A Deep Learning based Image Processing Technique for Early Lung Cancer Prediction

Nowshin Tasnim
Dept. of Computer Science & Engineering
Chittagong University of Engineering & Technology
Chittagong-4349, Bangladesh
tasnimnowshin95@gmail.com

Kazi Rifah Noor
Dept. of Computer Science & Engineering
United International University
Dhaka 1212, Bangladesh
rifahkazi@gmail.com

Mursalina Islam
Dept. of Computer Science & Engineering
United International University
Dhaka 1212, Bangladesh
mislam2320022@mscse.uiu.ac.bd

Mohammad Nurul Huda
Dept. of Computer Science & Engineering
United International University
Dhaka 1212, Bangladesh
mnh@cse.uiu.ac.bd

Iqbal H Sarker
Edith Cowan University
Perth, WA-6027, Australia
m.sarker@ecu.edu.au

2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS) | 979-8-3503-7222-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICETSIS61505.2024.10459494

***Abstract*—Lung cancer is the primary cause of cancer mortality all over the world due to the increase of tobacco consumption, and industrialization in developing nations. As the early-stage diagnosis can reduce the mortality rate significantly, early detection with the availability of high-tech Medical facilities is highly necessary. In this research, we used deep learning (DL) methods initially on patient's 1190 CT scan images from the Kaggle IQ-OTH lung cancer dataset, and after significant image preprocessing steps we found augmented images including normal, malignant, and benign cases to identify high-risk individuals to detect lung cancer and also predict the malignancy and thus, taking early actions to prevent long-term consequences. A thorough performance comparison between several classifiers, including the conventional CNN, Resnet50, and InceptionV3, has been presented. Here, affine transformation, gaussian noise, and other rigorous image preprocessing techniques are used. The contribution obtained a 98% validation accuracy while reducing the model's complexity with the previous preprocessing stage. The comparison method shows that the suggested preprocessing method yields a higher F1 score value of 97%, validating our suggested methodology.**

***Index Terms*—healthcare analytics, lung cancer prediction, CT scans, image processing, deep learning**

I. Introduction

Lung cancer was the most familiar reason for universal cancer deaths in 2018, demonstrating almost 9.6 million deaths from all forms of cancer [1]. The most common symptoms of lung cancer are relentless coughing with blood, shortness of breath, chest pain, and fatigue [2]. Lack of screening programs, limited rural health care facilities, late physician consultations after symptom appearance, and delayed diagnostic tests such as CT scans and biopsies are all liable for late-stage diagnosis of lung cancer. The diagnostic techniques include physical examination, imaging such as chest X-rays, computed tomography scans, bronchoscopy, histopathology examination, etc. [3]. To reduce these issues recently several works have been proposed on lung cancer detection using deep learning models on CT scan images.

Several Machine- Learning classifiers encompassing Support Vector Machine (SVM), Decision Tree, Multi-Layer Perceptron, Neural Network, and Naïve Bayes (NB) are discussed in [4] to predict early-stage lung cancer using the UCI repository dataset and also compared the results with ensembles such as Random Forest and Majority Voting where Gradient-boosted Tree exceeded all others' performance and ensemble classifiers and accomplished 90% accuracy. In another paper [5], authors predicted lung cancer using KNN, Naive Bayes, and SVM and achieved an accuracy of 81.25% with the RBF classifier using the above mentioned UCI dataset. Moreover, in this paper [6], researchers classified lung cancer into malignant and benign tumors using various image pre-processing along with feature extraction techniques. They also implemented decision trees, random forests, SVM, and neural network models and attained an accuracy of 95.2%. The authors in this paper [7], created a large-scale chest X-ray database, called ChestX-ray8, which contains 108,948 frontal-view X-ray images of 32,717 unique patients with eight common disease labels and implemented a weakly supervised multiple labels image categorization and disease localization framework to deal with the challenge of fully automated high-precision computer-aided diagnosis. This paper [8] proposed a lung cancer detection technique using image processing, segmentation techniques on CT images, and feature extraction techniques to differentiate between malignant and benign nodules. They used SVM, linear discriminate analysis, and artificial neural networks for classification achieving an accuracy rate of 90.12% in their experiment. Similarly, In this paper [9], researchers implemented Convolutional Neural Networks (CNNs) and evaluated their performance on chest X-ray and CT scan datasets where CNNs achieved high accuracy in diagnosing lung diseases, outperforming traditional machine learning

methods. They also discussed the constraints of applying deep learning techniques, such as the demand for an abundance of data and the high dimensionality of the data.

We are concerned about the difficulties and scope of our study, which is why we decided to use CT scan images for our experiment since they are more explicit in accurately detecting cancer. Several deep learning methods, such as CNN, Resnet50, and InceptionV3, were employed following a substantial image preprocessing stage. We used 1190 CT scan slices or images from 110 IQ-OTH lung cancer cases from Kaggle, which are divided into three classes: malignant, benign, and normal. Shortly, in this project pre-processing (removing noise if any), postprocessing (segmentation), and classification techniques have been used to categorize tumors into each of the three categories i.e. Normal, Malignant, and Benign where benign indicates a non-cancerous tumor, and the other areas are not affected by it. After augmentation newly generated dataset is used for further prediction steps. Our study aims,

- to develop an efficient image preprocessing technique on CT scan images.
- to show the detailed performance comparison among Resnet50, InceptionV3, and our proposed approach with an augmented dataset.
- to effectively predict early Lung Cancer using deep learning approaches.

II. Methodology

A. Dataset

The IQ-OTH lung cancer dataset is used here to verify the proposed model. The aforementioned dataset of Iraq-Oncology Teaching Hospital and National Center for Cancer Diseases (IQ-OTH/NCCD) was gathered during a three-month period in the fall of 2019. It includes patients' CT scans that have been diagnosed with lung cancer at numerous stages. In these two centers, radiologists and oncologists marked IQ-OTH/NCCD slides. A total of 1097 images, or CT scan slices from 110 cases, are included in the dataset. A sample image visualization of every class is presented in Fig. 1.

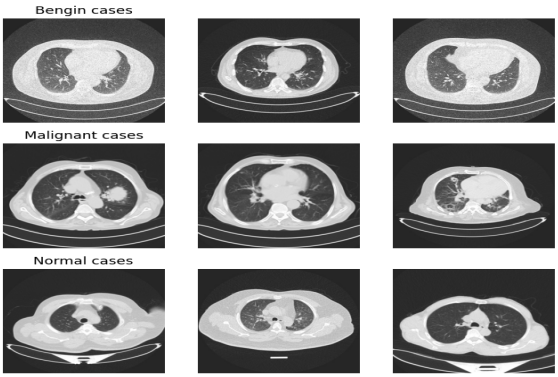

Fig. 1. Visulizing sample images from IQ-OTH dataset.

With this dataset, deep-learning image classification models may have difficulties due to the various dimensions of the pictures. Here, 120 images of benign cases have the dimension 512×512. 501 malignant cases and 415 normal cases have similar dimensions. But 31 malignant cases have a dimension of 512×623, whereas only one malignant case has a dimension of 404×511. Preprocessing the data is a very important step to ensure the effectiveness of the DL model.

B. Dataset Preparation

The three classes of these cases are malignant, benign, and normal. Out of these, 120 cases have been classified as benign cases, 561 as malignant cases, and 416 as normal cases. Here the imbalance in different classes is evident. We have balanced each class using the SMOTE approach, which lessens bias in the classification model. This procedure should be followed, particularly during disease diagnosis tasks, to prevent undue advantages to the majority class. Fig. 2 represents the bar chart of 3 class distribution.

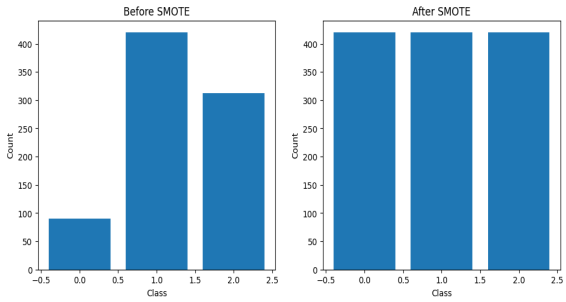

Fig. 2. Class distribution of IQ-OTH dataset.

In terms of accurate disease type or level prediction, medical data analysis and predictive research require the utmost sensitivity. To ensure the validity of our approach, we carried out several stages of data preprocessing, which encompassed image transformation and augmentation. Normalization is done in the first step of preprocessing data with the use of the mean and standard deviation values. To help the model learn more easily, normalization is a common preprocessing step in deep learning that brings the image pixel values to a standard scale. However, medical imaging is required for the early detection and diagnosis of various diseases, including cancer. In the realm of deep learning, the success of models commonly relies on the quality and diversity of the training data. Image augmentation techniques address this need by artificially enhancing the size of the training dataset through the application of various transformations. In this article, we delve into a series of image augmentation steps tailored for medical imaging, specifically focusing on CT scan data for early lung cancer prediction. Image augmentation incorporates applying a variety of transformations to the genuine images, creating diverse training samples. This diversification is crucial for training robust and generalized deep learning models. The augmentation steps discussed here are implemented using

the imgaug library, a powerful tool for creating complex augmentation pipelines. Augmentation Steps followed here,

- Rotation: The rotational augmentation involves variations in the orientation of the lung structures by rotating the CT scan images by 90, 180, or 270 degrees.
- Horizontal Flipping: In the context of CT scans, flipping the lung images provides the model with additional variations in anatomical arrangements.
- Affine Transformations: Affine transformations, such as translation and scaling, simulate variations in positioning and scanner settings. Translation introduces shifts in the image, while scaling simulates changes in the field of view. These transformations contribute to the model's adaptability to different clinical scenarios.
- Cropping and Padding: Randomly cropping or padding the images by a specified pixel range introduces spatial variations. This is particularly relevant in medical imaging, where patients' positions and the framing of scans can vary widely.
- Gaussian Noise: Adding Gaussian noise copies the inherent noise present in medical images. This step aids in making the model more resilient to noisy datasets and enhances its ability to generalize to real-world scenarios.
- Linear Contrast Adjustment: Adjusting the linear contrast of the images helps the model adapt to varying illumination conditions. This is essential for ensuring the model's robustness when faced with differences in brightness and contrast across different CT scans.

After augmentation newly generated image dataset is presented here in Fig. 3. The preprocessing steps are visualizing the following three major steps, original image, augmented image, and gaussian blur image.

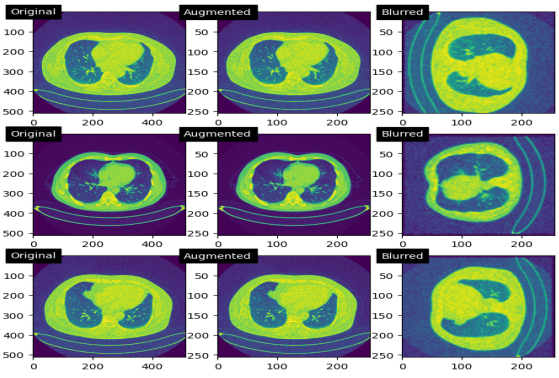

Fig. 3. Samples of preprocessed image.

C. Proposed Method

A general workflow diagram of the proposed method is presented in Fig. 4. In the very first stage of the method, we partitioned the dataset into test, train, and validation parts. Then in the next stage, we generated an augmented training dataset by applying a conscientious preprocessing step. With this newly generated dataset, we trained our model and then

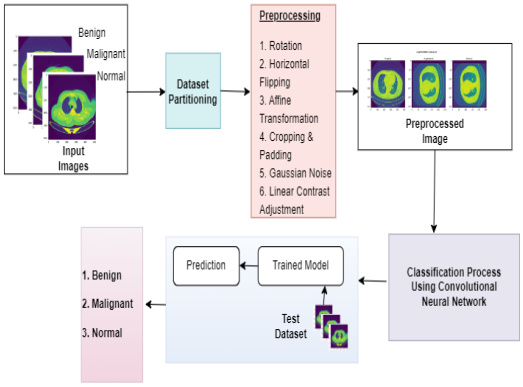

Fig. 4. Workflow diagram for the proposed method of Lung cancer prediction.

prediction is accomplished on the final model. The purpose of the deep convolutional neural network architecture is image classification. The main innovation of this architecture is the use of convolution modules, which allow the network to capture data at different scales. Because of its depth, it performs better than other models. As opposed to a deep network, the remaining models display a wider network. Though speed and time are not compromised in CNN. However, it is selected here due to its great efficiency and low computational cost. In Fig. 5 summary of the CNN model is presented. This

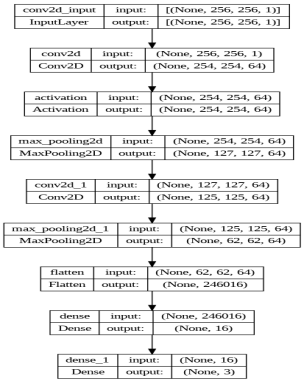

Fig. 5. Sumamry of the proposed CNN model.

designates that the first convolutional layer uses 64 filters with a typical filter size 3×3 and a rectified linear unit (ReLU). In the second activation layer, a ReLU activation function is employed to the previous layer's output. The next layer performs Max Pooling to downsample the spatial dimensions. After the next convolutional and max pooling layer the flatten layer works. This layer transforms the input into a 1D array. Finally, two fully connected dense layers help to represent the final output classes. In Fig. 6 the internal structure of the Convolutional neural network is shown. From this figure, it can be summarized that a convolutional neural network (CNN) architecture's first layer is made up of multiple convolutional layers, which are in charge of gathering data from input im-

ages. Later layers, such as max-pooling and activation, increase non-linearities and decrease spatial dimensions. Hierarchical features are captured through an iterative process. The output becomes a 1D array after the layer is flattened, creating dense layers that are fully connected. Three classes are produced in the classification output by the final deep neural layers. Because the model is sequentially designed, it is important to extract complex features using convolutional operations to ensure a reliable classification process.

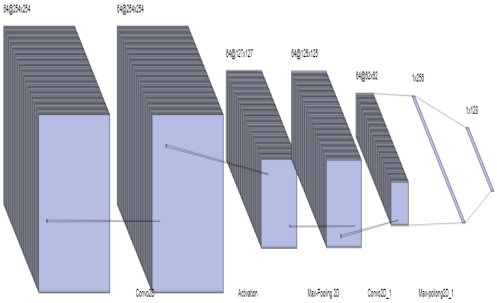

Fig. 6. Structure of the proposed CNN model.

One of the key equations in this module is the computation of the output image -

$$X_1 = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} \sum_{c=0}^{C-1} X(i+m, j+n, c).W(m, n, c) + b \quad (1)$$

Eq. (1) here works for evaluating the convolution layer. X_1 denotes the output for this layer. It takes the input as X(i,j) as the pixel value at position (i, j). W(m,n) is the weight value at position (m,n), b is the bias term and c is the number of input channels. In our case, the number of channels is 3.

$$X_2\,(i, j, c) = max_{m,n} X_1\,(i.s + m, j.s + n, c) \quad (2)$$

And the output of each max pooling layer is generated by Eq. (2). The output of the convo layer is used to analyze each max pooling level where X_1 is the input image, and s is the stride, which represents the step size when moving the pooling window. Here max(m,n) is used to find the maximum value within a specific region of the input feature map.

III. Results & Discussion

Initially, we implemented Resnet50 without any prior image preprocessing step, which shows poor training and validation accuracy. Fig. 7 represents the training and validation loss and accuracy for Resnet50. Here we get 86% training accuracy and 76% validation accuracy, which is evidently low.

We also conducted an experiment of lung cancer type classification with the help of InceptionV3 which shows a training accuracy of 97% and a validation accuracy of around 96%. Fig. 8 shows the accuracy and loss for the training and validation dataset using 20 epochs. After evaluating the

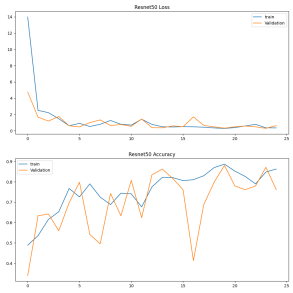

Fig. 7. Training and Validation loss and accuracy across 20 epochs for Resnet50.

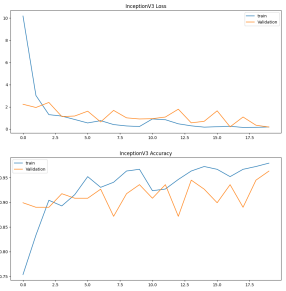

Fig. 8. Training and Validation loss and accuracy across 20 epochs for InceptionV3.

above mentioned two types of CNN, particularly Resnet50 and InceptionV3, we found that our proposed method shows the best result. The training and validation accuracy is more than 98% which is presented in Fig. 9. The experiment is processed across 20 epochs. The classification report of the proposed approach is represented in Fig. 10. This figure indicates that 97% of the real benign class instances were predicted correctly.

$$\text{PR} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

$$\text{RE} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

$$\text{F1-Score} = 2 \times \frac{\text{PR} \times \text{RE}}{\text{PR} + \text{RE}} \quad (5)$$

Furthermore, our approach successfully predicts the remaining class instances. Recall thus becomes a crucial metric in situations where it is costly to miss positive occurrences. For instance, when it comes to medical diagnosis, missing a disease is serious. Our method yields an ideal recall value. Therefore, a better performance is indicated by the F1 score, which is the mean value of precision and recall. Principles for measuring F1 score, precision and recall is given below in Eq. (3), Eq. (4) and Eq. (5). Here, PR, RE denotes precision and recall respectively. A model's precision is a measurement of how well its positive predictions come true. Furthermore, recall is a metric that assesses a model's capacity to identify all pertinent instances of the positive class.

As false positive values are important in the case of disease detection, it may force us to undergo a risky course of treatment due to wrong prediction. We find that our suggested method performs better than the approaches listed earlier. Using the F1 score, recall, and precision, we carried out a performance analysis step. Fig. 11 shows three metrics for each class to analyze the performance of the proposed approach.

IV. Conclusion

Lung cancer is an obvious public health concern, claiming a significant number of lives worldwide each year. Image processing and Deep learning have undoubtedly become excellent tools for medical healthcare, especially in early lung cancer diagnosis. Numerous algorithms have been proposed

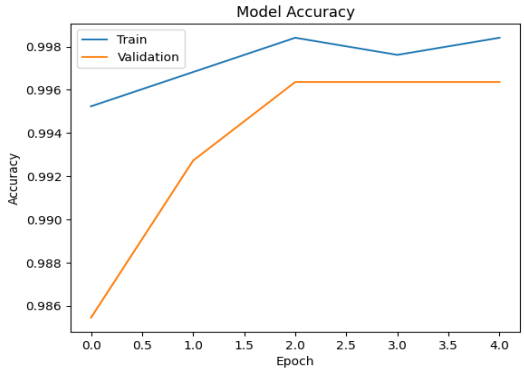

Fig. 9. Performance measure metrics for each class.

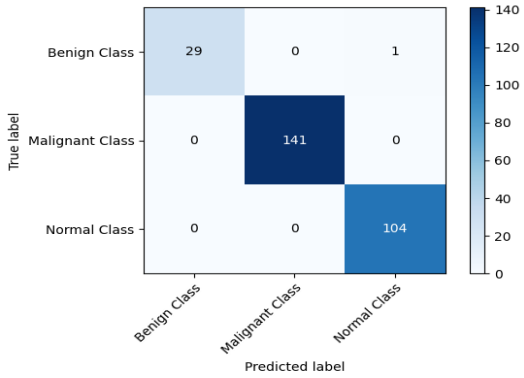

Fig. 10. Classification report of our proposed approach.

to reach the necessary level of accuracy and precision. By leveraging robust preprocessing with the CNN approach, the aim is to enhance the overall performance and reliability of the diagnosis process. We classified Lung cancer into Normal, Malignant, or Benign and then generated the classification report including accuracy, precision, recall, and F1-score with the assistance of a confusion matrix. Considering the epochs, we also plotted a validation accuracy and loss plot. The comparison of our proposed approach on the augmented dataset with Resnet50 and Inception V3 has been shown. Ultimately, our suggested approach demonstrates a high accuracy of more than 98 percent. Additionally, the improved F1 score value of 97% is achieved. In our future work, deep learning techniques can be used for Multiple Lung Disease detection. A wide range of CT Scan image data and histopathological data can be integrated in the future to detect multiple lung diseases more precisely.

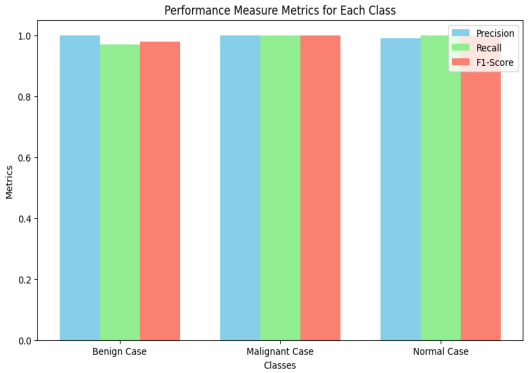

Fig. 11. Performance measure metrics for each class.

Acknowledgment

We would like to express our gratitude to Applied Science University in Bahrain (https://www.asu.edu.bh/) for offering free registration to the students who authored this academic paper.

References

[1] Romaszko, A.M. and Doboszyńska, A., 2018. Multiple primary lung cancer: a literature review. Adv Clin Exp Med, 27(5), pp.725-730.
[2] S.H. Bradley, M. P.T. Kennedy and R. D. Neal, Recognising lung cancer in primary care. Adv. Ther, 36, 19–30, 2019
[3] A. M. Romaszko,A. Doboszyńska, Multiple primary lung cancer: A literature review. Adv. Clin. Exp. Med., 27, 725–730, 2018
[4] Faisal, M.I., Bashir, S., Khan, Z.S. and Khan, F.H., 2018, December. An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In 2018 3rd international conference on emerging trends in engineering, sciences and technology (ICEEST) (pp. 1-4). IEEE.
[5] Patra, R., 2020. Prediction of lung cancer using machine learning classifier. In Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1 (pp. 132-142). Springer Singapore.
[6] Chaturvedi, P., Jhamb, A., Vanani, M. and Nemade, V., 2021, March. Prediction and classification of lung cancer using machine learning techniques. In IOP conference series: materials science and engineering (Vol. 1099, No. 1, p. 012059). IOP Publishing.
[7] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. and Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2097-2106.
[8] Pawar, V.J., Kharat, K.D., Pardeshi, S.R. and Pathak, P.D., 2020. Lung cancer detection system using image processing and machine learning techniques. Cancer, 3(2020), p.4.
[9] Ahmed, S.T. and Kadhem, S.M., 2021. Using Machine Learning via Deep Learning Algorithms to Diagnose the Lung Disease Based on Chest Imaging: A Survey. International Journal of Interactive Mobile Technologies, 15(16).