TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)
7-10 Dec 2021. Auckland, New Zealand

Detecting Lung Cancer from Histopathological Images using Convolution Neural Network

Dewan Ziaul Karim
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
ziaul.karim@bracu.ac.bd

Tasfia Anika Bushra
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
anika.cse@diu.edu.bd

***Abstract*— Lung cancer is one of the leading causes of mortality in both men and women throughout the world. That is why early identification and treatment of lung cancer patients bear a huge significance in the recovery procedure of such patients. A lot of time, pathologists use histopathological pictures of tissue biopsy from possibly diseased regions of the lungs to detect the probability and type of cancer. However, this procedure is both tedious and sometimes fallible too. Machine learning based solutions for medical image analysis can help a lot in this regard. The aim of this work is to provide a convolution neural network (CNN) model that can accurately recognize and categorize lung cancer types with superior accuracy which is very important for treatment. We propose a CNN model with 15000 images split into 3 categories: Training, validation, and testing. Three different types of lung tissues (Benign tissue, Adenocarcinoma, and squamous cell carcinoma) have been examined. 50 instances from every class were kept for testing procedure. The rest of the data was split as: about 80% and 20% for training and validation respectively. Eventually, our model obtained 98.15% training accuracy and 98.07% validation accuracy.**

Keywords—Lung Cancer, Histopathological Images, Deep Learning, CNN, Classification.

I. INTRODUCTION

Lung cancer is regarded as one of the most prominent cancers in the whole wide world. It makes up around 25% of all cancer related deaths [1]. The most common reason behind lung cancer is smoking. However, in the case of non-smokers, exposure to radon, second-hand smoking, air pollution, or certain other substances can all cause lung cancer [2]. Unfortunately, the mortality rate of lung cancer is on the rise and it is supposed to become about 17 million worldwide in the year 2030 [3]. In case of developing countries, at the current growth rate, people's odds of acquiring cancer during their lifespan may rise up to 50%-60% by 2050 [4].

There are many medical tests (CT scan, X-rays, biopsy, etc.) done to find out potential cancerous cells. In a biopsy, histopathology slides are evaluated by pathologists to establish the potential diagnosis [5,6,7] and determine the type of lung cancers [8]. But it is a time-consuming procedure and there is always a chance that cancer types could be misdiagnosed, which eventually results in incorrect treatment and puts a toll on patients' lives.

For the reason mentioned above, it is essential to implement an automated system for assisting doctors in the diagnosis of lung cancers as early as possible with high accuracy. Due to advancements in the technological sector, it is now possible to build such an automated system using artificial intelligence (AI) and machine learning (ML).

ML is considered as a branch of AI that concentrates on using algorithms and data to emulate the way that people learn and improve accuracy over time [9]. In recent years, many researchers considered combining different machine learning techniques with x-rays and CT images to provide a workable system for identifying types of lung cancer. These techniques involve Random Forest (RF), Support Vector Machine (SVM), Bayesian Networks (BN), and Convolution Neural Network (CNN) for detecting and recognizing lung cancers. Recently, some authors considered using histopathological images to differentiate between carcinomas and non-carcinomas images using CNN.

CNN is an approach under deep learning that is widely used in image recognition and classification [10,11,12]. It usually considers an input image, allocates biases and weights to the images and distinguishes one image from another. CNN is superior to other conventional approaches in a sense that it needs a very low amount of preprocessing. Meaning that in other traditional techniques, filters have to be set up manually, whereas the neural network obtains the information itself.

CNN is frequently used for image-related tasks including classification, segmentation, medical image analysis, recognition, etc. because it has numerous benefits over other methods. After providing the input images in CNN, they go through several convolution layers like flattening, pooling, and fully-connected (FC) layers. Some types of activation functions are also used in order to perfectly identify an image.

The primary aim of our research is to provide a feasible, efficient and accurate ML model to detect lung cancer from histopathological images by classifying benign tissue, adenocarcinoma, and squamous cell carcinomas using CNN architecture.

II. RELATED WORK

The authors Bijaya Kumar Hatuwal and Himal Chand Thapa [13] created a deep CNN model to identify benign tissue, adenocarcinoma, and squamous cell carcinoma where there were three hidden layers, one input layer, and one fully connected layer. A dropout value of 0.1 and max-pooling were

TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON) | 978-1-6654-9532-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/TENCON54134.2021.9707242

used in their research. They used "Adam" optimizer and eventually got 96.11% accuracy in training and 97.2% accuracy in validation.

Muayed S AL-Huseiny et al. [14] proposed the approach of deploying a transfer learning based deep neural network (DNN) to detect lung nodules that are malignant using CT images. They performed a fast pre-processing technique to find out the ROI (Region of Interest) from the images. In this work, GoogLeNet DNN was used and modified for their dataset. The code was run in a machine having a processor of 2.5 GHz (Core-i3) with 16 GBs of ram and eventually achieved an accuracy of 94.38%.

Another paper [15] described a lung cancer detection system using Alexnet CNN. This work only distinguished between malignant and benign lung tumors with the help of a model based on convolution neural network and AlexNet. It is to be mentioned that AlexNet is made up of 25 layers (with a scale of 227x227x3). SGDM optimization model and an initial learning rate of 0.0003 were used. MATLAB 2021a software was used to run the code and the proposed method achieved 96% accuracy in the end.

Ying Su et al. [16] proposed an approach for detecting lung nodules using Faster R-CNN. They experimented on the LIDC-IDRI dataset [17]. They used 0.001 as learning rate and 70000 as step size. Their attenuation coefficient, dropout rate, and batch size were 0.1, 0.5, and 64 respectively. The researchers achieved an accuracy of 91.2% with their optimized and improved Faster R-CNN method.

Mehedi Masud et al. [18] suggested a classification framework that differentiates among 5 different types of colon and lung tissues by analyzing their histopathological images. Among those 5 classes, 2 are benign and 3 are malignant. A total of 25000 pictures were included in the dataset. The authors used DFT and DWT techniques for feature extraction from images. Later they used a CNN based technique to identify cancer tissues with an accuracy of 96.33%. Satvik Garg et al. [19] conducted another research that demonstrated the results of various pre-trained CNN models.

Another work [20] suggested an automated system for detecting lung malignancies in WSI (Whole Slide Images) of lung tissues using two CNN architectures - ResNet and VGG16. The target was to identify image patches into normal and tumor cells. The authors used SGD as the optimizer. Binary crossentropy was assigned as the loss function and a learning rate of 0.0001 was chosen. Finally, it was observed that VGG16 (75.41%) outperformed ResNet (72.05%) in terms of patch level accuracy.

Albert Chon [21] et al. presented a Googlenet-based 3D CNN model for lung cancer detection. The dataset contained labeled data for 2101 patients, which the authors divided into training, validation and test set size of 1261, 420, and 420. A dropout with 0.3 probability was used after each convolution and inception layers during training. They used "Adam" optimizer with 0.0001 learning rate. It was seen that the suggested model achieved an accuracy of 75.1% with an AUC score of 0.757.

III. Dataset Description

The dataset used in this study contains 15000 lung histopathology images. This dataset is obtained from LC25000 Lung and colon histopathological image dataset [22]. Those 15000 images are divided into 3 different categories: benign tissue, adenocarcinoma, and squamous cell carcinoma. Among those 15000 images, 11850 were put into training, 3000 were used for validation and 150 were kept for testing purposes. The pictures were all in RGB format, with 256 X 256 pixel sizes. Some samples from different classes is shown below:

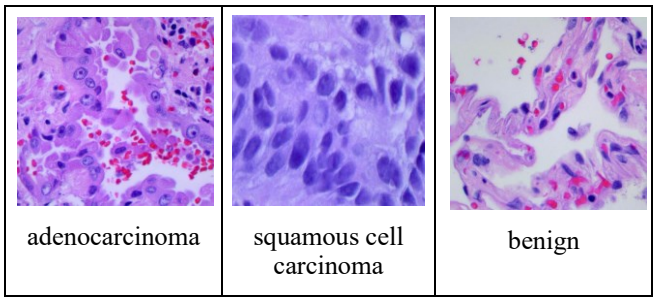

Fig. 1. Dataset Sample

IV. Methodology

In this study, a CNN model has been created to detect 3 classes of lung cancers. Fig. 2 indicates the complete workflow of this research.

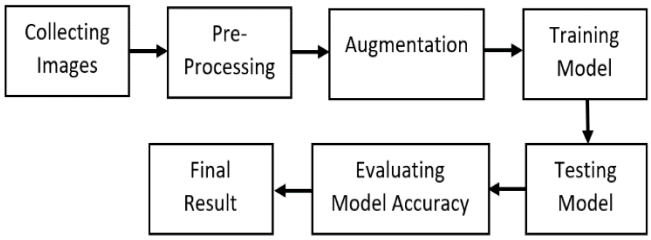

Fig. 2. Methodology of Detecting Lung Cancer

The procedure can be divided into 2 main steps: i) Preparation of dataset ii) Implementing CNN model

A. Dataset Preparation

To avoid getting a disappointing result, it is always better to pre-process the dataset to increase efficiency [23]. In our work, various steps were considered to prepare the training dataset.

- Outliers Removal: The dataset was examined rigorously for any outliers as outliers can affect the performance of our model.
- Resizing Images: All the images were scaled to a pixel size of 256 x 256 as CNN models tend to take a fixed dimension as inputs.
- Dataset Normalization: Normalized data can help deep learning based models gain more stability and provide a better chance of convergence. The range of pixel values in a picture is 0 to 255. So we used Minmax (1) normalizer to normalize the pixel values of our images.

$$Z = X - X_{min} / (X_{max} - X_{min}) \quad (1)$$

- Data Augmentation: Usually CNN models perform better with more images. Hence, we applied some data augmentation methods to expand our training data. Techniques such as shearing, rotating, shifting, flipping, etc. were applied to bring variety to the dataset and make the model more robust.

B. Proposed Model's Architecture

In this work, we suggest a multi-layered CNN model to classify different types of lung cancers from histopathological images. There are 6 convolution layers and 3 dense layers in our CNN model. There are 32,64,128,128,128, and 64 filters respectively in those 6 convolution layers with 3 x 3 kernel size. All the convolution operations are followed by Batch Normalization [24] operation (2) which helps to make the learning procedure faster. Following that, a Max-pooling [25] procedure with a pool size of 2 x 2 was performed.

$$\hat{X}_i = \frac{X_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \epsilon}} \quad (2)$$

Since convolution networks work better with ReLU [26], all the convolution layers use "ReLU" as activation function (3).

$$Y = max(0, X) \quad (3)$$

A flatten layer is designed just after those 6 convolution layers. It helps in the process of converting data into a one-dimensional array for usage in the next layer. After this, 3 consecutive dense layers are implemented with 512, 64, and 3 units respectively. There are 3 nodes in the last dense layer as we are trying to classify 3 different types of lung tissues. A softmax activation function (4) was applied in the last dense layer.

$$\frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_i}} \quad (4)$$

TABLE I. PROPOSED MODEL SUMMARY

Layers	**Shape of Output**
conv2d_0	(None,254,254,32)
batch_normalization_0	(None,254,254,32)
max_pooling2d_0	(None,127,127,32)
conv2d_1	(None,125,125,64)
batch_normalization_1	(None,125,125,64)
max_pooling2d_1	(None,62,62,64)
conv2d_2	(None,60,60,128)
batch_normalization_2	(None,60,60,128)
max_pooling2d_2	(None,30,30,128)
conv2d_3	(None,28,28,128)
batch_normalization_3	(None,28,28,128)
max_pooling2d_3	(None,14,14,128)
conv2d_4	(None,12,12,128)
batch_normalization_4	(None,12,12,128)
max_pooling2d_4	(None,6,6,128)
conv2d_5	(None,4,4,64)
batch_normalization_5	(None,4,4,64)
max_pooling2d_5	(None,2,2,64)
flatten_1	(None,256)
dense_0	(None,512)
batch_normalization_6	(None,512)
dense_1	(None,64)
batch_normalization_7	(None,64)
dense_2	(None,3)
activation	(None,3)
Total params: 631,299 Trainable params: 629,059 Non-trainable params: 2,240	

C. Parameters used in Training

For our proposed model, we tried to use multiple parameters e.g., optimizer, learning rate, metrics, batch size, epoch numbers, callbacks, etc. Table II indicates the various training parameters used in our model:

TABLE II. TRAINING PARAMETERS USED IN THE MODEL

Name of Parameter	**Value**
Used Optimizer	Adam
Learning Rate (Initial)	0.01
Learning Rate (Minimum)	.000001
Regularizer	L1 (0.000001)
Batch Size	20
Epochs	60
Steps per Epoch	593
Loss Function	Categorical Crossentropy
Metrics	Accuracy, Precision, Recall, Loss
Callbacks	ReduceLROnPlateau

D. Evaluation Tools

Python version 3.X was used for the whole experiment including dataset preparation, model implementation, and evaluation.

V. Result Analysis

From the whole dataset, 50 images from each class were kept aside for testing purposes. The remaining images were split in such a way that about 80% data went into training and 20% went into validation. The model finally achieved a training and validation accuracy of 98.15% and 98.07% respectively. Fig. 3 and 4 indicate accuracy and loss graphs for both training and validation respectively.

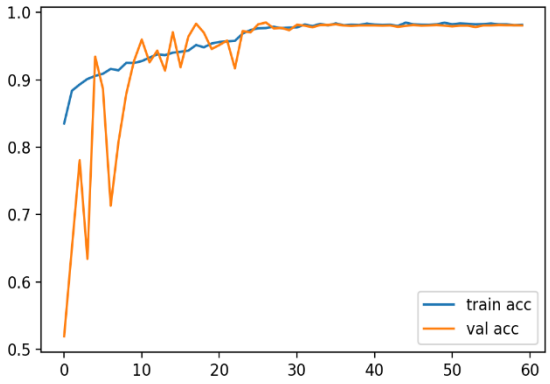

Fig. 3. Accuracy for Both Training and Validation

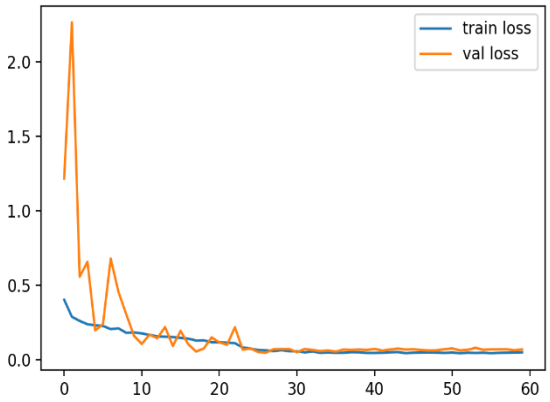

Fig. 4. Loss for Both Training and Talidation

Moreover, we also calculated the accuracy of different pre-trained CNN models for the same dataset along with same hyperparameters and compared with the result of our proposed CNN model. The different models that we tried out are DenseNet201, ResNet152V2, MobileNetV2, InceptionV3, Xception, InceptionResNetV2, VGG16, VGG19 and ResNet50. It was seen that all of those models performed poorer than our proposed model. Among the pre-trained models, DenseNet201 and MobileNetV2 achieved the highest training and validation accuracy of 95.41% and 95.03% respectively. Nevertheless, both of these are lower than the training and validation accuracy achieved by our proposed model. As a result, we came to the conclusion that compared to the different transfer learning approaches, our approach to lung cancer diagnosis has demonstrated better results with greater accuracy rates.

Table III shows the comparison between the accuracy rate of pretrained models and our suggested CNN model against the same dataset.

TABLE III. Training and Validation Accuracy Comparison of Proposed and Pre-trained CNN Models

Model	Training Accuracy	Validation Accuracy
Proposed Model	98.15%	98.07%
DenseNet201	95.41%	94.10%
ResNet152V2	94.55%	93.53%
MobileNetV2	94.23%	95.03%
InceptionV3	93.79%	93.20%
Xception	93.72%	92.30%
InceptionResNetV2	93.00%	92.60%
VGG16	91.91%	91.77%
VGG19	90.62%	82.50%
ResNet50	74.68%	51.50%

Fig. 5 and 6 indicate training and validation accuracy graphs for different models respectively.

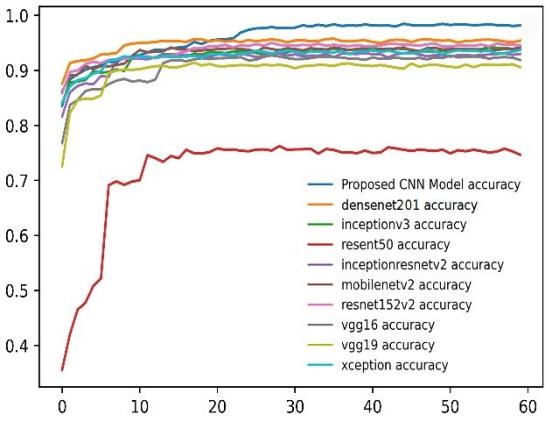

Fig. 5. Training Accuracy Comparison of Different Models

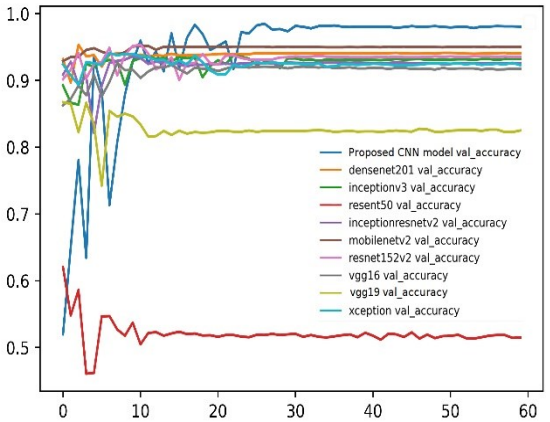

Fig. 6. Validation Accuracy Comparison of Different Models

To understand our results better, we noticed the confusion matrix based on test samples. It is important in the sense that it provides a clear overview of samples being classified correctly or incorrectly [27].

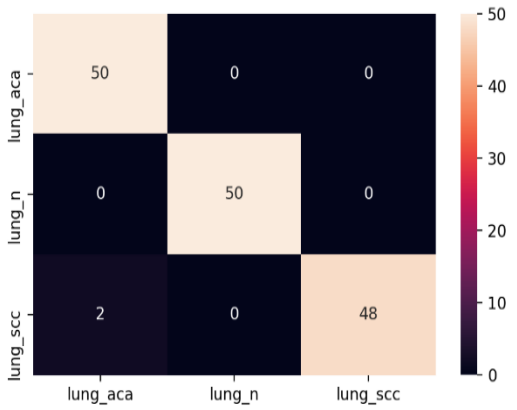

Fig. 7. Confusion Matrix on Test Samples

Fig. 7 exhibits the model's confusion matrix on our selected test(unseen) samples that include lung adenocarcinoma (lung_aca), lung squamous cell carcinoma (lung_scc), and lung benign tissue (lung_n).

If we look at the confusion matrix, we notice that our model identified all of the samples from lung adenocarcinoma and lung benign tissues with an accuracy of 100%. However, for the class squamous cell carcinoma, 48 instances were classified correctly and 2 were wrongly classified.

Observing the value of Recall (R), Precision (P), and F1 score on test samples is another good idea to check the reliability of any model [28]. The formula for F1 is = 2*((R * P) / (R + P)). Precision is calculated using the formula = TP / (TP + FP). Dividing TP by the addition of TP and FN gives us Recall. Here, TP is True Positive, FP is False Positive and FN is False Negative. Table IV illustrates precision, recall and F1 score for every category in our test sample dataset.

TABLE IV. CLASSIFICATION REPORT BASED ON TEST SAMPLES

Name of Classes	Precision	Recall	F1 Score	Support
Lung Adenocarcinoma	.96	1	.98	50
Lung Benign	1	1	1	
Lung Squamous Cell Carcinoma	1	.96	.98	
Accuracy	.99			50*3=150
Marco Average	.99	.99	.99	
Weighted Average	.99	.99	.99	

VI. FUTURE WORK

In the future, different CNN architecture with some hyperparameters tuning may result in better accuracy than the current one. This work may be extended to CT scan imaging problems too. It may also be possible to build a mobile application that will provide real time detection and eventually widen the utilization of our technique.

VII. CONCLUSION

This work represents a CNN model to detect lung cancer using histopathological images. The whole dataset consisted of 15000 images and our experimental findings indicated training and validation accuracy of 98.15% and 98.07% respectively. It is expected that this model will help pathologists to identify lung cancer (benign, adenocarcinoma, squamous cell adenocarcinoma lung tissues) with less time, effort and cost.

REFERENCES

[1] (2020) "American Cancer Society, Lung Cancer Statistics.[Online]".Available: https://www.cancer.org/cancer/lungcancer/about/key-statistics.html

[2] (2019) "American Cancer Society, Lung Cancer Causes. [Online]." Available: https://www.cancer.org/cancer/lungcancer/causes-risks-prevention/what-causes.html

[3] Nie, L., Zhang, L., Yang, Y., Wang, M., Hong, R. and Chua, T.S., 2015, October. Beyond doctors: Future health prediction from multimedia and multimodal observations. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 591-600).

[4] Kumar, P., Bhattacharyya, G.S., Dattatreya, S. and Malhotra, H., 2009. Tackling the cancer Tsunami. Indian journal of cancer, 46(1), p.1.

[5] Silvestri, G.A., Gould, M.K., Margolis, M.L., Tanoue, L.T., McCrory, D., Toloza, E. and Detterbeck, F., 2007. Noninvasive staging of non-small cell lung cancer: ACCP evidenced-based clinical practice guidelines. *Chest*, *132*(3), pp.178S-201S.

[6] Travis, W.D., Brambilla, E., Noguchi, M., Nicholson, A.G., Geisinger, K.R., Yatabe, Y., Beer, D.G., Powell, C.A., Riely, G.J., Van Schil, P.E. and Garg, K., 2011. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *Journal of thoracic oncology*, *6*(2), pp.244-285.

[7] Collins, L.G., Haines, C., Perkel, R. and Enck, R.E., 2007. Lung cancer: diagnosis and management. *American family physician*, *75*(1), pp.56-63.

[8] Yu, K.H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L. and Snyder, M., 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*, *7*(1), pp.1-10.

[9] Michie, D., Spiegelhalter, D.J. and Taylor, C.C., 1994. Machine learning, neural and statistical classification.

[10] O'Shea, K. and Nash, R., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

[11] Hijazi, S., Kumar, R. and Rowen, C., 2015. Using convolutional neural networks for image recognition. Cadence Design Systems Inc.: San Jose, CA, USA, pp.1-12.

[12] Sultana, F., Sufian, A. and Dutta, P., 2018, November. Advancements in image classification using convolutional neural network. In 2018 Fourth International Conference on Research in Computational

Intelligence and Communication Networks (ICRCICN) (pp. 122-129). IEEE.

[13] Hatuwal, B.K. and Thapa, H.C., 2020. Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images. *Int. J. Comput. Trends Technol*, *68*, pp.21-24.

[14] AL-Huseiny, M.S. and Sajit, A.S., 2021. Transfer learning with GoogLeNet for detection of lung cancer. *Indonesian Journal of Electrical Engineering and Computer Science*, *22*(2), pp.1078-1086.

[15] Agarwal, A., Patni, K. and Rajeswari, D., 2021, July. Lung Cancer Detection and Classification Based on Alexnet CNN. In *2021 6th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1390-1397). IEEE.

[16] Su, Y., Li, D. and Chen, X., 2021. Lung nodule detection based on faster R-CNN framework. *Computer Methods and Programs in Biomedicine*, *200*, p.105866.

[17] Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A. and Kazerooni, E.A., 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, *38*(2), pp.915-931.

[18] Masud, M., Sikder, N., Nahid, A.A., Bairagi, A.K. and AlZain, M.A., 2021. A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. *Sensors*, *21*(3), p.748.

[19] Garg, S. and Garg, S., 2020, December. 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* (pp. 38-45).

[20] Šarić, M., Russo, M., Stella, M. and Sikora, M., 2019, June. CNN-based method for lung cancer detection in whole slide histopathology images. In *2019 4th International Conference on Smart and Sustainable Technologies (SpliTech)* (pp. 1-4). IEEE.

[21] Chon, A., Balachandar, N. and Lu, P., 2017. Deep convolutional neural networks for lung cancer detection. *Standford University*.

[22] Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A. and Mastorides, S.M., 2019. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*.

[23] Pal, K.K. and Sudeep, K.S., 2016, May. Preprocessing for image classification by convolutional neural networks. In 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1778-1781). IEEE.

[24] Ioffe, S. and Szegedy, C., 2015, June. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.

[25] Scherer, D., Müller, A. and Behnke, S., 2010, September. Evaluation of pooling operations in convolutional architectures for object recognition. In International conference on artificial neural networks (pp. 92-101). Springer, Berlin, Heidelberg.

[26] Agarap, A.F., 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

[27] Ting, K.M., 2017. Confusion matrix. *Encyclopedia of Machine Learning and Data Mining*, *260*.

[28] Goutte, C. and Gaussier, E., 2005, March. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European conference on information retrieval* (pp. 345-359). Springer, Berlin, Heidelberg.