Lung Cancer Detection from CT Images Using Image Processing and Machine Learning Techniques

1st Deema Sohrab Sawalha
Computational Science Program
An-najah National University
Nablus, Palestine
deema_sawalha@outlook.com

2nd Adnan Salman
Computer Science Department
An-najah National University
Nablus, Palestine
aalshaikh@najah.edu

***Abstract*—Lung cancer is the most common type of cancer among males worldwide. It accounts for one of every five cancer-related fatalities and is prevalent in people aged 55 to 65. Detecting lung cancer in its earliest stages is a crucial step in the treatment process that can significantly increase the chance of survival. In this paper, we used image processing techniques with MATLAB on computed tomography (CT) images of lung cancer for multiple patients to determine the location and extent of cancerous spots. The stages included image analysis and segmentation, feature extraction, and candidate identification as distinct regions of interest (ROI). Algorithms based on machine learning were utilized to classify cancer from the ROIs of candidates by extracting the characteristics required for the classification of pathologic features from the annotated ROIs. Comparing the evaluated algorithms in order to identify the optimal algorithm for cancer detection.**

***Index Terms*—lung cancer, convolution neural networks, machine learning, image processing**

I. Introduction

Lung cancer is a disease characterized by the uncontrollable growth of abnormal cells into tumors. Lung cancer is by far the quietest disease in its early stages. There are no symptoms or warning signs, which makes it harder to detect before it actually develops to advanced stages. The 5-year survival rate is 14% for lung cancer detected at any stage. However, if detected at early stages, before it has spread to other parts of the body, the 5-year survival rate increases to 55%. Therefore, early detection of lung cancer is extremely important in the treatment process and can save many lives [1]. Periodically monitoring lung cancer is important in preventing the development of cancer in advanced stages.

In this study, we applied several image processing and machine learning techniques to detect lung cancer from CT images. The proposed system consists of four main stages: the preprocessing stage, the segmentation stage, the feature extraction stage, and the machine learning stage. In this system, the acquired CT images are passed through the preprocessing stage for image enhancement and noise removal. In the second stage, the image is segmented to define boundaries between different tissues. In the third stage, specific features of the extracted regions were identified. In the fourth stage, a machine learning process is used to classify these regions into benign tumors or malignant tumors by looking at the features that have been extracted.

Image processing techniques are promising procedures to detect the existence of lung cancer in early stages. However, the proposed system involves the application of several procedures and algorithms for each procedure. In image preprocessing stage, Gabor filter, Median Filter, and Weiner Filter algorithms are evaluated to get the most accurate result for image enhancement procedure. For the best binarization output, the Global thresholding, Local thresholding, and Otsu's Method are evaluated as further explained in the following sections. In Features Extraction Stage, it is necessary to extract all features and obtain the important ones to use them later in the classification stage which, in turn, undergoes many algorithms like Logistic Regression, Linear Discriminant Analysis, Classification and Regression Trees, K-nearest Neighbors, Support Vector Machine, and Convolutional Neural Network to detect if a nodule is cancerous or not. Another goal of this study is to look at how well and accurately the different algorithms are used to get the most accurate results.

II. Related Works

In most of the previous studies, the general four-stage system is used in automatic cancer detection [2], [3]. The system consists of four stages: image acquisition, image enhancement, image segmentation, and feature extraction and machine learning. In the image enhancement stage, several filters were used to enhance the quality of the image. These include the Gabor filter [2]–[5], FFT [3], Median and Wiener filters [4], [6], Gaussian Filter [6], Adaptive Histogram Equalization, and Laplacian operator [5]. In a comparative study, they reported that the Gabor filter outperforms the FFT [3].

The most frequently used image segmentation techniques were thresholding and marker-controlled watershed segmentation [3], [4], [6]. Thresholding was performed using Sobel's and Otsu's [5] methods. They found in [2] that watershed segmentation generates better results (accuracy of 85.27%) than the thresholding strategy, which had an accuracy of 81.24%. After ROIs are identified, the features of each region are extracted. Included in these features are the centroid, diameter, perimeter, area, eccentricity, and pixel mean intensity [6]. In [3] and [2], binarization and masking techniques are used for feature extraction as well. In [5], oriented gradient histograms were utilized. [7] Content-Based Image Retrieval

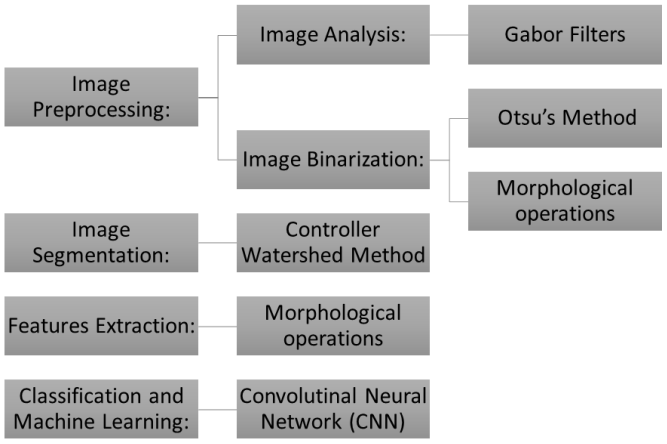


Fig. 1: System Architecture

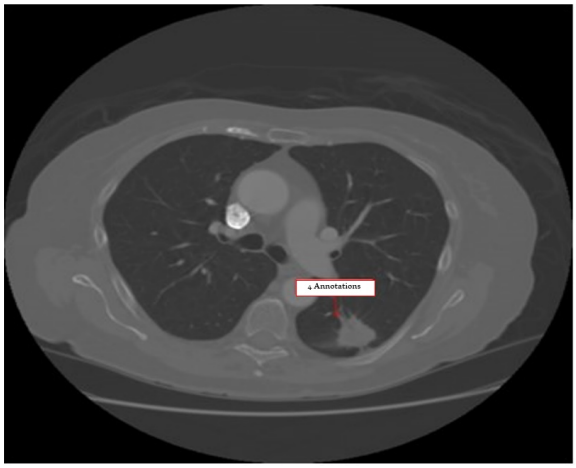


Fig. 2: Annotated region sample

(CBIR) is also used to extract characteristics such as contrast, intensity, texture, and shape.

The extracted features for each region are then classified into two classes: benign and malignant. Several classification algorithms are used in this stage. In [4], [6]–[8], the Support Vector Machine (SVM) algorithm reported classification accuracy. In [6], they reported a classification accuracy of 86.6%. They reported a detection accuracy of 92.4%. More recently, the convolution neural network (CNN) has been used for classification after image binarization with a reported accuracy of 94.34%. In [5], they used SVM, K-NN, decision trees, and artificial neural networks (ANN) for classification. Their result of the comparative study between SVM and ANN shows that both classifiers are effective. The proposed method produced a 98% prediction accuracy. In [9], CNN was used to develop a classification model. They have an accuracy of 90.47%. [10] developed a multi-classification deep learning model for detecting COVID-19, pneumonia, and lung cancer in chest X-ray and computed tomography (CT) pictures.

[8] Used machine learning techniques to early diagnose lung cancer. They applied SVR, LSTM, and Backpropagation on two groups: male and female dataset, then compared between the results of the three algorithms used. SVR performed outstanding prediction results compared to the others. They also managed to increase the accuracy of prediction by increasing the size of the training dataset. In [11] they used the blood count tests to investigate potential relationships between pathology tests, performed by General Partitioners, and cancer diagnosis. They used Decision trees models along with AdaBoost, LightGBM, and XGBoost models to increase the performance. They splitted the dataset used to depend on different ratios of non-cancer and cancer patients. They focused on detecting lung cancer with success in early indication of cancer diagnosis. [12] Proposed a system to solve the problem of imbalanced dataset using data minig models. They provided a new model called PoI to prune positive-class CARs from the dataset decreasing the FPR and FNR values when applied to diagnose breast cancer.

III. Method

The general system for lung cancer detection consists of four basic stages, as shown in Figure 1 and described in the following subsections.

In general, most of the related works used three-to-four stage systems. The difference was mainly in the algorithms used. The proposed system consists of a four-stage system similar to the mentioned methods in the related work. The goal of this study is to improve the design of some stages and to compare the performance of different machine learning algorithms.

A. Data set

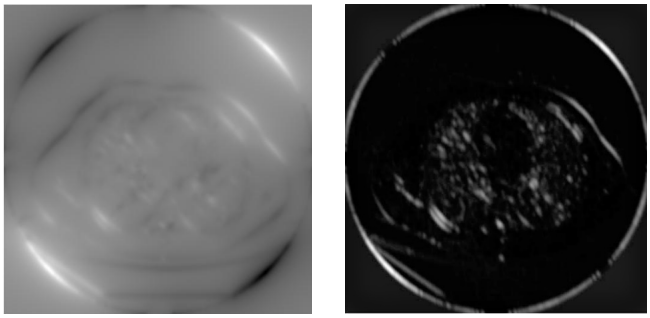
The dataset used in this study is the LIDC-IDRI dataset [13]. It comprises of 1018 thoracic CT scans with identified lesions that have been marked up. Each CT scan is accompanied by an XLM file that includes the results of a two-phase image annotation process carried out by four experienced thoracic radiologists. The data set contains nodule size reports and diagnosis reports. The pydicom library is used to read the DICOM images, and the pylidc library is used to extract the annotations. The radiologists' annotations include outlines of nodules $\geq$3mm in diameter on each CT slice along with the attributes needed for training the neural network model. The annotated pathologic features include: subtlety, internal structure, calcification, sphericity, margins, and malignancy. Malignancy feature is reduced from 5 classes to binary classes, where 1 indicates a cancer nodule and 0 indicates not-cancer nodule. Table I shows these features and the meaning of each annotation. Figure 2 shows a annotated region from an image.

B. Pre-processing

The original CT scans have Hounsfield Unit (HU) values. These values are used to propose candidate regions of interest (ROI) for the classification stage. After reading the images, a number of image-enhancing techniques are used to make the image smoother and remove the noise. The following image processing techniques were applied.

TABLE I: Pathologic features

Value	Subtlety	Internal Structure	Calcification	Sphericity	Margin	Malignancy
1	Extremely Subtle	Soft Tissue	Popcorn	Linear	Poorly Defined	Highly Unlikely
2	Moderately Subtle	Fluid	Laminated	Ovoid/Linear	Near Poorly Defined	Moderately Unlikely
3	Fairly Subtle	Fat	Solid	Ovoid	Medium Margin	Indeterminate
4	Moderately Obvious	Air	Non-central	Ovoid/Round	Near Sharp	Moderately Suspicious
5	Obvious		Central	Round	Sharp	Highly Suspicious
6			Absent			



(a) Gabor filter

(b) Boundaries detected after applying Gabor filter

Fig. 3: Image Enhancement

- Image enhancement: The Gabor filter enhances image quality by identifying local frequencies in certain directions surrounding the ROI. Because of its localization qualities in the spatial and frequency domains, the Gabor filter has the ability to detect edges. In this study, the Gabor filter is used to produce texture changes at the boundaries in order to partition the image into various regions. The images were filtered using a series of 28 Gabor filters with varied bandwidths and modulation frequencies. The used orientation angle on the positive y-axis (0°–180°) was changed by 45° to detect changes in each quarter of the frequency plane. The used wavelengths start at $2\sqrt{2}$ and grow in powers of 2 until they reach the hypotenuse length of the input image size. As a result, a batch of 28 filters was created. Then, for each filter in the set, a Fourier transform is applied. The obtained frequency characteristics are presented in Figure 3a, and the resulting Gabor magnitude responses are used as features to define ROI. Spatial information is added by smoothing the output with a low-pass Gaussian filter and estimating local energy. Finally, the features are molded into a 2D image of the same size as the input image, and the output features are normalized using the mean value of the pixels. Figure 3b depicts an image with detected boundaries.
- Image Binarization: In this step, the areas are separated into background and foreground, with the ROI located in the foreground. This is achieved by choosing a suitable threshold value. Various methods were evaluated, including Otsu's thresholding, global thresholding, and local (adaptive) thresholding. After multiple trials Otsu's method yielded an acceptable threshold value. Figure 4a shows the lung extraction as foreground objects. As shown in Figure 4b, the mask is then multiplied by the original image to provide the lungs with the pixel intensity values. Morphological procedures are applied to the binarized image to generate masks based on the forms of the region of interest and to ensure that pixels with identical features are contained within the same ROI, as illustrated in Figure 4c. As depicted in Figures 4d, and Figure 5, the subsequent phase eliminates any connected components with a total number of pixels less than the specified value of 5. This value was chosen because small blood arteries and lymph nodes can resemble malignancies.

C. Image segmentation

In this stage, watershed segmentation with the Watershed-Controller technique is applied to isolate each connected component and treat it as an independent ROI for feature extraction in the following stage. Figure 6 depicts a single ROI.

D. Features Extraction

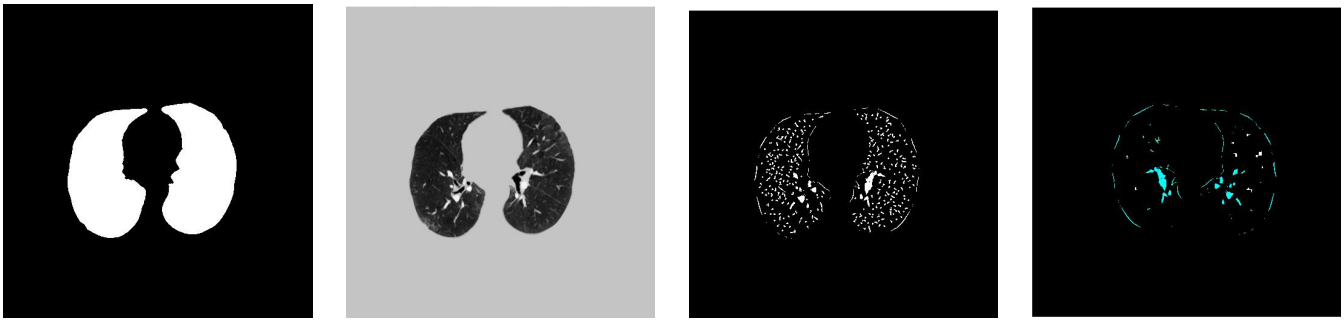
The segmented image is utilized to identify candidate regions and save each candidate as a distinct ROI. This stage is based on an examination of the ROI's geometrical properties, such as area, diameter, perimeter, centroid, and image. These details are saved in a csv file. Also, the images of the ROIs are saved in a separate matrix. Each row contains the intensity values of the pixels in the candidate ROI. Because these pixels have a HU unit, they are used as input in the classification stage. The ROI's images are stored at a resolution of 250×250 before being translated into a 1D of size 62501 ($250 \times 250 + 1 = 62501$), which includes the output labeled class.

E. Classification

Following the extraction of the candidate ROIs, these regions are classified into two classes: *malignant* and *benign*.

TABLE II: Annotation values for a ROI by 4 different radiologists

Feature	Radio I	Radio II	Radio III	Radio IV
Subtlety	5	5	5	5
Internal Structure	1	1	1	1
Calcification	6	6	6	6
Sphericity	3	4	3	5
Margin	3	4	2	4
Malignancy	2→ 0	5→ 1	5→ 1	4→1



(a) Extracted lungs (b) Multiplied with input image (c) identified ROI candidates (d) unwanted objects eliminated

Fig. 4: Image Binarization

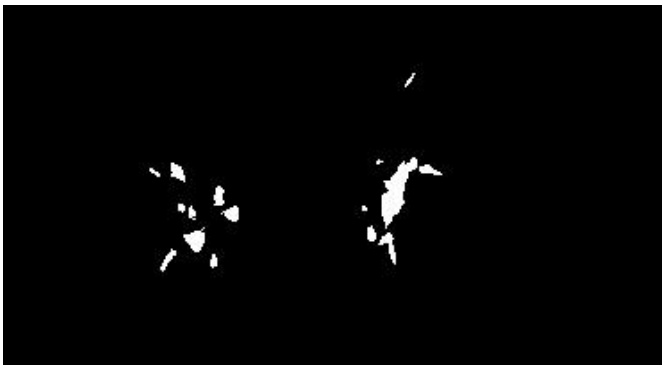


Fig. 5: ROI candidates

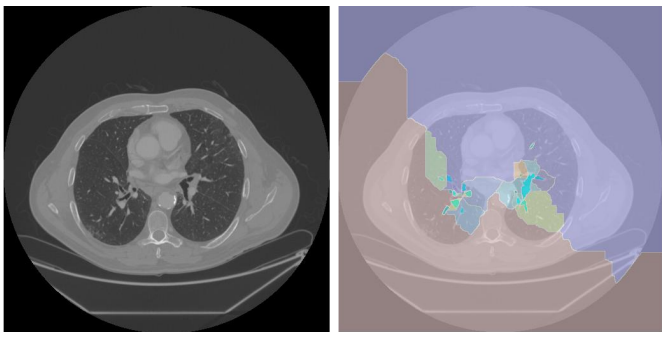
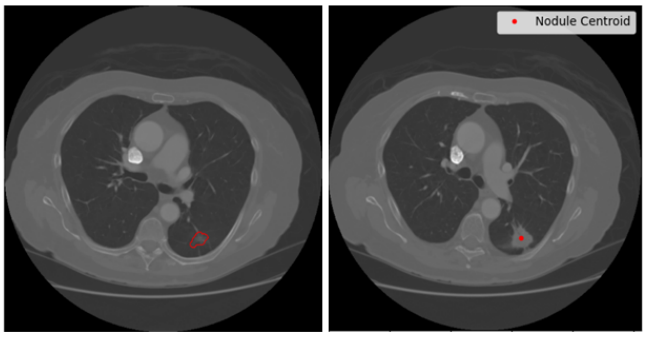


Fig. 6: Isolated ROI using Watershed-Controller



(a) (b)

Fig. 7: A ROI centroid

We used two approaches to classify these regions. In the first approach, we employed the geometrical parameters of each region as classification features. In the second technique, we fed the ROI images into a CNN and let the CNN learn the features. These approaches are discussed in further depth in the sections that follow.

1) Classification using ROI's geometrical features: The annotated data set contains various features as well as the contour of nodules with a diameter greater than 3 mm. Subtlety, internal structure, calcification, sphericity, and margins are the pathological characteristics utilized to classify a condition. A numerical scale from 1 to 5 represents subtlety, with 1 being the easiest to identify and 5 being the most difficult. The nodule's interior structure and internal makeup are identical. The calcification feature depicts the calcification pattern. Sphericity describes the nodule's three-dimensional shape in terms of how spherical it is. How effectively the nodule margin is described by the margin attribute. Malignancy is the target feature. The malignancy value is decreased from five classes to binary classes, where 1 denotes a cancer nodule and 0 denotes a non-cancer nodule. If the malignancy's value exceeds 3, it is given a 1; if it is equal to or less than 3, it is given a 0. Table II lists the annotation values for a specific region, while Figures 7 depicts the region's position, outline, and centroid. Each nodule's features are extracted based on the annotations the system has identified, and then the nodule is masked and saved as a separate ROI image for classification.

2) Classification using Convolutional Neural Networks (CNN): Using this method, a 2D image was extracted for each ROI. To accommodate the largest ROI image, an image dimension of 250×250 is selected. Each ROI's pixel values were maintained at their initial HU values from the original image. The matrix size of the input dataset is $33500 ROI \times 62501 pixels$, where each row corresponds to a 62500-pixel ROI image and 1 pixel is added for the output label. A convolutional neural network model, as shown in Figure 8, is then trained using these images. CNN would learn the features of each ROI automatically. The data set was divided into 70% training set, 20% test set, and 10% validation set. The CNN architecture consists of four convolution layers with Rectified

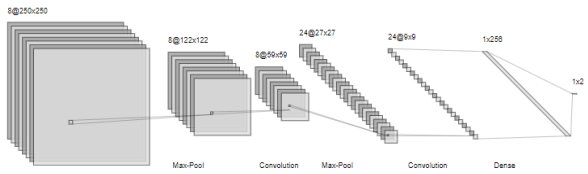


Fig. 8: CNN architecture

TABLE III: CNN layers

Layer	Output shape	No. parameters
Conv2D	$248 \times 248 \times 32$	320
Maxpooling	$124 \times 124 \times 32$	0
Conv2D	$122 \times 122 \times 64$	18496
Maxpooling	$61 \times 61 \times 64$	0
Conv2D	$59 \times 59 \times 128$	73856
Maxpooling	$29 \times 29 \times 128$	0
Conv2D	$27 \times 27 \times 256$	295168
Maxpooling	$9 \times 9 \times 256$	0
Flatten	20736	0
Flatten	256	4
Dense	256	4
Dense	1	4

Linear Units activation functions, four MAX pooling layers, a flattening layer, and a fully connected layer with a Sigmoid activation function to classify the image with a binary value; 1 for cancer and 0 for non-cancer, as explained in Table III. The loss function employed is the binary-cross entropy. The Adam gradient descent optimizer function was used.

IV. Results

The proposed technique is applied to multiple patient samples. The dataset contains 48 files with a total of 33542 ROIs across 8395 images. We utilized the MATLAB image processing toolbox for image enhancement and processing. The computational environment includes an Nvidia Geforce 1050Ti display adapter, a 3.30 GHz Intel Core i5 processor, 16GB RAM, and a 64-bit version of Windows 10 Professional. Utilizing parallel processing, specifically in loops, improved performance. A portion of the code was executed on the NVIDIA display adapter, which supports parallel processing, and the remainder was executed on the MATLAB parallel pool to accelerate code execution.

A. Preprocessing Stage:

The main procedures we used throughout the preprocessing stage are shown in Figure 9. Figure 9a displays the original image obtained from a CT medical scanning device. Figure 9b shows the image after using a series of 28 Gabor filters to eliminate noise during acquisition, then using Otsu's approach for binarization to find the best threshold value for differentiating foreground and background. In order to clear borders and locate pixels with similar intensity values in the same region of interest, a sequence of morphological operations were applied to the binarized image using this technique, as shown in Figure 9c. The lungs are depicted in Figure 9d following binarization. The ROIs that were found in the binary image after the border was erased and all undesirable areas were eliminated are shown in Figure 10.

B. Image Segmentation

In this stage, we applied the Water-Controller algorithm to separate each region as a single candidate for testing whether it indicates a cancer nodule or not, as shown in Figure 9d. The candidates are separated as shown in Figure 10.

C. Features Extraction Stage

This study examines three techniques for extracting ROI features. First, geometrical properties of each ROI are extracted and used as training features for the model. These features include area, centroid, and diameter. The second method consists of substituting the pixel values in each ROI image with their corresponding HU values from the preprocessed image. The binary ROI image is multiplied by the HU-valued image, and the resulting dataset is used in training a CNN. The third technique is identical to the second, with the exception that the original DICOM stored values (SV) were substituted for the pixel values.

The size of ROI images was based on the dimensions of the largest ROI. The dimension was set to 250×250 pixels. Due to performance concerns, a reduction in size was required. Comparing the nodules in the lungs to the labeled dataset, the accuracy of detecting ROIs with cancer is 100%. For a random sample extracted from the dataset containing nodules identified inside and near the boundaries, the accuracy of detecting suspicious areas of tumor near the lung's boundaries was 72.92%.

D. Classification stage:

This stage includes reading the annotations and extracting the desired attributes (pathologic features). The features were then classified using six different classification algorithms in one approach and the ROIs extracted images as input to a Convolutional Neural Network model in the other.

1) Reading annotations and extracting features
   After extracting the annotations and defining the nodules of interests, a boolean mask was taken for each nodule and multiplied by the original image to obtain its original pixels' intensity values. The output is stored in a csv file with size of 14000 annotations $\times$ 6 features.
2) Classification step The extracted annotations' csv file was used as the input dataset to determine whether a nodule is cancerous or not. The dataset consists of five input parameters and a single output class. The dataset was divided into an 80% training set and a 20% testing set. In the first approach, six classification algorithms were used to train and test the input dataset: Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Classification and Regression Trees, Naïve Bayes, and Support Vector Machine. The accuracy results using each algorithm are shown in Table IV. The Support Vector Machine returned the highest prediction accuracy of 85.43%.

(a) Original image

(b) After Gabor filters

(c) After binarization

(d) After morphological operations

Fig. 9: Image processing enhancements

TABLE IV: Cross validation using images of size 128×128

Algorithm	Precision	Recall	F1–score	Accuracy
Logistic Regression (LR)	79%	79%	79%	83%
Linear Discriminant Analysis (LDA)	79%	78%	78%	83%
Classification and regression trees (CART)	80%	77%	78%	83%
K-nearest neighbors (KNN)	77%	74%	76%	81%
Support Vector Machine (SVM)	78%	88%	83%	85%
Convolutional Neural Network (CNN)	100%	89%	94%	94%

Fig. 10: Final labeled ROI

In the second approach, CNN is used to automatically extract the features of each ROI using the images of the ROIs as input. These images consist of pixels' HU intensity values stored as row vectors for each ROI. Due to machine and memory limitations, the training dataset was reduced to 3500 samples × 62501 pixels. The accuracy obtained using this approach outperforms the earlier approach by 93.75%.

The dataset obtained from the image processing stage was used to validate the CNN model. This dataset was taken from 48 patients' with a total of 33540 ROIs extracted. The model predicted 179 ROIs as cancer nodules and the remaining regions were classified as not cancer nodules. This output is considered satisfying prediction as the dataset is greatly imbalanced considering the ROIs extracted are all the nodes' the system can define as candidates within the lungs from the 48 patients.

V. Discussion

In this paper, we propose a method for detecting lung cancer using CT scans. The proposed system had four major stages: preprocessing, segmentation, feature extraction, and machine learning and classification. In this system, CT images underwent preprocessing to improve image quality and eliminate noise. In the second step, the image was segmented to delineate the boundaries between various tissues. Important characteristics were extracted in the third phase. In the fourth step, machine learning was applied to the extracted features to identify locations where cancer was present.

The images were preprocessed sequentially using a set of Gabor filters to remove noise and improve the quality of ROI extraction for later stages. Comparing Gabor to Median and Weiner filters, Gabor accurately distinguished lung regions from those of other organs. For the binarization step, Otsu's Algorithm was used to determine the appropriate threshold value for background and foreground separation. The returned threshold value made this separation more practical than Global Thresholding or Local Thresholding. In the segmentation stage, the Watershed algorithm was used.

Features Extraction stage was performed on selected regions of interests producing a table of shape properties, and another matric of pixels' intensity value. Finally, various machine learning techniques were used to classify cancer using the pathologic aspects of the annotations that were collected from DICOM images. As noted in the Methodology section, six classification algorithms were contrasted for training the features that were derived from the annotations. When applied to the system's dataset, SVM demonstrated the best classification accuracy among them, scoring 85.43 percent. These steps were followed by a stage of testing the data extracted from image processing techniques with the Convolutional Neural Network model. Inserting the ROIs to the CNN model trained in this paper returned a detection accuracy of 93.75% and detected 179 cancerous nodules from the dataset used with this study.

The proposed system failed to detect tumors at the boundaries of the lungs because the threshold value obtained recognized them as background along with the chest wall. Otsu's binarization stage method did not provide the desired threshold value for each image to separate background from foreground. The sample size that the software can analyze was limited

due to hardware limitations. Increased dataset size for both categories to improve classification stage.

VI. Conclusion

In this study, two approaches for classifying the ROI as benign tumor or malign were evaluated. In the first approach, the geometrical characteristics of the ROI were used to classify them. The image of the ROI with HU unit pixel values was used in a CNN to automatically extract the features in the second approach. Our results show that the second approach outperforms the first, with an accuracy of 85.43% in the first approach compared to 93.75% in the second.

References

[1] A. C. Society, "Tests for non-small cell lung cancer," https://www.cancer.org/cancer/non-small-cell-lung-cancer/detection-diagnosis-staging/how-diagnosed.html, accessed: 2022-08-23.

[2] B. G. Patil and S. N. Jain, "Cancer cells detection using digital image processing methods," *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, vol. 3, no. 4, 2014.

[3] M. S. AL-TARAWNEH, "Lung cancer detection using image processing techniques," *Leonardo Electronic Journal of Practices and Technologies*, vol. 3, no. 20, 2012.

[4] M. Vijay, A.Gajdhane, and D. L.M, "Detection of lung cancer stages on ct scan images by using various image processing techniques," *IOSR Journal of Computer Engineering*, vol. 16, pp. 28–35, 2014.

[5] W. Rahane, H. Dalvi, Y. Magar, A. Kalane, and S. Jondhale, "Lung cancer detection using image processing and machine learning healthcare," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 2018, pp. 1–5.

[6] S. Makajua, P. Prasad, A. Alsadoona, A. K. Singhb, and A. Elchouemic, "Lung cancer detection using ct scan images," in *6th International Conference on Smart Computing and Communications, ICSCC 2017*, vol. 125. Procedia Computer Science, December 2018, p. 107–114.

[7] B. Muthazhagan, T. Ravi, and D. Rajinigirinath, "An enhanced computer-assisted lung cancer detection method using content-based image retrieval and data mining techniques," *J Ambient Intell Human Compute*, 2020.

[8] K. Tuncal, B. Sekeroglu, and C. Ozkan, "Lung cancer incidence prediction using machine learning algorithms," *Journal of Advances in Information Technology*, vol. 11, no. 2, pp. 91–96, 2020.

[9] A. Gupta, V. K. Manda, and B. I. Seraphim, "Lung cancer detection using image processing and convolutional neural network," *Annals of R.S.C.B*, vol. 25, no. 4, p. 3044–3048, 2021.

[10] D. M. Ibrahim, N. M. Elshennawy, and A. M. Sarhan, "Deep-chest: Multi-classification deep learning model for diagnosing covid-19, pneumonia, and lung cancer chest diseases," *Computers in Biology and Medicine*, vol. 132, p. 104348, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521001426

[11] G. Ristanoski, J. Emery, J. M. Gutierrez, D. McCarthy, and U. Aickelin, "Ai based cancer detection models using primary care datasets," *Journal of Advances in Information Technology*, vol. 13, no. 2, pp. 192–197, 2022.

[12] P. Liewlom, "Class-association-rules pruning by the profitability-ofinterestingness measure: Case study of an imbalanced class ratio in a breast cancer dataset," *Journal of Advances in Information Technology*, vol. 12, no. 3, pp. 246–252, 2021.

[13] C. I. Archive, "Cancer image archive," http://www.cancerimagingarchive.net/, accessed: 2022-08-23.