Lung Cancer Histopathological Image Classification Using Deep Learning Models

Abstract:

Lung cancer has the greatest morbidity and mortality rates among malignant tumours worldwide. Screening for lung cancer has been studied for decades to minimise lung cancer patient mortality rates, and treatment options have improved considerably in recent years. Pathologists use a variety of methods to diagnose the stage, type, and subtype of lung cancer, but one of the most prevalent is a visual examination of histopathology slides. Adenocarcinoma and squamous cell carcinoma are the most prevalent subtypes of lung cancer, and distinguishing between them requires visual inspection by a qualified pathologist. The goal of this paper is to create a deep learning model for categorising lung histopathology images by developing a neural network which will be able to classify the images into the different classes i.e., the types of lung cancer. This research uses the LC25000 Lung and colon histopathology image dataset, which contains 5,000 digital histopathology photos labelled as benign (normal cells), adenocarcinoma, and squamous carcinoma cells (both malignant cells). Using this dataset, the features are extracted from the images using various techniques and are fed to the neural network model which gives the highest accuracy of 96%. When compared to existing approaches, the findings obtained utilising the proposed method of classification based on Deep Learning models reveal that it can differentiate lung cancer variants with fewer characteristics and less computational complexity. They also demonstrate that using transformation methods to minimise features can provide a better interpretation of the data, hence enhancing the diagnosis procedure.

Keywords: Lung Cancer Diagnosis; Histopathology Images; Deep Learning; Neural Networks

Introduction:

Cancer develops because of the uncontrolled multiplication of aberrant cells in the organs or tissues of the body. Cancer cells can develop in a variety of organs or tissues throughout the body. Cancer is the major or secondary cause of death before the age of 70 in 112 countries, according to World Health Organization (WHO) estimates for 2019 [1,2]. Furthermore, according to a 2020 report by the International Agency for Research on Cancer

(IARC), cancer is the major or secondary cause of death in 134 countries. There are over 200 different forms of cancer [4]. According to statistical research conducted in the United States, lung and colon cancers are expected to rank among the top three most common cancer forms in 2020. According to the study, among all cancer diagnoses in the United States, Patients with lung and colon cancers have the greatest mortality rates in 2020 [5]. According to GLOBOCAN 2020 statistics, the rates of lung and colon cancer are 11.4% and 18.0%, respectively. Furthermore, the WHO estimates that approximately 4 million people worldwide may have colon or lung cancer by 2020. These cancers claimed the lives of around 2.7 million people [6]. These statistics show that lung cancer is common and dangerous diseases around the world. Along with that colon cancer is also as dangerous as lung cancer as determined in the above statements.

Lung cells become malignant when they mutate and begin to grow uncontrollably, forming a cluster known as a tumour. The worldwide increase in lung cancer has been attributed to a variety of variables, the most important of which are exposure to harmful or poisonous substances and an increase in the number of elderly individuals in the society. Yet, the symptoms are unlikely to be noticed until it has spread to other parts of the body, making treatment more difficult. Although lung cancer can occur in persons who have never smoked, it is usually more common in those who have. The most prevalent kinds of lung cancer are adenocarcinoma and squamous cell carcinoma, with other histological types including small and large cell carcinomas. Adenocarcinoma is a type of cancer of the lung that is most common in current or past smokers, but it can also occur in non-smokers. This is more common in women and young people, and it is found on the outer parts of the lung. Even before it spreads, it attacks the lungs. Squamous cell carcinomas are also linked to a smoking history. Small and large cell carcinoma, on the other hand, can develop in any section of the lung and has a proclivity to grow and unfurl quickly, making treatment more difficult.

Symptoms that can aid in the early detection of cancer are not direct markers of the disease. The most frequent symptoms, such as weariness, coughing, muscle discomfort, and so on, occur in conjunction with many types of disorders. Medical imaging instruments are the most important tool for detecting the existence of cancer. Mammography, histopathological imaging, computed tomography (CT), positron emission tomography (PET), magnetic resonance imaging (MRI), and ultrasound are all commonly utilised for cancer detection. Histopathology images with phenotypic information, for example, are essential for the diagnosis and evaluation of cancer illnesses in clinics. Experts must manually analyse such

medical photos, which is a delicate and challenging task. As a result, it takes time and requires intense concentration. Also, detecting cases is considerably more difficult in the case of early diagnosis because the symptoms are highly ambiguous and difficult to define at the start of the disease. It is too late for early treatment once symptoms appear. Because of improvements in the field of artificial intelligence (AI), AI-based medical image analysis approaches now serve as a decision support mechanism for both early diagnosis and doctor support.

The methods for autonomous diagnosis rely on AI technologies such as machine learning and deep learning. As a result, expert-based data analysis activities have evolved into expert-independent and totally automatic diagnostic systems. Many traditional machine-learning approaches and health-related applications have been used to solve medical problems. Unfortunately, these methods largely necessitate feature selection and feature extraction procedures, and they suffer from the drawbacks of not selecting the right feature extraction method and data loss during feature extraction. Deep learning (DL), on the other hand, has grown in popularity in medical diagnostic applications due to its capacity to eliminate these drawbacks as well as its excellent discrimination ability, because medical data are typically radiographic image data, the convolutional neural network (CNN) is a well-known DL architecture that is commonly utilised to evaluate medical images.

This paper proposes a framework based on numerous DL models and transformation approaches for the early diagnosis of lung cancer, which affects both men and women equally. Experiments are carried out using the five-class LC25000 dataset, which contains histological images of colon and lung cancer, the images of lung cancer are only used for the purpose of classification here, it contains of three classes. In contrast to earlier work, both precision and lower processing cost are used in this concept. The Convolutional Neural Network (CNN) which will be used for the development of the classification system takes input in form of image vectors or weights which will be extracted from the image by using pre-processing and feature extraction techniques, these features differentiate one image form the other image and gives the most accuracy. The following is how the paper is organised in the later sections: Section 2 discusses previously investigated research in the current domain. Section 3 offers' a quick overview of the methodology and the LC25000 dataset. Section 4 provides a brief introduction to CNN; Section 5 elaborates on the CNN architecture used in training both models. Section 6 summarises all the experimental findings and results. Finally, Section 7 finishes our experiment while providing some suggestions for further research.

Literature Survey:

7. This paper presents a machine learning-based approach to detect lung cancer. The authors used fuzzy neural networks to classify lung cancer into benign and malignant categories. The proposed method achieved an accuracy of 95.5% in detecting lung cancer. The authors also compared their method with other machine learning-based approaches and showed that their method outperformed other methods in terms of accuracy. However, the authors did not evaluate their method on a large dataset, which limits the generalizability of their results.

8. This paper presents a deep learning-based approach to classify lung cancer histology using CT images. The authors used a convolutional neural network (CNN) to extract features from CT images and classify them into different histological subtypes of lung cancer. The proposed method achieved an accuracy of 85% in classifying lung cancer histology. The authors also compared their method with other deep learning-based approaches and showed that their method outperformed other methods in terms of accuracy. However, the authors did not evaluate their method on a large dataset, which limits the generalizability of their results.

9. This paper presents recent achievements in lung cancer segmentation, detection, and classification using deep learning methods. The authors highlighted current state-of-the-art deep learning-based lung cancer detection methods and recent achievements, relevant research challenges, and future research directions. They also discussed the limitations of current deep learning-based approaches and suggested future research directions to overcome these limitations. However, the authors did not evaluate any specific deep learning-based approach in this paper.

10. This paper presents a deep learning-based approach to detect and classify lung cancer. The authors used novel deep learning methods such as Histogram of oriented Gradients (HoG), wavelet transform-based features, Local Binary Pattern (LBP), Scale Invariant Feature Transform (SIFT), and Zernike Moment to detect the location of the cancerous lung nodules. The proposed method achieved an accuracy of 95% in detecting lung cancer. However, the authors did not compare their method with other deep learning-based approaches.

11. This paper presents a deep learning-based algorithm to detect lung cancer on chest radiographs. The authors developed a deep learning (DL)-based model using the segmentation method and assessed its ability to detect lung cancer on chest radiograph images. The proposed method achieved an accuracy of 92% in detecting lung cancer. They also compared their method with other deep learning-based approaches and showed that their method outperformed other methods in terms of accuracy. However, the authors did not evaluate their method on a large dataset, which limits the generalizability of their results.

12. This paper presents a comparative study of deep learning-based approaches for lung cancer detection. The authors compared different deep learning-based approaches such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Hybrid Neural Networks (HNNs) for lung cancer detection. The proposed method achieved an accuracy of 94% in detecting lung cancer using CNNs. However, the authors did not compare their method with other deep learning-based approaches.

13. This paper presents a systematic review of deep learning techniques for lung cancer detection. The authors reviewed different deep learning-based approaches such as CNNs, RNNs, HNNs, and Generative Adversarial Networks (GANs) for lung cancer detection. They also compared different datasets used in these studies and evaluated their performance based on different evaluation metrics such as sensitivity, specificity, accuracy, etc. The authors concluded that deep learning-based approaches have shown promising results in detecting lung cancer.

14. This is a research paper that compares the performance of selected machine learning algorithms for lung cancer detection. The study analyses lung cancer prediction using classification algorithms such as Naive Bayes, SVM, Decision Tree, and Logistic Regression. The paper's key objective is to diagnose lung cancer early by examining the performance of these classification algorithms. The authors aim to provide insights into which algorithm performs best in detecting lung cancer and to contribute to the development of more accurate and efficient diagnostic tools.

15. This is a research paper that explores the feasibility of using deep learning algorithms for lung cancer diagnosis with cases from the Lung Image Database Consortium (LIDC) database. Three deep learning algorithms were designed and implemented, including Convolutional Neural Network (CNN), Deep Belief Networks (DBNs), and Stacked Denoising Autoencoder (SDAE). They compared the performance of deep learning algorithms with traditional computer-aided diagnosis (CADx) systems. The results showed that the accuracies of CNN, DBNs, and SDAE were 0.7976, 0.8119, and 0.7929, respectively; while the accuracy of the traditional CADx system was 0.7940, which is slightly lower than CNN and DBNs.

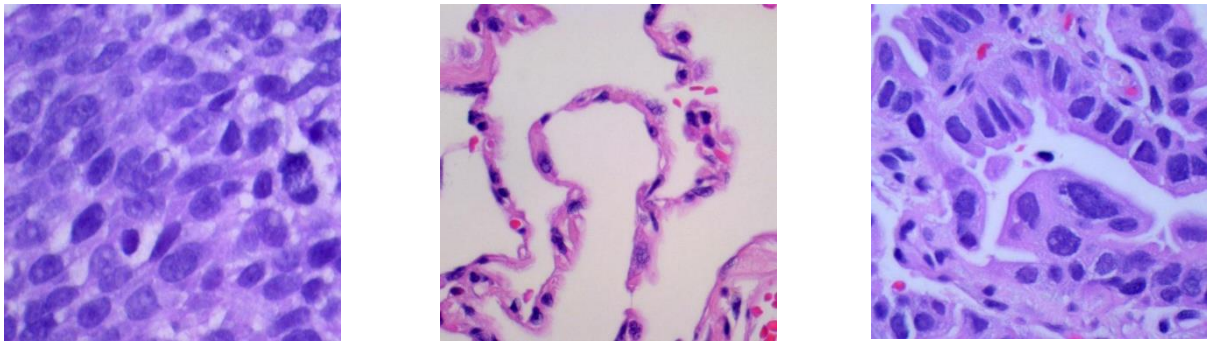
16. This is a research paper that proposes a novel neural-network-based algorithm, referred to as the entropy degradation method (EDM), to detect small cell lung cancer (SCLC) from computed tomography (CT) images. The authors used high-resolution lung CT scans provided by the National Cancer Institute and selected 12 lung CT scans from the library, 6 of which were for healthy lungs and the remaining 6 were scans from patients with SCLC. The authors trained their model using 5 scans from each group and tested it using the remaining two scans. Their algorithm achieved an accuracy of 77.8%. Overall, this paper provides valuable insights into the potential of using supervised machine learning algorithms for small cell lung cancer detection. It offers a novel approach to detecting SCLC from CT images and demonstrates its effectiveness through experimental results.

S. No	Paper Name	Dataset Used	Feature Extraction Technique	Model Used	Results
17.	Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning	Whole-slide images obtained from The Cancer Genome Atlas	Deep convolutional neural network (inception v3)	Deep convolutiona l neural network (inception v3)	The result of this method is comparable to that of pathologists, with an average area under the curve (AUC) of 0.97
18.	Artificial Intelligence in Lung Cancer Pathology Image Analysis	Whole slide imaging (WSI) in pathology	Convolutional Layers	Deep learning algorithms	Deep learning has shown great potential in pathology image analysis tasks such as tumour region identification.
19.	Detection of Lung Cancer Lymph Node Metastases from Whole-Slide Histopathologic Images Using a Two-Step Deep Learning Approach	349 whole-slide lung cancer lymph node images	Two-step deep learning algorithm	Two-step deep learning algorithm	Errors were reduced by 36.4% on average and up to 89% in slides with reactive lymphoid follicles.
20.	Robustness Fine-Tuning Deep Learning Model for Cancers Diagnosis Based on Histopathology Image Analysis	LC2500 dataset	Fine-tuning deep network for colon and lung cancers using regularization, batch normalization, and hyperparameter s optimization	Pre-trained ResNet101 network	The average precision, and accuracy were 99.84%, and 99.94%, respectively.
21.	Research on the Auxiliary Classification and Diagnosis of Lung Cancer Subtypes Based on Histopathologic al Images	121 LC histopatholo gical images	Relevant features (Relief) algorithm for feature selection	Support vector machines (SVMs) classifier	LUSC-ASC 73.91%, LUSC-SCLC 83.91% ASC-SCLC 73.67%

22.	Deep learning techniques for detecting preneoplastic and neoplastic lesions in human colorectal histological images	A dataset of human colon tissue images collected and labelled over a 10-year period by a team of pathologists	Direct labelling of raw images	Neural network comprising several convolutional and a few linear layers	Overall accuracy of >95%, with the majority of mislabelling referring to a near category. Tests on an external dataset with a different resolution yielded accuracies >80%
23.	Cancer diagnosis in histopathological image: CNN based approach	BreakHis database	Convolutional Layers	Convolutional Neural Network (CNN)	Prediction accuracy of 98.6%
24.	Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis	The Cancer Genome Atlas (TCGA) Dataset	Patch-based fully convolutional network (FCN) feature extraction	Random forest (RF) classifier	Accuracy of 97.3%, surpassing state-of-the-art approaches by a significant margin.
25.	A Deep Learning Approach for Breast Invasive Ductal Carcinoma Detection and Lymphoma Multi-Classification in Histological Images	Public datasets of digital histological images	Residual convolutional neural network feature extraction as part of a convolutional autoencoder network (FusionNet)	FusionNet, compared with UNet and ResNet	Improvement of 5.06% in F-measure score for the detection task, and an improvement of 1.09% in the accuracy measure for the classification task
26.	Prediction of lung and colon cancer through analysis of histopathological images by utilizing Pre-trained CNN models with visualization of class activation and saliency maps	LC25000 dataset	Pre-trained CNN-based model feature extraction with better augmentation techniques	VGG16, NASNetMobile, InceptionV3, InceptionResNetV2, ResNet50, Xception, MobileNet, and DenseNet169	All eight models achieved significant results ranging from 96% to 100% accuracy

Dataset:

A brief overview of the dataset is presented, followed by all data pre-treatment methods. Borkowski et al LC25000 Dataset provides microscopic pictures of the lung and colon. The dataset is divided into five categories: lung adenocarcinomas, lung squamous cell carcinomas, lung benign, colon adenocarcinomas, and colon benign, each with 5000 photos but for the classification point here only the classes of lung cancer are considered here. Figure 1 depicts various examples of photographs from the classifications of lung cancer. The original dataset only contained 750 photos of the lung and 500 photographs of the colon with pixel sizes of 1024x768, which were then transformed into squares of 768x768 pixels. With the use of rotation and flips, Augmenter was utilised to expand the dataset into 25000 images.



Lung Squamous Cell Carcinoma | Lung Benign Tissue | Lung Adenocarcinoma

Figure.1 Lung Cancer image samples from the Dataset.

It is pre-processed before being fed as augmented data to the model. Initially, data for each class was sampled into 4500 and 500 datapoints for training and test sets, respectively, using the random sampling provided. In addition, photos were scaled to 150x150 pixels, and some were randomised. Image shear and zoom transformations are followed by image normalisation.

Methodology:

The proposed approach is divided into three stages: histopathology image pre-processing, deep learning model training and feature extraction, and classification. Initially, the sizes of the histopathological images are analysed, and then these images are augmented. The photos are then split into sets using the image data generator, the size is set, and deep features are extracted. Finally, using the image data generator with the deep learning model, deep features are extracted. Eventually, these attributes are utilised to train a variety of deep learning models, including CNN. The three stages of the proposed approach are summarized in Figure.2.

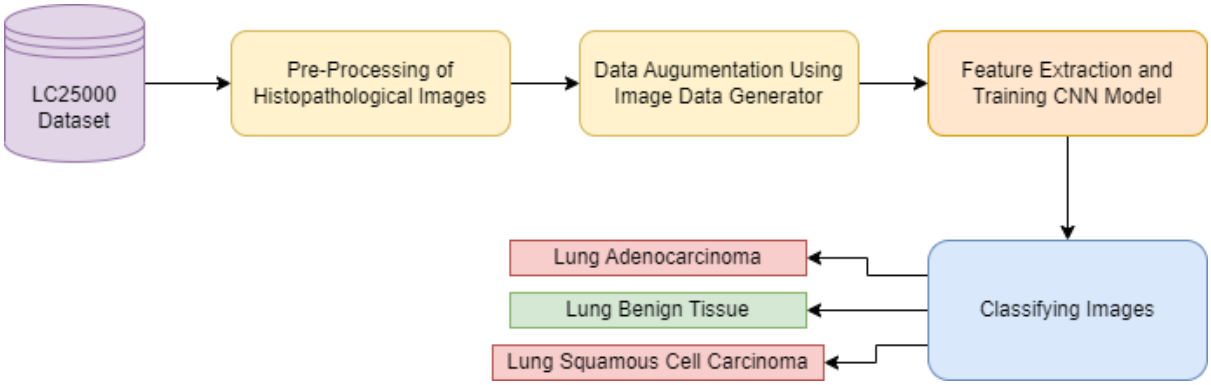


Figure.2 Stages of the Proposed Approach.

To begin the training phase, the input layer of each CNN receives a picture of a particular size. Hence, the histopathological image dimension of lung and colon cancer is originally changed to be like the deep learning models' input layer size. An augmentation approach is therefore required to improve the training performance of the deep learning models and minimise overfitting. Augmentation essentially increases the number of training images in a dataset, allowing training models to learn more successfully. In this paper, several augmentation methods are used, including scaling in x and y preferences, flipping in both x and y orientations, translation in both x and y orientations with an angle range, and shearing in both directions (x and y) within a range.

Here, Image data generator is used to get the attributes of the images, it is usually used to take in the original data and then randomly transform it before returning the outcome containing only the newly transformed data. It does not include the data. The Keras ImageDataGenerator class is also used for data augmentation, with the goal of increasing the overall generality of the model. In data augmentation, operations such as rotations, translations, shearing, scale changes, and horizontal flips are performed at random using an image data generator. Keras image data generator is used in the realm of real-time data augmentation to generate batches comprising data from tensor images. When we utilise Keras' image data generator, we can loop through the data in batches. The image data generator class has several methods and arguments that help determine the data generating behaviour.

In the next phase, the data extracted from the images using the image data generator are used to give the input to the deep learning model i.e., CNN, the data generated is usually the transformed data of the image which contains the augmented image in a matrix form. This matrix representation of data is used to train the deep learning models.

Algorithm:

Let us consider some parameters to derive the classification of lung cancer images through the convolutional neural network model. The working of the Convolutional Neural Network (CNN) is explained in the next section, the same implementation will be applied on the lung cancer dataset. To begin with, let's consider the raw image dataset of lung cancer images as **A** which consists of three classes which are determined as **A = {A_1, A_2, A_3}**. This dataset is first pre-processed, and the images present in it are sorted according to the classes such that the images belonging to the same class are together. Next, by using the ImageDataGenerator from the keras library the images are scaled down and resampled by using various parameters such as shear range, zoom range and rotation range. The data generator will split the data into training and validation sets also with parameters such as batch size, target size and colour mode. Let us consider the training dataset as **A'** and the validation dataset as **A".** The figure and table below represent the same, where Figure.4 represents the dataset and its classification, and Table.2 represents the algorithm of ImageDataGenerator for the lung cancer dataset.

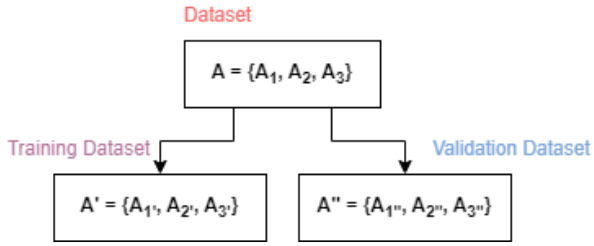


Figure.3 Dataset Classification

Algorithm 1 - Image Data Generator for raw Lung cancer images

1: INPUT: **A** = {A_1, A_2, A_3}

2: **repeat**

3: **repeat**

4: Obtain a single image from all the images of class **A_**

5: rescale image (1./255)

6: Augment the image (150 x 150)

7: **until:** All images in class A_

8: **until:** All classes in **A** i.e., A = {A_1, A_2, A_3}

9: OUTPUT: Augmented Images ~**A**

Table.2 Image Augmentation Algorithm

The above algorithm will work for both the training data and the validation data that is split from the raw image dataset of lung cancer classes hence, augmenting raw images.

In the next step, after the images are rescaled and augmented using the image data generator the images are passed as an input to the deep learning model which is the convolutional neural network, the images are in form of a matrix so, the conversion of the image occurs in numbers where each number determines the value of each pixel of an image the channel of the pixel will always be 3 because the image is in RGB format. The augmented image is in the shape of (150 x 150), this image is provided as input to the convolutional layers of the neural network. Let us consider the pre-processed and augmented image data as **~A'** for training dataset and **~A"** as validation dataset then, the convolutional matrix will be considered as **W'** and **W"** and features extracted from the images will be considered as **PW'** and **PW"** for training data and validation data respectively. The algorithm for extracting the features from the images through convolutional layers is given below in Table.3.

Algorithm 2 – Feature Extraction with convolutional layers
1: INPUT: ~**A'** = {~$A_{1'}$,~ $A_{2'}$,~ $A_{3'}$} **or** ~**A"** = {~$A_{1"}$,~ $A_{2"}$,~ $A_{3"}$}
2: Initialize the convolutional kernel matrices **W'** and **W"** with random values
3: **repeat**
4: Obtain a single image from all the images of ~A' or ~A"
5: Apply convolutional operation (*f', k'*), (*f", k"*)
6: Apply Max pooling operation (*p*)
7: Update the **W'** and **W"** matrix with the updated weights after the above two operations.
8: **until:** All images in ~**A'** or ~**A"**
9: OUTPUT: Updated Weight matrices i.e., **PW'** and **PW"**

Table.3 Feature Extraction Algorithm

The above algorithm will work for both the training data and the validation data that is split from the Augmented images of lung cancer classes hence, obtaining the extracted features in the matrices, these matrices are further minimized by applying more convolutional and max pooling operations by using the same algorithm hence, obtaining the final weight matrices which are flattened down for the classification purpose. The equation for convolutional and pooling operations for the lung cancer images are determined below:

$$\mathbf{PW'}[m, n] = (\sim\mathbf{A'} * \boldsymbol{k'})[m, n] = \sum_j \sum_k \boldsymbol{k'}[j, k] \sim\mathbf{A'}[m - j, n - k]$$

$$\mathbf{PW"}[m, n] = (\sim\mathbf{A"} * \boldsymbol{k"})[m, n] = \sum_j \sum_k \boldsymbol{k''}[j, k] \sim\mathbf{A''}[m - j, n - k]$$

In the above equations, **PW'** determines the weight matrices of the training data whereas **PW"** determines the weight matrices of validation data that is provided to the dense network for classification. The other parameters such as [m, n] determine the pixel range in an image to which the convolutional and pooling operations are applied to get the feature map matrices where, **~A', ~A"** are the images and the ***k'***, ***k"*** are the initial kernel matrices. After obtaining the weight matrices by applying the convolutional and pooling operations on each image the matrices are fed to the dense layer which performs the classification of cancer for lungs. In the next step, the extracted features are fed to the dense neural network which comprises of neurons which help in classifying the images according to the classes specified. Let us consider the initial input to the dense layers as the transformed version **PW'** which is denoted as **~PW'** this is the flattened values from the matrix which are considered as the input neurons of the dense layer, the dense layer consist of weights which are randomly initialized and updated accordingly by forward and backward propagation in numerous iterations, let us consider this randomly initialized weights as **S'**. These weights are forwarded to next dense layer and updated. Finally, the weights determine the probability of each class from where the maximum probability of class is chosen to classify that image. The algorithm for this process is specified in Table.4 below.

Algorithm 3 – Training and classification with dense layers
1: INPUT: **PW'** or **PW"**
2: Transforming the matrices i.e., Flattening them into one-dimensional Neurons
3: Initializing the weight matrices randomly for the dense layers **S'**
4: **repeat**
5: Apply $X\rightarrow$ **(DL1)**
6: Apply $Y\rightarrow$ **(DL2)**
7: Apply $Y\leftarrow$, $X\leftarrow$ and update values **(~DL2 & ~DL1)**
8: Update the **S'** and **S" w**eights
9: **until:** Iteration Convergence
10: OUTPUT: Trained model **(M)** and final updated weights **(~S')**

Table.4 Training and Classification Algorithm

The above algorithm works for training data which gets transformed into input layer for the dense layers and the training process occurs accordingly, the validation set of values is used to validate the training of the model by classifying the values in that iteration. The output of the above algorithm will give the completely trained model **M** and the final updated weights **~S'.** These weights determine the probability of each class when an input image is given.

Convolutional Neural Network (CNN):

Due to the high convolution of inter-intraclass relationships, image classification is a difficult issue for visual content, particularly microscopic pictures such as histopathology images. Because of similar structural morphological textures, the underlying structures are complicated and interconnected. Figure 1 depicts some of the most complex textures found in histopathology photographs. Deep learning is popular because of its capacity to learn features directly from input, allowing us to forgo time-consuming feature extraction techniques. One of the fundamental advantages of deep learning is the ability to uncover abstract level features and then deep dive into the feature map to extract structural semantics. Deep learning, particularly CNNs, has shown to be an efficient method for categorising and diagnosing medical pictures in recent years. CNN is made up of numerous trainable layers that may be combined with a supervised classifier to learn feature mappings from a given input data flow. Digital or signal data, such as music, video, pictures, and time series, could be used as input data stream. Consider a coloured image, which is a 3D tensor feature map, with a 2D tensor for each colour channel. CNN designs are made up of three layers: the convolution layer, the max pooling layer, and the fully connected layer or dense layer. These layers could be built in various ways to create a CNN. Figure.3 depicts an example of a conventional CNN.

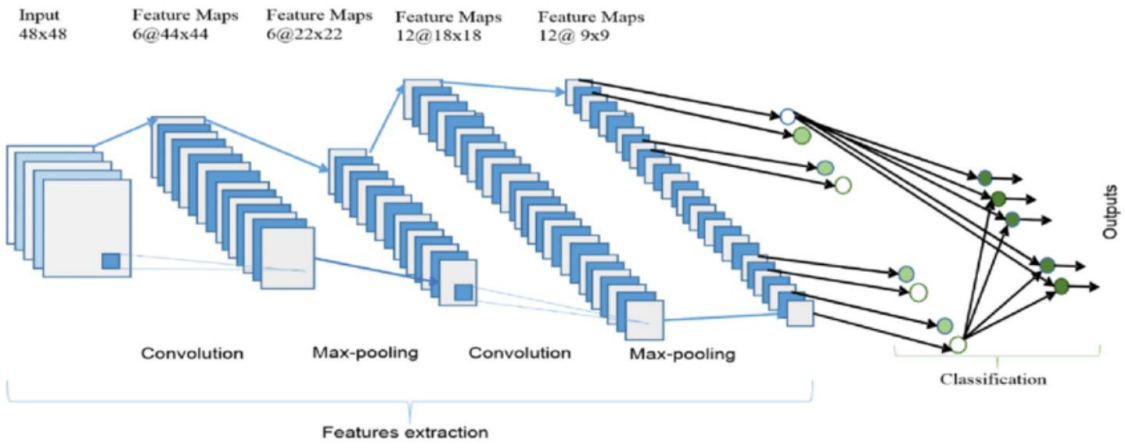


Figure.4 Commonly used CNN Architectures.

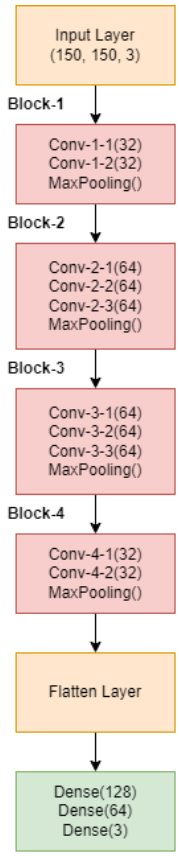
The convolution layer is a critical component of any given CNN architecture. Convolution layers calculate the dot product of the weights and the input signal related to that local region. The kernel or filters are the set of weights that are twisted along the input vector. Each filter is tiny yet covers the entire depth of the input volume. A typical size of filter for visual inputs is typically (3x3, 5x5, 8x8). These weights are shared throughout neurons,

allowing filters to learn all the geometrical structures in the image. Stride is the distance between these filters' applications. When the hyperparameter stride is less than the filter size, overlapping convolutions are applied to the picture.

It is usual practise to insert a pooling between two convolution layers to down sample the image along the volume component. This is critical for gradually reducing the spatial size of the representation. As a result, limiting the number of parameters and calculations needed by the network aids in overfitting management. Pooling resizes photos along height and breadth while ignoring activation. Researchers found that the max pooling method, which gives a window for selecting the maximum value from an input patch among neighbours, produced better results.

Full connection is created between input activations and their activation in a completely linked layer or dense layer. The computation is carried out using matrix multiplication and consecutive bias offset. The last dense layer contains the final output such as probability density or logit values for the classification of data.

Custom Developed CNN Model for Lung cancer detection:




```
# Custom CNN-Model
model = Sequential()

# Input Layer
model.add(Input((150,150,3)))

# Block-1 of Custom CNN
model.add(Conv2D(filters=32,kernel_size=(3,3),padding="same", activation="relu"))
model.add(Conv2D(filters=32,kernel_size=(3,3),padding="same", activation="relu"))
model.add(MaxPooling2D())

# Block-2 of Custom CNN
model.add(Conv2D(filters=64, kernel_size=(3,3), padding="same", activation="relu"))
model.add(Conv2D(filters=64, kernel_size=(3,3), padding="same", activation="relu"))
model.add(Conv2D(filters=64, kernel_size=(3,3), padding="same", activation="relu"))
model.add(MaxPooling2D())

# Block-3 of Custom CNN
model.add(Conv2D(filters=64, kernel_size=(3,3), padding="same", activation="relu"))
model.add(Conv2D(filters=64, kernel_size=(3,3), padding="same", activation="relu"))
model.add(Conv2D(filters=64, kernel_size=(3,3), padding="same", activation="relu"))
model.add(MaxPooling2D())

# Block-4 of Custom CNN
model.add(Conv2D(filters=32, kernel_size=(3,3), padding="same", activation="relu"))
model.add(Conv2D(filters=32, kernel_size=(3,3), padding="same", activation="relu"))
model.add(MaxPooling2D())

model.add(Flatten(name="flatten"))

model.add(Dense(128, activation="relu", name="fc1"))
model.add(Dense(64, activation="relu", name="fc2"))

model.add(Dense(3, activation="softmax", name="output"))
model.compile(optimizer='RMSprop', loss='categorical_crossentropy', metrics=['accuracy'])

model.summary()
```

Figure.5 Lung Cancer Image Classification Architecture.

Result Analysis:

Two models have been compared for analysing the results which are VGG16 and the custom developed CNN model, the results for each model have been determined below:

1. VGG16 Model:

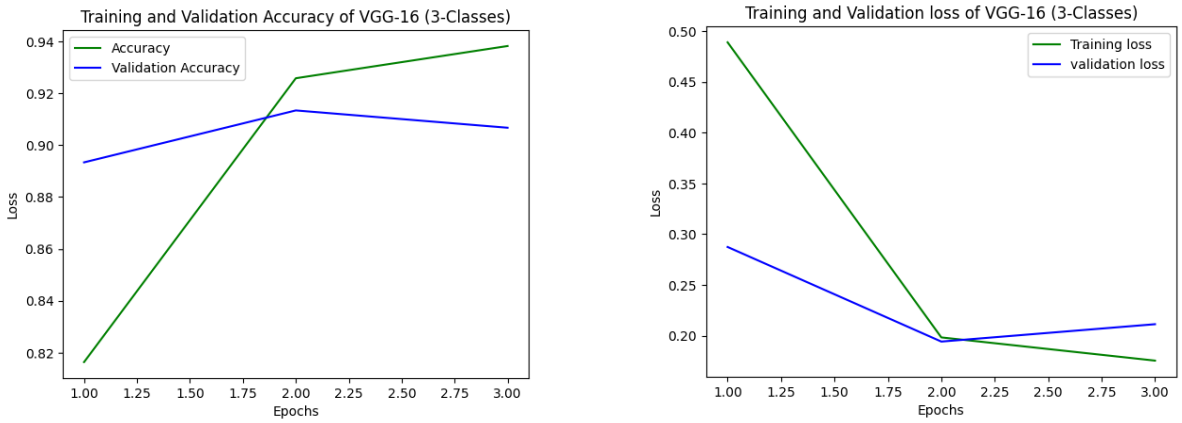


Figure.6 Training and Validation Curves for VGG-16 Model.

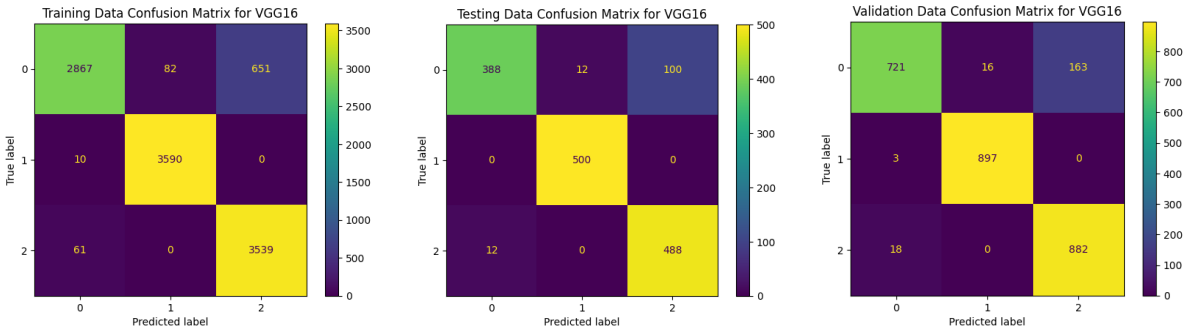


Figure.7 Confusion matrices for VGG-16 Model.

```
VGG16 Accuracies :

Training Accuracy :  0.9255555555555556

Testing Accuracy :  0.9173333333333333

Validation Accuracy :  0.9259259259259259
```

Figure.8 Accuracy of VGG-16 Model.

The above figures represent the results that are achieved by the VGG-16 model for classification of lung cancer images, figure.6 represents the training and validation accuracy achieved by the VGG-16 model for the number of iterations i.e., 20 and the figure.7 represents the confusion matrices of training, testing and validation data respectively. Finally, the figure.8 represents the accuracy of the sets. The accuracy achieved by this model is **92%** overall.

2. Custom - CNN Model:

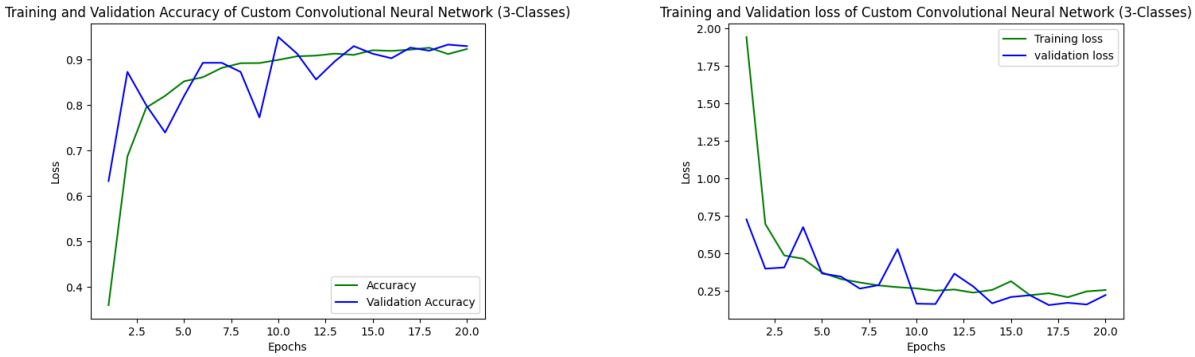


Figure.9 Training and Validation Curves for Custom-CNN Model.

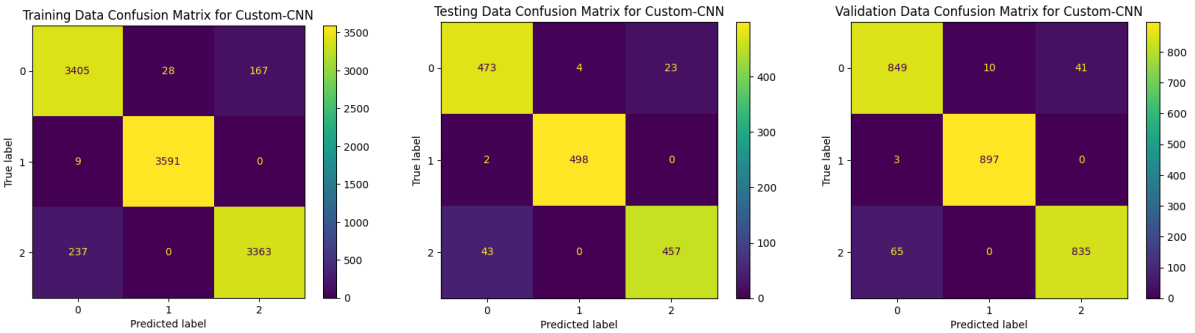


Figure.10 Confusion matrices for Custom-CNN Model.

```
Custom CNN Accuracies :

Training Accuracy :  0.9591666666666666

Testing Accuracy :  0.952

Validation Accuracy :  0.955925925925926
```

Figure.11 Accuracies of Custom-CNN Model.

The above figures represent the results that are achieved by the Custom-CNN model for classification of lung cancer images, figure.9 represents the training and validation accuracy achieved by the Custom-CNN model for the number of iterations i.e., 20 and the figure.10 represents the confusion matrices of training, testing and validation data respectively. Finally, the figure.11 represents the accuracy of the sets. The accuracy achieved by this model is **96%** overall. The models are compared accordingly in the below table where the type of data is specified in the first column and the accuracies of each model i.e., VGG-16 and Custom-CNN models are specified in the next two columns for training as well as testing data. This comparison helps achieve the overall best model based on accuracy and loss.

Type	VGG-16		Custom-CNN	
	Accuracy	**Loss**	**Accuracy**	**Loss**
Training	0.925	0.198	0.959	0.265
Validation	0.925	0.194	0.955	0.163

Table.5 Performance Metrics of Both the Models on Lung Cancer Images

To guarantee that classifiers generalise effectively, the data was divided into two groups, with 80-20 of the data divided into training and validation sets, respectively. This procedure was used independently for pictures of lung cancer. By comparing both the VGG-16 and the Custom developed CNN model, the accuracy of the Custom developed CNN model is higher than VGG-16 model for the Lung cancer histopathological image dataset. The difference between the accuracy of both the models is minute but the classification metrics are a bit better in the Custom developed-CNN model.

Conclusion:

This paper presented a collection of experiments conducted on the LC25000 Lung Cancer Image dataset using a deep learning approach. We demonstrated that we could adapt a shallow CNN architecture built for identifying colour images of objects to categorise lung histopathology images. We also suggested a training and assessment technique for training the CNN architecture, which allows us to cope with high-resolution textured pictures without reducing them to low-resolution images. Our experimental results on the LC25000 Lung Cancer Images demonstrated that Custom developed CNN outperformed classical machine learning models and deep convolutional neural network models employing transfer learning trained on the same dataset but with state-of-the-art texture descriptors. Further research could look into different CNN designs and hyperparameter optimisation. Additionally included are ways for using neural style transfer to generate interclass images for various types. Furthermore, generative models could be utilised to produce histopathological pictures for visualising and understanding mutations across several ontologies.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 2021, 71, 209–249.
2. World Health Organization (WHO). Global Health Estimates: Life Expectancy and Leading Causes of Death and Disability.
3. International Agency for Research on Cancer. Available online: https://www.iarc.who.int
4. Yadav, A.R.; Mohite, S.K. Cancer-A silent killer: An overview. Asian J. Pharm. Res. 2020, 10, 213–216.
5. Rl, S.; KD, M.; Jemal, A. Cancer statistics, 2020. CA Cancer J Clin 2020, 70, 7–30.
6. World Health Organization (WHO). Cancer. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer
7. Dr. K. Batri, P. Pretty Evangeline, 2019, Detection of Lung Cancer by Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 08, Issue 09 (September 2019)
8. Chaunzwa, T.L., Hosny, A., Xu, Y. et al. Deep learning classification of lung cancer histology using CT images. Sci Rep 11, 5471 (2021). https://doi.org/10.1038/s41598-021-84630-x
9. Wang, Lulu. 2022. "Deep Learning Techniques to Diagnose Lung Cancer" Cancers 14, no. 22: 5569. https://doi.org/10.3390/cancers14225569
10. Asuntha, A., Srinivasan, A. Deep learning for lung Cancer detection and classification. Multimed Tools Appl 79, 7731–7762 (2020). https://doi.org/10.1007/s11042-019-08394-3
11. Shimazaki, A., Ueda, D., Choppin, A. et al. Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. Sci Rep 12, 727 (2022). https://doi.org/10.1038/s41598-021-04667-w
12. S. Das and S. Majumder, "Lung Cancer Detection Using Deep Learning Network: A Comparative Analysis," 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Bangalore, India, 2020, pp. 30-35, doi: 10.1109/ICRCICN50933.2020.9296197.

13. Mattakoyya Aharonu, R Lokesh Kumar, " Systematic Review of Deep Learning Techniques for Lung Cancer Detection" International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 14, No. 3, 2023 https://thesai.org/Downloads/Volume14No3/Paper_84-Systematic_Review_of_Deep_Learning_Techniques.pdf

14. R. P.R., R. A. S. Nair and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2019, pp. 1-4, doi: 10.1109/ICECCT.2019.8869001.

15. Wenqing Sun, Bin Zheng, Wei Qian, "Computer aided lung cancer diagnosis with deep learning algorithms," Proc. SPIE 9785, Medical Imaging 2016: Computer-Aided Diagnosis, 97850Z (24 March 2016); https://doi.org/10.1117/12.2216307

16. Q. Wu and W. Zhao, "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm," 2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC), Budapest, Hungary, 2017, pp. 88-91, doi: 10.1109/ISCSIC.2017.22.

17. Coudray, N., Ocampo, P.S., Sakellaropoulos, T. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 24, 1559–1567 (2018). https://doi.org/10.1038/s41591-018-0177-5

18. Wang, Shidan, Donghan M. Yang, Ruichen Rong, Xiaowei Zhan, Junya Fujimoto, Hongyu Liu, John Minna, Ignacio Ivan Wistuba, Yang Xie, and Guanghua Xiao. 2019. "Artificial Intelligence in Lung Cancer Pathology Image Analysis" Cancers 11, no. 11: 1673. https://doi.org/10.3390/cancers11111673

19. Hoa Hoang Ngoc Pham, Mitsuru Futakuchi, Andrey Bychkov, Tomoi Furukawa, Kishio Kuroda, Junya Fukuoka, “Detection of Lung Cancer Lymph Node Metastases from Whole-Slide Histopathologic Images Using a Two-Step Deep Learning Approach” The American Journal of Pathology, Volume 189, Issue 12, 2019. https://www.sciencedirect.com/science/article/pii/S0002944019307187

20. El-Ghany, Sameh Abd, Mohammad Azad, and Mohammed Elmogy. 2023. "Robustness Fine-Tuning Deep Learning Model for Cancers Diagnosis Based on Histopathology Image Analysis" Diagnostics 13, no. 4: 699. https://doi.org/10.3390/diagnostics13040699

21. M. Li et al., "Research on the Auxiliary Classification and Diagnosis of Lung Cancer Subtypes Based on Histopathological Images," in IEEE Access, vol. 9, pp. 53687-53707, 2021, doi: 10.1109/ACCESS.2021.3071057.

22. Sena, P., Fioresi, R., Faglioni, F., Losi, L., Faglioni, G., Roncucci, L."Deep learning techniques for detecting preneoplastic and neoplastic lesions in human colorectal histological images". Oncology Letters 18, no. 6 (2019): 6101-6107. https://doi.org/10.3892/ol.2019.10928

23. Sumaiya Dabeer, Maha Mohammed Khan, Saiful Islam, “Cancer diagnosis in histopathological image: CNN based approach”, Informatics in Medicine Unlocked, Volume 16, 2019, 100231, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2019.100231.

24. X. Wang et al., "Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis," in IEEE Transactions on Cybernetics, vol. 50, no. 9, pp. 3950-3962, Sept. 2020, doi: 10.1109/TCYB.2019.2935141.

25. N. Brancati, G. De Pietro, M. Frucci and D. Riccio, "A Deep Learning Approach for Breast Invasive Ductal Carcinoma Detection and Lymphoma Multi-Classification in Histological Images," in IEEE Access, vol. 7, pp. 44709-44720, 2019, doi: 10.1109/ACCESS.2019.2908724.

26. Satvik Garg and Somya Garg. 2021. Prediction of lung and colon cancer through analysis of histopathological images by utilizing Pre-trained CNN models with visualization of class activation and saliency maps. In 2020 3rd Artificial Intelligence and Cloud Computing Conference (AICCC 2020). Association for Computing Machinery, New York, NY, USA, 38–45. https://doi.org/10.1145/3442536.3442543