B.Sc. Thesis Proposal

Lung Cancer Detection using Machine Learning for Histopathological Images

Submitted By

Sabahat Tabbassum

18-EE-01

Azib Farooq

18-EE-43

Syed Irtaza Hussain

18-EE-46

Javeria Noor Tariq

18-EE-53

Supervisor

Dr. Gulistan Raja

Professor

Department of Electrical Engineering

Faculty of Electronics & Electrical Engineering

University of Engineering & Technology

Taxila

November 2021

Table of Contents

Problem Statement

Cancer is among the deadliest diseases and is the second largest cause of death among individuals. In order to reduce the death risk due to cancer, early and valid finding of the cancerous cells (carcinoma) is required, which is really troublesome for histologist. Large number of cases makes the report to be more vulnerable to human error and wrong prescriptions. In the case, if the histologist is not prepared then the outcomes will be hazardous for the patient, and it can cost his/her life. Other the hand, huge data collection capability has enabled to store the dataset of histopathology images. Machine learning techniques has influenced individuals from all walks of professions for betterment by deploying robust techniques to automate tedious and tiresome methods.

Detection of lung cancer is not an easy task for histologists. This project describes automated detection of lung cancer. Machine learning will be used for this detection. Further we will classify the type of lung cancer. The obtained results will be validated and model performance will be accessed by numerous comparative metrics

Aims & Objectives

Aims

- Detection of lung cancer.
- Classification of lung cancer into its type.
- Obtain acceptable accuracy on detection and classification.

Objectives

- We will be using Computer Aided Design approach to detect lung cancer from histopathology images.
- Implementation of ResNet50 transfer learning models on dataset and identify lung cancer and its subtypes.
- Apply four commonly used accuracy metrics on the model.

Literature Review

Lung cancer is the leading cause of death among many individuals. Lung cancer contribute 25% of the total cancer death. The primary cause of lung is the due to smoking, exposure to air pollution, second-hand smoke, and other factors. Medical

professionals spend most of the time classifying the type of the cancer before starting medical treatment. There are various methods for detection of lung cancer.

Conventional Methods

Following are conventional methods used for detecting the lung cancer

a. Imaging tests: An x-ray image can be used to detect cancerous cells in lungs. CT scan can be used to detect details which is not possible with X -ray imaging.
b. Sputum cytology: Sputum produced in coughing can be analyzed under microscope to confirm lung cancer.
c. Tissue sample (biopsy): Cells sample taken in process called biopsy is analyzed under microscope. Methods for biopsy can include bronchoscopy, Mediastinoscopy, needle biopsy. Analysis of cell will confirm what type of cancer does patient have. Which will help prognosis and guide doctor to adopt the treatment.

Bronchoscopy can be used for obtaining tissue sample from the body. A lightning tube is inserted in into lungs passing form throat into lungs. The tube is used to study abnormal cells in lungs. In mediastinoscopy, a cut is made at the base of the neck, surgical tools are inserted at the back of breastbone and sample is taken from lymph nodes. In Needle Biopsy, X-ray and CT guide needle through chest wall to take suspicious cell from the lungs and analyzed under microscope.

After detection of lung cancer, we have to determine the stage of cancer that vary from 0-IV stage. Staging tests include MRI, Bone Scan, CT and PET. These tests will help the doctors to determine spread of cancerous cell and guide them to adopt method for treatment [1].

Evaluation of histopathological slides by pathologists is necessary for diagnosis and classify the type of lung cancer. For pathologist and other medical professional diagnosing process is time consuming. Due to cumbersome process, the cancer type can be detected wrongly and directing toward wrong treatment can cause risking of patient's life. To reduce the risk, machine learning algorithms can be used to detect lung cancer [2]. Some methods used to detect lung cancer are being reviewed in this section. Although we have to take the sample same as conventional methods and take images of those histopathological slides. New methods involve the analysis of these histopathological slides using machine learning algorithm.

Modern Approaches for detecting Lung cancer

Number of techniques are being deployed in the literature for the classification of the lung cancer through application of algorithms on images obtained from different sources like X-RAY and histopathology images. In this section, we will discuss techniques and approaches on both type of images, and also outlined the results being obtained from the techniques.

Prediction of lung cancer using chest x-ray through transfer learning is being applied by the W. Ausawalaithong et.al.(2018) Images of size 224x224 with 121-layers densely connected CNN (DenseNet-121) is used. Single sigmoid node is applied in FC layer. The result obtained by this techniques had 74.43% mean accuracy with tolerance of almost 6%. The model has the same mean sensitivity for the dataset of the different images [1]. T. Atsushi et.al. (2017) used DCNN (Deep Convolutional Neural Networks) to automate lung cancer classification on cytological images. The primary structure of the DCNN include 3 convolutional and pooling layers with 2 fully connected layers, with 0.5 dropout for the DCNN. The model has accuracy of 71.1% [2].

W. Rahane et.al. (2018) detected lung cancer on CT images using image processing and SVM (support vector machine). Primary image processing techniques applied were grayscale conversion, noise reduction and conversion to binary image. Followed by the feature extraction, which in turn fed to the SVM. The features obtained was area, perimeter, and eccentricity from the segmented image region of interest [3]. M. Saric et.al. () applied deep learning techniques of VGG and ResNet for lung cancer detection. The dataset for the architecture is the whole slide histopathology images. The output obtained from the algorithm was related using the receiver operating characteristic (ROC) plot. The patch accuracy of the VGG16 obtained was 75.41% and 72.05% for ResNet50, respectively. The author also provide justification for low accuracy as the huge pattern diversification in the dataset [4].

S. Sasikala et.al. (2019) examined CT scan images to detect and classify lung cancer using CNN. The programming tools used in this research was MATLAB. The training process was divided into two phases i.e. valuable volumetric feature extraction was the first phase followed by the classification in the next and final phase. The proposed architecture can make classification of cancerous and non-cancerous cells with 96% accuracy [5]. SRS Chakravarty et.al. () used gray level co-occurrence matrix (GLCM) along with chaotic crow search algorithm (CCSA) for the purpose of feature

selection, which are being computed on the CT scan images. The extracted features are then supplied to the probabilistic neural network (PNN), which is responsible for classification. The final accuracy of the model on the features extracted from the CCSA was 90% [6].

Hatuwal et.al (2020) use convolutional neural network to detect lung cancer on histopathology images. Three types of carcinomas are considered in the study i.e. Adenocarcinoma, squamous cell carcinoma and benign carcinoma. The analysis and experimentation is based on LC25000 dataset. In which each class of the carcinoma has 5000 images. Data augmentation techniques like image flipping (horizontal, vertical), zooming, etc. is applied to increase the dataset to avoid overfitting of the model. Stacked layers of CNN (CovNets) is used for recognition and classification. The model obtain validation accuracy of 97.2% [7].

Zhu Y et.al. (2010) applied methods of texture feature extraction of solitary pulmonary nodules, followed by the application of the genetic algorithm to choose the most significant features and make classification using support vector machines [8]. S. QingZeng et.al. () used a CNN, DNN and stacked auto-encoder to classify CT images of benign and malignant classes of the lung cancer. After training the CNN achieved the accuracy of 84.15% [9]. Laksmanaprabu et.al. () designed and optimal DNN for the analysis of the CT images and extracted features from these images and classify them as malignant or benign. After training the accuracy of the model was 92.2% [10].

S. Garg et.al. (2021) utilizes the pre-trained CNN models to identify lung and colon cancer using LC25000 dataset. The dataset is also augmented using different techniques and fed to eight different pre-trained models, which are VGG16, NASNetMobile, InceptionV3, InceptionResNetV2, ResNet50, Xception, MobileNet and DenseNet169. Multiple evaluation parameters of all these models are monitored and examined i.e. precision, recall, f1score, accuracy, and auroc score. The range of results of all these pre-trained models are from 96% to 100%, on the particular dataset [11].

SDP Methodology

Theoretical Studies

Deep Learning

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to “learn” from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy. There are different types of neural networks to address specific problems or datasets.

Convolutional Neural Network

It is used primarily in computer vision and image classification applications, can detect features and patterns within an image, enabling tasks, like object detection or recognition. In 2015, a CNN bested a human in an object recognition challenge for the first time.

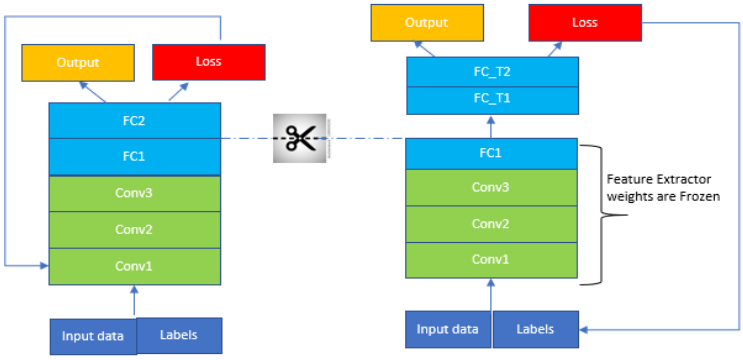
Recurrent Neural Network

They are typically used in natural language and speech recognition applications as it leverages sequential or times series data.

Feature Extraction through Transfer Learning

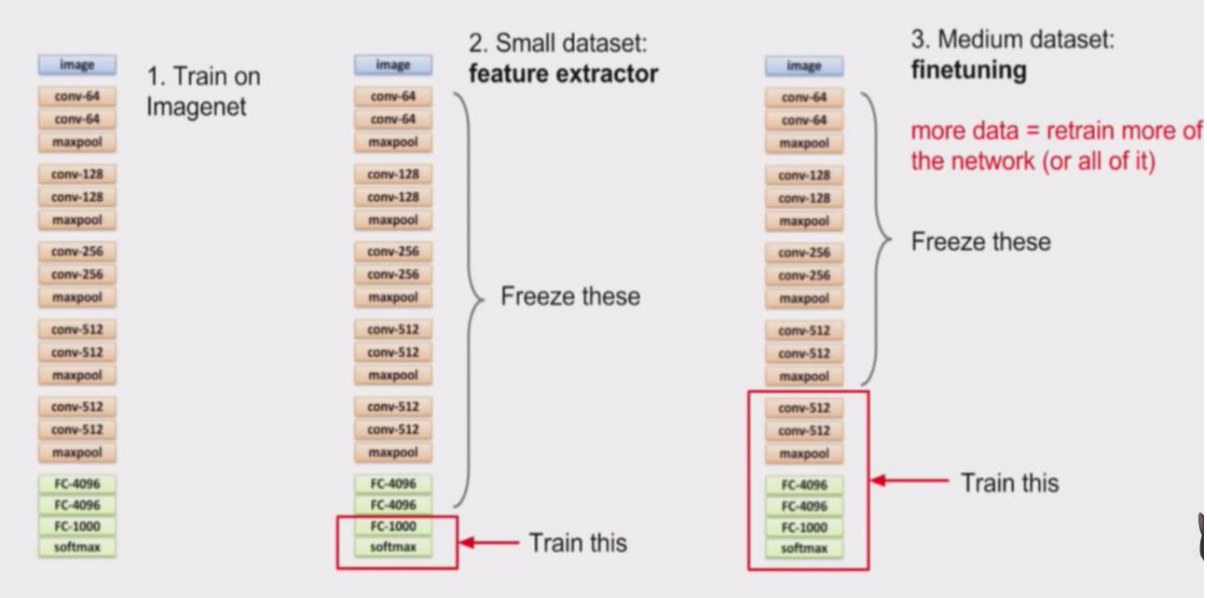
We use ResNet50 deep learning model as the pre-trained model for feature extraction for Transfer Learning.

- To implement Transfer learning, we will remove the last predicting layer of the pre-trained ResNet50 model and replace them with our own predicting layers. FC-T1 and FC_T2 as shown below
- Weights of ResNet50 pre-trained model is used as feature extractor
- Weights of the pre-trained model are frozen and are not updated during the training.



Transfer Learning

Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. For example, knowledge gained while learning to recognize cars could apply when trying to recognize trucks.



Pre-Trained Models

A pre-trained model has been previously trained on a dataset and contains the weights and biases that represent the features of whichever dataset it was trained on. Learned features are often transferable to different data. For example, a model trained on a large dataset of bird images will contain learned features like edges or horizontal lines that you would be transferable your dataset.

Pre-trained models are beneficial to us for many reasons. By using a pre-trained model you are saving time. Someone else has already spent the time and compute resources to learn a lot of features and your model will likely benefit from it.

Experimental Setup

COLAB

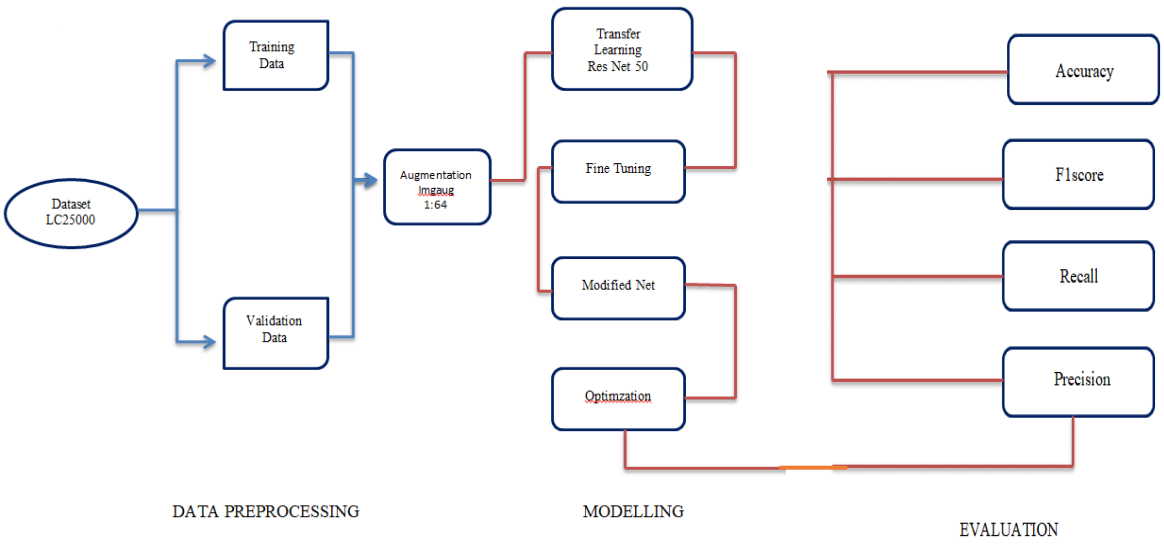
Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

With Colab we can import an image dataset, train an image classifier on it, and evaluate the model, all in just a few lines of code. Colab notebooks execute code on Google's cloud servers, meaning you can leverage the power of Google hardware, including GPUs and TPUs, regardless of the power of your machine. All you need is a browser.

Colab is used extensively in the machine learning community with applications including:

- Getting started with TensorFlow
- Developing and training neural networks
- Experimenting with TPUs
- Disseminating AI research
- Creating tutorials



1. Data Processing

Data Split

The dataset used in this proposal is LC25000. The dataset contains five classes, three for lung cancer and two for colon cancer. The scope of this study is limited to the lung cancer detection, so only 15000 images of three classes of lung cancer will be used in this study. The three classes of lung cancer include adenocarcinoma (aca), squamous cell carcinoma (scc), and benign (n). Each class has 5000 images. For the purpose of training, and validation the dataset is being divided into two chunks. The bigger chunks has 80% of the data from all three classes and is called training data. Remaining data (20%) is divided for validation and in all classes.

Dataset	Training (80%)			Validation (20%)		
	ACA	SCC	Ben	ACA	SCC	Ben
LC25000	4000	4000	4000	1000	1000	1000
Total (15000 images)	12000			3000		

Table 2: Data split

Data Augmentation

It is generally accepted that more data trains the model better. If there is less data then model either start overfitting or give unexpected results. In medical imaging we can leverage the advantage of data augmentation techniques without worrying about skewness or introducing error in the dataset. The LC25000 dataset is already augmented for 750 lung images with left and right rotation (upto 25 degrees) and horizontal and vertical flips as well. We can use built-in complex data augmentation library “imgaug” to augment data upto desired images. 1:64 augmentation is applied in our case.

2. Modelling

Fine Tuning of pre-trained model

Pre-trained feature extraction will be applied in this study. The concept of transfer learning is that weight of pre-defined models are trained on generic dataset and we input our data and the weight are adapt corresponding to our input data. There are huge benefits of using pre-trained models and they give exceptional results in wide range of cases, along with requirement of less computational capability.

In our case, we used ResNet50 for feature extraction and the only task at this step is the fine tuning of the pre-trained model i.e. ResNet50. Fine tuning of the model requires setting the hyper-parameters. And, wide research in literature is still unable

to figure out a proper way to come up with appropriate values of hyper-parameters. So, we are only left with the hit and trial method to set the values of ResNet50's hyper-parameters.

Modified CNN

After fine tuning of the model hyper-parameters, we will add customized layers at the end of the transfer-learning. Three layers are added Max-pooling2D, Average-pooling2D and flatten. The output of the flatten layer will give feature vector, which is supplied to the output layers having sigmoid activation. Beside all that a dropout layer with dropout rate of 0.5 is also applied.

Optimization

In the end, optimizer is also required to reduce computational complexity and optimum training of the model. There are wide range of optimizers being used in the domain deep learning for wide range of applications. Some of the most common among them are gradient descent, stochastic gradient descent, adadelta, RMSProp, ADAM, etc. In our case, we will be using ADAM, as it is the standard and gives brilliant output in wide range of deep learning models.

Method of Analysis

The model designed and applied in the above heading will be evaluated on four metrics i.e. precision, recall, f1 score, and accuracy. Relevant percentages of these models will evaluate the performance of the model. In order to better understand these evaluation metrics, we should look into the below formulas.

$$Tp = Correct\ positive\ class\ prediction$$

$$Tn = Correct\ negative\ class\ prediction$$

$$Fp = Wrong\ positive\ class\ prediction$$

$$Fn = Wrong\ negative\ class\ prediction$$

From the above definition of the terminology, the metrics of evaluations can be derived as follow:

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + F_n}$$

$$F1score = 2\frac{Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

Result Expected

The model is expected to achieve performance more than 50% on all the above metrics. Statistical models based on the extracted features will stratified NSCLC patients into high-risk and low-risk groups and will also the type of lung cancer

SDP Result's Utilization

There are wide avenues of research and medicine in which the outcomes can served. The significant and primary area of utilization is the histology. Where our can assist in following services.

- Classifying the lung cancer histolopathological images.
- Expert reviews can be made based on the pre-identified histology images.
- Healthcare sector can utilize results to manage increasing number of lung cancer cases.
- Give histologist more visibility of the area of interest in histological image.
- Faster and better treatment and medicine prescription for serious patients.

Work Schedule Plan

Work Schedule is as under:

Collection of literature	Two Weeks
Study of Literature	Two Weeks
Analysis of Proposed Scheme	Three Weeks
Preparation of Scheme/Model	Three Weeks
Implementation of Scheme/Model	One Month
Analysis and Simulation	Three Weeks
Result Formulation	Two Weeks
Final Write-up & Thesis Submission	Two Weeks

Proposed Time Schedule

Activity	**Time Schedule**
Collection of Literature	Dec 13 – Dec 21 2021
Study of Literature	Dec 22 – Jan 5 2021
Analysis of Proposed Scheme	Jan 6 – Jan 26 2021
Preparation of Schemes / Model	Feb 08 – Feb 27 2021
Implementation of Schemes/Model	Feb 28 – March 27 2021
Analysis & Simulation	March 27 – April 17 2021
Result Formulation	April 18 - May 1 2021
Final Write-up & Thesis Submission	May 2 – May 15 2021

Table 3: work schedule plan

Ethical Issues

Medical ethics has a closed relationship with law. Ethical principles such as respect for the persons, informed consent and confidentiality are basic to the patient-physician relationship.

Autonomy: Patient has freedom of thought, intention and action. Patient should know all the risks, benefits of the procedure and likelihood of success before making the decision.

Beneficence: The main aim of the procedure is to do good to the patient. Considerate the patient's welfare.

Confidentiality: Personal, medical and treatment information should be kept confidential. If the information is crucial for the patient, only in that case it can be revealed.

Non-Maleficence: Making sure that the procedure doesn't harm the patient or others in the society.

Justice and Equity: Fair and equal distribution of scarce health resources and decision of who gets what treatment.

Scientific digital image acquisition and manipulation guidelines

1. Scientific digital images are data that can be compromised by inappropriate deceptions or manipulations.
2. Digital image manipulation should always be done in the data copy of the unprocessed image.
3. Simple adjustments throughout an image are generally acceptable.

4. Use of software filters to improve image quality is usually not recommended for biological images.
5. Manipulations that are specific to one area of an image and are not performed on other areas are questionable.
6. Comparable digital images should be obtained under the same conditions, and any post-acquisition image processing should be the same.
7. Cloning or copying objects into a digital image from other parts of the same image or from a different image, is very questionable.
8. Avoiding use of lossy compression.
9. Intensity measurements should be performed on uniformly processed image data, and the data should be calibrated to a known standard.
10. Image cropping is acceptable.

Our work entirely satisfies all the ethical consideration of pathology being referred in [12].

Budget Description

The principal entity which demanding monetary expenditure is the training of the Deep learning model on the training dataset. Dedicated hardware is required to be purchased for training of the model, namely, abbreviated as GPU.

GPU is basically a Graphic Processing Unit, originally designed to accelerate graphics rendering. GPUs can process many pieces of data simultaneously, making them useful for machine learning, video editing, and gaming applications. GPUs may be integrated into the computer's CPU or offered as a discrete hardware unit. GPU accelerates display rendering, zooming, and navigation, but the bulk of the actual processing happens on the CPU. Range of GPU hardware is also available within different price brackets starting from $850 and goes as high as $2000.

AMD RX 6800 is the right and most feasible option for our project in terms of purchasing physical equipment for training the model.

Other than that, there are also cloud based GPU options which are cheap as well as equally efficient. Range of Nvidia cloud based GPU provide significant amount of memory in different variants starting from 8GB upto 40GB with small price of $0.80 to $3.0 per GPU. Google colab also provide GPU for training model upto 6 hours and kaggle gives 40GB of GPU for training without any cost.

Following table summarizes the priority of hardware/software for training the model.

Name	**Price**	**Memory**	**Time**
Google Colab	Free	6GB	6 hours
Kaggle	Free	40GB	N/A
Nvidia A100	$0.80/month	8GB	N/A
Colab Pro	$9.99/month	12GB	6 hours
AMD RX 6800	$850	4GB	N/A

Table 4: GPUs Priorities

References

[1] A. T. S. M. a. T. W. Ausawalaithong, "Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach.," in *11th Biomedical Engineering International Conference (BMEiCON),*, Chiang Mai, 2018.

[2] T. T. K. Y. a. F. H. T. Atsushi, "Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks," in *BioMed Research International.* , 2017.

[3] H. D. Y. M. A. K. a. S. J. W. Rahane, "Lung Cancer Detection Using Image Processing and Machine Learning HealthCare," in *International Conference on Current Trends towards Converging Technologies (ICCTCT),*, Coimbatore, 2018.

[4] M. R. M. S. a. M. S. M. Šarić, "CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images.," in *4th International Conference on Smart and Sustainable Technologies (SpliTech), Split.*, Croatia, 2019.

[5] M. B. B. R. S. S. Sasikala, "Lung Cancer Detection and Classification Using Deep CNN.," 2019.

[6] S. C. a. H. Rajaguru, "Lung Cancer Detection using Probabilistic Neural Network with modified Crow-Search Algorithm.," *Asian Pacific Journal of Cancer Prevention,* vol. 20, no. 7, pp. 2159-2166, 2019.

[7] B. Hatuwal and H. Thapa, "Lung Cancer Detection Using convolutional Neural Network on Histopathological Images," *International Journal of Computer Trends and Technology,* vol. 68, no. 10, pp. 21-24, 2020.

[8] Y. T. Y. H. Y. W. M. Z. G. &. Z. J. Zhu, "Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography.," *Journal of digital imaging,* vol. 23, no. 1, pp. 51-65, 2010.

[9] Q. Z. L. L. X. &. D. X. Song, "Using deep learning for classification of lung nodules on computed tomography images.," *Journal of healthcare engineering,* 2017.

[10] S. K. M. S. N. S. K. N. &. R. G. Lakshmanaprabu, "Optimal deep learning model for classification of lung cancer on CT images.," *Future Generation Computer Systems,* vol. 92, pp. 374-382, 2019.

[11] &. S. G. S. Garg, "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," *arXiv,* pp. 1-12, 2021.

[12] M. Cocks, "Ethical Considerations in Pathology," *AMA Journal of Ethics,* 2016.

Undertaking

I certify that SDP work titled “Lung Cancer Detection using Machine Learning for Histopathological Images” is my own work. The work has not, in whole or in part, been presented elsewhere for assessment. Where material has been used from other sources it has been properly acknowledged/referred.

Signature of Student
Sabahat Tabbassum
18-EE-01

Signature of Student
Azib Farooq
18-EE-43

Signature of Student
Syed Irtaza Hussain
18-EE-46

Signature of Student
Javeria Noor Tariq
18-EE-53

Supervisor’s Comments

Signature of Supervisor

Dr. Gulistan Raja

Professor