Non-Small Cell Lung Cancer Classification from Histopathological Images using Feature Fusion and Deep CNN

Nur-A-Alam, Md. Mahbubur Rahman, Khandaker Mohammad Mohi Uddin, Al Bashir, Jahanara Akhtar

Abstract: Lung cancer is the overgrowth of cells in digestive organs. Identifying different types of lung cancer (squamous cell cancer, large cell carcinoma and adenocarcinoma) from lung histopathological images is outrageous works that shorten the chance of infected with lung cancer in the future. This research proposes an accurate diagnosis scheme using various neural network features and fusion of contourlet transform from lung histopathological image. This lesson has used several pre-train models (Alexnet, ResNet50, and VGG-16) in addition to divers scratch models while the pre-train Resnet50 model works better. The two reduction techniques (Principle Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (MRMR)) have used to classify the type of lung cancer with the extraction of the most significant properties. In Convolution Neural Network (CNN) based lung cancer detection, the reduction approach PCA performs better. This proposed methodology is performed on ordinary datasets and establishes comparative better performance. The accuracy of this paper is 98.5%, sensitivity 96.50, specificity 97.00%, which is more effective than other approaches.

Keywords: Convolutional Neural Network (CNN), Principle Component Analysis (PCA), Contourlet Transform (CT), Histopathological Image, Minimum Redundancy Maximum Relevance (MRMR).

I. INTRODUCTION

The second most affected cancer is lung cancer for gender (both male and female) in the USA. Each year, almost 135,720 people deaths from lung cancer than other cancer in the USA (American Cancer Society, 2020). Statistic says people are dying for both small cell lung cancer (SCLC) and Non-small cell lung cancer (NSCLC). The conclusive stage of lung cancer is Non-Small Cell Lung Cancer (NSCLC). This year, 228,820 people have diagnosed with Non-Small Cell Lung Cancer in USA. Every day, the number of affected people in lung cancer are greatly increased. If there is an effective system that can detect lung cancer at an early stage, then the molarity rate can be greatly reduced. This proposed system can detect lung cancer automatically and reduce the death rate of people.

The histopathological image refers to the biological structures of tissues and cells of organisms. Pathologists analyze the histological images using microscope and govern the morphological features of cells which indicate the availability of disease. The diagnosis of lung cancer disease in histopathology images to identify histological structures (cancer cells, glands, and lobule) [19]. But some problem occurs in histopathology images include complicated clinical features, inadequate training data and large size of per images (Daisuke Komura, 2018). The pre-train CNN activation features perform for classification which solves these complex clinical features the histopathological image refers to the biological structures of tissues and cells of organisms. Pathologists analyze histological image using microscope and govern the morphological features of cells which indicate the availability of disease. The diagnosis of lung cancer disease in histopathology images to identify histological structures (cancer cells, glands, and lobules). Problem to solve the second third problem, we are used large datasets that is available in academic torrent and preprocess the image to get small image in size [20].

For many years, scientists have been working to quickly detect lung cancer. For this many researchers recommend multiple approaches for automation lung cancer detection. This paper Vaishnavi. D. (2019) proposed the automatic system to classify lung cancer tissues from CT scan images. The system consists of three procedures such as preprocessing on desire image set, extract of features and finally classify the extract data set. For feature extraction, the DTCWT and PNN techniques used and to classify the normal and abnormal lung tissues by using PNN. To clarify the lung cancer from CT image, Wasudeo Rahane, (2018) proposed a smart system using distinguishable data analysis procedures and classification methods. In this system, first of all, preprocess the CT image and classification by using support vector machine (SVM). In this study S. Sasikala. (2018) deal with the detection of lung cancer using CNN classification techniques. It provides 96% accuracy.
This paper Mr. Sandeep A. Dwivedi. (2018) extract 13 different features to classify lung cancer by using multivariate multinomial Bayesian. And the performance of this system is to check FCM algorithm. The ANN-based architecture used for lung cancer classification (S. Senthil, 2019), FU Ling developed the standard approach to identify lung cancer by using 3D deep CNN. This paper used a multiscale 11 layers of FCN and achieves better performances in the LUNA16 challenge. In those papers, they used CNN, wavelet methods for extract features and CNN, SVM and ANN used to classification. 

The key exploit of this research are:

- An improved feature extraction method has been formed to support the intelligent system for lung cancer detection.
- A powerful reduction technique has used to extract the most significant features from the fusion feature vector.
- The proposed detection technique of lung cancer from a large database of lung histopathological images is better performance and results are also satisfactory.

The proposed system takes input of histopathological images and processed the output of lung cancer types. Each input histopathological image generates the training features through the contourlet transform and CNN technique. The contourlet transform and CNN approaches are extracted features from an image and stored in a vector. The fusion will perform further classifications to reach the final result.

The paper is structured like: research methods and materials are described in section 2. Sections 3, analyze the results and articulate the implementation methods. Finally, the conclusion of the whole study and future scopes are mentioned in Section 4.

II. PROPOSED METHODOLOGY

This section narrates a synopsis of the detection method of propounding lung cancer. Fig. 1 shows the complete process of detection of lung cancer types as we mentioned before. This paper has conducted the standard dataset of the histopathological image. First of all, the preprocessing part established to resize the image. Then the system has used extract features by using contourlet transform and CNN. The extracted features are formed into a vector. Afterward, dimensionally reduction techniques, PCA abate the features by taking only related to un-correlated features. At the detection point, CNN categorizes features into adenocarcinoma or squamous cell cancer or large cell carcinoma classes. Algorithm 1 represents the overall proposed methodology in this research paper.

![Algorithm 1: Approach for solving lung cancer type detection problem](image)

**Algorithm 1: Lung cancer types detection proposed algorithm**

**Input:** Lung cancer histopathological image dataset (D) with resize image (M)

**Extraction:** Extract Feature Matrix (F).

**CNN Feature Vector (Fc).**

Step 1: Initialize \( F_i = M(i, 1) \).

Step 2: Extract each image feature \( D(i, 25, 000) \).

Step 3: \( F_C(i, 1) = M(i, 1) + F_C(i, 1) \).

Step 4: \( F_C \) = overall CNN features.

**Contourlet Transform Vector (CT).**

Step 1: Initialize \( H_L = \text{Low pass output} \).

Step 2: \( CT(i, 1) = H_L(i, 1) + H_H(i, 1) \).

Step 3: \( CT \) = overall contourlet transform features.

**Fusion of features in Vector (V).**

\[
V = F_C + CT
\]

**PCA**

\[
PCA(i) = PCA(V), \text{PCA}(i) = \text{reduce feature vector result}(i) = \text{predict (classifier, feature vectorPCA)}
\]

**Output:**

\[
\text{result}(i) = \text{adenocarcinoma or squamous cell cancer or large cell carcinoma.}
\]

2.1 DATASET USED

When a histopathological image is compiled into the introduced automatic system, it is considered as ready for an experiment. This Histopathological image of the patient is acquired and stored in a commonplace. The images are classified as squamous cell cancer, adenocarcinoma and large cell carcinoma then used for reference to check the performance of measurement matrices. In this experiment, we used the standard datasets from Academic torrents. The dataset consists of RGB 15,000 images with three classes of 5000 images each which are publicly available in: (https://academictorrents.com). All the brain histopathological images are 768×768 pixel resolution in size. Fig. 2 shows some samples of the dataset.
2.2 DATA PREPROCESSING

Most of the images have not processed when it gets from a dataset provider. For Medical images preprocessing part is essential. Each image contains many pixels include noise, distortion and hence. To achieve accurate results, redundant pixels and distortion pixels are eliminated from images by using some approaches. At first, the entire images are converted into RGB to grayscale images using MATLAB tools and resize the entire images within 224×224 in size.

2.3 Feature Extractors

This section describes two feature extraction techniques and also describes the appropriate process of features of extraction. This research paper has applied two powerful feature extraction approaches, firstly Contourlet Transformation and another is CNN. The Contourlet Transform approach describes in Section 2.3.1 and the CNN architecture demonstrate in Section 2.3.2.

2.3.1 Contourlet Transformation

This part describes the proposed contourlet transform. The Laplacian pyramid decomposition and directional filter bank parts describe in the contourlet transform (Do and Vetterli, 2005). Two parts take the multi-scale transform of image related to wavelet and another one captures the contour segment which associates with the local directional transformation (Paygude and Vyas, 2019). Fig. 3 illustrates the accurate architecture of CT.

Laplacian Pyramid:

\[ g_l = g_{l+1} - \sum_{i=1}^{3} g_{l+1} \]  

Here, \( g_l \) is the \( l^{th} \) level of the Gaussian Pyramid.

2D image basically has chosen to perform contourlet transform, each RGB image splits into three channels according to the proposed system model. We have established sixteen several multi-scale and multi-directional image. Computing gray level co-occurrence matrix (GLCM) on every image, the contourlet transform calculates 13 variant features for each image such as homogeneity entropy, correlation, energy and others. By comparing with wavelet images, the properties (textures, capture edges, lines and contours) are better in Contourlet transformed images (Hamdi, 2012). The category finding process in lung cancer, it is mandatory to detect edges of the image and smooth contours of each one. Effective localization, visual representation and directionally of images are achieved by CT image.

2.3.2 CONVOLUTIONAL NEURAL NETWORK

This section describes the fine-tuned VGG19 pre-trained model that classifies lung cancer types. This experiment takes the lung cancer image dataset with the size of 224 × 224 raw pixel input, where width 224, height 224. The first layer of convolution has used padding 0 with stride 4, and 96 filters of size 11 × 11. The layer of convolution has set padding as 2 with stride 1, and 256 filters of size 5 × 5. A rectifier function is following to each layer. A max-pooling layer with size 3 × 3 and stride of 2 is set in each convolutional layer. Again, the third convolutional layer contains 384 filters while the fourth, and fifth layers contain 256, and 128 filters, respectively. These three are followed by the ReLU function. After these three, there is a max-pooling layer of size 3 × 3 and stride 3. There are three fully connected layers in total. The first and second fully connected layer has 4096 neurons each. The third fully connected layer is the output layer with three neurons. Finally, CNN provides 4096 features. Fig. 4 shows the proposed VGG19 architecture.

Fig. 4: Illustrate the proposed VGG19 architecture.

2.4 Feature Reduction

A feature vector is formed from two feature extraction processes. The feature vector contains a number of features that are complex and redundant features. To get better performance features we have used feature reduction techniques. Though the MRMR method has great accuracy but comparatively the accuracy of PCA is better than MRMR. Besides, PCA provides better performance than MRMR.

Minimize Redundancy:

\[
\min W_1,
\]

Maximize Relevance:

\[
\max V_1,
\]

Here, \( S \) is the collection of features, \( I(x,y) \) is Mutual information(MI) between features \( x \) and \( y \). To find the Mutual information(MI) of two variables \( x \) and \( y \) are calculated using the following formula:
Non-Small Cell Lung Cancer Classification from Histopathological Images using Feature Fusion and Deep CNN

\[ I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \, dx \, dy \] (4)

Dimensional reduction or feature extraction of data is done with the PCA technique. Basically it is used on medical image processing for several purposes like segmentation, image fusion, registration, compression, feature extraction and classification (Nandi et al., 2015). As a feature extractor we have used PCA for dimensionality reduction on previously extracted features which may hold noisy information. This procedure chose only uncorrelated features by uniting correlated features (Chao et al., 2019). The covariance between features are calculated as follows by eliminating uncorrelated features,

\[ \text{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)} \] (5)

III. RESULT AND ANALYSIS

This research has used three measurement schemas like as accuracy, sensitivity, specificity that check the performance of this system. To calculate three matrices for several parameters are used. These four parameters are known as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

Table 1 presents the performance metrics equation. This system has trained on the lung cancer database and computed four parameter values. Then we have calculated the three measurement schemas and got standard values that represent the performance of the proposed system.

In Table 2, describes a comparative study on measurement matrices after apply various feature extraction techniques. First of all we have calculated three measurement matrices values based on the original image. Then we have applied the feature extraction method namely wavelet and contourtlet transform and got better performance from the wavelet transform in this research.

We have also used another feature extraction and classification techniques for lung cancer detection. The CNN architecture contains many pre-trained models. We have trained the entire lung cancer dataset based on various pre-trained models but fine-tuned VGG19 performance is better than others. Table 3 shows the performance measurement of various models. In this system the ResNet50 pre-trained model gives a greater sensitivity point than VGG19 fine-tuned model in as much as VGG19 gained greater specificity and accuracy.

We have applied two feature reduction techniques. The PCA reduction techniques achieved a better score for three performance matrices than MRMR techniques. Table 4 illustrates the comparative result of these methods.

Table 1: Equations of performance matrices.

| Metrics         | Computing equation |
|-----------------|--------------------|
| Accuracy(ACC)   | \( \frac{TP + TN}{TP + TN + FP + FN} \) (6) |
| Specificity(SPEC)| \( \frac{TN + FP}{TN + FP} \) (7) |
| Sensitivity(SEN)| \( \frac{TP + FN}{TP + FN} \) (8) |

Table 2: Comparison of various methods using different properties of GLCM.

| Transformed Methods | Accuracy | Specificity | Sensitivity |
|---------------------|----------|-------------|-------------|
| Original Image      | 0.7873   | 0.7452      | 0.7373      |
| Wavelet Transformed | 0.7634   | 0.7812      | 0.7607      |
| Contourlet Transform| 0.8357   | 0.8545      | 0.7968      |

Table 3: Different Classification CNN model performance measure

| CNN Architecture | Accuracy | Specificity | Sensitivity |
|------------------|----------|-------------|-------------|
| Alexnet          | 0.9127   | 0.9528      | 0.8543      |
| VGG-16           | 0.9266   | 0.9533      | 0.9182      |
| ResNet50         | 0.9465   | 0.9469      | 0.9465      |
| VGG19            | 0.9619   | 0.9860      | 0.9370      |

Table 4: Performance measures of feature reduction techniques

| FR Method | Accuracy | Specificity | Sensitivity |
|-----------|----------|-------------|-------------|
| PCA       | 0.9959   | 0.9944      | 0.9974      |
| MRMR      | 0.9876   | 0.9857      | 0.9897      |

Table 5 represents a comparative study on the existing methodology and proposed methodology. S. Sasikala (2018) used to extract feature by using CNN and classify Lung Cancer based on Convolutional Neural Network (CNN) from CT scan images. Mr. Sandeep A. Dwivedi (2014) studied this system with the Gray Level Co-occurrence Matrix (GLCM) with previously trained Alexnet model. CNN and GLCM methods are trained and tested on CT scan images. We have used fusion of two features extraction methods and CNN for classification. The reduction techniques PCA have used to extract significant features that give better performance result.

Table 5: A Comparison study with existing methods

| Methods           | Dataset                  | Methodology                | Feature Extractor | Classifier | Results          |
|-------------------|--------------------------|----------------------------|-------------------|------------|------------------|
| S. Sasikala (2018)| LIDC+ID R1 (1000 CT scan images) | CNN                       | (Alexnet)        | CNN        | Accuracy 91%     |
| Mr. Sandeep A. Dwivedi (2014)| Kaggle | GLCM (Alexnet) | SVM                | Accuracy 92% |
| Proposed method   | Academic Torrents (25000 histopathological image) | Contourlet transform (VGG19 + PCA) | CNN           | Accuracy 96.50% |
IV. CONCLUSION

This work has designed an intelligent model for lung cancer type’s classification with high accuracy and low complexity. To detect lung cancer automatically the appropriate features selection and classification are more important.

This experiment has combined the potency of CT features and the strength of CNN. In this system, histopathological image is used as input which generates the output of lung cancer types. The proposed system shows better accuracy from the ROC analysis than the previous ones. The raised method has established with 96.50% accuracy, 95.50% sensitivity and 97.00% specificity. Comparing with the other approaches, this proposed method has better accuracy. So, we will focus on developing more efficient methods in future to improve the existing result. Besides, CNN features and the fusion of CT images will be used for the analysis of three-dimensional volume also.

ACKNOWLEDGEMENT

We are very thankful to Dr. Shibendu Majumder, Professor, Department of Radiology and Imaging, Sher-e-Bangla Medical College and Hospital for his rich confirmation, productive suggestions and consultancy.

REFERENCES

1. Daisuke Komura, Shumpei Ishikawa. “Machine Learning Methods for Histopathological Image Analysis”, “Computational and Structural Biotechnology Journal” 16 (2018) 34-42, DOI: https://doi.org/10.1016/j.csbi.2018.01.001.

2. Vaishnavi, A., Arya, K. S., Devi Abirami, T. M. N. Kavitha, 2019. “Lung Cancer Detection using Machine Learning, International Journal of Engineering Research & Technology (IJERT)”; ISSN: 2278-0181,RTICCT - 2019 Conference Proceedings.

3. Wasudeo Rahane, Himadri Dalvi, Yamini Magar, Anjali Kalane, Satyajeet Jondhale, 2018. “Lung Cancer Detection Using Image Processing and Machine Learning”, “HealthCare Proceedings of 2018 IEEE International Conference on Current Trends toward Converging Technologies”, Coimbatore, India.

4. H. The American Cancer Society. Available at https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/stati cs [Last accessed on December 5, 2019]

5. S. Saskała, M. Bharathi, B. R. Sowmiya, 2018. “Lung Cancer Detection and Classification Using Deep CNN”, “International Journal of Innovative Technology and Exploring Engineering” (IJITEE) ISSN: 2278-3075, Volume-8 Issue-28 December, 2018. 259.

6. Mr. Sandeep A. Dwivedi, Mr. R. P. Borse, Mr. Anil M. Yametakar, 2017. “Lung Cancer detection and Classification by using Machine Learning & Multinomial Bayesian”, “IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)” ; e-ISSN: 2278-2834, p-ISSN: 2278-8735 Volume 9, Issue 1, Ver. III (Jan. 2014), PP 69-75.

7. S. Senthil. B. Ahyawarya, Shubha, 2019. “Improving the Performance of Lung Cancer Detection at Earlier Stage and Prediction of Recurrence using the Neural Networks and Ant Lion Optimizer”, “International Journal of Recent Technology and Engineering (IJRTE)”, ISSN: 2277-3878, Volume-8 Issue-2, July 2019.

8. S. Senthil. B. Ahyawarya, Shubha, 2019. “Improving the Performance of Lung Cancer Detection at Earlier Stage and Prediction of Recurrence using the Neural Networks and Ant Lion Optimizer”, “International Journal of Recent Technology and Engineering (IJRTE)”, ISSN: 2277-3878, Volume-8 Issue-2, July 2019.

9. K.Mohanambal, Y. Nirosha, E.Oliviya Rosshini, S.Prunitha, M.Shammi, 2019. “Lung Cancer Detection Using Machine Learning Techniques”, “International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering” Vol. 8 Issue 2, February 2019.

10. Albert Chen, Niranjan Balachandar, Peter Lu. “Deep Convolutional Neural Networks for Lung Cancer Detection”.

11. Do, M.N., Vetterli, M., 2005. The contourlet transform: an efficient directional multiresolution image representation. Image Process IEEE Trans. https://dx.doi.org/10.1109TIP.2005.859376.

12. Chao, G., Luo, Y., Ding, W., 2019. Recent advances in supervised dimension reduction: a survey. Mach. Learn. Knowl. Extr. 1, 341–358. https://doi.org/10.3390/make1010020.

13. Hamdi, M., 2012. “A comparative study in wavelets, curvelets and contourlets as denoising biomedical images” Image Process. Commun. https://doi.org/10.2478/v1248-012-0007-1.

14. Karkaus, S.A., Iakovides, D.K., Maroulis, D.E., Karras, D.A., Tzivras, M., 2003. “Computer-aided tumor detection in endoscopic video using color wavelet features” Inf. Technol. Biomed IEEE Trans. https://doi.org/10.1109/TITB.2003.813794.

15. Nandi, D., Ashour, A.S., Samanta, S., Chakraborty, S., Salem, M.A.M., Dey, N., 2015. “Principal component analysis in medical image processing: J. a study” Image Int Min Int. https://doi.org/10.1504/jmi.2015.070024.

16. Payyut, S., Vyas, V., 2019. “Unified dynamic texture segmentation system based on local and global spatiotemporal techniques” J. Reason. Syst Int. https://doi.org/I.O. 1504/1JR1S.2019.0998855.

17. Zhang, X., Chen, F., Yu, T., An, J., Huang, Z. L., Liu, J., Hu, W., Wang, L., Duan, H., Si, J., 2019a. “Real-time gastric polyp detection using convolutional neural networks” PLoS one. https://doi.org/I.O. 1371journal.pone.0214133.005.

18. Zhang, Y., Ma, Y., Yang, X., 2019. Multi-label feature selection based on mutual information, in: ICNC-FSDK 2018 — 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. https://doi.org/10.1109/FSDK.2018.8687290.

19. A. D. Belsare and M. M. Mushrif, “Histopathological Image analysis using image processing techniques: an overview”, Signal & Image Processing : An International Journal (SIPJ), Vol.3, No.4, August 2012. DOI:10.5211/sipj.2012.3403

20. Daisuke Komura, Shumpei Ishikawa, “Machine Learning Methods for Histopathological Image Analysis”, Computational and Structural Biotechnology Journal 16, 2018. DOI: https://doi.org/10.1016/j.csbi.2018.01.001.

AUTHORS PROFILE

Md. Nur-A- Alam is an academic researcher in the Department of Computer Science and Engineering, Dhaka International University. He completed his B.Sc. (Engg.) degree in 2017 from the Department of Computing Science and Engineering at Mawlana Bhashani Science and Technology University, Tangail, Bangladesh. Now he is pursuing his M.Sc. (Engg.) degree/Research from Mawlana Bhashani Science and Technology University.

His research interests include Image processing, Machine learning, Data mining, Wireless communication and Deep learning. He has published 4 research papers in different international journals and conferences. Mr. Alam is currently working as a faculty member of Dhaka International University in Computer Science and Engineering department. He is a Member of Bangladesh Computer Society.

Md. Mahbubur Rahman is a smart sensing based researcher at Department of Computer Science and Engineering, Dhaka International University. He completed his B.Sc. (Engg.) degree from department of computer science and engineering, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh. Now he is pursuing M.Sc. (Engg.) degree/ Research from Mawlana Bhashani Science and Technology University. His research interests on the field of Artificial Intelligence, Internet of Things (IoT), Machine Learning and Smart Sensing. Now Mr. Rahman serving as a faculty member at Dhaka International University in department of computer science and Engineering.

Khandaker Mohammad Mohi Uddin has done his B.Sc. degree in Computer Science and Engineering from Jagannath University, in 2016 and MSc. (Research) in Computer Science and Engineering from the same university, in 2018.
His research interests in the field of Wireless Networking, Software Defined Networking, Artificial Intelligence, and Machine Learning. Now he is working as an academic researcher and a lecturer in the Department of Computer Science and Engineering, Dhaka International University, Dhaka.

Al Bashir is a researcher and faculty member of Dhaka International University in the Department of Computer Science and Engineering. He completed his B.Sc. (Engg.) degree from the Department of Computer Science and Engineering, Jagannath University, Bangladesh. Now he is pursuing his M.Sc. degree (Research) at the same University and Department. His research interests on Image Processing, Artificial Intelligence, Machine learning, Deep learning and Smart Technology. His first research paper based on Image Processing has been published on Springer journal. Mr. Bashir is an IEEE member and also a mentor of IEEE Student Branch, Dhaka International University.

Mst. Jahanara Akhtar is currently serving as an Associate Professor of Computer Science and Engineering department, Dhaka International University, Dhaka, Bangladesh. She is a Research Fellow (PhD) in the department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh. Jahanara completed B.Sc. in Electronics & Computer Science and M.Sc. in Computer Science and Engineering from Jahangirnagar University. She has some publications in national and international conferences and journals. Her research interest includes Cryptography, Secure Wireless Sensor Network, Image Processing and Artificial Intelligence.