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. Globally, 9.6 million deaths accounted are due to cancer. Lung cancer contributed 2.06 million cases in the above figure. 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. Their energy also wipes out due to large number of cases, which 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. Artificial intelligence techniques has influenced individuals from all walks of professions for betterment by deploying robust techniques to automate tedious and tiresome methods.

In our problem,

- We will be using deep learning approach to detect lung cancer from histopathology images.
- The obtained results will be validated and model performance will be accessed by numerous comparative metrics.

Aims & Objectives

Following are the primary goals in accordance to the problem statement.

- Deep Learning Approach
    - Implement ResNet50 transfer learning models on dataset and identify lung cancer and its subtypes.
- Validation
    - Apply four commonly used accuracy metrics on the above two models.
    - Obtained more than 50% accuracy on all the metrics.

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. 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].

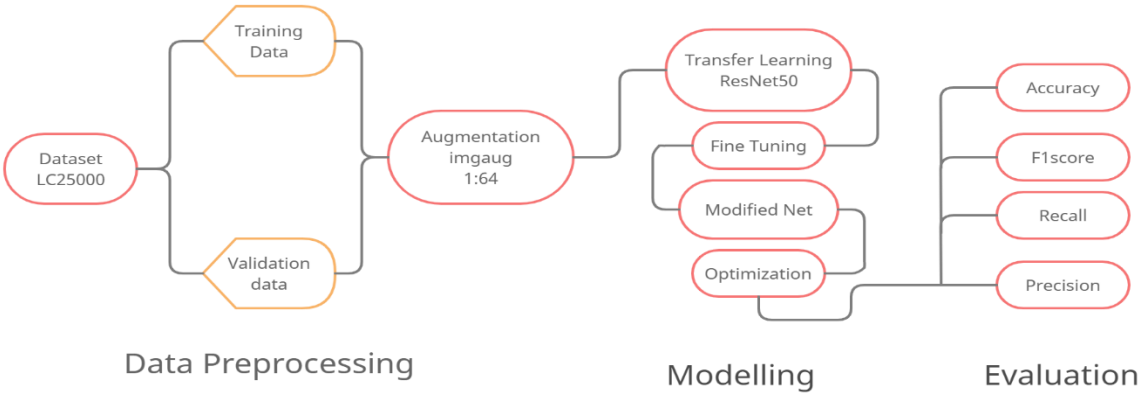
Model	Precision	Recall	F1-score	Accuracy	Auroc
VGG16	0.975	0.975	0.98	0.98	0.999
ResNet50	0.965	0.965	0.96	0.96	0.999
InceptionV3	1.0	1.0	1.0	1.0	1.0
InceptionResNetV2	1.0	1.0	1.0	1.0	1.0

MobileNet	1.0	1.0	1.0	1.0	0.999
Xception	1.0	1.0	1.0	1.0	1.0
NASNetMobile	0.965	0.965	0.97	0.97	0.997
DenseNet169	1.0	1.0	1.0	1.0	0.999

Table1: Evaluation of eight Pre-trained CNN models identifying type of lung cancer [11]

SDP Methodology

The below schematic summarizes the entire methodology and it consists of three primary steps, which are also further divided into sub-steps.



Data Preprocessing

- 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.

Tasks	Training data 80%			Test data 20%		
	ACA	SCC	Ben	ACA	SCC	Ben
Lung cancer	2000	2000	4000	500	500	1000
Subtype	4000	4000	-	1000	1000	-

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.

Modeling

- 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.

Evaluation and Expectation

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.

$$T_p = Correct\ positive\ class\ prediction$$

$$T_n = Correct\ negative\ class\ prediction$$

$$F_p = Wrong\ positive\ class\ prediction$$

$$F_n = Wrong\ negative\ class\ prediction$$

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

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

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

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

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

The model is expected to achieve performance more than 50% on all the above metrics.

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

Ethical Issues

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

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