[1]Rajkumar Maharaju
[2]Rama Valupadasu

Deep Learning-Based Identification and Classification of Histological Growth Patterns in Lung Cancer

Journal of Electrical Systems

Abstract: - The most recent statistics from the WHO underscore the worldwide impact of cancer, a dangerous ailment claiming 10 million lives annually. In 2020, lung cancer emerged as a predominant contributor to this toll, responsible for nearly 2.21 million deaths. Recognizing the escalating threat of lung cancer, early detection assumes paramount significance for timely intervention and patient survival. This study leverages histopathological images, derived from the microscopic analysis of biopsies, to categorize distinct types of lung cancer. The primary focus is identifying Histological Growth patterns (including Acinar, any papillary, any Solid, and Lepidic patterns) for each lung cancer type, and predicting disease severity scores through Advanced Signal Processing Techniques. Subsequently, a Deep Learning algorithm, EfficientNetB7, is employed to classify three categories: Adenocarcinoma, benign, and Squamous cell carcinoma. These outcomes give clinicians the means to discern between benign and malignant classifications, facilitating tailored treatment initiation. Performance metrics such as classification Accuracy, F1-Score, Precision, and Recall are meticulously evaluated. The proposed methodology significantly enhances classification accuracy, achieving a notable rise from 97.5% to 99.8%, surpassing existing benchmarks.

Keywords: Histopathological images, CNN, EfficientNetB7, disease severity score, Adenocarcinoma, Squamous cell carcinoma, Benign

I. INTRODUCTION *(HEADING 1)*

Cancer manifests when ordinarily healthy cells undergo malignant transformations, proliferating uncontrollably. In the USA, the second leading cause of deaths is attributed to cancer and within the spectrum of this ailment, lung cancer stands as the second most prevalent form [1]. Primary contributors to lung cancer is use of tobacco, secondhand smoke exposure, and also environmental air pollution are the major risk factors [2]. The life expectancy of individuals diagnosed with lung cancer hinges on various determinants. Approximately, 40% of patients who are suffering from lung cancer endure the disease for one year, while only 15% manage a five-year survival, and around 10% surpass a decade with the illness. These survival rates are contingent on factors such as cancer subtype and stage [3].

Numerous techniques, including imaging methods such as MRI, CT, X-ray, and histopathology images, are employed to detect abnormalities in the lungs. However, the manual interpretation of these images by medical professionals is time-consuming, spanning hours for a comprehensive diagnosis [4], [5], [6]. This manual process is susceptible to human error. According to [7], deep learning models exhibit superior performance compared to human pathologists, particularly in the diagnosis of metastatic cancer. So, the need for computer-aided diagnostic (CAD) systems is rising to facilitate rapid analysis and mitigate the risk of misdiagnosis, ultimately ensuring the timely and accurate treatment of patients.

Artificial Intelligence enables machines to learn independently without explicit programming by supplying input data along with corresponding labels [8]. The machine learning model acquires the ability to learn specific tasks and can subsequently classify new data accurately [9], all without direct intervention from the human brain. Furthermore, extending beyond initial classification, the data can undergo manipulation and categorization based on growth patterns, structural changes, and the identification of abnormalities across the entire dataset, leveraging advanced signal-processing techniques. This all-encompassing method greatly enhances the model's training accuracy.

The application of advanced machine learning techniques, including Convolutional Neural Networks (CNN), Random Forest (R.F.), Batch Normalization (B.N.), and Support Vector Machines (SVM), has drawn the attention of numerous researchers, particularly in the context of MRI, CT, and X-ray images [10]. However, it has been

[1] Research Scholar Dept. Electronics & Communication Engineering National Institute of Technology Warangal Telangana, India
rm712110@student.nitw.ac.in
[2] Assistant Professor Dept. Electronics & Communication Engineering National Institute of Technology Warangal Telangana, India
agni@nitw.ac.in

observed that the information derived from CT/MRI images alone may not be sufficient for effectively categorizing different types of lung cancers. These imaging modalities primarily provide structural details of the lungs and can detect abnormalities such as tumors or nodules. In contrast, the Biopsy test involves the examination of cells in tissues to assess structural changes caused by the disease process. Hence, there is a growing emphasis among researchers and medical professionals focus on histology-based diagnosis, emphasizing manual categorization for the prompt initiation of treatment.

This paper follows a structured outline. Section II presents an extensive literature review, summarizing a few contributions in recent research on classification techniques with detailing their outcomes. Section III delves into the proposed works and outlines the methodology employed. Subsequently, Section IV elucidates the results derived from the proposed classification and segmentation techniques, featuring tables and essential plots. Finally, Section V offers a conclusion, providing a concise overview of the contribution, a comparative analysis with ongoing research, and prospects for future work.

II. Literature Review

Gertych A et al. [11] conducted research using various CNN architectures trained on images (tiles = 19,924, slides = 78) from CSMC and MIMW datasets. The most successful CNN achieved notable F1 scores: Acinar 74%, Solid 91%, micropapillary 76%, Cribriform 60%, and non-tumor 96%, resulting a total accuracy is 89.24% in distinguishing the five growth pattern classes.

Hatuwal et al. [12] utilized the algorithm called Convolution Neural Network for the LC25000 dataset [14] and proposed a lung cancer detection classification model, which comprises 15,000 lung and 1,000 colon cancer images. Focusing on lung cancer images only, The model achieved training and validation efficiencies of 96.11% and 97.20%.

Subhankar Roy et al. [13] introduced a Reg-STN+SORD pipeline network to classify the disease from COVID-19 to normal. Their study focused on video-based prediction of the score on Lung Ultrasound Images. LUS Dataset collected from several Italian hospitals, achieving an accuracy is 65.10% in performance metrics.

M. Saric et al. [14] used VGG and ResNet on images of lung histopathology with results being evaluated using the plot of ROC. The accuracy of VGG was reported as 75%, while ResNet achieved an accuracy of 72%. The author attributed the lower accuracy to the presence of many parameters in the given dataset.

S. Mehmood et al., [15] implemented Class selective image processing on histopathology images (LC25000). The results were compared with previous work, demonstrating an impressive accuracy of 98%. The author remarks that accuracy may improve using advanced signal processing techniques on given Data.

III. Methodology

The pipeline of the proposed methodology included the acquisition of data and preprocessing to ensure compatibility with the architecture. Subsequently, several training techniques were employed, with adjustments to hyperparameters, to facilitate the model's learning process and minimize loss for effective classification. To further enhance classification accuracy, the histologic growth patterns are identified, and the areas affected are determined through segmentation techniques. Following this, assigned disease severity indexing in a score ranging from 0 to 3 (low to high levels respectively), and its foundation lies solely in the recognized patterns and the degree to which they have proliferated.

A. Data Acquisition

The Lung and Colon Histopathology Images (LC25000) dataset [16], curated by Andrew A. Borkowski et al. and available at www.academictorrents.com, was employed in this study. The dataset includes a total of 25,000 samples, of which 15,000 are related to Lung Cancer and 10,000 to Colon Cancer. The Lung Cancer category is further subdivided into three types: benign, squamous cell carcinoma, and adenocarcinoma, each comprising 5,000 samples. All images in the dataset are anonymized, adhere to HIPAA regulations, have been thoroughly validated, and are accessible for free to AI researchers.

B. Data Formatting

Originally the dataset was in the .JPG and RGB file format with the (728, 728) pixel size. Subsequently, they were scaled down to (224, 224) pixels for maintaining a consistent aspect ratio. Following this, the values of pixel in the images were scaled to 0 to 1 for efficient convergence. To augment the dataset, various techniques such as, rotation,

zooming, flipping in horizontal and vertical, cutting and pasting were applied. These augmentation methods were employed to increase the dataset size, introduce diversity, and mitigate the risk of overfitting [17]. The dataset was partitioned into training and validation subsets, allocating 80% of the images for training purposes and the remaining 20% for validation. Sample data of lung cancer histopathology images in three types, namely benign, squamous cell carcinoma, and adenocarcinoma, are illustrated in Figure 1.

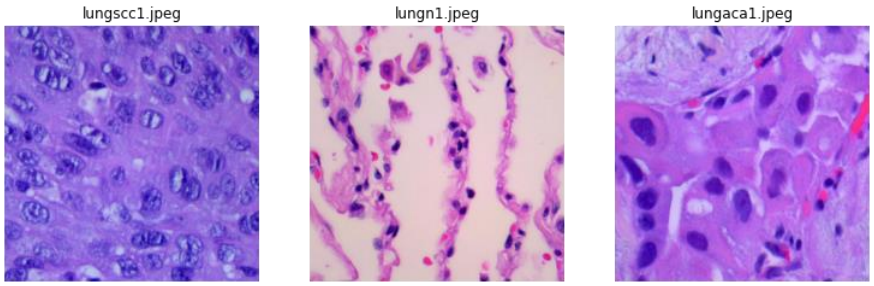


Fig.1 Three types of lung cancer images from the LC25000 dataset.

C. Identification of Growth Patterns

As it is shown in Fig. 2 with different color boundaries Lung carcinoma has five histologic growth patterns. i.e. Lepidic (Black): Tumor cells grow along existing structures; Acinar (Green): Tumor cells form glandular structures with glandular or papillary architecture; Papillary (Red): Tumor cells form finger-like projections or papillae; Micropapillary (White): Tumor cells grow in small clusters resembling inverted papillae; Solid (Yellow): Tumor cells form cohesive sheets or nests without recognizable structures [18].

For the identification of Histologic Growth Patterns, a variety of segmentation techniques, including, Contour Detection, Otsu Thresholding, K-means, and color masking, are employed. These techniques aid in recognizing different cells with normalizing with color values, grouping them the same type of cells with similar colors. The subsequent step involves identifying the area based on the shape and growth style of the cell.

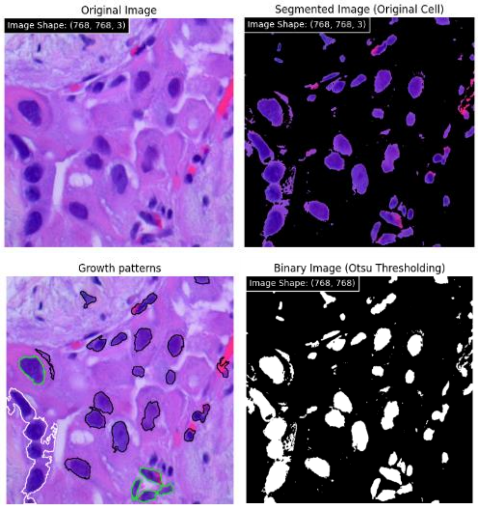


Fig. 2 Identification of growth patterns and segmented area.

TABLE I. Number of Histological Growth Patterns

Cancer type Image No.	**Histologic Growth Pattern**				
	Lepidic	***Acinar***	***Papillary***	***Micro-papillary***	***Solid***
lungaca1.jpeg	18	2	0	1	0
lungaca100.jpeg	8	7	2	2	0
lungaca1000.jpeg	14	5	4	1	1
lungaca5000.jpeg	6	1	0	0	2

This information is then matched with Histological patterns such as Lepidic, Papillary, Micro-papillary, and Acinar. The cell's shape plays a major role in identifying the type of histological growth pattern. Similar growing cells are identified, and boundaries are assigned to distinguish various diseased growth patterns.

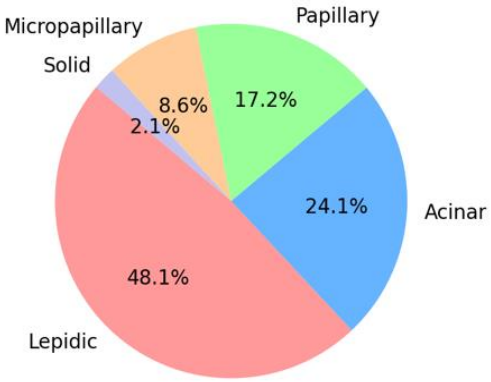


Fig. 3. Histologic growth patterns in Lung_Adenocarcinoma in LC25000 dataset.

As presented in Table I, examples of some images are provided alongside the corresponding count of growth patterns in lung adenocarcinoma. In Figure 3, the pie chart illustrates the percentage distribution of each growth pattern across the entire dataset (LC25000) for lung adenocarcinoma.

D. Severity level Indexing

The Disease Severity Score was designated on a scale of 0 to 3 levels, reflecting a risk in low to high. This assignment was grounded in the count of abnormal cells exhibiting growth patterns in the image and the extent of the region they occupied, particularly noteworthy in Squamous Cell Carcinoma. In Fig. 5, segmentation masks with the 4-level categorization are presented, offering insights into the degree of disease severity for patients and facilitating prioritization, especially in emergencies. Fig. 4 further illustrates the percentage distribution of images across severity levels for the entire LC25000 dataset, segmented by each cancer type. This provides a comprehensive overview of disease severity distribution within the dataset.

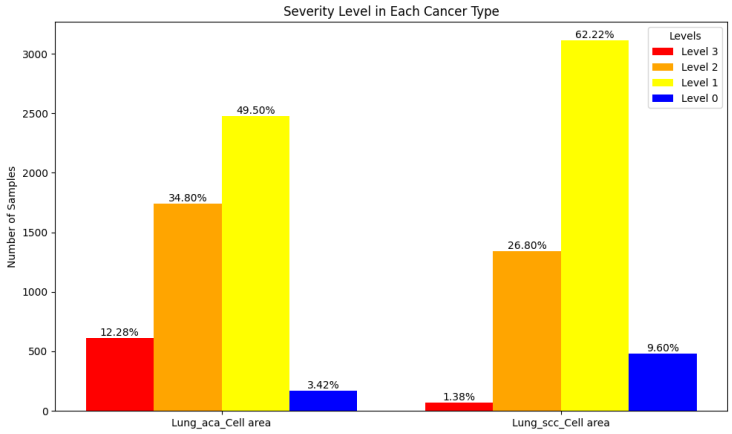


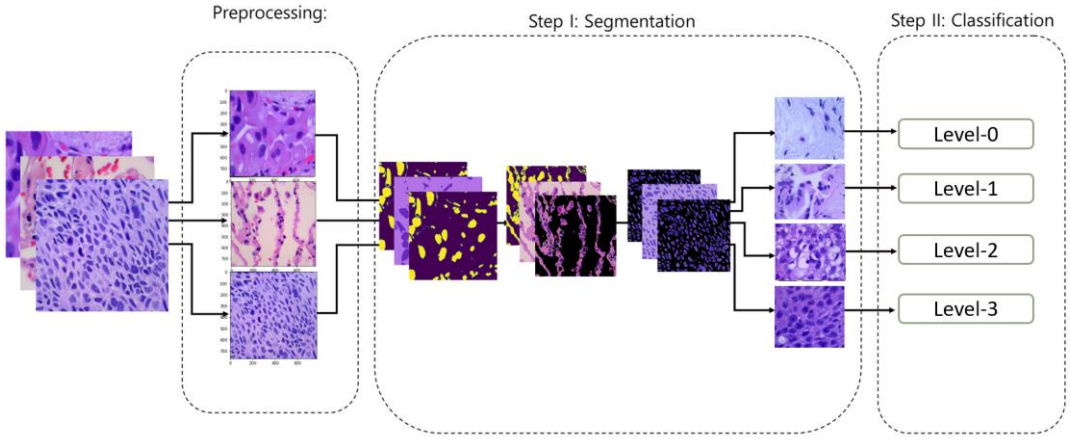
Fig. 4. Severity level in each cancer type.

Fig. 5. further illustrates the percentage distribution of images across severity levels for the entire LC25000 dataset, segmented by each cancer type. This provides a comprehensive overview of disease severity distribution within the dataset.

E. Severity Based Classification

The proposed architecture adaptive fine-tuned EfficientNetB7 was used for classification and input images taken based on Severity level indexing. Here we introduced Severity Based Classification (SBC), the highest level of severity images (levels 2 & 3) passed through the architecture for training and validation by adjusting the balanced count in the input data for reducing the time consumption and complexity. This may help to deploy the model in other applications. Google Colab is used as an environment for executing the model for all the operations.

EfficientNet is a convolutional neural network (CNN) architecture notable for its application of a compound coefficient factor, which ensures uniform scaling across the network's width, depth, and resolution. As illustrated in Figure 6 [19], this approach ensures a well-balanced network by applying fixed coefficients to these dimensions. To increase the computational task by a factor of 'n', the network is scaled accordingly: depth = αn, width = βn, and size = γn, where α, β, and γ are constants. EfficientNet further incorporates a factor 'ϕ,' known as the compound coefficient, to ensure consistent scaling across all dimensions of the network.



(a) (b) (c)

Fig. 5. Corresponding Fig. 5. Corresponding Lung Histopathology Images with Segmented Masks: a. Preprocessed images representing three types of lung cancers. b. Segmented images for Adenocarcinoma, Benign tissue, and Squamous Cell Carcinoma, respectively. c. Classification of disease severity score ranging from level 0 to level 3 (low to high). (low to high).

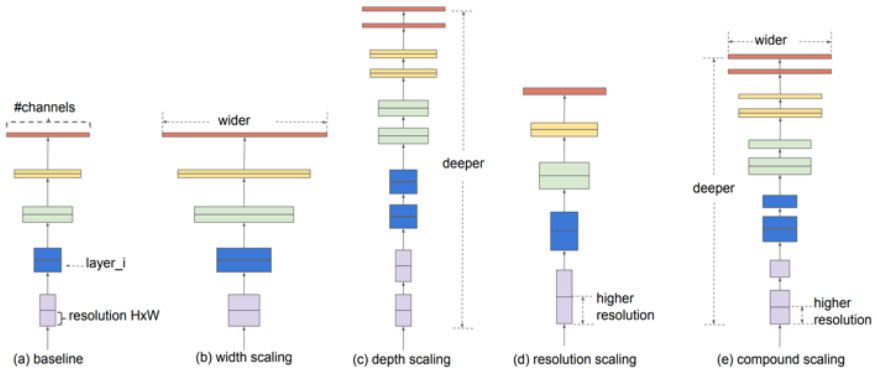


Fig. 6. Various scaling dimensions of efficientnet architecture.

In the proposed network, both stem and final layers were utilized, encompassing the input layer, rescaling, normalization, zero padding, convolution, batch normalization, and activation processes. For feature map extraction, Separable Convolution (depth-wise) was employed with filter matrices of sizes (3,3) and (5,5). Nonlinearity was introduced through the ReLU function (ReLU(x) = max (0, x)), and global average pooling with a size of (2,2) was applied to improve computational efficiency. To prevent overfitting, a fixed Dropout rate of 0.2 was implemented. For the classification into three categories, a sigmoid activation function was used in the output layer.

In the presented research, the compound coefficient was omitted, and manual tuning and scaling were implemented to optimize performance across all dimensions. The architecture consists of 5 individual modules, and the interconnected structures of these modules are termed sub-blocks. In the methodology, blocks were rearranged and linked to the subsequent block after a two-fold multiplication. Furthermore, pre-trained weights from the 'ImageNet' model [22], [23] were employed to attain the highest accuracy.

The optimizer called Adam employed for the network. Categorical Cross-Entropy (CE) is utilized to measure the error (the value of deference between predicted and output) to reduce the loss for training the model efficiently. The formula for (CE) is as follows:

$$CE = -\log\left(\frac{e^{Sp}}{\sum_{j}^{C} e^{Sj}}\right) \quad (1)$$

Sj - net score for each class of C.
C - no. of output labels,
Sp - positive class from Convolution Layer,

The proposed model was evaluated with following metrics, along with confusion Matrix and key evaluation indicators including parameters like recall, F1-score, precision, and accuracy. The respective equations for these metrics are presented below:

$$Recall = \frac{TP}{(TP + FN)} \quad (2)$$

$$F1score = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (4)$$

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (5)$$

In the formulas mentioned earlier, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are essential metrics used to calculate various performance measures. various performance measures. Moreover, the trained model's weights are saved in the HDF5 format, ensuring that the model and its learned parameters are preserved effectively. A comparative analysis was conducted with previous research with architectures including InceptionNetV2, CNN, ResNet50, DenseNet, and BreastNet for classifying normal and abnormal images. The Results were presented in Table III in Section IV. Our approach demonstrated superior

IV. Result and Discussion

The images with Identified growth patterns were augmented and then fed into the network for training purpose. The batch size is fixed at 128 for each epoch over 35 times. The model demonstrated outstanding performance, achieving an accuracy of 99.85% and 99.9% on the validation set after the training. Figures 6 and 7 illustrate the plots for accuracy and loss versus epochs, depicting the performance on both train and valid performance.

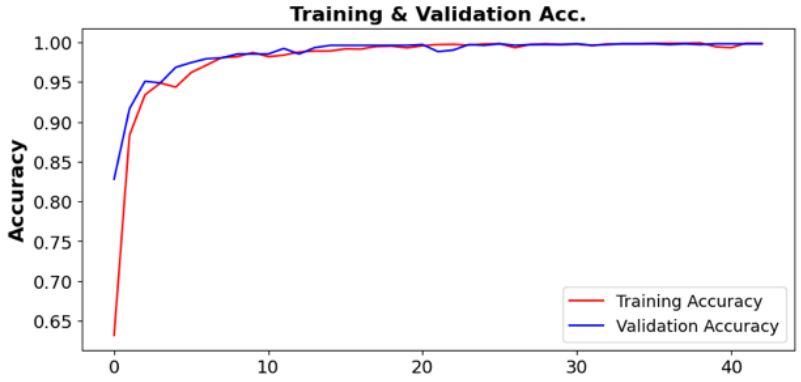


Fig. 7. Plot for model accuracy vs. Epoch for training and validation images.

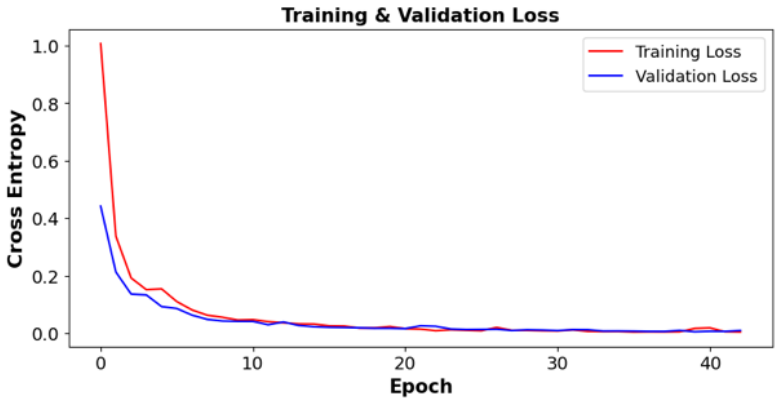


Fig. 8. Plot for model loss vs. Epoch for Training and validation images.

TABLE II. Recall, Precision, and F1-Score of Model for Different Categories

Cancer type	**Performance Metrics**		
	Precision	***Recall***	***F1-Score***
Adenocarcinoma	99	99	99
benign tissue	100	100	100
Squamous Cell Carcinoma	99	99	99

TABLE III. Recall, Precision, and F1-Score of Model for Different Categories

Labeled Category	**Performance Metrics**		
	Precision	***Recall***	***F1-Score***
Accuracy	-	-	99
Macro average	99	99	99
Weighted average	99	99	99

Table II offers a comprehensive overview of the performance metrics for the proposed model, including recall, precision, and F1-score for different cancer categories. Furthermore, Table III details the labeled categories, including macro average, weighted average, and overall accuracy. The formulas for calculating these performance metrics are outlined in Section III-C.

TABLE IV. Comparison of Performance Metrics with Other Network Architectures

Author	**Performance Metrics**			
	Precision	***Recall***	***F1-Score***	***Accuracy***
B.K.Hatuwal *et.al* [12]	0.96	0.95	0.96	0.96
S.Mehmood *et.al* [15]	0.97	0.94	0.97	0.98
S. Mangal *et.al* [24]	-	-	-	0.96
M. Masud *et.al* [25]	0.96	0.96	0.96	0.96
Proposed Model	**0.99**	**1.00**	**0.99**	**0.99**

Table IV provides an overall view of the evaluation metrics in comparison to existing research on the same dataset. The metrics including parameters like recall, F1-score, precision, and accuracy for different cancer-type categories. In this comparison, the proposed model demonstrates superior performance compared to existing models.

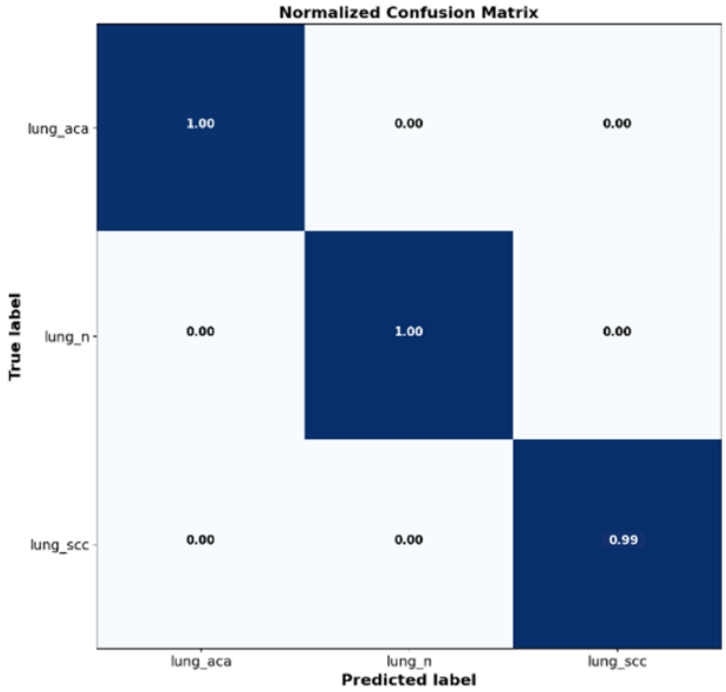


Fig. 9. True label vs predicted label in confusion matrix for different image categories for validation images.

The Figure 9 presents the confusion matrix, which visually contrasts the predicted labels with the actual labels for the validation data across various categories.

V. Conclusion

This paper introduces an automated system for classifying lung cancer severity based on histopathological images using deep learning, achieving a remarkable classification accuracy of 99.8% across three categories: Adenocarcinoma, Squamous Cell Carcinoma, and Benign. The research utilizes sophisticated signal processing methods to assess disease severity scores., which are categorized into four levels based on Histological Growth Patterns. It identifies five distinct histology patterns within specific tumor groups: I. Solid (n=5111; 2.1%), II. Papillary (n=10,221; 25.8%), III. Lepidic only (n=92,115; 48.1%), and IV. Acinar only (n=15,610; 24.1%), where ‘n’ indicates the number of groups within each pattern. The proposed classification approach shows significant potential for early detection across various cancer types. Future work could focus on applying different deep learning frameworks to predict cancer stages. The model demonstrates strong performance, with training and validation accuracies of 99.8% and 99.9%, respectively. Furthermore, the evaluation metrics including recall, precision, and F1-score are computed, and a confusion matrix plot is presented to gauge the model's effectives. These results may helpful to the doctors to treat cancer patients as early as possible to save their life.

References

[1] “American Cancer Society, Lung Cancer Causes) [Online].” Available: https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/what-causes.html

[2] “American Cancer Society, Lung Cancer Statistics. [Online]”. Available: https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html.

[3] Yu, Kun-Hsing, Ce Zhang, Gerald J. Berry, Russ B. Altman, Christopher Ré, Daniel L. Rubin, and Michael Snyder. "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." Nature communications 7, no. 1 (2016): 12474

[4] Silvestri, Gerard A., Michael K. Gould, Mitchell L. Margolis, Lynn T. Tanoue, Douglas McCrory, Eric Toloza, and Frank Detterbeck. "Noninvasive staging of non-small cell lung cancer: ACCP evidenced-based clinical practice guidelines." Chest 132, no. 3 (2007): 178S-201S.

[5] Travis, William D., Elisabeth Brambilla, Masayuki Noguchi, Andrew G. Nicholson, Kim R. Geisinger, Yasushi Yatabe, David G. Beer et al. "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, no. 2 (2011): 244-285.

[6] Collins, Lauren G., Christopher Haines, Robert Perkel, and Robert E. Enck. "Lung cancer: diagnosis and management." American family physician 75, no. 1 (2007): 56-63.

[7] Wang, D., Khosla, A., Gargeya, R., Irshad, H. and Beck, A.H., 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

[8] Ristanoski, Goce, Jon Emery, Javiera Martinez Gutierrez, Damien McCarthy, and Uwe Aickelin. "AI based cancer detection models using primary care datasets." Journal of Advances in Information Technology Vol 13, no. 2 (2022).

[9] Michie, Donald, David J. Spiegelhalter, Charles C. Taylor, and John Campbell, eds. Machine learning, neural and statistical classification. Ellis Horwood, 1995.

[10] Tuncal, Kubra, Boran Sekeroglu, and Cagri Ozkan. "Lung cancer incidence prediction using machine learning algorithms." Journal of Advances in Information Technology Vol 11, no. 2 (2020).

[11] Gertych, A., Swiderska-Chadaj, Z., Ma, Z., Ing, N., Markiewicz, T., Cierniak, S., ... & Knudsen, B. S. (2019). Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Scientific reports*, *9*(1), 1483.

[12] Hatuwal, Bijaya Kumar, and Himal Chand Thapa. "Lung cancer detection using convolutional neural network on histopathological images." Int. J. Comput. Trends Technol 68, no. 10 (2020): 21-24.

[13] Roy, Subhankar, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben et al. "Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound." IEEE transactions on medical imaging 39, no. 8 (2020): 2676-2687

[14] Šarić, Matko, Mladen Russo, Maja Stella, and Marjan Sikora. "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, 2019.

[15] S. Mehmood *et al*., "Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning With Class Selective Image Processing," in *IEEE Access*, vol. 10, pp. 25657-25668, 2022, doi: 10.1109/ACCESS.2022.3150924.

[16] Borkowski, Andrew A., Marilyn M. Bui, L. Brannon Thomas, Catherine P. Wilson, Lauren A. DeLand, and Stephen M. Mastorides. "Lung and colon cancer histopathological image dataset (lc25000)." arXiv preprint arXiv:1912.12142 (2019).

[17] Hinton, Geoffrey E., Alex Krizhevsky, and Ilya Sutskever. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25, no. 1106-1114 (2012): 1.

[18] Solis LM, Behrens C, Raso MG, Lin HY, Kadara H, Yuan P, Galindo H, Tang X, Lee JJ, Kalhor N, Wistuba II, Moran CA. Histologic patterns and molecular characteristics of lung adenocarcinoma associated with clinical outcome. Cancer. 2012 Jun 1;118(11):2889-99. doi: 10.1002/cncr.26584. Epub 2011 Oct 21. PMID: 22020674; PMCID: PMC3369269.

[19] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In International conference on machine learning, pp. 6105-6114. PMLR, 2019.

[20] Sasikala, S., M. Bharathi, and B. R. Sowmiya. "Lung cancer detection and classification using deep CNN." International Journal of Innovative Technology and Exploring Engineering 8, no. 25 (2018): 259-262.

[21] SR, Sannasi Chakravarthy, and Harikumar Rajaguru. "Lung cancer detection using probabilistic neural network with modified crow-search algorithm." Asian Pacific journal of cancer prevention: APJCP 20, no. 7 (2019): 2159.

[22] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Communications of the ACM 60, no. 6 (2017): 84-90

[23] Maharaju, R. and Valupadasu, R., 2023, March. Lung Cancer Classification and Prediction of Disease Severity Score Using Deep Learning. In *2023 6th International Conference on Information and Computer Technologies (ICICT)* (pp. 100-104). IEEE.

[24] S. Mangal, A. Chaurasia, and A. Khajanchi, ‘‘Convolution neural networks for diagnosing colon and lung cancer histopathological images,’’ 2020, arXiv:2009.03878.

[25] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, ‘‘A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework,’’ Sensors, vol. 21, no. 3, p. 748, Jan. 2021, doi: 10.3390/s21030748.