[Date]

Name

Classification of lung cancer using CNN and Resnet 50

# Abstract

Lung cancer has been considered as one of the most dangerous diseases in recent time where lung cancer are diagnosed in latest stages and most of the prognosis is poor which gives an overall 5 years survival rate. There are different Diagnostic procedures of lung cancer which are time consuming and expensive which is why Deep learning techniques are used to find out the patterns behind adeno carcinoma and squamous cell carcinoma from histopathological images. We have used a Convolution Neural Network with 5 convolution layers and ResNet 50 model with pre trained weights to classify normal lung tissue, adeno carcinoma and squamous cell carcinoma where we found an overall 54% accuracy with the CNN model but remarkable 98% accuracy with the ResNet 50 model. The preprocessing of the images are done with the help of data augmentation techniques which are imported from Tensorflow and keras. The models are evaluated with the help of confusion Matrix, precision, Recall and F1- score and also the model loss and accuracy is visualized at each epochs. From the observation of the research, it is found that ResNet 50 model outperforms the CNN model in terms of accuracy and gives 98% accuracy in predicting the lung cancer.

Keywords: CNN, ResNet 50, data augmentation, confusion matrix, precision, recall, F1-score, etc.

Table of Contents

[Abstract 1](#_Toc120029525)

[List of Figures 2](#_Toc120029526)

[List of Tables 2](#_Toc120029527)

[1. Introduction 4](#_Toc120029528)

[1.1 Background 4](#_Toc120029529)

[Etiology 4](#_Toc120029530)

[Adenocarcinoma 4](#_Toc120029531)

[Histopathology 4](#_Toc120029532)

[Squamous Cell Carcinoma 5](#_Toc120029533)

[Histopathology 6](#_Toc120029534)

[Epidemiology 6](#_Toc120029535)

[1.2 Motivation 7](#_Toc120029536)

[1.3 Problem Statement 7](#_Toc120029537)

[1.4 Proposed Solution 7](#_Toc120029538)

[1.5 Aims and Objectives 8](#_Toc120029539)

[1.6 Research Questions 8](#_Toc120029540)

[1.7 Project Outline 8](#_Toc120029541)

[2. Literature Review 9](#_Toc120029542)

[2.1 Literature Gap 12](#_Toc120029543)

[3. Methodology 13](#_Toc120029544)

[3.1 Values of Agile Manifesto 13](#_Toc120029545)

[3.2 Principles of Agile Manifesto 13](#_Toc120029546)

[3.3 KDD approach 13](#_Toc120029547)

[3.4 About the Data 14](#_Toc120029548)

[3.5 Materials Required 16](#_Toc120029549)

[3.6 Data Processing 16](#_Toc120029550)

[3.7 Convolution Neural Network 16](#_Toc120029551)

[ Convolution Layers 17](#_Toc120029552)

[ Pooling Layers 18](#_Toc120029553)

[ Activation Function 19](#_Toc120029554)

[ Fully connected layer 20](#_Toc120029555)

[ Loss Functions 20](#_Toc120029556)

[ Regularization to CNN 21](#_Toc120029557)

[ Benefits of Using CNN 23](#_Toc120029558)

[3.8 ResNet-50 23](#_Toc120029559)

[ Skip Connection 25](#_Toc120029560)

[ Benefits of using ResNet 50 26](#_Toc120029561)

[3.9 Model Evaluation 26](#_Toc120029562)

[4. Implementation 28](#_Toc120029563)

[4.1 CNN implementation 28](#_Toc120029564)

[ Data Exploration 28](#_Toc120029565)

[ Data Augmentation 28](#_Toc120029566)

[ Building of Models 29](#_Toc120029567)

[ Evaluation of Models 31](#_Toc120029568)

[4.2 ResNet 50 Implementation 31](#_Toc120029569)

[ Data Exploration 32](#_Toc120029570)

[ Data Augmentation 32](#_Toc120029571)

[ Building of models 33](#_Toc120029572)

[ Model Evaluation 34](#_Toc120029573)

[5. Results and Analysis 35](#_Toc120029574)

[5.1 CNN model performance 35](#_Toc120029575)

[5.2 ResNet-50 model 37](#_Toc120029576)

[5.3 Parameters Comparison 40](#_Toc120029577)

[6. Discussion 41](#_Toc120029578)

[6.1 Model Implementation 41](#_Toc120029579)

[6.2 Research Investigation 42](#_Toc120029580)

[7. Conclusion 43](#_Toc120029581)

[7.1 Limitations of the Research 44](#_Toc120029582)

[7.2 Future Works 44](#_Toc120029583)

[References 45](#_Toc120029584)

## List of Figures

[Figure 1 Samples of Adenocarcinoma 8](#_Toc120030663)

[Figure 2 Samples of Squamous cell carcinoma 10](#_Toc120030664)

[Figure 3 KDD approach 20](#_Toc120030665)

[Figure 4 Normal Lung tissue 21](#_Toc120030666)

[Figure 5 Adencarcinoma Samples 22](#_Toc120030667)

[Figure 6 Squamous samples 22](#_Toc120030668)

[Figure 7 CNN network 25](#_Toc120030669)

[Figure 8 Convolution layer 27](#_Toc120030670)

[Figure 9 Pooling layer 27](#_Toc120030671)

[Figure 10 Fully connected layer 29](#_Toc120030672)

[Figure 11 Regularization to CNN 31](#_Toc120030673)

[Figure 12 Working of ResNet 35](#_Toc120030674)

[Figure 13 Architecture of ResNet 50 37](#_Toc120030675)

[Figure 14 Skip connection 37](#_Toc120030676)

[Figure 15 Stages of ResNet 50 38](#_Toc120030677)

[Figure 16 Model loss CNN 48](#_Toc120030678)

[Figure 17 Model accuracy CNN 49](#_Toc120030679)

[Figure 18 Confusion matrix of CNN 50](#_Toc120030680)

[Figure 19 Classification report of CNN 51](#_Toc120030681)

[Figure 20 Model loss ResNet-50 52](#_Toc120030682)

[Figure 21 Model accuracy ResNet 50 52](#_Toc120030683)

[Figure 22 Confusion matrix of ResNet 50 53](#_Toc120030684)

[Figure 23 Classification report of ResNet 50 54](#_Toc120030685)

## List of Tables

[Table 1 Parameters of CNN model 42](#_Toc120030686)

[Table 2 Parameters of ResNet 50 model 46](#_Toc120030687)

# Introduction

## 1.1 Background

For many years now, lung cancer has emerged as the most lethal form of the disease. The rate of new cases climbed by 37% between 2007 and 2017, and in 2022, lung cancer was responsible for nearly 131,000 fatalities in the United States alone (US). Most new occurrences of lung cancer are diagnosed in the latter stages of the disease, when treatment options are limited, and the prognosis is poor (20.5% overall 5-year survival rate). Because of its widespread availability, reasonable price tag, and superior spatial resolution, computed tomography (CT) has become the gold standard for the diagnosis as well as screening of lung cancer. Despite the inherent use of ionizing radiation in the procedure, low dose computed tomography (CT) for lung cancer screening has been demonstrated to be accurate with an average effective dosage of roughly 1.5 mSv.

The American Cancer Society makes an annual population-based estimate of the total number of new cancer cases and deaths in the United States. There will be approximately 1,898,160 new cases of cancer in the United States in 2021 and an estimated 608,570 fatalities, with 85% of these instances being non-small cell lung cancer (NSCLC). Radiofrequency (RF) excision as well as stereotactic body radiotherapy (SBRT) procedures can be used to diagnose NSCLC. Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCC) are the two most common forms of lung cancer (SCLC). These two forms of cancer have different modes of dissemination and hence require different therapies. In contrast to small cell lung cancer (SCLC), which is strongly linked to smoking and spreads rapidly, forms tumors, and can affect the entire body, NSCLC develops relatively slowly and stays localized. Smoking causes small cell lung cancer, and fatalities from it are directly correlated with cigarette consumption.

## Etiology

This research is commonly addressing two types of lung cancer which is adenocarcinoma and squamous cell carcinoma.

### Adenocarcinoma

Adenocarcinoma is a form of glandular cancer. These are the cells responsible for the secretion and excretion of chemicals throughout the body. The prognosis, therapy, and overall survival for adenocarcinoma vary on the tumor's location, size, as well as stage, as well as personal characteristics such as the patient's general health (Vecerzan and Mihaila, 2016).

Adenocarcinoma in the lungs can give rise to various symptoms including blood mucus, weight loss, weakness, exhaustion and coughing.

There are various risk factors that are related to a adenocarcinoma and some of the primary risk factors include smoking tobacco products or exposure to harmful environment which includes toxins and chemicals and also exposure to Radiation therapy around the lungs (Vecerzan and Mihaila, 2016).

### Histopathology

The histological study of a lung biopsy will indicate a tumor emerging from the bronchial glands. The formation of mucus is also extremely visible. The new categorization of the World Health Organization (WHO) sub classifies adenocarcinomas as originating from Acinar, Papillar, Bronchoalveolar and Mucus-secreting.

Adenocarcinoma has a significantly inferior prognosis than squamous cell cancer, with the exception of lung cancer in Stage 1.

The samples of adenocarcinoma extracted from the dataset is illustrated below.

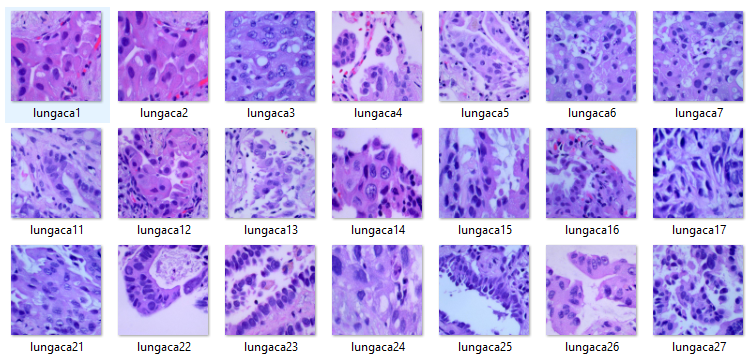


Figure Samples of Adenocarcinoma

### Squamous Cell Carcinoma

Non-small cell lung cancer includes squamous cell carcinoma (SCC) of the lung, commonly referred as squamous cell lung cancer (NSCLC). Adenocarcinoma is the most prevalent subtype of NSCLC, trailed by the squamous cell carcinoma of the lungs, particularly in women. This is ascribed to the shift in cigarette smoking patterns, but there is no conclusive proof (Bargotya et al., 2019).

The change of the squamous cells covering the airways causes SCC lung disease. Squamous cells are slender, flat cells that line numerous organs within the human body. Typically, squamous cell lung cancers develop in the middle portion of the lung or the major airways, such as the left or right bronchial tubes. Tobacco smoke, consisting contains over 300 toxic chemicals and forty probable carcinogens, is the leading cause of cellular change. Transformed squamous cells are distinguished by keratinization including intercellular bridges as well as frequently display a high mutation frequency.

NSCLCs can cause cough, difficulty breathing, breathlessness, bleeding in sputum, asthma, sore throats, repeated chest infections such as pneumonia and bronchitis, weight loss, loss of appetite, and weariness. In the initial stages of the disease, however, NSCLC patients frequently exhibit no symptoms. Metastases may develop in advanced illness, with symptoms including bone pain, spinal cord compression, and cognitive symptoms such as headache, weakness or numbness of limbs, dizziness, seizures (Bargotya et al., 2019).

### Histopathology

An accurate histologic diagnosis is growing in importance as it can predict therapeutic response and toxicity. A diagnosis of SCC is verified when at least 10% of the tumor mass of resected samples demonstrates transformation characteristics including keratinization. If the squamous part of the tumor is minimally differentiated, a diagnosis of SCC with poor differentiation is made. A combination of an immunohistochemistry (IHC) board with a mucin stain can help to identify variants of squamous cell carcinoma. The expression of squamous biomarkers, such as p63 and p40 proteins, is also prominent in SCC.

The World Health Organization updated categorization in 2015 to distinguish three variations of SCC depending on histological testing, including Keratinizing, Non-keratinizing, and basaloid. If the basaloid component exceeds fifty percent of the tissue and squamous differentiation is negligible (Hassan, Mozayani and Wong, 2015).

The samples of squamous cell carcinoma is extracted from the data is illustrated below

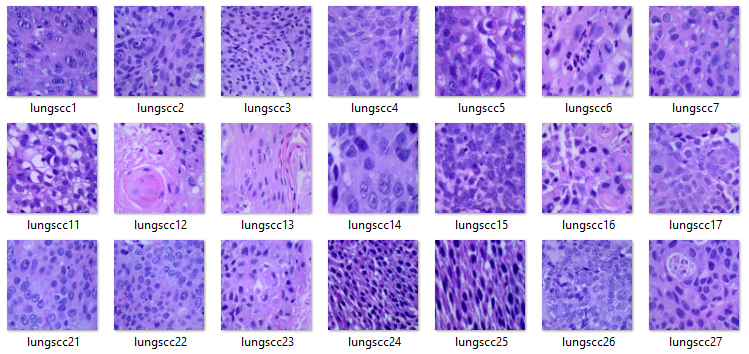


Figure Samples of Squamous cell carcinoma

## Epidemiology

Worldwide, lung cancer is the most prevalent cancer and leading cause of cancer-related mortality. In 2018, an estimated 2,093,876 instances of lung cancer were diagnosed worldwide, representing 11.6% of the worldwide cancer burden. In 2018, lung cancer was estimated to have caused 1,761,007 deaths worldwide, or around 18.4% of all cancer-related deaths.

In 2018, lung cancer incidence was distributed as follows: Asia (58.5%), Europe (22.4%), North America (12.1%), Latin America and the Caribbean (4.3%), Africa (1.9%), and Oceania (0.81%).

NSCLCs comprise around 85 percent of all lung malignancies. The most prevalent subtypes of NSCLC are adenocarcinoma as well as squamous cell carcinoma, representing for 50% as well as 30% of cases, respectively.

## 1.2 Motivation

Conventional chest radiography, lung biopsy, computed tomography, magnetic resonance imaging, positron emission tomography, and biopsies all play a role in the diagnosis of lung cancer. As a result of the disease being so difficult to detect until it has spread extensively, lung cancer has a dismal prognosis even when caught early.

Localized lung cancer has a 5-year survival rate of about 54%, while incurable, advanced lung cancer has a 4% survival rate. Surgery, radiation therapy, and chemotherapy are only some of the options that may be used to treat cancer, depending on its stage and type. The diagnostic, prognostic, and monitoring capabilities of computed tomography (CT) for many diseases are all well developed.

Convolutional neural network and other deep learning techniques had been useful in detection of diseases from medical images which saves time and also gives better accuracy compared to traditional techniques. Such techniques also extract the patterns and goes on improving by studying on more updated images

Radiography and CT scan techniques require medical experts to diagnose and also they are expensive which cannot be afforded by most of the patients. This is why CNN and other deep learning techniques can be useful in different rural Healthcare organizations where it can be used to detect lung cancer with more accuracy and in less time.

## 1.3 Problem Statement

Most of the lung cancers are undetected at an early stage that led to severe consequences when the lung cancer develops at a later stage. Histopathological images recorded from the infected patients require at least 10 days for proper diagnosis by medical experts. Lung cancer can be divided into adeno carcinoma and squamous cell carcinoma where histopathological images is required to study the disease of the tissues. For such time consuming and complex process deep learning techniques can be used to automatically consider all the inputs in the form of histopathological images and form a convolution neural network that can help in extracting the patterns from the images similar to the observations done by a medical expert (Petrella, 2021).

The observations of medical experts can contain errors and it is time consuming but deep learning technique can give better accuracy and can detect the lung cancer from histopathological images within minutes. This is why the sole motive of this research is to perform and evaluate a CNN model trained on histopathological images that can be helpful in various Healthcare organizations for accurate predictions of lung cancer.

## 1.4 Proposed Solution

The solution of this research is a deep learning model which will extract the patterns from the histopathological images of lung cancer where CNN technique and a transfer learning method such as ResNet 50 algorithm will be applied to classify various types of lung cancer. The input images will contain three classes where one class belongs to normal lung tissue, the other class belongs to lung adeno carcinomas and the other class belongs to squamous cell carcinomas. The images are perfectly balanced and augmented where the original images are transformed using various transformations with the help of Augmentor package and the size of the images are increased compared to the original one.

For evaluation of the solution, we have implemented confusion Matrix to determine the mis classification of the model in the prediction data and we also used several other metrics such as precision, recall and F1 score.

## 1.5 Aims and Objectives

The aim of the research is to detect lung cancer using deep learning methods like CNN and ResNet 50. The objectives which will be implemented to achieve the following aim are discussed below

* For improving the accuracy data augmentation will be implemented which will manipulate the images and increase the total number of samples using various transformations
* For deep learning algorithms like CNN and ResNet 50, Tensorflow and Keras will be used in Python that will be helpful for model building process
* For evaluation of lung cancer, we will use Evolution metrics such as confusion Matrix, Precision, recall and F1 score which will analyse the results achieved in this research
* We will also discuss the limitations observed during the prediction and various steps of improvement that will be required to improve the performance of our models

## 1.6 Research Questions

1. What techniques make deep learning models really work in lung cancer detection?
2. Which parameters are effective in accurate predictions of lung cancer using deep learning methods?

## 1.7 Project Outline

The outline of the project is summarized below

1. Introduction- This part discusses about the introduction of lung cancer and aims, and objectives implemented in our research.
2. Literature review - This section represents the related research about Deep learning algorithms that are used to predict various types of lung cancers.
3. Methodology – This section discusses about methods, tools and resources implemented in this research.
4. Implementation – This section discusses about the implementation and the procedure implemented for lung cancer prediction including CNN and ResNet 50 architecture.
5. Results and analysis – This section represents the results achieved in the research including evaluation of CNN and ResNet 50 models.
6. Discussion – This section includes the investigation of the research and achievements of the research
7. Conclusion – This section includes the overall performance achieved in the results along with limitations and future work required in the study

# Literature Review

There are several approaches and researchers which are conducted to predict the lung cancer from various types of images with the help of the deep learning techniques. We are going to review some of them and understand the advantages behind each study.

Methods based on data on gene expression were not cheap in 2021 and 2022, but they are extremely precise. However, there is a radiometric approach that saves money without sacrificing quality. This research (Aonpong et al., 2021) demonstrated the feasibility of a low-cost, high-accuracy genotype-guided radiomics (GGR) approach. Pre-processing, radiomics feature extraction as well as selection of input characteristics, and prediction are the steps that make up this technique. GGR is a method of forecasting that takes advantage of two models and involves a two-stage process. Gene estimation is used in the first model, while recurrence prediction is made in the second. The CT scans and gene expression data that make up the basic NSCLC radiography data set are used in this procedure. It has been shown experimentally that the proposed GGR greatly improves prediction accuracy over both the previous radiometric approach and ResNet50, reaching 83.28 percent.

According to the data, late-stage diagnosis is the leading cause of illness and death from lung cancer. Author (Strauss et al., 2015) suggests MLP as a means of determining whether or not Epidermal Growth Factor Receptor (EGFR) mutations are pathogenic. The nodule, the lung with the primary nodule, and both lungs are evaluated for EGFR activity. There are two primary stages to this proposed strategy. Phase one entails an effort to learn features. The next step is a comprehensive classification model that makes use of transfer learning strategies. The LIDC-IDRI as well as NSCLC-Radiogenomics datasets are used in this methodology. The experimental evidence demonstrates its superior predictive power.

Another writer (Yu et al., 2020) developed a method for the automatic diagnosis of lung cancer; they dubbed it the "Adaptive Hierarchical Heuristic Mathematical Model (AHHMM)". There are five phases to this process. Obtaining the image is the first stage. In the second step, the preprocessing step is carried out.  Binarization along with Thresholding and categorization are the last steps of the study. At long last, a Deep Neural Network can extract and detect features (DNN). Pre-classification photos were also clustered using a modified version of K-means. The experimental results reveal that the approach is 96.67 percent accurate on the lung cancer dataset.

Screening is widely used by radiologists for reliable examination of a variety of CT scans. The link involving algorithmic solutions and physicians has its own set of difficulties, yet automatic algorithmic solutions may aid. Low-dose CT scans are a solution that has been proposed by an author (Kadouri et al., 2019). This technology performs on-the-spot analysis of CT scans and returns standardized ratings. In this case, we're dealing with 3D convolutional neural networks trained on data from across the whole system. Each step of the process—pre-processing, the CADe module for segmentation, and the CAD diagnostics module—is carried out in turn (CADx). CADx is designed and configured concurrently since its efficacy is tied to that of CADe. The LIDC-IDRI and LUNA-16 and Kaggle datasets are used in this technique. In practical applications, this system performs better. The proposed approach has a 96.5% success rate in diagnosing lung cancer.

According to the author (Zhang & Kong, 2020), early detection of lung cancer can save many lives. Radiologists have a difficult, time-consuming, and repetitive task when looking for nodules that could indicate lung cancer. To address these issues, they suggested a "Multi-Scene Deep Learning Framework (MSDLF)" equipped with a "vesselness filter" to reduce false positives and improve accuracy. The primary goal of this study is to identify nodes with diameters greater than 3 mm. An artificial intelligence model developed using a four-channel CNN. Data set preparation, parenchyma segmentation and repairing of lung shape, vascular removal, data set standardization, CNN construction, segmentation, and classification, normalized spherical sampling, and normalized spherical sampling are all parts of this method. This strategy employs the LIDC-IDRI data collection.

A radiology expert once described the process of manually sketching lung nodules as slow and laborious. The technology by (Alakwaa et al., 2017) offers a 3D Deep Convolutional Neural Network (3DDCNN) to aid radiologists. Their setup outperforms cutting-edge alternatives. Accurate detection of lung nodules is achieved by combining deep learning with cloud computing. The mRPN, or Multi-Regional Proposal Network, was incorporated into their design. The procedure involves a cloud-based 3DDCNN CAD system, training datasets, feature extraction, pre-processing, a proposed model architecture, and a training methodology. Specifically, this technique makes use of the ANODE09, LUNA-16, LIDC-IDRI, and SHANGHAI Hospital datasets. The experimental results evaluation demonstrates that the provided model can detect lung cancer with 98.5% accuracy.

One author (van de Kamp et al., 2019) has developed a network for analysing knowledge to predict mortality, called KAMP-Net, which might be used to assess the likelihood of a patient dying from lung cancer. Data augmentation is employed to train a Convolutional Neural Network in this approach (CNN). They hypothesised that by adding more data to the mix, they could boost CNN's efficiency. Both the CNN as well as SVM results are used to produce a risk of death, with the clinical measures used to train the SVM classifier. Manual data collection was used to get the clinical measures. These are the steps that make up the method: Coding multi-channel images; planning and executing a network; using deep learning and clinical expertise. This approach employs data from the National Lung Screening Trial (NLST).

A study (Pang et al., 2020) conducted at the Shandong Province Hospital employed deep learning to classify CT pictures of patients with lung cancer. Image preparation techniques including "rotation, translation, and transformation" were utilised to increase the size of the training set and address the problem of insufficient patient data. In order to classify lung cancer images, the authors trained "densely linked convolutional networks (DenseNet). At last, they combined the results of several classifications with the adaptive boosting (adaboost) algorithm. This strategy makes use of data from the Shandong Provincial Hospital. Statistical analysis of the experimental data demonstrates that the suggested model can identify lung cancer with 89.85% precision.

In order to automate lung node analysis, it is necessary to have a reliable method for classifying malignant lung nodes, as well as a reliable method for calculating feature score regression. For a fully automated evaluation of pulmonary nodules, another author (Ferguson et al., 2018) has proposed the MTMR-Net model. It also serves as a blueprint for the Siamese network's physical layout. The three primary components of the proposed architecture method are: The first part is a module that extracts features. It had one Res Block A, three Res Block B, and a convolution layer. The second part is a categorization section. One single layer made up the entirety of it. The final part is a regression analysis section. The structure had two layers that were joined together at all points. Multi-Task Learning for Lung Nodule Analysis, Margin Ranking Loss for Distinguishing Marginal Nodules, and Joint Training of MTMR-Net are all components of the MTMR-Net model. The experimental results demonstrate that the suggested model can detect lung cancer with 93.5% accuracy. The "MTMR-NET" method outperformed other cutting-edge approaches in terms of accuracy, sensitivity, and specificity.

The difficulty of making an early diagnosis has been identified as a major contributor to lung cancer-related illness and mortality, according to another study. The early identification of lung cancer requires the risk level of pulmonary nodules in adenocarcinoma. The researchers (B, 2020) used STM-Net where they introduced a deep Convolutional Neural Network (CNN) that includes a scale transfer module (STM) in order to improve the performance of pulmonary adenocarcinoma. The following are the steps involved in the procedure. To begin, four convolution layers are applied to the input bronchial nodule images. Second, standardizing feature map sizes with max-pooling and STM. The third step is to employ channel fusion to arrive at a conclusive categorization. This approach employs data from the Zhongshan Hospital Fudan University dataset. A risk phase prediction network was developed as well as validated. The experimental results demonstrate a sensitivity of 95.455% for identifying lung cancer using this method.

Another researcher (P, 2021) had used Mask R-CNN technique to detect the pulmonary nodule. They had conducted three modules where the first module is the preprocessing module. The second is the segmentation module and the third is the reconstruction module. Also, the proposed model in the research contains four parts where the initial part is the ResNet 50 model with multi-layer neural network. The second part of their approach was a pyramid network followed by the third part which is the regional proposal Network, and the final part is the function branches. ResNet 50 was used for feature extraction from the original features of the images, and they used multi scale feature map using feature pyramid network. They had applied the model in LUNA-16 data set and with their proposed approach they achieved a sensitivity of 88.7% in detecting lung cancer.

In 2019, the early detection of lung cancer relies heavily on the correct classification of lung nodules as benign or malignant. In a multi-perspective, collaboratively created body of knowledge, another author suggests using a deep neural network approach to identify lung nodules as either benign or malignant (MVKBC). As a first step in understanding the attributes of each 3D lung node, they divide it into nine fixed perspectives. Then, they (Du Parcq et al., 2019) develop a knowledge-based collaborative (KBC) sub-model for each perspective, tailoring the extraction of patches based on their overall appearance (OA), another approach known as heterogeneity in voxel values (HVV), and heterogeneity in shapes (HS) for lung nodules in each slice. Finally, a consistent weighting technique is applied to nine KBC sub-models for classifying nodes. Since this is the case, the "MV-KBC model" provides comprehensive training for them. This technique makes use of the LIDC-IDRI data set. With a 91.60% detection rate for lung cancer, the proposed approach is quite promising.

The use of deep learning was offered as a solution for thoracic MR imaging. Since there weren't many ways to detect anomalies in MR pictures, they weren't used very often. The suggested method (Su et al., 2020) accepts the entire image as input and does not rely on candidate extraction, which is a common step in other nodule recognition algorithms for CT scans. It's because the lung nodules range in size from extremely small to quite large. With the goal of avoiding candidate extraction and reducing the need for a scale, a faster R-CNN has been developed for the detection of lung nodules. There are two parts to the Faster R-CNN. The first component, dubbed a Region Proposal Network, is capable of proposing regions for each image (RPN). A second part, a Fast R-CNN detector, sorts the proposed regions. This technique is implemented on the First Affiliated Hospital of Guangzhou Medical University dataset. The experimental evaluation demonstrates that the suggested model has a sensitivity of 85.2% for detecting lung cancer.

Small cell lung cancer (SCLC) survival rates can be dramatically improved via rapid and precise detection. Doctors can make more informed diagnoses thanks to precise cancer segmentation. Manual segmentation, while possible, is a laborious process. According to one author, a "hybrid segmentation network (HSN)" built on a "convolutional neural network (CNN)" is the way to go. The suggested network approach by (Peng et al., 2022) is built around two primary components necessary for precise tumour segmentation. The first is an efficient lightweight 3D convolutional neural network for understanding distant 3D context. The second is a two-dimensional convolutional neural network (CNN) that can record granular semantic details. They suggested combining the 2D and 3D features using a "hybrid features fusion module" (HFFM) and a "multiscale separable convolution" (MSC) Block that is similar to architecture with S3D convolution. These techniques are applied to data from the Shandong University Affiliated Hospitals dataset. The experimental findings show that the average accuracy and sensitivity of this approach are 0.909 and 0.872, respectively.

## 2.1 Literature Gap

Most of the researchers focused on CT scan images and they have used advanced neural network techniques to detect lung cancer. However carcinoma and squamous cell carcinoma are the most common types of lung cancer which is why we have attempted to predict these two types of cancer with traditional CNN technique and a ResNet 50 model. We have used ResNet 50 model as it uses a skip connection that can train the data with the help of 50 deep convolution layers unlike other algorithms.

# Methodology

During implementation of the research, we have followed the Agile methodology where we have followed the principles of Agile Manifesto and we also followed the KDD approach during experimentation with the data.

For work guidance, the four values and 12 principles of Agile Manifesto had been followed which are also known as the “Manifesto for Agile Software development”. These values where produced by 17 developers that can be used for different software development processes (Denning, 2015).

### 3.1 Values of Agile Manifesto

* Individuals and interactions over processes and tools
* Working software over comprehensive documentation
* Customer collaboration over contract negotiation
* Responding to change over following a plan

### 3.2 Principles of Agile Manifesto

* The highest priority is to satisfy the customer through early and continuous delivery of valuable software.
* The project team welcomes changing requirements, even late in development. Agile processes harness change for the customer’s competitive advantage.
* Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.
* Business people and developers must work together daily throughout the project.
* The process builds projects around motivated individuals, giving them the environment and support they need, and trusts them to get the job done.
* A face-to-face conversation is the most efficient and effective method of conveying information to and within a development team.
* Working software is the most important measure of progress.
* Agile processes promote sustainable development. The sponsors, developers, and users should maintain a constant pace indefinitely.
* Pay continuous attention to technical excellence, and good design enhances agility.
* Simplicity is essential. This is the art of maximizing the amount of work not done.
* Self-organizing teams produce the best architectures, requirements, and designs.
* At regular intervals, the team reflects on how to become more effective and adjusts its behavior accordingly.

## 3.3 KDD approach

KDD approach is specifically used during data mining and data cleaning process. KDD in data mining is a planned as well as analytic method for modelling database data in order to extract usable and practical 'knowledge' Data mining is the foundation of KDD and is therefore essential to the entire technique.

It employs a number of self-learning algorithms to extract meaningful trends from the dataset. The process is a closed-loop, constant-feedback one in which many iterations occur between the multiple processes in accordance with the requirements of the algorithms and interpretations of the patterns.

A typical illustration of KDD approach is given below

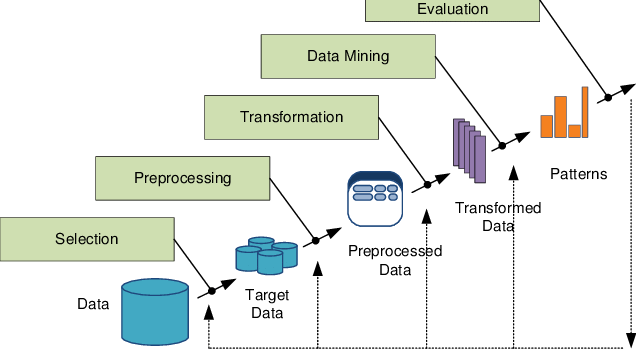


Figure KDD approach

## 3.4 About the Data

The data contain 15000 augmented histopathological images that are divided into 3 classes which are benign lung tissues, adenocarcinoma and squamous cell carcinoma. The original images are in the size of 768 by 768 and are generated from original sample of HIPAA complaint as well as validated sources that consist only 750 images where 250 images belong in benign tissues, 250 images belong in adenocarcinoma and 250 images belong in squamous cell carcinoma. With the help of Augmentor package, these 750 images are extended up to 15000 images.

The benign lung tissue samples from the data is visualized below

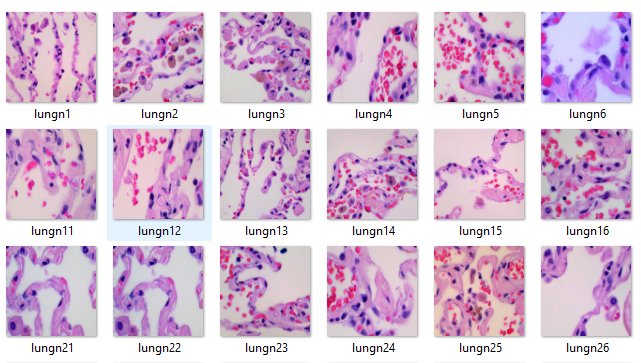


Figure Normal Lung tissue

The adenocarcinoma samples from the data is visualized below

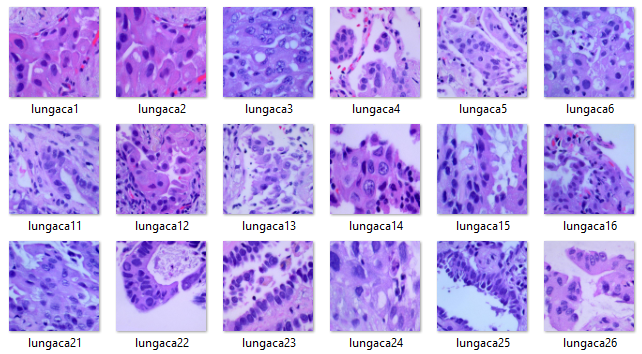


Figure Adencarcinoma Samples

The squamous cell carcinoma samples from the data is visualized below

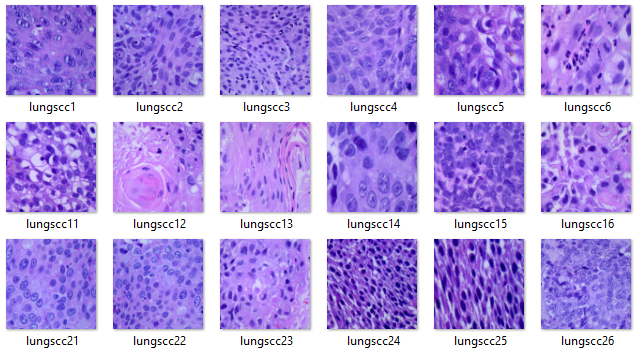


Figure Squamous samples

## 3.5 Materials Required

The materials collected for this research are top journals and articles which are updated and are relevant to lung cancer. We have searched all the supporting journals using the keywords like **‘deep learning techniques for lung cancer prediction’, ‘CNN technique for lung cancer’, ‘systematic review for lung cancer’, ‘types of lung cancer’,** etc. With the help of these keywords, we have found out 15 relevant papers from which we have taken only 7 papers and conducted the research.

The other material is the data which is collected from the Kaggle which contain 15000 histopathological images of benign lung tissue, adeno carcinoma and squamous cell carcinoma.

## 3.6 Data Processing

Data augmentation is the process of generating new data samples from existing ones. Data augmentation is often used in computer vision to artificially increase the amount of training data by creating new samples from existing ones. Data augmentation can be used to create new images by randomly transforming existing ones, or by adding noise to them.

Some common methods of data augmentation include:

* Randomly cropping images
* Randomly flipping images horizontally
* Randomly adding noise to images
* Randomly changing the brightness or contrast of images

Data augmentation is a powerful tool that can be used to improve the performance of deep learning models. However, it is important to use data augmentation judiciously, as it can also lead to overfitting if used excessively.

## 3.7 Convolution Neural Network

Convolutional neural networks are deep learning algorithms that are particularly well-suited for analyzing images. They are made up of a series of layers, where each layer is made up of a series of neurons. The first layer is the input layer, which is where the image data is fed into the network. The next layer is the convolutional layer, where the neurons are connected to each other in a way that allows them to learn to recognize patterns in the image data. The next layer is the pooling layer, where the neurons are connected to each other in a way that allows them to learn to recognize patterns in the image data. The final layer is the output layer, where the neurons are connected to each other in a way that allows them to learn to recognize patterns in the image data.

To this day, the Convolutional Neural Network (CNN) remains the most well-known and widely-used DL algorithm. CNN's key advantage over its forerunners is that it can detect important elements automatically, without any help from a human. Computer vision, audio processing, facial recognition, etc. are just a few of the many areas where CNNs have been put to use. CNNs, like traditional neural networks, take their structural cues from the neurons found in the brains of humans and other animals. More specifically, the visual cortex in a cat's brain is formed by a complicated sequence of cells, and this sequence is what the CNN attempts to emulate. To maximize the usage of 2D input-data formats like image signals, the CNN uses shared weights as well as local connections rather than typical fully connected (FC) networks.

The training process is greatly streamlined and the network's speed is increased due to the incredibly low number of parameters used in this operation. All of this mirrors what happens in the brain's visual cortex. Notably, these cells only perceive localized parts of a scene rather than the entire picture (i.e., they act like local filters over the input, spatially extracting the local correlation accessible in the input).

One popular kind of CNN, analogous to the MLP, features many convolution layers followed by sub-sampling (pooling) levels and finishing with FC layers.

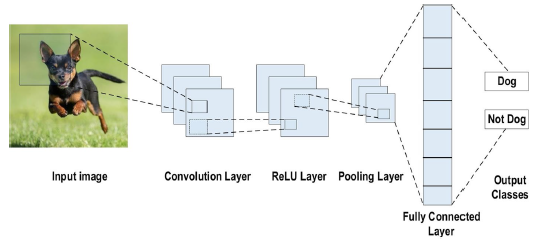


Figure CNN network

The above architecture is an illustration of CNN network, and the layers are explained below that helps to make the entire CNN algorithm.

### Convolution Layers

The convolutional layer is the primary building block of CNNs. A set of convolutional filters makes up the whole thing (so-called kernels). These filters are convolved with the input image (represented as N-dimensional metrics) to produce a feature map.

* **Kernel:** The kernel is described as a grid of numbers or values. Kernel weights are used to describe each individual numerical value. At the outset of the CNN training process, random numbers are chosen to serve as the kernel's weights. And the weights can be "set" in a number of ways. The weights are then fine-tuned at each training iteration, allowing the kernel to gradually extract more and more relevant characteristics (Alhussainy, 2020).
* **Convolution Operation:** At first, the CNN input format is explained. In contrast to the multi-channeled image that serves as the input for a CNN, the vector format is used by classical neural networks. Grayscale images, for example, are in the single-channel format, while RGB images are in the three-channeled format. Let's look at an example of a 4x4 grayscale image using a 2x2 random weight-initialized kernel to better grasp the convolutional operation. The kernel first moves horizontally and vertically across the entire image. Moreover, the dot product between both the input image as well as the kernel is calculated, where the values of both are multiplied by one another and then added to produce a single scalar result. Repeat this procedure until additional sliding is impossible (Alhussainy, 2020).

The picture provides a visual representation of the fundamental computations performed at each stage. The 2x2 kernel is depicted in this figure by the light green hue, whereas the similar-sized region of the input image is depicted by the light blue colour. After multiplying, the product values (highlighted in light orange) are added together to yield an input value for the output feature map.

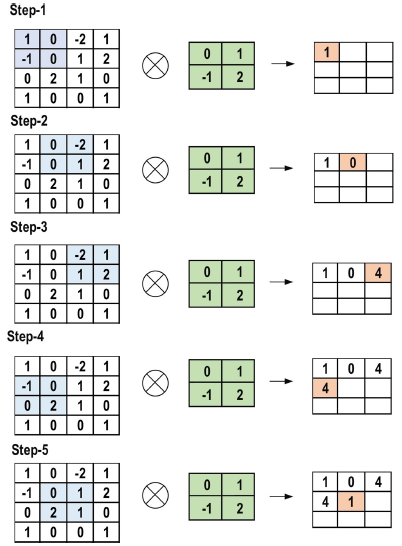


Figure Convolution layer

## Pooling Layers

Pooling layers are responsible for subsampling feature maps. It's the convolutional procedures that lead to these maps being created. To put it another way, this method reduces the size of existing feature maps. Also, it keeps most of the most important data (features) during the entire pooling process. Size assignments are made to the stride and kernel before the pooling procedure is carried out, much like they are during the convolutional operation. When it comes to pooling layers, there are a few different options. Methods such as global average pooling (GAP) and global max pooling (GMP) are also included in this category. Maximum, minimum, and GAP pooling are the three most common and well-known approaches to dividing up a resource (Kaur, 2020).

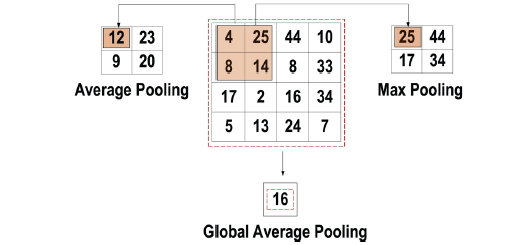


Figure Pooling layer

## Activation Function

For every given neural network, the primary role of the activation function is to map the inputs to outputs. The input value is calculated as the bias-adjusted weighted sum of the neuron's inputs (if present). This implies that the activation function generates the output that tells a neuron whether or not to fire in response to a certain input. After all weighted layers (so-called learnable layers, including such FC layers as well as convolutional layers) in a CNN design, non-linear activation layers are used. Because of the activation layers' non-linear performance, the mapping from input to output will also be non-linear, and the CNN will be able to learn more complex tasks because of it. It is crucial that the activation function be capable of differentiation so that error back-propagation may be utilised to train the network (XU, GUAN and CAI, 2017). CNNs and other forms of deep learning typically employ the following activation functions.

* **Sigmoid**

This activation function takes in real values as input, and returns a value between 0 and 1. The mathematical representation of the sigmoid function curve is an S-shape.



* **Tanh**

It takes real values as input, like the sigmoid function, but its output is only between -1 and 1. The following equation gives a symbolic representation of this.



In the context of CNN, this is by far the most popular feature. The supplied entire numbers are made positive. The primary advantage of ReLU over the others is the reduced computing load it requires. The equation for this is given below



## Fully connected layer

This layer typically appears last in CNN architectures. The commonly applied Fully Connected (FC) method involves linking every neuron in this layer to those in the one below it. It's the classifier that the convolutional neural network uses. As a feed-forward artificial neural network, it is based on the same fundamental principles as the more common multilayer perceptron. The output of the last pooling as well as convolutional layer feeds into the FC layer. Following flattening, the feature maps provide this input in the shape of a vector. According to the following diagram, the result of the FC layer is the final CNN output.

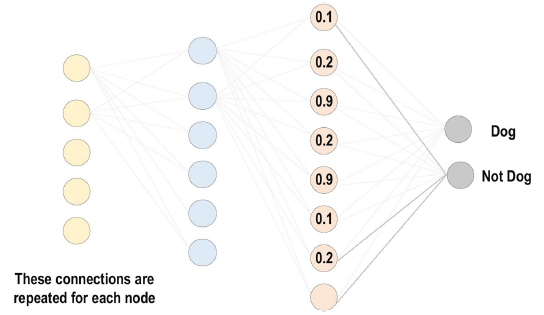


Figure Fully connected layer

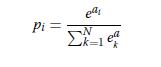
## Loss Functions

The loss functions are implemented the output units to determine the expected error created throughout the training in the data points present in the CNN model. This mistake displays the discrepancy between the actual result and the projected one. Afterwards, it will be optimized through the CNN process of learning (Liu et al., 2020).

However, two components are employed by the loss function to determine the error. The CNN predicted output is the initial parameter. The actual output is considered as the second parameter. Several kinds of loss function are applied in distinct problem categories. The following simply explains several of the loss function kinds.

* **Cross Entropy or Softmax**

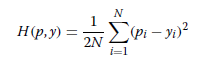
This metric is frequently used to evaluate the efficacy of a CNN model. The log loss function is another name for this metric. It gives the probability p = 0 | 1 | 2. It also often replaces the mean square loss function in multi-class classification tasks. The softmax activations are used in the output layer to produce a probabilistic output. This equation provides a mathematical expression for the likelihood of each output class (Kouretas and Paliouras, 2020).



N is the number of neurons in the output layer, and eai is the unnormalized output from the layer above it.

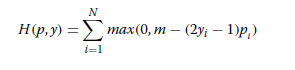
* **Euclidean Loss Function**

In the context of regression problems, this function is ubiquitous. The so-called mean square error is another term for this. In order to quantify the predicted Euclidean loss, we can use the formula below.



* **Hinge Loss function**

This operation is frequently used in binary classification issues. For SVMs, which employ the hinge loss function, wherein the optimizer endeavors to maximize the margin around dual objective classes, this issue is particularly pertinent because of the importance of maximum-margin-based classification. In the following equation, we can see its mathematical formula (Xing and Ji, 2018).



### Regularization to CNN

Over-fitting is the main problem when trying to achieve good generalization in CNN models. When a model performs exceptionally well on training examples but fails on testing set (unseen data), as will be detailed in greater detail below, it is said to be over-fit. In contrast, an under-fitted model is one that does not pick up enough information from the training data. If the model performs well on both the training data and the testing data, we say that it is "just-fitted." The figure below depicts these three categories.

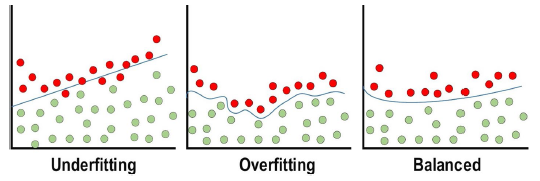


Figure Regularization to CNN

* **Dropout**

This method of generalization is commonly employed. Neurons are selectively eliminated at random throughout each training iteration. This forces the model to acquire a variety of independent features while also equitably distributing the feature selection power among the entire set of neurons. The removed neuron will not participate in back- or forward-propagation as training progresses. Contrarily, during testing, predictions are carried out using the full-scale network (Poernomo and Kang, 2018).

* **Drop-weights**

Dropout is a method that is very similar to this one. The main distinction underlying drop-weights as well as dropout is whether or not the neurons themselves are removed from the network during each training session.

* **Batch Normalization**

The efficacy of the final activations is guaranteed by this procedure. The results are consistent with a Gaussian with a width of one. To standardize the results at each level, we subtract the mean and divide by the standard deviation. This can be viewed as a preprocessing task at each network layer, and it can be differentiated from and integrated with other networks. It's also used to lessen the activation layers' "internal covariance shift." The internal covariance shift is defined by the difference in activation distribution between layers (Liang et al., 2020).

This variation becomes extremely high as a result of constant weight update during training, which may occur if training data samples are collected from a wide variety of unrelated sources (for example, day and night images). As a result, the model's convergence time will increase, and so will the training period. To address this problem, the CNN design incorporates a layer that stands for the batch normalization process.

Using batch normalization has the following benefits:

* As a result, the vanishing gradient issue is avoided.
* When properly implemented, it can prevent the bad initialization of weight from causing more problems.
* It drastically cuts down the time needed for a network to converge, which is very helpful for large-scale datasets.
* It has difficulty reducing hyper-parameter training dependence.
* **Optimizer**

The first important challenge is the choice of learning algorithm (optimizer), and the second key issue is the usage of several upgrades (such AdaDelta, Adagrad, and momentum) in conjunction with the learning algorithm to improve the output.

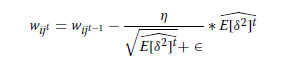
The goal of any supervised learning algorithm is to minimize an error measure (the difference between the actual as well as anticipated output) that is based on a loss function, which is itself based on a large number of learnable factors (such as biases, weights, etc.). The gradient based learning methods seem to be the go-to when building a CNN network. In order to minimize error, the network must search for the locally optimum solution across all training epochs and maintain constantly updated network parameters.

The rate of learning is described as the size of the updates to the parameters. The training epoch is a single repetition of the entire training dataset used for the parameter update. Although the learning rate is a hyper-parameter, it still needs to be carefully chosen so as not to have a negative impact on the learning process (Chen et al., 2020).

We have used Adam Optimizer to reduce the losses of the model.

It's yet another popular optimization method or learning process. Adam is the cutting edge of optimization strategies for deep learning. The Hessian matrix, which uses a second-order derivative, is used to express this. Adam is an approach to learning developed for the purpose of educating neural networks. Adam's benefits include reduced computational complexity and greater memory efficiency. Adam works by determining the value of adaptive LR for each model parameter. Combining the best features of Momentum and RMSprop. It is similar to momentum in that it takes a moving average of the gradient to determine the learning rate (RMSprop), but it uses squared gradients instead of a constant gradient.

Adam's equation can be written as follows.



## Benefits of Using CNN

Here are some of the advantages of employing CNNs in computer vision over more conventional neural networks:

* For one, CNN's weight-sharing feature makes it easier to train the network with fewer parameters, which in turn improves generalization and prevents overfitting.
* Second, the model's output is both well-structured and heavily reliant on the extracted features when they are learned alongside the classification layer simultaneously.
* Third, compared to other types of neural networks, CNN is significantly simpler to deploy in a network of a large scale.

## 3.8 ResNet-50

The 2015 ILSVRC champion, ResNet (Residual Network), was created by an author (Faiz Nashrullah, Suryo Adhi Wibowo and Gelar Budiman, 2020). The team's goal was to create a new type of ultra-deep network that will solve the vanishing gradient problem once and for all. Depending on the number of layers required, different ResNet variants were created (with layer counts ranging from 34 to 1202. ResNet50 was the most popular choice, and it has 49 convolutional layers and one FC layer. There were 25.5 million network weights and 3.9 million media access controls (MACs). ResNet's innovative approach is the usage of the bypass pathway concept, as depicted in the figure below, which was first used in 2015 by Highway Nets to solve the difficulty of training a deeper network. The figure depicts the basic ResNet block diagram, which demonstrates this. You can think of this as a regular feed forward network with an extra residual link. The (l-1)th outputs from the previous layer (xl-1) are the outputs from the residual layer. F (xl-1) is the final result after doing various processes [such convolution with variable-size filters or batch normalization before implementing an activation function such ReLU on (xl-1)].

The results of the model will be xl which can be mathematically deduced as



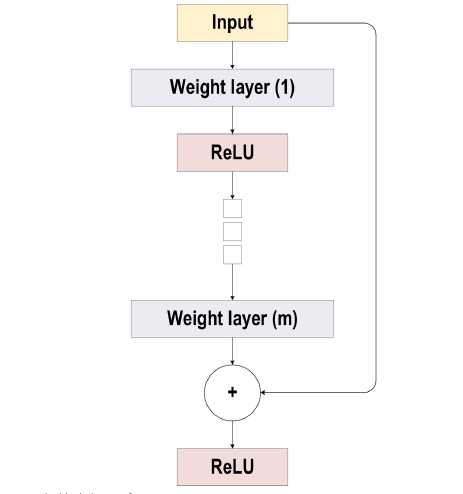


Figure Working of ResNet

ResNet offered shortcut connections inside layers to allow cross-layer communication, which do not contain any parameter and data-independent, in contrast to a highway network.

When a gated shortcut in a highway network is closed, the layers characterize non-residual functions. In contrast, in ResNet, the residual information is always transferred while the unique shortcuts are never closed. Since ResNet's shortcut connections (residual links) hasten the deep network's convergence, it may also be able to avoid gradient diminishing's pitfalls.

The 2015-ILSVRC champion network, ResNet, had 152 hidden layers, making it eight times as deep as VGG and twenty times as deep as AlexNet. Its computational complexity is less than that of VGG, even when its depth is doubled.

ResNet-50's architecture is inspired on the one seen up top, with one key modification. ResNet 50 is a 50-layer network with a bottleneck architecture. One way to limit the amount of parameters as well as matrix multiplications in a residual block is to utilize a "bottleneck," also known as 1x1 convolutions. With this, training time for each layer can be drastically reduced. Rather than the more common two levels, this one make use of a three-layer stack.

His 50-layer the elements of ResNet architecture are listed in the table below.

* Convolution of a 7x7 kernel with 64 additional kernels, each of which has a stride of 2.
* Maximum pooling layer with a stride length of 2.
* There are 9 more layers, including a 33, 64 kernel convolution, an 1x1, 64 kernel convolution, and an 1x1, 256 kernel convolution. In total, there are three iterations of these three levels.
* An additional 12 layers were added, each with a different size kernel: 1x1, 128 kernels, 3x3, 128 kernels, and 1x1, 512 kernels, each of which was repeated four times.
* There are now 18 layers, each with 1x1, 256 cores, plus 2 cores with 3x3, 256 and 1x1, 1024, for a total of 36 cores.
* Additional 9 levels were added, each with 1x1, 512, 3x3, 512, and 1x1, 2048 cores respectively.

(There are currently 50 levels in the network structure)

* A pooling average was created, and then a fully connected layer of 1000 nodes was used, using a softmax activation function.

The detailed architecture of the ResNet 50 layers is given below

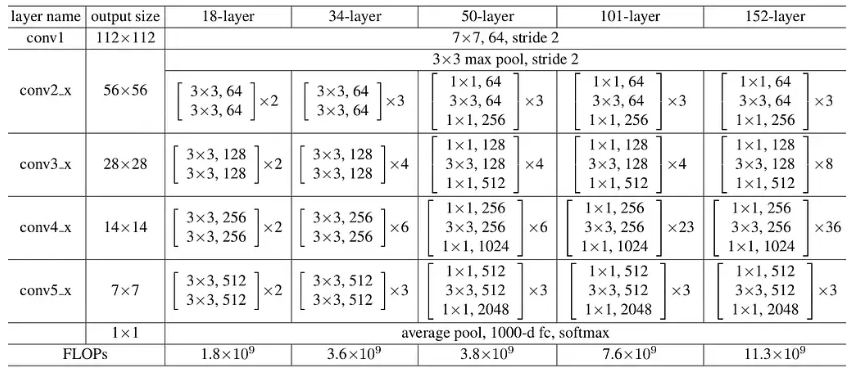


Figure Architecture of ResNet 50

## Skip Connection

This skip connection idea was pioneered by ResNet. The following diagram shows an example of a skip connection. It can be seen on the left that convolution layers are being stacked upon one another. It's still a stack of convolution layers on the right, but we're also including the original input in the final product. We refer to this as "skip connection."



Figure Skip connection

The benefits of skip connection of ResNet 50 is given below

1. As a result, they reduce the occurrence of the vanishing gradient problem by opening up this shortened gradient flow route.
2. They permit the model to acquire an identity function that guarantees the superior layer will perform no worse than the inferior layer.

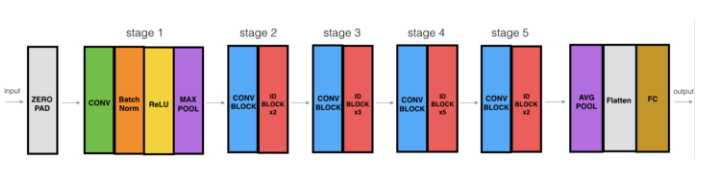


Figure Stages of ResNet 50

Each of the 5 stages in the ResNet-50 model is made up of a convolution as well as Identity block. There are three convolution layers in each convolution block, and the same number of convolution layers in each identity block. The number of adjustable settings in the ResNet-50 is far over 23 million.

## Benefits of using ResNet 50

To the output of a series of convolutional layers, the input is added via a skip connection in a residual block. Given the difference in dimensions between input and output, a skip connection is not possible when the layers employ pooling or strided convolutions.

Despite having 50 deep convolutional layers ResNet 50 still out performs other techniques as because the weights are already trained and it perform skip connection skips some of the layers and reduces the training time.

## 3.9 Model Evaluation

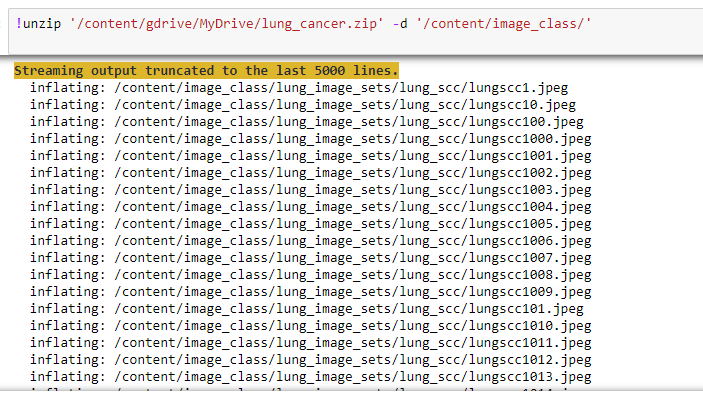
The model evaluation depends upon the nature of the data where the outcome is either a continuous or a discrete one. Since our research contain a discrete outcome which we need to classify three types of images, we will evaluate the model with the help of confusion Matrix. A confusion matrix is a diagonal Matrix that sum of with the correctly classified in which classified labels between the prediction and the actual outcome. The diagonal of the matrix represents the correct classification and all the samples other than the diagonals represent the misclassified samples. The other Evaluation metrics that will be used are precision, recall and the F1 score metrics (Shankar, 2019). The Precision metrics is used to determine the rate of false positive samples and recall metrics is used to determine the rate of false negative samples. The combination of Precision and recall is the F1-score which will be the final evaluation metric used in our research.

# Implementation

## CNN implementation

For implementation of CNN algorithm we hand used tensor flow and Keras library and implemented both CNN and ResNet 50 algorithms with the help of Google Colab pro. Google Colab pro is used as it offers higher GPU which takes less time in training of the algorithms.

## Data Exploration

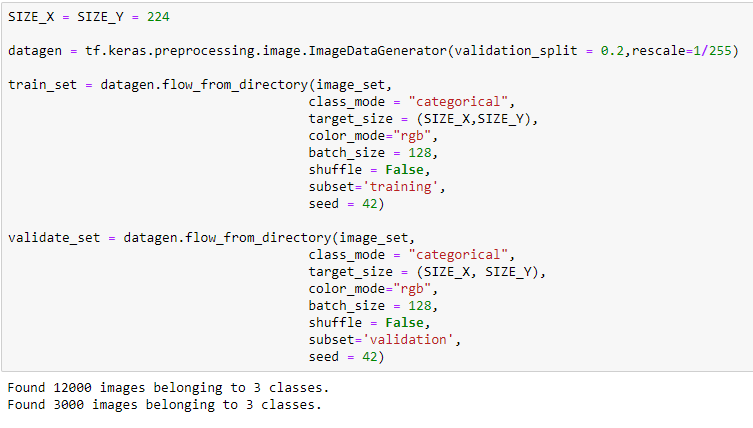


We have first uploaded the data to Google drive and then we have mounted the Google drive with Google Colab pro. After that we have imported all the normal images along with adeno carcinoma and squamous cell carcinoma.

## Data Augmentation

Data augmentation is a process of using transformation to the images where images are increased compared to the original samples. We have used Imagedatagenerator from tensor flow to augment the images and we have normalized the images by rescale and taken the validation split of 0.2 which indicates 20% of the data is taken for validation.

The images are there defined into a definite shape of 224 by 224 where we have found 12000 images belonging to 3 classes and 3000 images belonging to 3 classes in the validation set.



The colour mode of the images of set to ‘rgb’ which indicates all the images in the data are colored images and different subset of training and validation set is taken which indicates train\_set will take the training data a validate\_set will take the validation data.

## Building of Models

We have used sequential library from Keras and activation, dropout, flatten, Dense, Conv2D, MaxPooling2D and BatchNormalization from Tensorflow and Keras is imported to perform different layers of the model.



The architecture of the model looks like the following



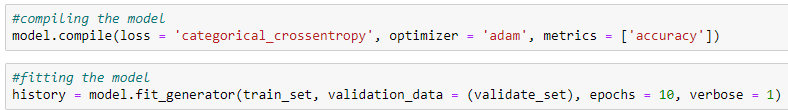
The summary give the following number of parameters in the given CNN model.

|  |  |
| --- | --- |
| Total parameters | 1,813,379 |
| Trainable parameters | 1,811,907 |
| Non-trainable parameters | 1,472 |

Table Parameters of CNN model

The non-trainable parameters are less compared to the trainable parameters.

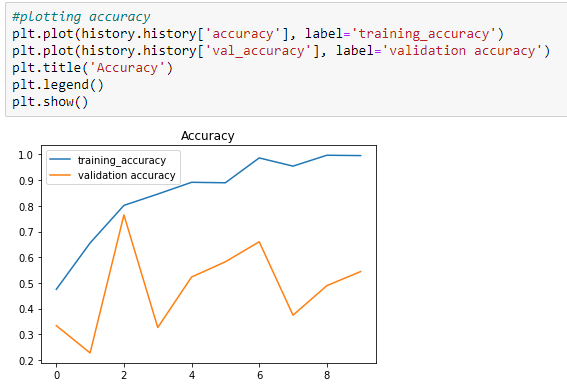
The model is compiled with the help of ‘categorical\_crossentropy’ as the outcome is categorical in nature. Also we have used Adam Optimizer and accuracy metrics to monitor the accuracy in training and validation data. We then fit the model with the help of fit\_generator where we set epochs to 10 so that the model iterates over 10 times.



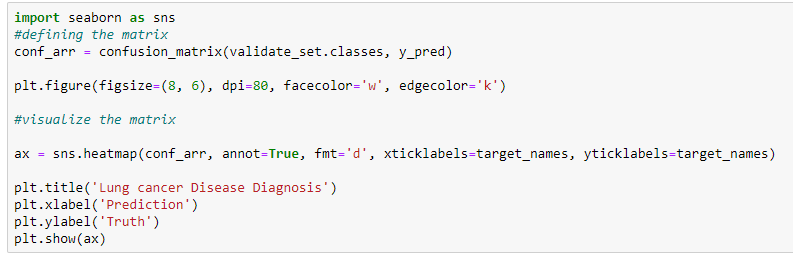
## Evaluation of Models

The model is evaluated with the help of model loss and accuracy plots and using confusion Matrix as well as classification report.

We have used matplotlib library to reproduce model loss and model accuracy in training data to look into the effect of accuracy over different epochs.



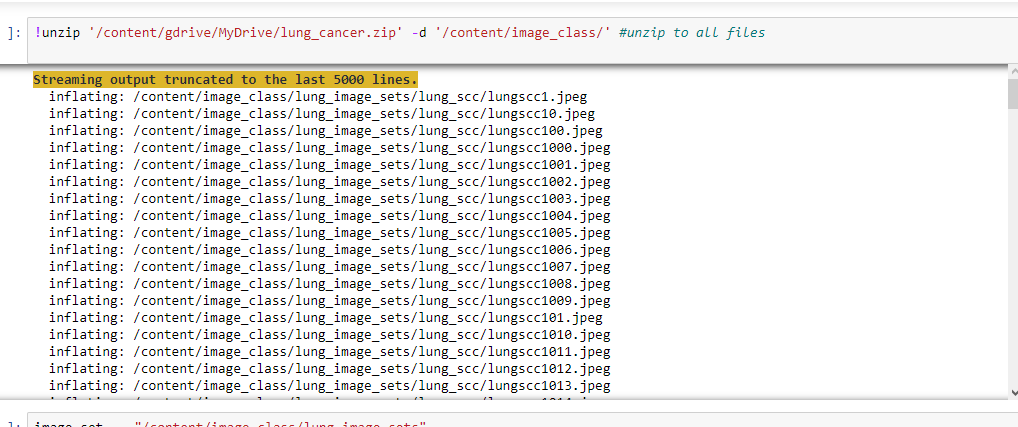
For confusion Matrix we have predicted the test data outcome with the help of our model and used seaborn library to visualize the confusion matrix where we made a comparison between the actual and the predicted outcome of the test data.



## ResNet 50 Implementation

## Data Exploration

The data collected for the ResNet 50 model is similar to the technique compared to the CNN model. We have imported the data in Google Drive and then mounted the Google drive with Google Colab Pro to import the data.



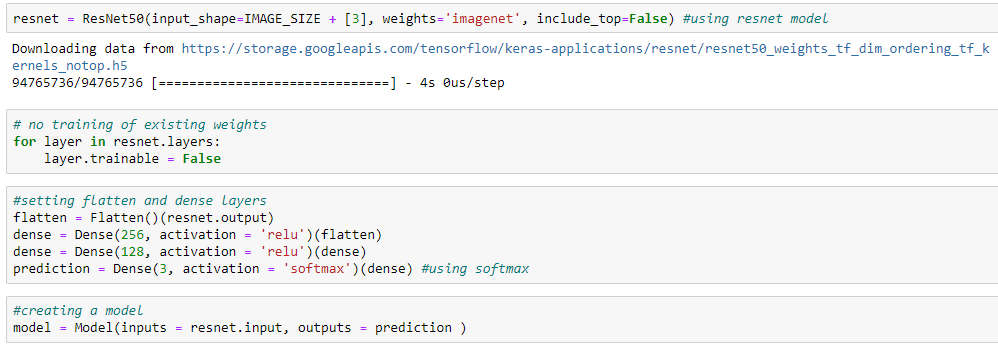
## Data Augmentation

The data augmentation is performed with the help of ImageDataGenerator where the validity data is prepared from the original data at a ratio of 0.2 which indicates 20% of the validation data is splitted. However data augmentation process is similar to the CNN process and the image shape taken is also similar to the CNN model.



The training data contain a sample of 12000 images and the test data contain a sample of 3000 images that belongs to 3 classes.

## Building of models

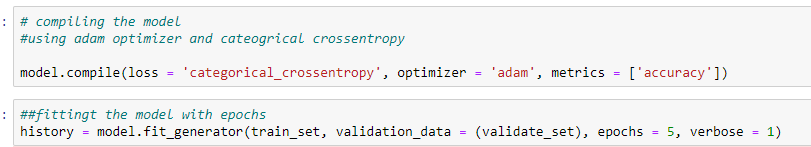


The ResNet 50 model is built where the weights are pre reloaded from the ‘imagenet’ function. The existing weights are not trained as the layer is said to False which will reduce the training time. There are two dense layers added with an outcome layer with softmax activation function as the outcome is categorical in nature.

|  |  |
| --- | --- |
| Total params | 49,311,363 |
| Trainable params | 25,723,651 |
| Non-trainable params | 23,587,712 |

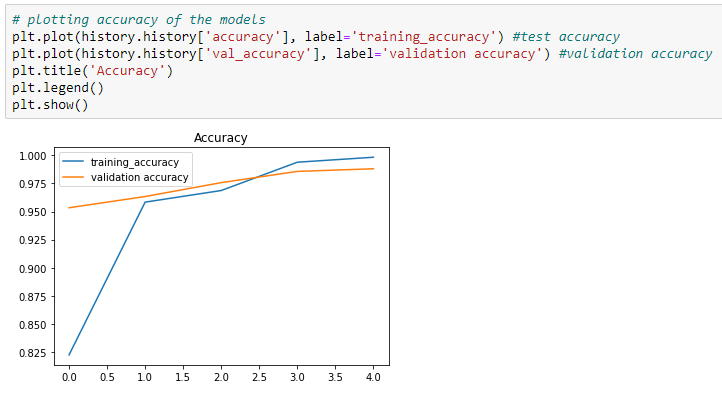
Table Parameters of ResNet 50 model

The parameters of ResNet 50 model is much higher compared to the CNN model but there are large number of non-trainable parameters that are not required to be trained which will reduce the training time compared to the CNN model.

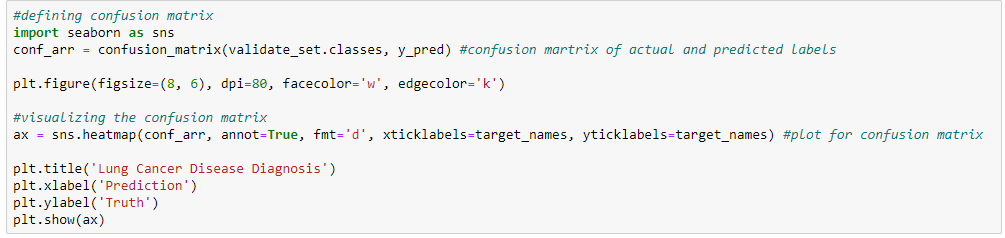


The model is compiled with a loss function of **categorical\_crossentropy** and Optimizer of Adam to reduce the loss function. Also the model is fitted with an epoch of 5 whereas in the CNN model is trained with 10 epochs.

## Model Evaluation



The model loss and model accuracy is plotted with the help of matplotlib library where we can evaluate the model at different epochs. We can see that the accuracy increases after Second epochs but the loss stayed stable after second epochs.



The model is evaluated using a confusion Matrix where it is predicted with a predict function and the confusion matrix is visualized with the help of seaborn library. A heatmap is used that distinguish between the mis classification of true value and the predicted value.

# Results and Analysis

The results of the lung disease prediction of both CNN and ResNet 50 model had been evaluated using confusion Matrix, model accuracy and loss and classification report. Let us analyze the results generated by both the models.

## 5.1 CNN model performance

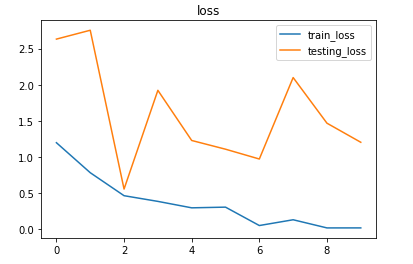


Figure Model loss CNN

The above plot indicates the training and validation loss of the CNN model where the model is trained with 10 epochs and we can see that the testing loss is minimum in second epoch and it increased from second epoch. So we can say that training the model with 2 epochs is an ideal solution to get better accuracy. Also the train loss after second epochs started to decrease.

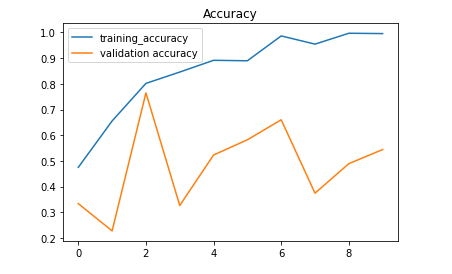


Figure Model accuracy CNN

The above plot indicates the accuracy where we can see that the validation accuracy decreased after two epochs and the training accuracy increased after 2 epochs. This shows that after training the CNN model with 2 epochs, the model starts to give an over fitting result where there is a high training accuracy but low validation accuracy.

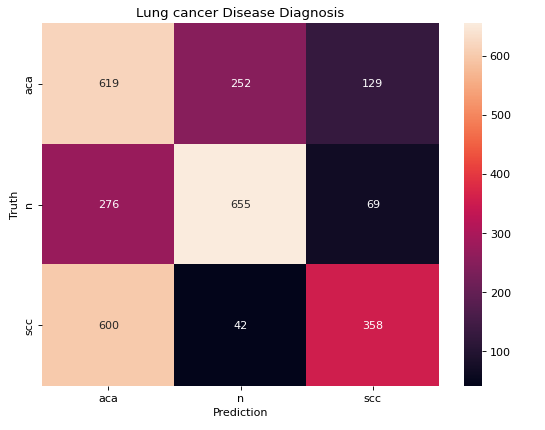


Figure Confusion matrix of CNN

The confusion Matrix above gives an idea of the mis classification given by the CNN model. From the model we can see that there are 619 adenocarcinoma images which are correctly classified, 655 normal images which are correctly classified and 358 squamous cell carcinoma which are correctly classified.

In the test data there were total of 1000 samples in each classes which indicates that normal images gave the least mis classification by the CNN model and there is a high mis classification observed in predicting squamous cell carcinoma images.

The mis classification also tells that there are a total of 600 squamous cell carcinoma images which are predicted as adeno carcinoma images.

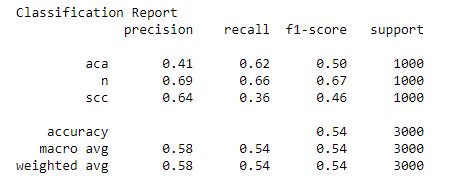


Figure Classification report of CNN

The above classification report indicates that the model is tested on thousand adenocarcinoma images, 1000 normal images and 1000 squamous cell carcinoma images. The average score is 0.54 where the model gives the highest accuracy in predicting normal images which is 0.67.

* The precision rate of normally images is high and then the recall which indicates that there is lower false positive rates compared to the false negative.
* The lowest accuracy is achieved in predicting squamous cell carcinoma images where the precision is much higher which indicates that there are higher false negative rates compared to the false positive.
* The recall rate in predicting adeno carcinoma images is higher than the Precision which indicates that there is a lower false negative rates compared to the false positive.
* The average Precision is higher than the recall which indicates the model gives higher false negative rates compared to the false positive in predicting the images in 3 classes.
* The overall accuracy is 0.54 by the CNN model which also indicates that the model can predict squamous cell carcinoma with 46% accuracy and adeno carcinoma with 50% of accuracy.

## 5.2 ResNet-50 model

The ResNet 50 model had also been evaluated similar to the evaluation metrics used in predicting the images by the CNN model. We have used confusion Matrix, classification report, model loss and model accuracy plots to evaluate the model.

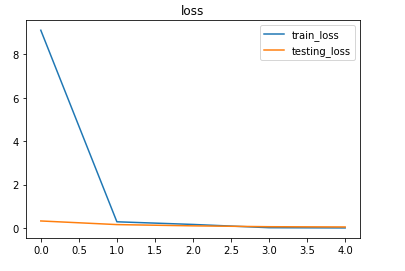


Figure Model loss ResNet-50

The model lose plots indicates that the train loss was quite high during the first epochs but from the second epochs, the train loss decreased and became similar with the test loss. The model did not improve after two epochs so it indicates that training the model with 2 epochs is sufficient in achieving the same accuracy. This will reduce the training time and also minimize the chance of over fitting at later epochs.

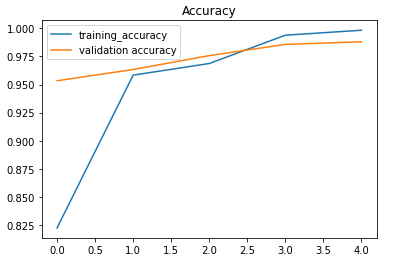


Figure Model accuracy ResNet 50

The model accuracy plots also indicates that the training accuracy increased after two epochs and also the validation accuracy tends to increase after 2 epochs. Since the model performance did not improve much after 2 epochs which is indicated from the loss plot. Also we can say that training the model with two epochs is sufficient as the validation accuracy again decreased after 3 epochs that indicates a good chance of over fitting.

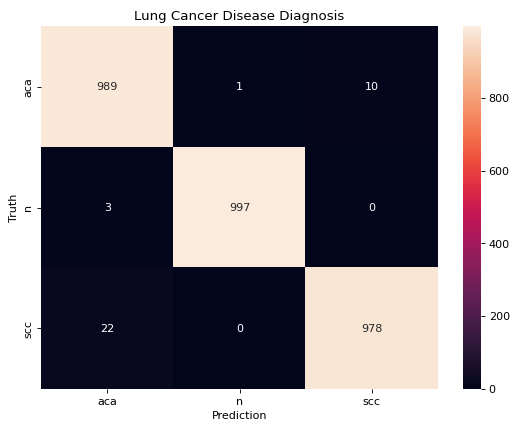


Figure Confusion matrix of ResNet 50

The model is predicted on 1000 carcinoma images, 1000 normal images as well as 1000 squamous cell carcinoma images. We can see that the model gave correct classification in predicting 989 adenocarcinoma images, 997 normal images and 978 squamous cell carcinoma images. So the highest accuracy is given in predicting normal images and the lowest accuracy is achieved in predicting squamous cell carcinoma images.

All the mis classified squamous cell carcinoma images are predicted as adeno carcinoma images similar to the scenario observed in predicting by the CNN model. This indicates that there are some patterns which resemble in both adeno carcinoma and squamous cell images in some of the samples which is why the model fails to detect such patterns and mis classifies squamous cell carcinoma and adeno carcinoma.

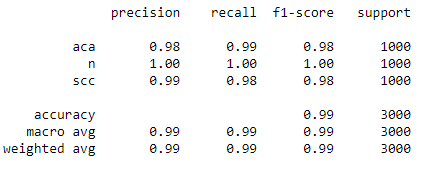


Figure Classification report of ResNet 50

The classification report above indicates that the model is tested on 1000 images present in each class and the average F1 score of each image is 0.99 which indicates the model is 99% accurate in predicting lung cancer.

* In case of Carcinoma images the recall rate is higher than Precision which indicates that false negative rate is lower than the false positive rate.
* In case of normal images the Precision and the recall rate is almost the same
* In squamous cell carcinoma images the precision rate is higher than recall rate which indicates lower false positive than false negative images
* The highest accuracy given by the model is predicting normal images and almost similar F1 score is achieved in predicting adenocarcinoma and squamous cell images.
* The performance indicates that ResNet 50 is much better model compared to CNN where the CNN gave almost 50% accuracy and ResNet 50 is giving an accuracy of around 99% which is a good improvement seen in using residual network in predicting lung cancer images.

## 5.3 Parameters Comparison

The parameter comparison indicates that the total parameters to train CNN model is 1813379 and the total parameters to ResNet 50 model is 49311363 which is much higher compared to the CNN model. So the training time taken by the ResNet 50 model will obviously be higher as it contains higher trainable parameters compared to the CNN model.

The performance indicates that CNN technique is not suitable approach to predict lung cancer as it gives such lower accuracy in predicting adeno carcinoma and squamous cell images.

# Discussion

This research is focused on building two Deep learning network such as CNN model and ResNet 50 model in predicting lung cancer where the image tissues are collected from adeno carcinoma and squamous cell carcinoma tissues.

For improving the accuracy we have implemented data augmentation techniques and to avoid over fitting we have also used batch normalization and dropout techniques in CNN while training the model.

We have gathered histopathological lung cancer images from Google where we augmented the total number of 12000 training images belonging to 3 classes and 3000 test images belonging to 3 classes.

With the help of Tensorflow and Keras libraries we have implemented the CNN model where we have developed a CNN model with 5 convolution layer, 5 pooling layers and additional batch normalization layers with two drop out layers where also we used 3 dense layers and Adam Optimizer to reduce the loss function.

For implementation of ResNet 50 model we have updated the weights with the Imagenet library where the image sizes are defined with the definition shape of 224 by 224 and trained on the augmented data which is performed with the help of preprocessing ImageDataGenerator library of Tensorflow.

For optimizing the loss function, we have used similar Adam function compared to the CNN model and categorical cross entropy as the loss function is the outcome which are categorical in nature.

Also we applied different activation function such as rectified linear unit and Softmax and had been applied in CNN model and in case of ResNet 50 model we have applied the same activation function with dense layer of size 256 and 128 in case of ResNet 50 model.

## 6.1 Model Implementation

The models a run in Google Colab pro where we have used high GPU and RAM to execute both the models faster but it took around 4 hours to build both models to train in a total of 12000 lung cancer images belonging to three classes. At first we have experimented with the help of Anaconda navigator in local system but the estimated time of training both the models was much higher which is why we have opted to choose Google Colab pro that offers high RAM and GPU to execute the models faster.

There was another problem which was encountered during training the model in Google Colab pro where the file do not save automatically if the internet connection gets cut during training the model. This is why we have ensured high level internet connection to run the model in Google Colab pro as frequent interruption in the middle of the training delayed the experimentation of the model using CNN and ResNet 50.

We have repeatedly trained the CNN model we have achieved much lower accuracy as we have trained the model with higher epochs. However after using drop out and batch normalization, we have achieved better accuracy but it was not enough as there was high mis classification in adeno carcinoma and squamous cell carcinoma images. This is why we have opted to use a Resnet 50 model where the weights are already trained from Imagenet library and it was a winner in most of the competition and was helpful in building various benchmark models.

## 6.2 Research Investigation

We conducted two research questions to determine the techniques that make deep learning models really work in lung cancer detection. For this purpose we have developed data augmentation techniques to improve the accuracy and manipulate the images with the help of several transformations. After that, we have used drop out and batch normalization techniques to prevent over fitting and increase the accuracy in prediction of Cancer images. There are some other technique such as using of different optimizers to reduce the loss function, using of different activation function and using of different Nobel architecture models that will help increase of the models that can be helpful in lung cancer detection.

We achieved an overall 54% accuracy with the help of CNN model which is why we have trained a residual network like ResNet 50 model where the weights are already pre trained and we have found that ResNet 50 really work in lung cancer detection as it gives 99% accuracy in predicting the test images.

The second Research question was to evaluate the parameters which are effective in accurate prediction of lung cancer using deep learning methods. So we have experimented with different parameters but we have seen the most effective parameters are the optimizers that responsible to minimize the Loss function as well as the use of dropout technique that is helpful to deactivate the neurons and prevent over fitting. Also in order to reduce the computation time, batch normalization techniques is found to be effective and using high end system with high RAM and GPU also tend to decrease the training time of the model.

In data augmentation techniques that are several parameters that are really effective such as zooming the images, adjusting brightness of the images and several transformations that makes a copy of the original images and increase the size of the images that really work in accurate prediction of lung cancer.

The image shape also play a pivotal role as cropping the image shape into minimum size will also decrease the importance of the images as there is a possibility of cropping some important patterns from the images that will reduce the performance of the model.

# Conclusion

For many years now, lung cancer was by far the most lethal form of the disease. The rate of diagnosis climbed by 37% between 2007 and 2017, and in 2022, lung cancer was responsible for almost 131,000 fatalities in the United States alone (US). However, most new occurrences of lung cancer are identified in the latter stages of the disease, when treatment options are limited and the prognosis is dismal (20.5% 5-year survival rate). Computed tomography (CT) is the gold standard for diagnosing and screening for lung cancer because it is widely accessible, relatively inexpensive, and produces images with the highest possible spatial resolution

There are various diagnosis procedure to detect lung cancer where CT Scan is the most common used to technique to detect lung cancer. Also there are other techniques of achieving pathological images of lung cancer which are expensive for most of the patients cannot afford such procedure. This is why deep learning techniques can be important in detecting lung cancer from the image data where they can learn the patterns to distinguish normal and malignant lung tissues. This research is aimed to classify normal and two types of lung cancer which are adeno carcinoma and squamous cell carcinoma with the help of Convolution Neural Network and ResNet 50 deep learning method. We have luckily obtained the images from Kaggle that contain a total of 15000 images where there are equal samples of images divided into 3 classes.

During preprocessing of the model we have performed data augmentation technique where we have applied transformations to the training images to increase the size of the original images. The data augmentation technique was done with a validation split of 20% and all the images are normalized by dividing all the pictures with 255. The image shapes are obtained with the definite size of 224 by 224 and a batch size of 128 is applied to train the model.

The first model was a CNN model that contain 5 convolution layers with 5 Max pooling layers and 5 batch normalization layers. We also applied dropout to the last two layers where 10% of the neurons are deactivated to prevent over fitting and improve the performance. We also applied 3 dense layers with a softmax activation function and Adam Optimizer to reduce the loss function. The models are trained and then are predicted in the test data with 10 epochs where we can see that the test loss is lower in two epochs and then it increased after two epochs.

The model is then evaluated with the help of confusion Matrix where three types of lung images are classified and we can see that the model give an overall 54% accuracy in predicting various types of lung cancer images. The model gave high miss classification of squamous cell carcinoma into adeno carcinoma which indicates there is a similar pattern observed between these two types of cancers.

The second model was a ResNet 50 model which was a residual network where the weights are already trained on Imagenet library. The model is used and evaluated on the test data where the model is trained with 5 epochs and tested in the validation data. The model is evaluated with the help of confusion Matrix and we have seen the average score of the model is 99% where only 22 squamous cell images are miss classified as adeno carcinoma and 10 adeno carcinoma images are miss classified as squamous cell carcinoma. The rate of mis classification in these two types indicates that the model cannot predict the patterns of these miss classified samples which is why there is a need to update the data of these two classes with different patterns that can be easily distinguished.

The effect of batch normalization and dropout cannot be observed as the CNN model gave much lower accuracy which cannot be considered in predicting lung cancer in real life situation. So the ResNet 50 model is the best model to predict lung cancer as it gives 99% accuracy which is almost difficult to defeat by any other baseline models.

The CNN model was first applied without using drop out and batch normalization which give much lower accuracy to around 35% which was further increased with the help of dropout techniques. So we can easily interpret that dropout and batch normalization techniques are really effective in building CNN models where the accuracy can be improved after deactivating a part of the neurons in CNN model.

## 7.1 Limitations of the Research

We have encountered the following limitations during implementation of the research.

1. With the help of ResNet 50 model, we have achieved an accuracy of 99% but there is a high chance of over fitting if the test accuracy decreases than the train accuracy at higher epochs. This is why it is advisable do not to train the model at high epochs or we can set a call back function that will deactivate the training at similar accuracy in two epochs.
2. Data augmentation is effective with balanced classes but in case of imbalanced samples after updating the data, there is a high chance of underperformance as the model will give lesser accuracy after retraining the model on the updated data.
3. The model trained on these images is only suitable to predict adeno carcinoma and squamous cell carcinoma in histopathological images in real situations. The model is limited to predict lung cancer in any other type of images as there is a need of updating and retraining the model with other kind of images if they are needed to be predicted in real life applications.

## 7.2 Future Works

The following steps of improvements are recommended as a part of the future work of this research

* The CNN model gave worse performance which is why there is a need of updating the data in future only in case of adeno carcinoma and squamous cell carcinoma with different distinguishing patterns.
* For better performance in CNN model we can use other techniques such as avoiding pooling layers to extract the original and raw features from the images. This might help in improving the accuracy in predicting lung cancer images.
* The model made by ResNet 50 algorithm can be deployed and additional requirements must be added to state that the model is only suitable in detecting lung cancer from pathological images.
* The other works include training the model with lightweight architecture such as Inception and VGG19 to see if the training time improves over CNN and ResNet 50 model.

# References

Alhussainy, A.M.H. (2020). A New Pooling Layer based on Wavelet Transform for Convolutional Neural Network. Journal of Advanced Research in Dynamical and Control Systems, 24(4), pp.76–85. doi:10.5373/jardcs/v12i4/20201420.

Bargotya, M., Das, P., Bhardwaj, M. and Aeron, T. (2019). Basaloid Squamous Cell Carcinoma of Oesophagus: Definitely Rare but does it Differs Significantly from the Conventional Squamous Cell Carcinoma. International Journal of Contemporary Medical Research [IJCMR], 6(10). doi:10.21276/ijcmr.2019.6.10.7.

Chen, Y., Wen, Z., Tang, K. and Li, W. (2020). SSE Composite Index Forecasting Model via BP Neural Network with ADAM Optimizer. International Journal of Computer Applications Technology and Research, 9(1), pp.008-014. doi:10.7753/ijcatr0901.1002.

Faiz Nashrullah, Suryo Adhi Wibowo and Gelar Budiman (2020). The Investigation of Epoch Parameters in ResNet-50 Architecture for Pornographic Classification. Journal of Computer, Electronic, and Telecommunication, 1(1). doi:10.52435/complete.v1i1.51.

Denning, S. (2015). Updating the Agile Manifesto. *Strategy & Leadership*, 43(5). doi:10.1108/sl-07-2015-0058.

Hassan, U., Mozayani, B. and Wong, N.A.C.S. (2015). Primary combined neuroendocrine carcinoma (small-cell type) and squamous cell carcinoma of the colon. Histopathology, 68(5), pp.755–758. doi:10.1111/his.12786.

Kaur, K. (2020). Improved Methodology for Mammography Images Classification by Convolution and Pooling Layers with SVM Kernel base Classifier. Journal of Advanced Research in Dynamical and Control Systems, 12(SP4), pp.163–173. doi:10.5373/jardcs/v12sp4/20201478.

Kouretas, I. and Paliouras, V. (2020). Hardware Implementation of a Softmax-Like Function for Deep Learning. Technologies, 8(3), p.46. doi:10.3390/technologies8030046.

Liang, S., Huang, Z., Liang, M. and Yang, H. (2020). Instance Enhancement Batch Normalization: An Adaptive Regulator of Batch Noise. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), pp.4819–4827. doi:10.1609/aaai.v34i04.5917.

Liu, P., Shi, L., Miao, Z., Jin, B. and Zhou, Q. (2020). Relative Distribution Entropy Loss Function in CNN Image Retrieval. Entropy, 22(3), p.321. doi:10.3390/e22030321.

Petrella, F. (2021). Diagnosis and Treatment of Primary and Secondary Lung Cancers. Cancers, 13(3), p.448. doi:10.3390/cancers13030448.

Poernomo, A. and Kang, D.-K. (2018). Biased Dropout and Crossmap Dropout: Learning towards effective Dropout regularization in convolutional neural network. Neural Networks, [online] 104, pp.60–67. doi:10.1016/j.neunet.2018.03.016.

Shankar, P.M. (2019). Pedagogy of Bayes’ rule, confusion matrix, transition matrix, and receiver operating characteristics. Computer Applications in Engineering Education, 27(2), pp.510–518. doi:10.1002/cae.22093.

Vecerzan, L. and Mihaila, R.G. (2016). Ovarian Adenocarcinoma with Serous Effusions Appeared after a Mammary Adenocarcinoma: A Case Report. Journal of Adenocarcinoma, 01(01). doi:10.21767/2572-309x.100004.

Xing, H.-J. and Ji, M. (2018). Robust one-class support vector machine with rescaled hinge loss function. Pattern Recognition, 84, pp.152–164. doi:10.1016/j.patcog.2018.07.015.

XU, D., GUAN, Y. and CAI, P. (2017). Periodic Function as Activation Function for Neural Networks. DEStech Transactions on Computer Science and Engineering, (aita). doi:10.12783/dtcse/aita2016/7565.

Alakwaa, W., Nassef, M., & Badr, A. (2017). Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN). *International Journal of Advanced Computer Science and Applications*, *8*(8). https://doi.org/10.14569/ijacsa.2017.080853

Aonpong, P., Iwamoto, Y., Han, X.-H., Lin, L., & Chen, Y.-W. (2021). Genotype-Guided Radiomics Signatures for Recurrence Prediction of Non-Small Cell Lung Cancer. *IEEE Access*, *9*, 90244–90254. https://doi.org/10.1109/ACCESS.2021.3088234

B, K. S. (2020). Prediction of Lung Cancer Using Convolutional Neural Network (CNN). *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3), 3361–3365. https://doi.org/10.30534/ijatcse/2020/135932020

Du Parcq, P., Harper, J., & Khorashad, J. (2019). Collaborative genomics for improved patient care; a lung cancer case report. *Lung Cancer*, *127*, S2–S3. https://doi.org/10.1016/s0169-5002(19)30049-2

Ferguson, S., Mansoor, W., & Talbot, D. (2018). The incidence, diagnostic pathway and management of pulmonary carcinoid tumours in the UK: results from the National Lung NET pathway (“LEAP”) Project. *Lung Cancer*, *115*, S17–S18. https://doi.org/10.1016/s0169-5002(18)30069-2

Kadouri, L., Rottenberg, Y., Zick, A., Hamburger, T., Lipson, D., Peretz, T., & Nechushtan, H. (2019). Homologous recombination in lung cancer, germline and somatic mutations, clinical and phenotype characterization. *Lung Cancer*, *137*, 48–51. https://doi.org/10.1016/j.lungcan.2019.09.008

P, S. (2021). Lung Cancer Detection in Radiology Images using CNN. *International Journal for Research in Applied Science and Engineering Technology*, *9*(VII), 2471–2475. https://doi.org/10.22214/ijraset.2021.36886

Pang, S., Zhang, Y., Ding, M., Wang, X., & Xie, X. (2020). A Deep Model for Lung Cancer Type Identification by Densely Connected Convolutional Networks and Adaptive Boosting. *IEEE Access*, *8*, 4799–4805. https://doi.org/10.1109/access.2019.2962862

Peng, T., Wang, C., Zhang, Y., & Wang, J. (2022). H-SegNet: hybrid segmentation network for lung segmentation in chest radiographs using mask region-based convolutional neural network and adaptive closed polyline searching method. *Physics in Medicine & Biology*, *67*(7), 075006. https://doi.org/10.1088/1361-6560/ac5d74

Strauss, G., Flores, J. P., & Dominioni, L. (2015). Computed Tomography (CT) and Chest X-ray (CXR) Screening for Lung Cancer (LC): Mortality (MORT), Survival (SURV), and Randomized Population Trials (RPTs) - Analysis of the Mayo Lung Project (MLP) and the National Lung Screening Trial (NLST). *Chest*, *148*(4), 554A. https://doi.org/10.1378/chest.2264144

Su, Y., Li, D., & Chen, X. (2020). Lung Nodule Detection based on Faster R-CNN Framework. *Computer Methods and Programs in Biomedicine*, 105866. https://doi.org/10.1016/j.cmpb.2020.105866

van de Kamp, H. J., Molder, M. te, Schulkes, K. J. G., van Elden, L. J. R., & Hamaker, M. E. (2019). Impact of non-small cell lung cancer treatment on cognitive functioning. *Clinical Lung Cancer*. https://doi.org/10.1016/j.cllc.2019.06.006

Yu, H., Zhou, Z., & Wang, Q. (2020). Deep Learning Assisted Predict of Lung Cancer on Computed Tomography Images Using the Adaptive Hierarchical Heuristic Mathematical Model. *IEEE Access*, *8*, 86400–86410. https://doi.org/10.1109/access.2020.2992645

Zhang, Q., & Kong, X. (2020). Design of Automatic Lung Nodule Detection System Based on Multi-Scene Deep Learning Framework. *IEEE Access*, *8*, 90380–90389. https://doi.org/10.1109/access.2020.2993872