
Google Colab	Pandas
RAM 8GB	TransorFlow Backend Engine

#### 3.3 Data collection procedure



Dataset collection for Lung cancer detection is very complex, as we used histopathological image which can be only collected from medical diagnosis lab. The Kaggle dataset, named lung-and-colon-cancer-histopathological-images [2], serves as our primary source of data. The total number of data 15000 divided into three classes.

### 3.4 Statistical Analysis

#### 3.4.1 Data

We used a dataset from Kaggle. It contains 3 type of data. Every class includes 5000 of data [2].

Table 3: List of dataset

<b>Class Name</b>	<b>No. of Data</b>	<b>Type of Data</b>
lung_n	5000	Histopathological images
lung_aca	5000	Histopathological images
lung_scc	5000	Histopathological images

#### **lung\_n**

Belongs to normal class. This type of tissue represents healthy, non-cancerous cells in the lung tissue. These cells exhibit normal growth, morphology, and function

#### **lung\_aca**

Belongs to Adenocarcinoma type. This represent a type of non-small cell lung cancer (NSCLC) that develops in the glandular cells of the lungs. These cells are responsible for producing mucus and other fluids. Reason behind this type of cancer are smoke or who used to smoke, but it's also the common type of lung cancer seen in who don't

smoke. Women and young people are more affected by this type of lung cancer. Adenocarcinoma is usually found in the outer parts of the lung and is more likely to be found before it has spread [22].

#### **lung\_scc:**

Belongs to Squamous Cell Carcinoma type. Lung cancer arising from squamous cells, the flat lining within the airways, exhibits a strong association with smoking and a preferential occurrence near central bronchi [22].

### **3.4.2 Dataset Preprocessing**

Here we don't need to use any pre-processing technique as data augmentation has already been extensively applied to generate these class-specific image sets, utilizing a base of 250 original images. Therefore, further augmentation on these datasets would be redundant and unnecessary.[2]

#### **Data Augmentation**

The process of increasing the amount and diversity of data is called data augmentation. In this technique we do not need to collect new data, rather we transform the present data. The most commonly used operations are-

- Rotation
- Shearing
- Zooming
- Cropping
- Flipping
- Changing the brightness level [23]

### **3.4.3 Data visualization**

We have 3 classes of data in our dataset. Here in this part we just view the classes of our dataset and some images for every classes

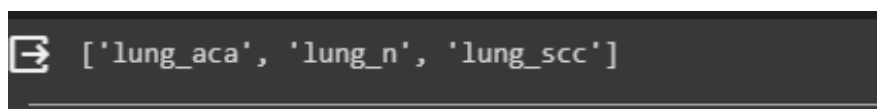


Figure 1: Dataset classes

Images for lung\_aca category . . . .

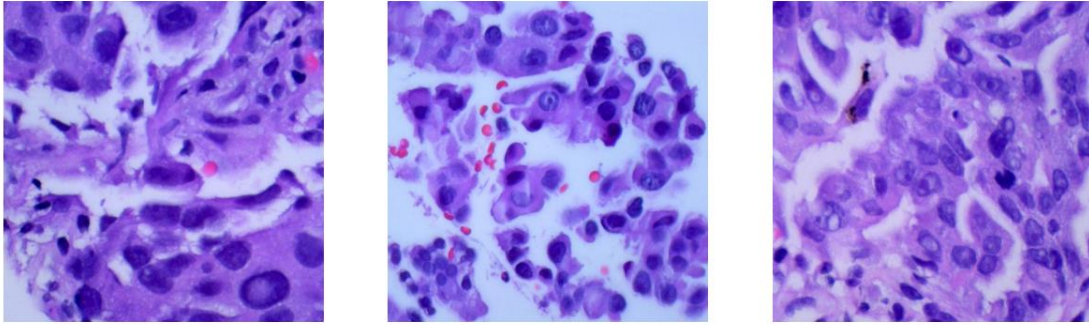


Figure 2: Data visualization for lung\_aca

Images for lung\_n category . . . .

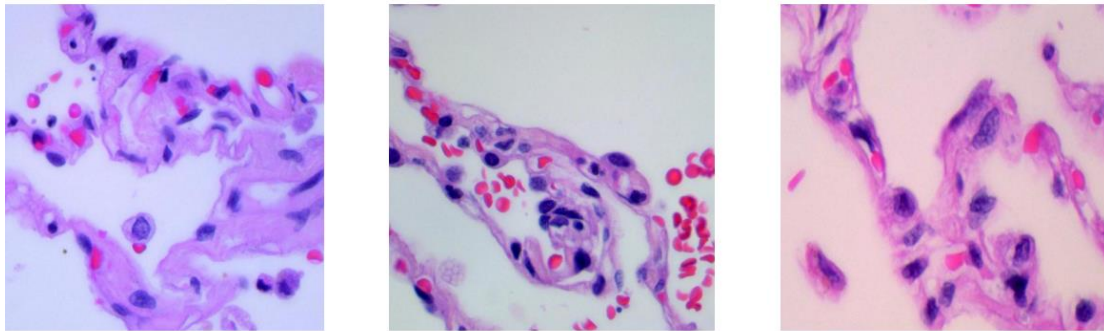


Figure 3: Data visualization for lung\_n

Images for lung\_scc category . . . .

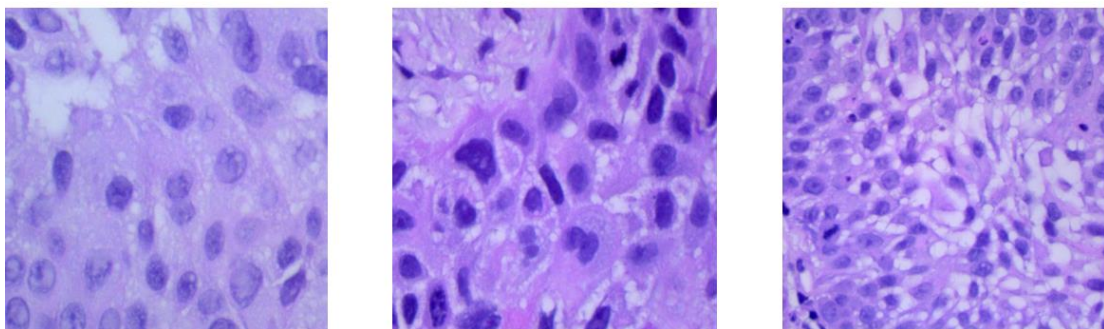


Figure 4: Data visualization for lung\_n

### 3.4.4 Data Preparation

#### 3.4.4.1 Image Resizing

It is needed to resize our image to train our model. we converted given images through NumPy arrays. Resizing images because training a Deep Neural Network on large-size images is not efficient cause it requires computational cost and time. That's why, we use the OpenCV library and NumPy library of python to resize our image Most of the ML model needs specifics type of image size. Smaller size help to increase the speed of model. In my paper we used 224 size of image, so that model requirements and training speed balanced.

#### 3.4.4.2 Data split

Image splitting is a essential preprocessing step of automated lung cancer cell detection. This technique involves with dividing large histopathological images into smaller and more manageable sections. This technique enhances computational efficiency, smooth the analysis process, and reduce time and cost for training models. We split our dataset into training and validation data so, that we can evaluate the performance of our model. We split our dataset as 0.2, that means 80% of data for train the model and 20% for test.

Here the split Dataset summary

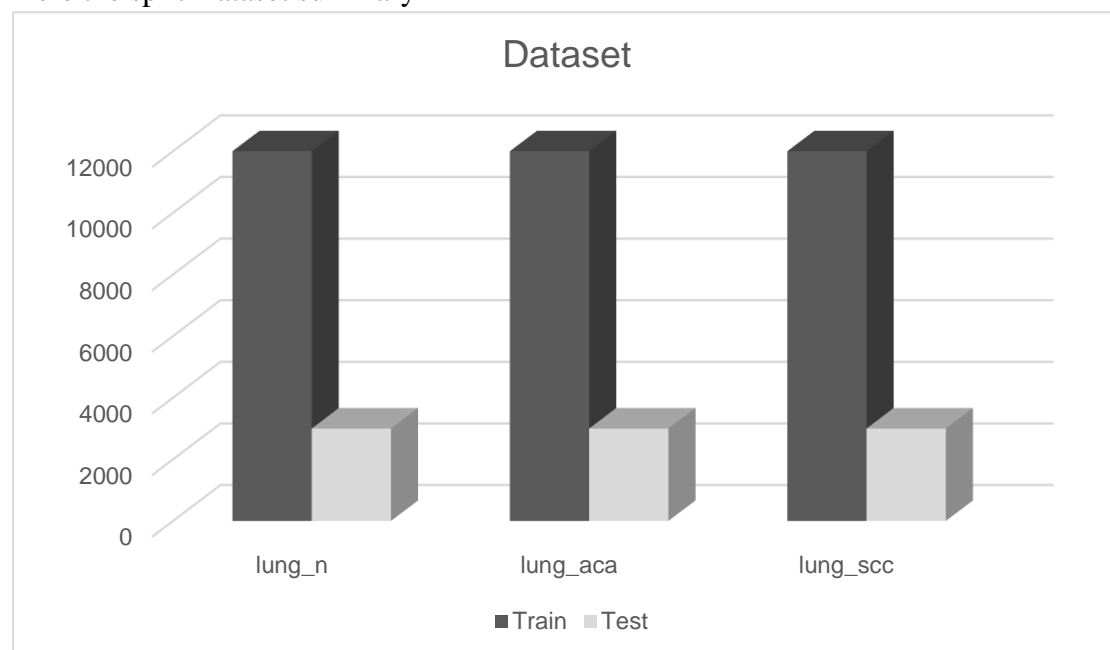


Figure 5: Split dataset

### 3.5 Proposed Methodology

Here given the flow chart for entire work methodology:

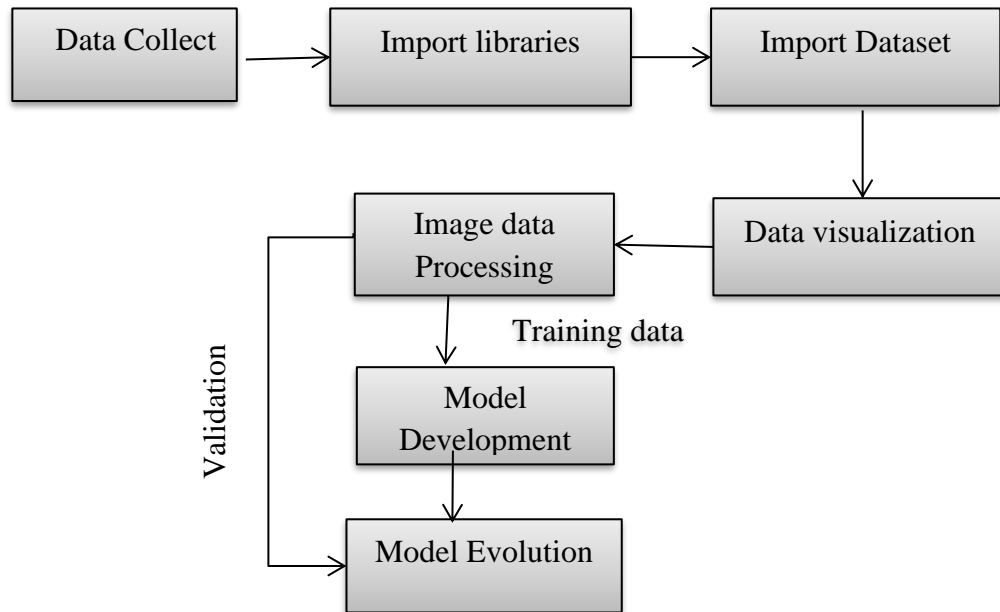


Figure 6: Methodology flowchart

#### 3.5.1 Model Development

##### 3.5.1.1 Algorithms

- CNN

Convolutional Neural Networks (CNNs), a specialized form of deep learning algorithms tailored for image processing and recognition tasks. Unlike alternative models, CNNs necessitate minimal preprocessing, autonomously acquiring hierarchical feature representations from raw input images. CNNs excel at discerning important objects and features within images through convolutional layers, which employ filters to identify local patterns. Inspired by the visual cortex, their connectivity pattern allows

CNNs to adeptly capture spatial relationships and patterns. The layer stacking process, incorporating multiple convolutional and pooling layers, empowers CNNs to learn intricate features, thereby achieving high accuracy in image classification, object detection, and segmentation type of tasks.

Architecture diagram of CNN models are in below

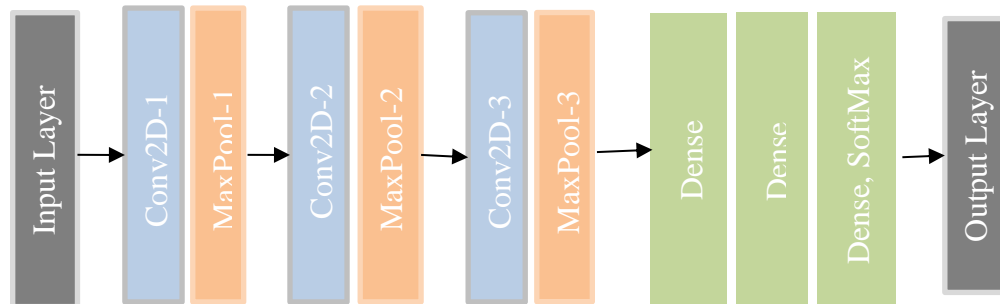


Figure 7: CNN algorithm architecture

From figure 7, we constructed a sequential model with use convolutional layer sequence and each followed by a max pooling layer in our CNN algorithm. Then added a flatten so that the output of the convolutional layer would be flatten. Next two fully connected layer was added followed by final output layers which has 3 neurons for three classes. Here we also included BatchNormalization layer and a dropout layer to make training fast, stable and avoid overfitting possibilities. The final layer was output layer which included softmax activation for three different classes.

- **Xception**

The Xception model represents a significant advancement in Convolutional Neural Network (CNN) architecture, designed to address limitations associated with traditional convolutional layers. Introduced as a part of the Google Net architecture, the exception model employs a unique approach by utilizing multiple convolutional filter sizes within a single layer. Unlike conventional CNNs that rely on fixed-size filters, the Xception model strategically integrates filters of varying sizes, allowing the network to capture

features at different scales. This enables the model to extract intricate details from both fine and coarse structures within the input data, enhancing its ability to recognize complex patterns and objects. The exception model's innovative design has inspired subsequent CNN architectures and continues to serve as a foundational concept in the evolution of deep learning models. Its success lies in the balance it strikes between computational efficiency, feature diversity, and effective training strategies, making it a pivotal milestone in the development of advanced convolutional neural networks.

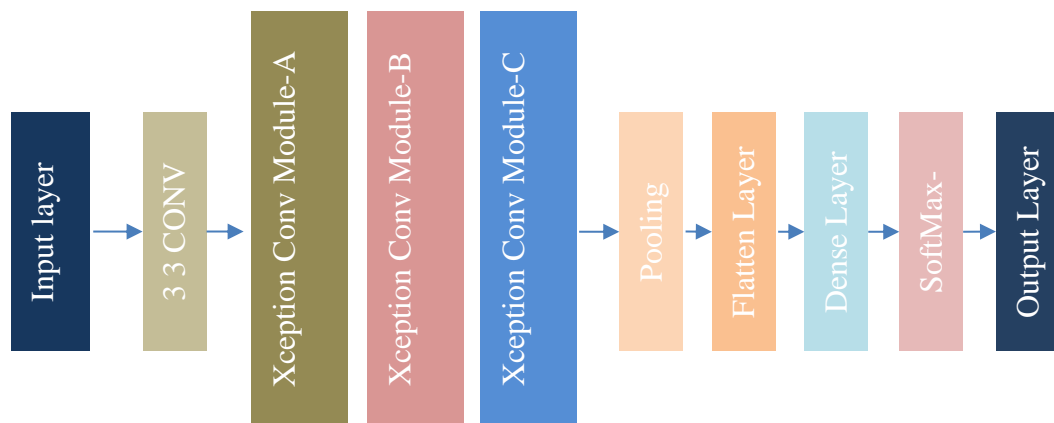


Figure 8: Xception algorithm architecture

Figure 8 shows the architecture of Xception model. In our model we used pre-trained Xception model which has excellent image classification performance. Keras used to import the model. Then we set the dimensions of images at  $224 \times 224$  pixels. To reduce the computational cost and training time, we froze most of the layer of pretrained model so that previous classification knowledge was preserved. Then added a set of top layers to adapted pre-trained model for a new task. Global average pooling also added

to reduce computational cost and training time. Three dense layers with relu activation was added and lastly a softmax output layer was added for classification task.

- **VGG16**

The VGG16 model, a pioneering architecture in Convolutional Neural Networks (CNNs), was introduced by the Visual Geometry Group (VGG) at the University of Oxford. Renowned for its simplicity and effectiveness, VGG16 gained widespread recognition and became a benchmark in image classification tasks. The model builds of 16 weight layers, including 13 convolutional layers and 3 fully connected layers (figure-9). The convolutional layers predominantly employ small 3x3 filters with a stride of 1, emphasizing depth and enabling the network to learn intricate hierarchical features. The use of multiple 3x3 filters in lieu of larger filters with the same receptive field allows for a deeper representation of non-linear transformations. VGG16's straightforward architecture, with its stacked convolutional layers and pooling operations, facilitates feature extraction at different scales. The model's simplicity has contributed to its versatility, making it applicable to various computer vision tasks beyond image classification. However, its depth comes at the cost of increased parameter count and computational resources. Despite subsequent advancements in CNN architectures, VGG16's impact endures, serving as a foundational reference for understanding deep neural networks and inspiring further developments in the quest for more efficient and powerful models.

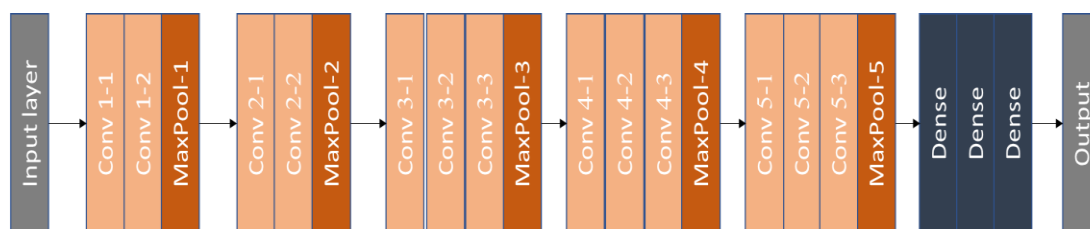


Figure 9: VGG16 algorithm architecture



VGG16 algorithm architecture are shown in figure 9. In our research we used a pre-trained model which known for its proficiency in image classification task. First we import our pretrained model by using Keras, then we specify the images dimension for the critic of examine. We customize our model by removing top layers. For pre-serving of previous knowledge of our model we set all the pretrained layers as not trainable. By following transfer learning scenario, added some task specific top layers followed global average pooling used to reduce time and cost. Three dense layers with ReLU activation added to empower model. Finally a output layer with softmax activation used to grants the model the ability of classification of three classes.

- **ResNet50**

ResNet50, an integral member of the ResNet (Residual Network) family, stands out in the realm of Convolutional Neural Networks (CNNs) for its groundbreaking architecture. Introduced by Microsoft Research, ResNet50 addresses the challenge of training very deep networks by incorporating residual learning blocks. These blocks enable the model to skip connections, allowing the direct flow of information across layers. The introduction of residual connections helps alleviate the vanishing gradient problem, facilitating the training of exceedingly deep networks with 50 weight layers. This depth allows ResNet50 to capture intricate hierarchical features, enhancing its performance in various computer vision tasks, including image classification and object detection. The ResNet architecture, with its profound impact on the stability and efficiency of training deep neural networks, has become a cornerstone in the development of advanced CNN models. ResNet50 has become a widely adopted and influential architecture in the field of deep learning.

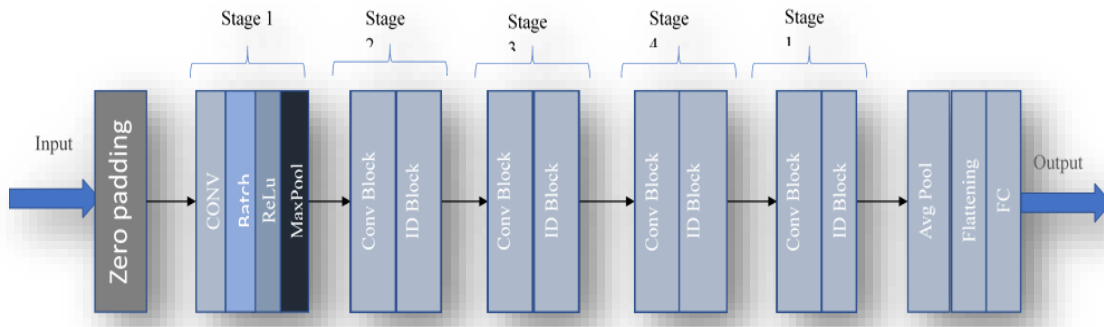


Figure 10 : ResNet50 algorithm architecture

Here we used pretrained ResNet50 as a base model. By Keras application we called the pre-trained model which is expert for image understanding and classifying. Then removed the original top layer for reserved the space for customized based on specific task. Declared the size of images and number of classes. For preserving the pre-trained understanding and visual insight we froze most of the layer. Then a flatten layer added for preprocessing data for further process in dense layer. All connected layers a dense layer with activating Relu are used. Lastly a softmax output layer was added as a final classification layer.

- **InceptionV3**

Inception V3, an evolution of the Inception model, represents a significant milestone in Convolutional Neural Network (CNN) architecture. Developed by the Google Research Brain Team, Inception V3 focuses on achieving a balance between computational efficiency and high performance in image recognition tasks. The model incorporates innovative features such as factorized convolutions, batch normalization, and aggressive regularization techniques. Factorized convolutions, involving the use of smaller filters, contribute to a more efficient use of computational resources, enhancing the model's speed without compromising accuracy. Batch normalization aids in stabilizing and accelerating the training process. Inception V3 also introduces the concept of auxiliary classifiers, enhancing gradient flow during training. The architecture's inception modules utilize various filter sizes to capture features at

multiple scales, enabling effective extraction of complex visual patterns. Inception V3's versatility and robustness make it a popular choice for image classification, object detection, and transfer learning applications, reflecting its pivotal role in the evolution of CNNs.

InceptionV3 model start with multiple ( $1 \times 1$ ,  $3 \times 3$ ,  $5 \times 5$ ) convolutional layer and pooling layers. Factorization help to reduce the cost without sacrificing the performance of the model. This model has 31 convolutional layer which used for feature extraction of images using filters. Besides this 5 pooling and 2 fully connected layer are used in this model. As well as 2 immediate classifiers used to improved convergence and generalization on model.

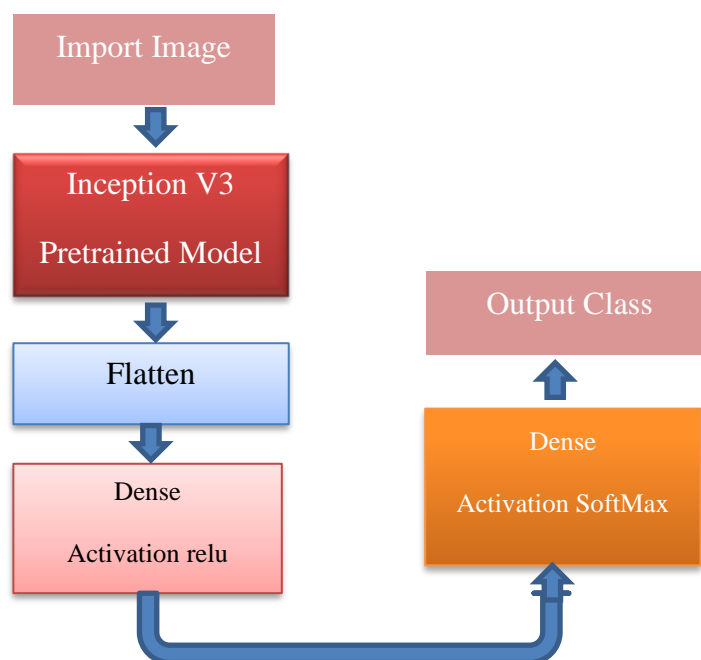


Figure 11: InceptionV3 algorithm architecture

In figure 11 explained the structure of used InceptionV3 model. As how we used transfer learning method to train our model , here Keras application used to call the pre-trained inceptionv3 model. Customized model by removed pre-trained top layer and made no trained layers so that the previous knowledge of model preserved for further

use. Then added Customized top layer as the task specific. Added flatten layer then a dense layer with 512 neurons and relu activation. And output layer was a softmax activation layer

### Algorithm comparison

Here given all the algorithms structural comparisons for my research:

Table 4: Algorithm Comparison Table

Algorithm	Avg Accuracy	Total params	Trainable params	Non-Trainable params
CNN	0.90	25819971	25819203	768
ResNet50	0.99	24638339	3674115	23587712
VGG16	0.98	16815939	2101251	14714688
InceptionV3	0.85	22853411	1050627	21802784090
Xception	0.85	24535595	3674115	20861480

### 3.5.2 Model Training

In this research, all Machine Learning Algorithm models- CNN, Xception, VGG16, ResNet50 and InceptionV3 trained in Google Colab platform using GPU and TPU hardware acceleration. All the dataset's image resized in 224×224 pixels. Deep learning Library like TensorFlow and Keras API used for development and implementation of algorithm models. Categorical cross-entropy function used to train models which help to measuring the performance of the model. The definition of categorical cross-entropy loss function is

$$\text{Categorical cross-entropy loss} = \sum_{c=1}^N y_{i,c} \log(p_{i,c}) \dots\dots\dots (1)$$

Where,  $N = \text{Classes}$

$y_{(i,c)}$  = Ground truth for each image

$p_{(i,c)}$  = Prediction Probability for each image

Here Mention Loss and Accuracy graph for every model used in this research:

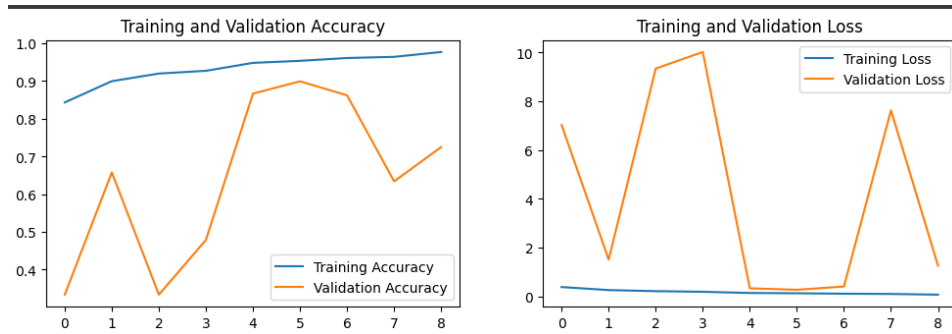


Figure 12: Training and validation accuracy and loss for CNN

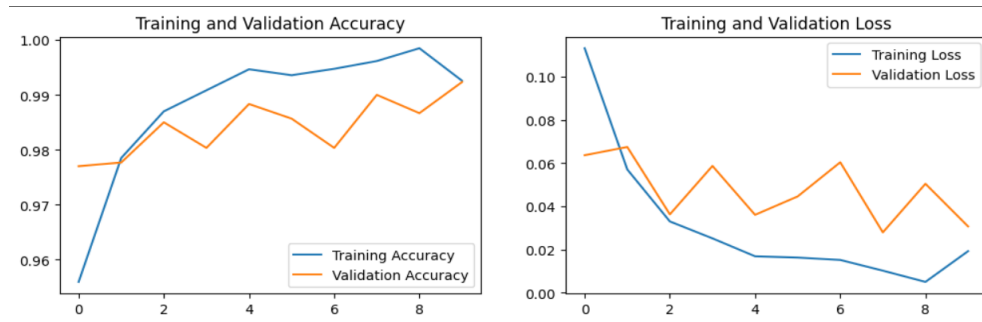


Figure 13: Training and validation accuracy and loss for ResNet50

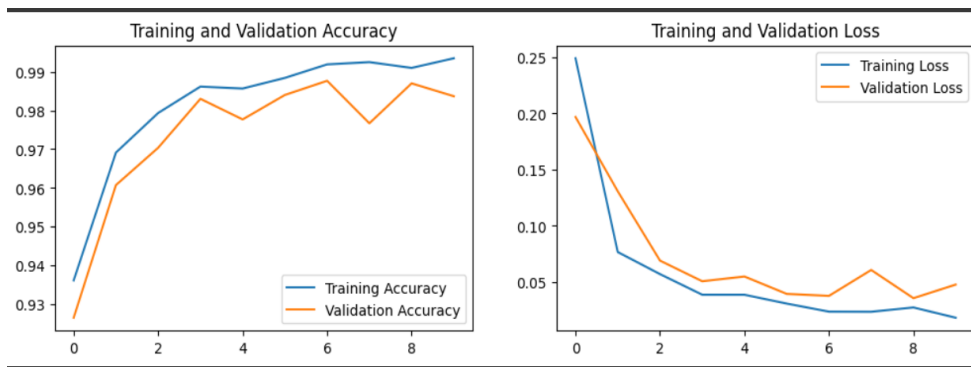


Figure 14: Training and validation accuracy and loss for VGG16

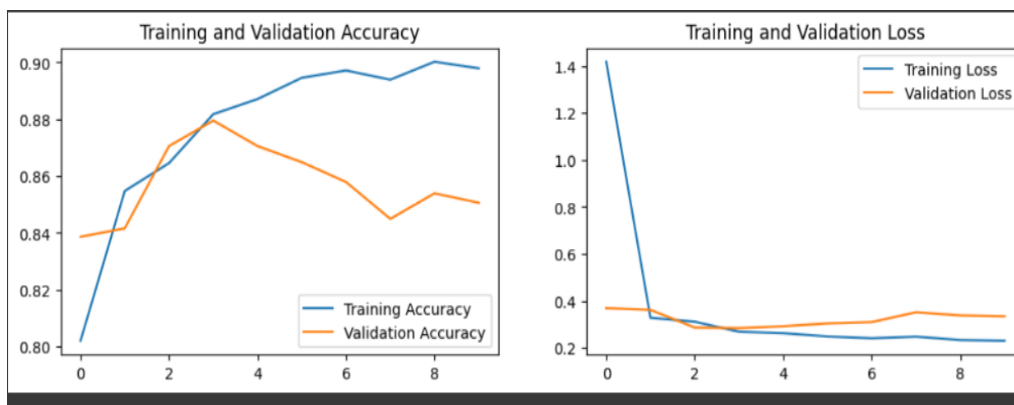


Figure 15: Training and validation accuracy and loss for Xception

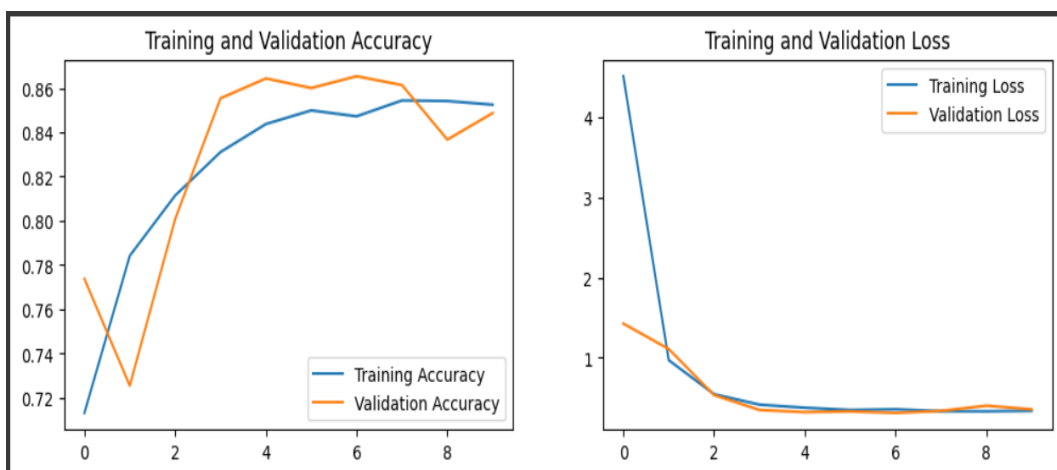


Figure 16: Training and validation accuracy and loss for InceptionV3

### 3.5.3 Model Evolution

#### Confusion Matrix

A confusion matrix is a tabular representation used in machine learning and classification tasks to evaluate the performance of a model by displaying the counts of true positive, true negative, false positive, and false negative predictions. It is particularly useful for binary and multiclass classification problems. This matrix provides a clear overview of the model's performance, allowing for the calculation of various evaluation metrics such as precision, recall, accuracy, and F1 score. It is a valuable tool for understanding where a model excels or falls short in its predictions and aids in refining the model based on these insights. Confusion Matrix generate deepens on number of class. M number of classes matrix will be  $M \times M$ . To gauge a machine learning model's accuracy, we hold up a mirror called a confusion matrix. It reveals how closely its predictions align with reality, exposing any blind spots or distortions. Table are given for confusion matrix:

Table 5: Confusion Matrix

Confusion Matrix		
	Actual Class	
Prediction Class	TP	FP
	FN	TN

## Performance Metrics

When the training of the models is completed, they were tested on test to evaluate the model accuracy. We test our Models on 3004 images from 3 different classes. For Evaluating the performance of every models, the metrics supported overall accuracy, Precision, Re-call, and F1 score. Let's explain all the metrics:

### Accuracy

Accuracy measures how many times a classifier correctly predicts. Accuracy is the ratio between the number of correct predictions and the total number of predictions. It is not suited for those classes which are imbalanced. If the data is imbalanced, then the model predicts that each point belongs to the majority class label. It can be misleading when there are remarkably more data points in class than others. In his region, a model can achieve high accuracy by predicting only the majority class all the time. In other hand, it's very useful to measure the overall percentage of valid prediction of a model. In simple words, it tells us how many of the predictions are actually positive. Accuracy is a valid choice for assessing classification problems that are well-balanced, are not skewed, or do not result in class imbalance.

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN} \dots\dots\dots (2)$$

### Precision

Precision used to measure of correctness. From total number of positive predicted, the number of actual positive are known by precision. The ratio between the total number of correctly classified positive classes and total number of predictive positive classes are called precision. It's useful when False positive is higher concern than false negative.

$$\text{Precision} = \frac{TP}{TP+FP} \dots\dots\dots (3)$$

### Recall



The heart of accuracy in positive predictions. Simply put, it reveals how many of our predicted "positives" were actually correct. It's a ratio: the number of true positives divided by the total number of predicted positives. In essence, it tells us how often we hit the bullseye when aiming for "positive" outcomes. Ideally, we want a high precision – a perfect score of 1 meaning every "positive" prediction was spot-on. This metric shines when false positives, mistakenly predicting positive outcomes, carry greater weight than false negatives, missing actual positives.

$$\text{Re-call} = \frac{TP}{TP+FN} \dots\dots\dots (4)$$

**F1-Score**

The F1 score is the harmonic mean of precision and recall between 0 and 1. Unlike simple averages, harmonic means are not sensitive for immensely large values. A balance between classifier precision and recall are balanced by F1 score. If your precision is low, your F1 score is low, and if recall is low again, your F1 score is low. If there is any cases where there is no clear segregation between whether precision is more important or recall, then we try to increase the precision of our model, the recall goes down and vice-versa. The F1-score captures both trends in a single value.

$$\text{F1 score} = \frac{2 \times (\text{Re-call} \times \text{Precision})}{(\text{Re-call} + \text{Precision})} \dots\dots\dots (5)$$

Here,

- TP (True Positive): Model accurately predict the positive data point
- TN (True Negative): Model accurately predict negative data point
- FP (False Positive): Model incorrectly predicts a positive data point
- FN (False Negative): Model incorrectly predict a negative data

All the required value of TP, TN, FP, FN find from confusion matrix. Trained model and then evaluate it by performing confusion matrix and other metrics. All the matrices value justifies a model's accuracy and ability to detected specific classification. When the confusion matrix has no FN and FP value then the situations, we said that our model correctly trained and catch the pattern of all classes. F1, Recall, Precision value also indicate the models training and validation capacity.

## CHAPTER 4

### EXPERIMENTAL RESULTS AND DISCUSSION

#### 4.1 Introduction

This chapter delves into the outcomes of our experimentation, providing a comprehensive analysis of the experimental results obtained in the pursuit of automated lung cancer cell detection. Through a systematic approach, we present the findings, conduct a descriptive analysis, and summarize the key insights derived from the experiments. This chapter delves into the outcomes of our experimentation, providing a comprehensive analysis of the experimental results obtained in the pursuit of automated lung cancer cell detection. Through a systematic approach, we present the findings, conduct a descriptive analysis, and summarize the key insights derived from the experiments.

#### 4.2 Experimental Result

It is quite impossible to obtain 100% accurate value from a model. But it can provide real time output. In my proposed model used 5 different type of machine learning algorithm. Here we used 80% data to train the model. 12000 data are used to train each model. We used 10 epoch, and 224×224 size of images. Then we evaluate our model accuracy and loss. From fig 12,13,14,15,16 shows the training and validation accuracy as well as training validation loss. Accuracy measures the proportion of correctly classified data points. Error measures the proportion of incorrectly classified data points. Training accuracy indicates how models learned the patterns from training dataset. On the other hand, Training loss indicates the amount of model's prediction deviate from actual level of training dataset. Validation accuracy demonstrates how well models generalize to unseen data, and validation loss reveals the potential model overfitting or underfitting. High training and validation accuracy indicate a good well-trained model. From the above figure we can see that ResNet50 and VGG16 both model's training and validation accuracy is high. VGG16 obtained 99.35% training

accuracy and 98.37% validation accuracy, on the other hand ResNet50 achieved 99.26% training accuracy and 99.23% validation accuracy. Others also scored good. CNN model overfitting occurs as its training accuracy 97.72% is much higher than the validation accuracy 72.50%. that's means this model memorization of training data rather than generalization.

Evaluate models accuracy by using confusion matrix. Here given the confusion matrix for used algorithms.

Table 6: Confusion Matrix Score

Algorithm	Confusion Matrix			
		lung_aca	lung_n	lung_scc
CNN	lung_aca	843	65	131
	lung_n	0	1002	0
	lung_scc	105	1	857
Xception	lung_aca	696	4	339
	lung_n	33	964	5
	lung_scc	68	0	895
ResNet50	lung_aca	1030	0	13
	lung_n	0	1005	0
	lung_scc	10	0	943
	lung_aca	1009	0	30

<b>VGG16</b>	lung_n	1	1001	0
	lung_scc	18	0	945
<b>InceptionV3</b>	lung_aca	721	168	158
	lung_n	23	981	1
	lung_scc	98	10	845

As we know that a perfect model only has TP and TN values in confusion matrix. Here in my work ResNet50 and CNN model's for lung\_n classes there is no FN and TN value. Others classes TP value is appreciable. All applied algorithm makes a good score in confusion matrix except Xception and InceptionV3.

### **True positive (TP)**

When a model correctly predicts a positive outcome for a case that is actually positive. Uses to measure the ability to correctly identify actual positives. In our work CNN, ResNet50 and VGG16 models are 95% plus scores to detect positive prediction from actual positive.

### **True Negative (TN)**

When The actual result is Negative and model predicts the negative value. Use for measuring the ability to correctly identify actual negatives. As our confusion matrix is multiclass so it's not directly represented in matrix. But it's easy to calculate as the sum of all values outside of classes row and column

### **False Positive (FP)**

Model predict positive value but actual value is negative. Incorrectly Prediction are belonging to this classes. The value of FP lies outside TP. In My work VGG16, ResNet50 acquires FP value is very low

### False Negative (FN)

When Model Predict negative value but the actual value is positive. This type of prediction called type-2 error. The rate of accuracy and F1 score identified that the appearance of FN is low.

### Classification report Analysis

Classification report for all applied algorithm for lung cancer dataset in tabular form:

Table 7: Classification Report

Algorithm	Classes	Precession	Re-call	F1	Accuracy %
CNN	lung_aca	0.89	0.81	0.85	97.72
	lung_n	0.94	1.00	0.97	
	lung_scc	0.87	0.89	0.88	
ResNet50	lung_aca	0.99	0.99	0.99	99.26
	lung_n	1.00	1.00	1.00	
	lung_scc	0.99	0.99	0.99	
VGG16	lung_aca	0.98	0.97	0.98	99.35
	lung_n	1.00	1.00	1.00	
	lung_scc	0.97	0.98	0.98	
Xception	lung_aca	0.87	0.63	0.76	89.79
	lung_n	1.00	0.96	0.98	
	lung_scc	0.72	0.93	0.81	
InceptionV3	lung_aca	0.86	0.69	0.76	84.87
	lung_n	0.85	0.98	0.91	
	lung_scc	0.84	0.89	0.86	

## **F1 Score**

From the table, if we see the F1-score 0.76 to 1.00, most of the models scores above .90%. As we know if any models f1-score achieve 1.00, it will be the perfect model. Here ResNet50 model achieved 1.0 for lung-n class and 0.99 for others two classes which proved that ResNet50 is almost perfect model for lung cancer detection by using histopathological images. Another model VGG16 also achieved 0.98-0.99 which also indicated towards a perfect model.

## **Precision**

As we know that precision 1.0 indicated that model produces no False Positive (FP). From the evolution of our used model we observed that precision scores 0.72 o 1.0. ResNet50, VGG16, Xception achieved 1.0 for lung\_n classes which was a remarkable achievement for this research.

## **Recall**

Recall identify the part of actual positives which were correctly classified. Recall achieved 1.0 when model produces no false negative value. In our proposed models recall 1.0 achieved in CNN, ResNet50 and VGG16 algorithm.

it's very clear that ResNet50 and VGG16 model Works good for providing dataset, and its more applicable for lung cancer detection. Accuracy of any models indicate the model's prediction correct level. In fig we see the VGG16, ResNet50 and CNN obtain above of 95% accuracy, 99.35% accuracy achieved by ResNet50 model.

## **Avg Accuracy**

Average accuracy" is used to describe the overall performance of a classifier or model across multiple predictions. It's calculated by taking the mean of the accuracy scores achieved for each individual prediction. Our model achieved 99% avg accuracy in some cases. ResNet50 algorithm obtained 99% avg accuracy, followed by that VGG16 achieved 98%. High rate of accuracy indicates the models are classifying correctly and can produce almost accurate prediction result.

Accuracy chart for all model:

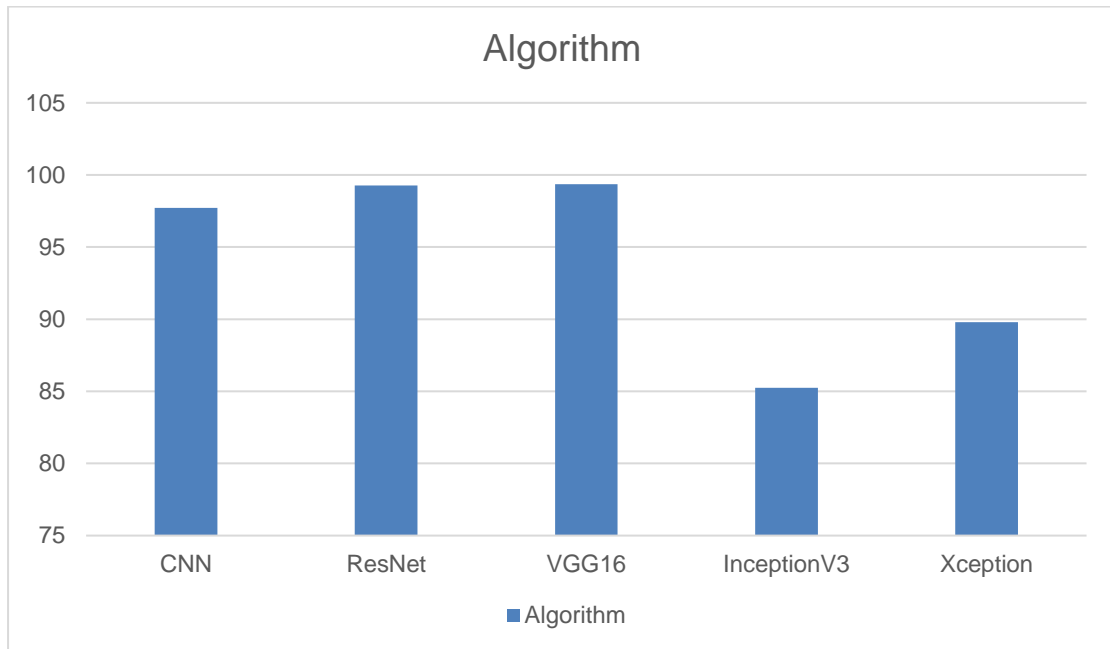


Figure 17: Models accuracy

### 4.3 Descriptive Analysis

This Research provides different type of algorithms accuracy so that we can choose the best result. I used 15000 large datasets for this paper which is enough to train a model and provides almost accurate result. Here I used CNN, ResNet, VGG16, InceptionV3 and Xception model to train the dataset. All of the image Classification model can help us to provide the almost 100% accuracy(99.36%)

Our model algorithms ResNet50 scores accuracy 99.26%. Here given the ResNet50 accuracy and loss Analysis graph plot. The ResNet50 classifier achieved the best performance: 99.26% high accuracy & validation accuracy is 99.23%. The proposed systems are compared with other recent systems, and the obtained results showed the excellent performance of the proposed systems. The ResNet50 model is tested on data from lung cancer patients, and the observed results are shown. It is calculated using validation and training. A loss indicates Resnet-50 Provide a very good performance in This Research. VGG16 scores highest accuracy 99.35% and validation accuracy

72.50%. Which is good accuracy rate for the research. Here given graphical representation of accuracy and loss. CNN obtained accuracy 97.72% and validation accuracy 98.37%, Which provides good accuracy rate for the research. Here given graphical representation of accuracy and loss. Xception obtained accuracy 89.79% and validation accuracy 85.05%. Here given graphical representation of accuracy and loss InceptionV3 scores accuracy 85.25% and validation accuracy 84.87%, Which loss 33.91%.

#### **4.4 Summary**

As there is some limitation of the automated method of lung cancer detection, our model also failed in some case to detect 100% accurately. We noticed that although our used algorithms perform well in detect cancer cell but it also shown some false positive and false negative value. The chapter concludes with a comprehensive summary, encapsulating the main takeaways from the experimental results and descriptive analysis. This synthesis of quantitative and qualitative findings aims to distill the essence of our research, offering a nuanced perspective on the effectiveness of advanced machine learning algorithms for automated lung cancer cell detection. The summary sets the stage for the subsequent chapters, guiding the reader towards a deeper comprehension of the implications and contributions of our study.



## **CHAPTER 5**

### **IMPACT ON SOCIETY & SUSTAINABILITY**

#### **5.1 Introduction**

Lung cancer is a deadly type of cancer and the rate of its growth is simultaneously high. As we know Lung cancer is most fatal among the cancer related diseases globally, significantly impacting life expectancy and being a cause of premature loss of life. It's also a burden on mental health, stress, family faces emotional and financial stress. Early detection can make a change to reduce the mortality rate in a significant way. At first stage cancer can be overcome by taking proper treatment and medicine. Moving beyond the nuts and bolts: This chapter examines the wider consequences of using automation to find lung cancer cells. It looks at how this technology might affect society, the ethical issues it raises, and whether it can be realistically adopted in healthcare settings

#### **5.2 Impact on Society**

According to the LAFC, one person takes almost 23,040 breaths in a day, constantly exposing our lungs to the surrounding environment [24]. Several lung diseases, including lung cancer has many identified risk factors. Some of these risks can be controlled, some less so. Smoking, second-hand smoking, family history, air pollution, radon gas are responsible for lung cancer risk factors [24]. 80% lung cancer occurs due to smoking. Low survival rates plague lung cancer globally despite diverse spread. While cases of lung cancer and its deadly impact differ internationally, the 5-year survival rate remains saddeningly low at 10-20% worldwide, largely due to late diagnoses [25]. So, it's urgent to take proper steps to reduce the epidemic effect of lung cancer. The need for early and easy-to-use detection methods keep in mind this research work on automated lung cancer detection. The proposed model will reduce the cost and time for medical diagnosis. In this paper we used histopathological data which may come from pathological tests like biopsy. By using manual ways to analyze

pathological report is not easy and required much time which may impact on further treatment.

As we talk about Bangladesh people are not interest to regularly checkup process. The rate of using tobacco and air pollution rate are enough high to effected lung cancer. Lack of knowledge, apartheidness for pathological test can be the season of high death rate in LC. But this research proposed a ML technology which help possible to make a computer-aided diagnosis system where people diagnosed their disease in time and go for further step. AI-powered lung cancer screening makes early detection affordable and efficient, boosting healthcare access in resource-scarce areas. The main impact of this study on our society will be the cost-effective automated scans democratize early lung cancer detection, especially in regions struggling with healthcare access.

### **5.3 Ethical considerations**

Ethical practices are essential to safeguard patient rights, promote fairness, and foster trust in healthcare Computer Aided systems. Protecting patient privacy is paramount. Medical data, especially sensitive information related to cancer diagnosis, must be handled with the utmost confidentiality. No one forced to others to share their personal information and data. Data are used in this paper collected from authorized source and take care of people's privacy. Our data collected from publicly open dataset, which open for researcher or other who can use them for novel work. Ensure algorithms learn from data that mirrors real-world diversity to reduce discriminatory outcomes and achieve greater equity. Predicting anything perfectly is always out of reach, even for machine models. This project, built on one machine learning model, faces current uncertainty, but we anticipate significant accuracy gains as the database expands to millions of data points. In this paper carefully follow all the moral for collecting and sharing data. When we developed this model, we try to strictly follow the ethics of people's privacy, physical and information security.

## 5.4 Sustainability

Ensuring sustainability in automated lung cancer detection involves a holistic approach that considers environmental, economic, and social aspects. From an environmental standpoint, optimizing the computational resources and energy consumption during model training and inference contributes to the eco-friendliness of the technology. Long-term viability encompasses regular updates and adaptability to evolving medical practices, minimizing the need for frequent overhauls. Integration with existing healthcare infrastructure ensures efficiency and reduces the environmental footprint associated with system implementation. A sustainable deployment also requires continued training and education for healthcare professionals in order to ensure the long-term viability of the technology. In essence, sustainability in automated lung cancer detection involves a balanced consideration of environmental impact, economic feasibility, and societal benefits to create a lasting and responsible solution in the realm of healthcare. This research unlocks a treasure trove of possibilities. Combining deep learning, AI, and IoT in future studies can enhance reliability and foster a wave of sustainable development initiatives. In Future a user-friendly Website may be published. Focused in working on more dataset, added more algorithm , updates instrument may make this a ultimate system for lung cancer detection

## **CHAPTER 6**

### **FUTURE SCOPE & CONCLUSION**

#### **6.1 Summary of the study**

The report presents a novel approach to automate the detection of lung cancer using machine learning networks and histopathological tissue images for lung cancer. The report utilized a dataset containing 15000 photos of histopathological lung cancer tissue images obtained impressive accuracy values, including 97.09% accuracy, 96.89% precision, 97.31% recall, 97.09% F-score, and 96.88% specificity. The images underwent annotation by skilled pathologists to pinpoint the presence of lung cancer cells. Originating from various sources, the dataset offers diversity, featuring different lung cancer cell types. This inclusiveness is essential for training effective deep learning models. Leveraging such an extensive and varied dataset is pivotal in advancing the development of automated systems for accurate and efficient lung cancer detection. The deep convolutional neural network was trained to extract important features for efficient and accurate lung cancer cell detection. I also discussed the theoretical basis of digital image processing and medical imaging, highlighting the significance of machine learning in extracting crucial features for cancer detection.

The report emphasized the potential impact of the proposed machine learning models in providing fast, almost accurate, and low-cost cancer detection for both pathologists and patients. I also suggested the extension of the project to create a user interface for easy utilization and recommended regular screening for individuals at risk.

Furthermore, the report come up with extensive review of various methods and techniques used for the prediction and classification of lung cancer. It covers a range of approaches including machine learning, image processing, feature extraction, and classification algorithms. The review discussed about the importance of early detection in improving patient outcomes and highlights the increasing role of computer aided diagnosis (CAD) systems in this domain. The report also presents a comparison of

different methodologies and their results, showcasing the effectiveness of various machine learning techniques in lung cancer detection. Overall, the report aims to provide a valuable resource for researchers and practitioners in the field of medical technology and healthcare.

## **6.2 Implication for Further Study**

The implications for further study based on the findings include the exploration of additional deep learning architectures and algorithms for enhanced lung cancer detection, the investigation of more substantial and diverse datasets to validate the performance of the proposed methods across different populations, and the development of user-friendly interfaces for seamless integration of the automated detection system into clinical practice. Furthermore, future research could focus on the ethical considerations and challenges associated with implementing such systems in healthcare settings, as well as the potential for personalized risk assessments and treatment planning for lung cancer patients. Additionally, there is scope for exploring the integration of other modalities such as genetic data to further improve the accuracy and efficiency of early-stage lung cancer detection.

## **6.3 Recommendation**

Further exploration of deep learning architectures and algorithms for improved lung cancer detection. Expansion of the research to encompass larger and more diverse datasets to validate the proposed methods across different populations. Development of user-friendly interfaces to facilitate the seamless integration of the automated detection system into clinical practice. Ethical considerations and challenges associated with the implementation of such systems in healthcare settings should be thoroughly addressed. Investigation of the potential for personalized risk assessments and treatment planning for lung cancer patients. Integration of other modalities, such as genetic data, to enhance the accuracy and efficiency of early-stage lung cancer detection. These recommendations aim to guide future research endeavors in the field of medical

imaging and cancer detection, building upon the valuable insights provided in the report.

## **6.4 Conclusion**

In conclusion, this report offers insightful analysis on the use of machine learning and image processing techniques for detecting lung cancer. The report highlights the potential of machine learning to increase classification accuracy for low population, high dimensional lung cancer datasets without the need for hand-crafted features. The dataset used 15000 photos of histopathological lung tissue images for cancer cell detection. The dataset was obtained from the publicly available Kaggle database. The images were collected from various sources and were annotated by experienced pathologists to identify the presence of lung cancer cells. The use of such a large and diverse dataset is crucial for the development of accurate and efficient automated systems for lung cancer detection. The report also highlights the significance of machine learning strategies in improving cancer characterization and detection. As there are many ways to detection lung cancer cell and thousands of researchers works on it, they give different type of ideas, methodology. Many works on varies deep learning algorithms, but here I used 5 different ML algorithm, used transfer learning techniques, used pretrained model so that the procedure of lung cancer detection is not require long time. Our ML model achieved 99% highest accuracy that's means our proposed model can identify almost accurate type of lung cancer. Others type of parameters like f1 score, precision, recall achieved 90% plus which means that our model's prediction correct above of 90% of time. Overall, this report can be a useful resource for researchers and practitioners in the field of medical technology and healthcare, providing a comprehensive review of various methods and techniques used for the prediction and classification of lung cancer.

## **APPENDIX**

As we select one of the vital topics “Automated lung cancer detection”, we faced many hardships to complete this research. Firstly, finding a methodology takes lot of dedication, as already many researchers works on it. We might not be done a standard work but a approached to find a way to help others who are interested to work on this field. In future we also try work on it with more dedication by updating hardware and software, added more algorithm, find best approach to detect lung cancer. On field data set collection was not easy for me because of luck of resources and peoples are not aware of this matter that much also have some ethical issues. It’s was a starting journey for me as a researcher ,so I was try to find a analytical prediction to make lung cancer detection’s best approach and try to find best strategic way to build a computer aided diagnosis system which may help peoples to detect lung cancer early and save their lives.

## REFERENCES

- [1] who.int. (n.d). cancer, Retrieved January 21,from <https://www.who.int/news-room/fact-sheets/detail/cancer>.
- [2] <https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images>
- [3] Sundaresh Ram, Wenfei Tang, Alexander J. Bell, Ravi Pal, Cara Spencer, Alexander Buschhaus, Charles R. Hatt, Marina Pasca diMagliano, Alnawaz Rehemtulla, Jeffrey J. Rodríguez, Stefanie Galban, Craig J. Galban. "Lung cancer lesion detection in histopathology images using graph-based sparse PCA network", *Neoplasia*, Volume 42,2023
- [4] Jie Ji, Weifeng Zhang, Yuejiao Dong et al. Automated Lung Cancer Detection using Histopathological Images, 02 July 2023.
- [5] Faria, Nelson, Sofia Campelos, and Vítor Carvalho. 2023. "A Novel Convolutional Neural Network Algorithm for Histopathological Lung Cancer Detection" *Applied Sciences* 13, no. 11: 6571
- [6] Radical, Rakhman, Wahid., Chilyatun, Nisa., Rahayu, Prabawati, Amaliyah., Eva, Yulia, Puspaningrum. "Lung and colon cancer detection with convolutional neural networks on histopathological images." *Nucleation and Atmospheric Aerosols*, null 2023.
- [7] Jehangir, B.; Nayak, S.R.; Shandilya, S. Lung Cancer Detection Using Ensemble of Machine Learning Models. In *Proceedings of the Confluence 2022—12th International Conference on Cloud Computing, Data Science and Engineering*, Noida, India, pp. 411–415, 27–28 January 2022;
- [8] Aayush Rajput, Abdulhamit Subasi, Chapter 2 - Lung cancer detection from histopathological lung tissue images using deep learning, Editor(s): Abdulhamit Subasi, In *Artificial Intelligence Applications in Healthcare&Medicine, Applications of Artificial Intelligence in Medical Imaging*, Academic Press, Pages 51-74,2023.
- [9] Manop Phankokkruad. 2021. Ensemble Transfer Learning for Lung Cancer Detection. In *2021 4th International Conference on Data Science and Information Technology (DSIT 2021)*. Association for Computing Machinery, New York, NY, USA, 438–442.
- [10] Coudray, N., Ocampo, P.S., Sakellaropoulos, T. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med* 24, 1559–1567 (2018)
- [11] Javier, Civit-Masot., Alejandro, Bañuls-Beaterio., Manuel, Domínguez-Morales., M., Rivas-Perez., Luis, Muñoz-Saavedra., José, María, Rodríguez, Corral. "Non-small cell lung cancer diagnosis aid with histopathological images using Explainable Deep Learning techniques." *Computer Methods and Programs in Biomedicine*, 226 (2022)



- [12] Sethy, Prabira Kumar et al. 'Lung Cancer Histopathological Image Classification Using Wavelets and AlexNet'. 211 – 221. 1 Jan. 2023
- [13] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, B. Yener, Histopathological image analysis: a review, *IEEE Rev. Biomed. Eng.*, 2 (2009), pp. 147-171
- [14] J. Shi, J. Wu, Y. Li, Q. Zhang, S. Ying." Histopathological image classification with color pattern random binary hashing-based PCANet and matrix-form classifier", *IEEE J. Biomed. Health Inform.*, 21 (5) (2017), pp. 1327-133
- [15] S. Ram, J.J. Rodriguez," Size-invariant detection of cell nuclei in microscopy images", *IEEE Trans. Med. Imaging*, 35 (7) (2016), pp. 1753-1764
- [16] S. Reis, P. Gazinska, J.H. Hipwell, T. Mertzaniidou, K. Naidoo, N. Williams, S. Pinder, D.J. Hawkes "Automated classification of breast cancer stroma maturity from histological images", *IEEE Trans. Biomed. Eng.*, 64 (10) (2017), pp. 2344-2352
- [17] H. Kong, M. Gurcan, K. Belkacem-Boussaid," Partitioning histopathological images: an integrated framework for supervised color-texture segmentation and cell splitting" *IEEE Trans. Med. Imaging*, 30 (9) (2011), pp. 1661-1677
- [18] S. Alinsaif, L. Jochen, "Partitioning histopathological images: an integrated framework for supervised color-texture segmentation and cell splitting", *BMC Med. Inform. Decis. Mak.*, 20 (14) (2020), pp. 1-19
- [19] A.B. Tosun, C. Gunduz-Demir, "Graph run-length matrices for histopathological image segmentation", *IEEE Trans. Med. Imaging*, 30 (3) (2011), pp. 721-732
- [20] Bangladesh - Global Cancer Observatory: <https://gco.iarc.fr/today/data/factsheets/populations/50-bangladesh-fact-sheets.pdf>
- [21] Report: Lung cancer on the rise in Bangladesh: <https://www.dhakatribune.com/bangladesh/health/236728/report-lung-cancer-on-the-rise-in-bangladesh>
- [22] <https://www.cancer.org/cancer/types/lung-cancer/about/what-is.html>
- [23] <https://www.geeksforgeeks.org/python-data-augmentation/>
- [24] <https://lcfamerica.org/lung-cancer-info/lung-cancer-risks/>
- [25] World Cancer Report: Cancer Research for Cancer Prevention. Lyon, France: International Agency for Research on Cancer, WHO.,; 2020. Available from: <http://publications.iarc.fr/586>.
- [26] H. Lin, H. Chen, S. Graham, Q. Dou, N. Rajpoot and P. -A. Heng, "Fast ScanNet: Fast and Dense Analysis of Multi-Gigapixel Whole-Slide Images for Cancer Metastasis Detection," in *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1948-1958
- [27] Alakwaa, Wafaa, Mohammad Nassef, and Amr Badr. "Lung cancer detection and classification with 3D convolutional neural network (3D-CNN)." *International Journal of Advanced Computer Science and Applications* 8.8 (2017).

sm

ORIGINALITY REPORT

22%

SIMILARITY INDEX

18%

INTERNET SOURCES

9%

PUBLICATIONS

13%

STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	5%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
3	www.mdpi.com Internet Source	1%
4	Submitted to University of Hertfordshire Student Paper	1%
5	www.geeksforgeeks.org Internet Source	1%
6	"Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications", Springer Science and Business Media LLC, 2023 Publication	<1%
7	Jie Ji, Weifeng Zhang, Yuejiao Dong, Ruilin Lin, Yiqun Geng, Liangli Hong. "Automated Lung Cancer Detection using Histopathological Images", Research Square Platform LLC, 2023 Publication	<1%