Enhancing lung lesion localization in CT-scans: a novel approach using FE_CXY and statistical analysis

Nurul Najiha Jafery[1], Siti Noraini Sulaiman[1,2,3], Muhammad Khusairi Osman[1,2], Noor Khairiah A. Karim[4], Mohd Firdaus Abdullah[1], Iza Sazanita Isa[1,2], Zainal Hisham Che Soh[1]

[1]Electrical Engineering Studies, College of Engineering, Universiti Teknologi MARA, Cawangan Pulau Pinang,
Permatang Pauh Campus, Permatang Pauh, Penang, Malaysia
[2]Advanced Rehabilitation Engineering in Diagnostic and Monitoring Research Group (AREDiM), Electrical Engineering Studies,
College of Engineering, Universiti Teknologi MARA, Cawangan Pulau Pinang,
Permatang Pauh Campus, Permatang Pauh, Penang, Malaysia
[3]Integrative Pharmacogenomics Institute (iPROMISE), Universiti Teknologi MARA, Puncak Alam Campus,
Puncak Alam, Malaysia
[4]Department of Biomedical Imaging, Advanced Medical and Dental Institute, Universiti Sains Malaysia, Kepala Batas,
Penang, Malaysia

Article Info

Article history:

Received Nov 30, 2023
Revised Feb 16, 2024
Accepted Feb 22, 2024

Keywords:

Diagnosis
Feature extraction
Lesion localization
Lung cancer
Stastical approach

ABSTRACT

Intelligence algorithm systems rely on a large dataset to effectively extract significant features that can recognize patterns for classification purposes and extensively utilized to assist the physicians in diagnosis of lung cancer. Extracting valuable features from the available dataset is crucial, especially in cases where additional real data may not be readily accessible. In this context, we propose a novel method called feature extraction based on centroid (FE_CXY) for lesion localization, utilizing a statistical approach. The approach begins with a segmentation process that employs image processing techniques to extract features of interest which is data centroid. This extracted data is then used to compute statistical measurements, revealing hidden patterns that contribute to distinguishing between lesion and non-lesion locations. The method's efficiency is reflected in the development of robust models with improved performance in localizing lung lesions. The study's statistical findings strongly indicate that FE_CXY plays a crucial role as an important feature for detecting lesion localization supported by a student's t-test, which identifies a statistically significant difference in the patterns between lesion and non-lesion localization ($p<0.05$). By incorporating this method into lung cancer detection systems, we anticipate improved accuracy and efficacy, thereby benefiting early diagnosis and treatment planning.

Corresponding Author:

Siti Noraini Sulaiman
Electrical Engineering Studies, College of Engineering, Universiti Teknologi MARA
Cawangan Pulau Pinang, Permatang Pauh Campus
13500 Permatang Pauh, Penang, Malaysia
Email: sitinoraini@uitm.edu.my

1. INTRODUCTION

Lung cancer remains the leading cause of cancer-related deaths in men aged 40 and above, as well as women aged 60 and above [1], [2]. Recent data indicate a concerning increase in its incidence, with over 350 daily deaths due to the disease [2]. Lung cancer, also known as bronchogenic carcinoma, encompasses tumours that originate in the lung parenchyma or bronchi. It is typically classified into two types: small-cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). SCLC tends to grow more rapidly than NSCLC but exhibits better response to chemotherapy. SCLC is further classified as “limited stage" (usually confined to the chest) or "extensive stage" (cancer that has spread beyond the chest) [3]. Lung nodules, clusters of cells found in the lungs, are common, often resulting from scarring due to previous lung infections. Although lung cancer symptoms are typically rare. Previous studies have associated various symptoms with lung nodules, such as shortness of breath, back discomfort, weight loss, weakness, and fatigue. Lung nodules are often incidentally detected during chest X-rays or computed tomography (CT) scans conducted for other purposes. While some lung nodules are cancerous, others are benign [4]. Early detection of lung nodules is crucial for improving survival rates. Smoking remains the primary cause of lung cancer, with male smokers at the highest risk. Interestingly, the number of packs smoked per year does not exhibit a straightforward correlation with lung cancer due to the complex interplay between smoking, environmental factors, and genetics [3], [5].

The complexity of medical data poses challenges for physicians in extracting relevant information, leading to longer screening, diagnosis, and prognosis times. Manpower shortages and human errors, such as eye fatigue and tiredness, can result in misdiagnosis and false interpretations. computer-aided diagnosis (CAD) systems can serve as a valuable second opinion for physicians, aiding in accurate diagnosis and improving the effectiveness of treatment. In recent times, researchers have focused on developing CAD systems that employ machine learning and deep learning techniques for early diagnosis by automatically learning and extracting features [6]–[8]. CAD systems for lung diagnosis typically consists of process such as segmentation, feature extraction, and classification. CAD systems can provide valuable insights for medical practitioners. But an accurate segmentation is crucial for efficiently lung cancer diagnosis, enabling the identification of lung regions and nodules [9]. CAD systems often used a two-step pattern recognition approach, combining feature extraction through neural networks or statistical classifiers. Researchers have made significant progress in training and classifying large datasets using feature extraction for pattern recognition [6], [10]. Medical image classification, a specialization that merges machine learning and computer vision, utilizes machine learning approaches to automate visual model acquisition, signal translation, and trainable image processing system construction [11]. These classifiers possess the ability to predict and classify diseases based on the employed machine learning methodologies, contributing to significant advancements in medical diagnostic knowledge using feature extraction and machine learning classifiers in medical imaging applications.

Elwahsh *et al.* [7] study involved the utilization of a deep neural learning cancer prediction model (DNLC) that consist of three stages. In the first stage, a deep network (DN) was used to select the best set of features from the datasets. Then training genomic or clinical data samples with a deep neural network (DNN). Finally, the DNLC model's ability to predict cancer at earlier stages was evaluated. The study found that when a traditional neural network was used, the number of hidden layers would increase, and weight matrices in the initial hidden layers close to the input layer would remain unchanged. Consequently, traditional neural networks were deemed unsuitable, highlighting the need for feature extraction and reduction techniques such as principal component analysis (PCA) and DNLC. Experimental results demonstrated that the proposed model achieved higher accuracy compared to earlier convolutional neural networks (CNN) and recurrent neural network (RNN) models, with an average accuracy of 93%, outperforming other methods in all scenarios [7].

Suresh and Mohan in their study in 2022 [6] trained DCNNs using samples from the nodule region of interest (NROI) and further classified them into non-cancerous, benign, or malignant categories based on tumour patterns. They manually extracted a total of 26 features from traditional hand-crafted methods, defining discriminative features such as area, perimeter, eccentricity, contrast, correlation, energy, entropy, homogeneity, sum average, and sum entropy. These features were combined and trained using support vector machine (SVM) classifiers. Yang *et al.* [12], in 2021 introduced a generative adversarial network (GAN)-based framework to generate visually normal-looking CT slices from CT slices with COVID-19 lesions. They developed a feature-matching strategy to enhance the realism of generated images by guiding the generator to capture the complex texture of chest CT images. By subtracting the output image from its corresponding input image, the localization map of lesions could be easily obtained. Kuwil [11] presented a new approach called feature extraction based on region of mines (FE_mines) that combined feature selection, reduction, and extraction. Three methods, namely FE_AM, FE_CM, and FE_UM, were used for different types of images. The statistical methodology relied on data distribution and employed measures of central tendency (MCT) and dispersion to investigate the distribution of data. The results demonstrated that the FE_mines approach achieved higher accuracy ranges (1 to 13%) within the three methods [11].

There is still a shortage of research on real-time data, even though the current CAD system has good classification accuracy [13]. Other than that, adding to many features into classifier would harm its performance. Some researchers had stated that finding meaningful patterns is more challenging when there are numerous weakly informative features compared to when there are only a few from strong informative features. Many practical research design often results in suboptimal patterns, making it less likely to achieve statistically significant validation results [14]. Table 1 present an additional previous study on feature fusion of handcrafted features for CT scan related CAD task in similar domain with different type of features.

Table 1. Previous studies on feature fusion of handcrafted features for CT scan related CAD task

Reference	Feature extracted	Task
[15]	Geometrical features (area, perimeter, solidity, centroid, equivalent diameter, convex area, eccentricity and roundness) Statistical features (contrast, correlation, variance and homogeneity)	Nodule vs non-nodule
[16]	Delta radiomics features (volume, diameter, boundary sharpness, shape, and texture)	Lung cancer classification
[17]	Local binary pattern (LBP) based features	Nodule vs non-nodule
[18]	Gabor features	Lung cancer diagnosis
[19]	Wavelet features, texture features and histogram features	Lung cancer prediction
[20]	Super pixels features	Tumour vs non-tumour vs fundus (segmentation)
[21]	Morphological features, genomic features, and molecular features	Tumour vs non-tumour

This research employed an appropriate processing, training, and validation procedure tailored to the specific features of the data. This benchmarking process enhances the understanding of the relationship between proposed features and lung lesion detection and localization. The contributions of this paper can be summarized as follows:

- Introduction of a novel method based on statistical analysis to identify patterns in lung CT scan images features for lesion localization.
- Utilization of appropriate statistical analysis techniques to extract hidden features in lung CT scan images.
- Employment of a quantitative approach to enhance the performance and efficiency of machine learning and deep learning models.

The structure of this paper is as follows: section 2 describes the methodology used for analysis. The experiments and results are presented in Section 3. Section 4 discusses the outcomes, while section 5 presents the conclusion and future work.

2. METHOD

In this section, an overview of the dataset used will be provided, along with the data processing and statistical approach employed. The statistical measurements employed in this method is to explain the strength of the selected features and reveals hidden patterns inside these features. The experimentation was conducted using MATLAB (R2022a) on a notebook equipped with an AMD Ryzen 7 Pro 5850U CPU @ 1.9 GHz processor, 16 GB RAM, and the Windows 11 64-bit operating system. Based on Figure 1, the proposed work is divided into three (3) main stages, which are data collection, pre-processing, and the introduction of an innovative approach known as FE_CXY, which is used to extract features for lesion localization coupled with statistical analysis.

2.1. Data collection

The data used in this study consisted of Axial-cut Lung CT Scan Images obtained from the Imaging Unit at the Advanced Medical and Dental Institute (AMDI), USM. The ethical application was approved for the collection of data with the study code: USM/JEPeM/21110721. There are 990 CT scan images from with a resolution of 512×512 were acquired with slice thickness 1.25 mm from thorax regions have been used in this experiment. These images were in digital imaging and communications in medicine (DICOM) format and were imported into MATLAB software after being sorting them into files. Then the images were converted from DICOM to grayscale in bmp format because it is easier to handle in MATLAB software.

2.2. Pre-processing

In the initial phase of data processing, the procedure involves two critical segments: one is focused on isolating the lungs, and the other is dedicated to extracting lesions within the lungs. The forthcoming section will provide a detailed breakdown of the steps involved in each of these stages, illustrated in Figure 2. This process of segmentation holds immense significance. It’s responsible for pinpointing objects or delineations within the image, a crucial step in identifying the specific area of interest. By dividing the image into distinct regions, it becomes possible to extract meaningful information, specifically valuable features [13]. The application of a CT scan yields intricate views of the body’s soft tissues, encompassing detail of blood vessels, muscle tissue and organs [22]. Hence, the extraction of the lung region from CT images plays a pivotal role. It significantly narrows down the scope of search when identifying lung lesions, making the process more efficient and accurate.

In this study, the process of isolating the lung region involved a technique based on thresholding [23]. Initially, the image undergoes thresholding, which is a fundamental method in image segmentation [24]. It offers a simple yet remarkably effective way to separate an image into a foreground and background. This is achieved by converting grayscale images into binary format. Next, to precisely locate the lung region in a CT scan image, we employ a method known as connected component analysis (CCA) [25]. This technique operates on a binary image, where ‘False’ values represent background pixels and ‘True’ values represent foreground or object pixels. Once region boundaries of regions are identified, it proves beneficial to extract regions that are contiguous and not divided by a boundary. These sets of connected pixels form cohesive units, enabling the image to be partitioned into distinct segments. Segmenting lung lesions and non-lesions involves several steps. The initial lung segmentation process yields images with low contrast, making it challenging to distinguish between lesions and non-lesions. Therefore, we begin by enhancing the image quality using local contrast stretching. Once the image quality is improved, areas of high contrast are identified to represent potential lung lesion regions. Subsequently, the enhanced image is converted into a binary format, and Otsu’s thresholding is applied to extract potential lesions and non-lesions. Despite these efforts, there are still challenges that require additional computational resources for image segmentation. To simplify lung lesion detection and classification in subsequent tasks, the number of non-lesions needs to be minimised. This is achieved by extracting several geometrical features using the equations provided in Table 2. By considering the diameter and roundness ranges of lung lesions, we reduce the number of non-lesions in the segmented images [26]. The output images from these processes will be used as input for the next step of the main proposition of this paper.

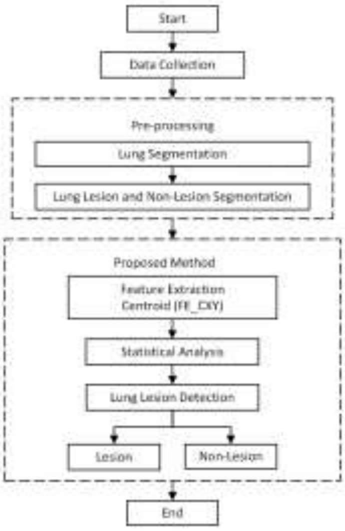


Figure 1. The workflow of proposed method

Table 2. Geometrical features used with equations

Geometric features used	Equation and function	
Diameter (*f1*)	mean ([stats.MajorAxisLength stats.MinorAxisLength],2);	(1)
Roundness (*f2*)	$\frac{4\pi \times \text{Area}}{\text{Perimeter}^2}$	(2)
Centroid (*f3*)	regionprops(bw,"Centroid")	(3)

2.3. Statistical approach for enhanced detection

This section holds paramount importance in our research. Feature extraction plays a pivotal role in enabling CAD systems to accurately identify true-positive lesions [27]. The good features can differentiate between lesion and non-lesion. Furthermore, we want to highlight the contribution of centroid (*f3*) as a valuable feature that enhances the precision of lung lesion localization. In the preceding step, centroids (FE_CX, FE_CY) for both lesions and non-lesions were derived using a MATLAB function. Figure 3 provides a visual representation of the proposed innovative approach, which encompasses three main processes: feature extraction, statistical analysis, and lung lesion detection.

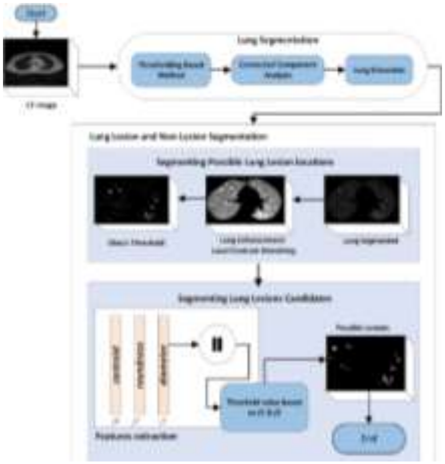

Start
CT image
Lung Segmentation
Thresholding Based Method
Connected Component Analysis
Lung Extraction
Lung Lesion and Non-Lesion Segmentation
Segmenting Possible Lung Lesion Locations
Otsu's Threshold
Lung Enhancement Local Contrast Stretching
Lung Segmented
Segmenting Lung Lesions Candidates
Possible Lesions
Features extraction
End

Figure 2. Block diagram representing outlining the pre-processing

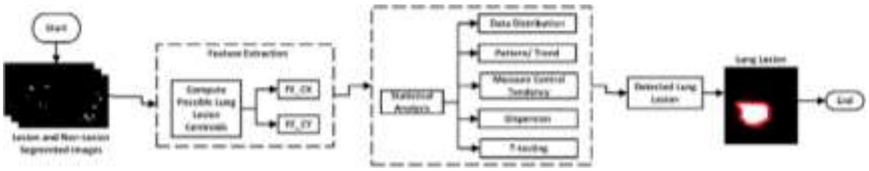


Figure 3. Block diagram representing statistical approach for enhanced detection

At the outset, we included a set of segmented images, comprising lesion samples and non-lesion samples, for this study. Subsequently, we extracted centroid values for both the lung lesion and non-lesion samples. The centroid of a circle, often referred to as the circle's central point, coincides with its radius measured from the circle's edges [28]. It's represented by two coordinates, known as X and Y. In our representation, FE_CX signifies the X-coordinate, while FE_CY represents the Y-coordinate, as illustrated in Figure 3. To delve deeper into the information provided by centroids and unearth their hidden insights, we will conduct a statistical analysis.

Sample standard deviations (s) were employed to describe the centroid feature pattern for both lesion and non-lesion samples. These sample standard deviations were calculated using (1).

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}} \tag{1}$$

Where: s = sample standard deviation, N = the number of observations for each lesion/non-lesion, x_i = the centroid value of lesion/non-lesion and $\overline{x}$ = the centroid mean value of lesion/non-lesion.

The analysis of residuals plays a critical role in validating the accuracy of a model. These residual values, derived from data at hand, serve as approximations of the model's margin of error. The residual calculated using (2).

$$r = x - x_0 \tag{2}$$

where r = residual, x = the centroid value for lesion/non-lesion and x_0 = the centroid mean value for lesion/non-lesion. The distance between two points can be described as the measurement of the straight line that connects these two points in a two-dimensional plane. The formula to find distance between two points is usually given by:

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \tag{3}$$

where d = distance, $(X_2 - X_1)$ = coordinates of the first point and $(Y_2 - Y_1)$ = coordinates of the second point.

Student’s t-test or also known as T-test was used for a quantitative comparison of centroid FE_CXY values between lesions and non-lesions. Student’s is defined in the (4).

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \tag{4}$$

where $\bar{x}$ = mean, s = standard deviation and n = number of values in each group.

In this study, the standard deviation, residual, and distance were utilized to measure the extent of motion of lesions and non-lesions within the CT slices based on the centroid coordinates (X, Y). With the information from the standard deviation of FE_CXY, student's t-test will be conducted to detect the difference between centroid data of lesion and non-lesion. This analysis is to prove that FE_CXY is a valuable feature for intelligent algorithm system in detecting the lesion of lung cancer.

3. RESULTS AND DISCUSSION

This section focuses on result of the proposed approach, which is based on a simple idea of obtaining different formulas of lung CT slice and determining the FE_CXY utilizing statistical analysis to discriminate between lesions and non-lesions. To avoid overwhelming readers with numerous equations and mathematical formulas, a concise analysis will be provided. There are two parts inside this section: analysis of potential lung lesion segmentation and result of further analysis of FE_CXY.

3.1. Analysis of potential lung lesion segmentation

As previously described, unwanted non-lesion was removed by applying diameter and roundness features as filters. This resulted in a significant reduction in false lesions. The counts of both lesions and non-lesions before and after this filtering process are presented in Table 3 and illustrated in Figure 4. Table 3 highlights five specific images where a noticeable decrease in both lesion and non-lesion counts is evident. This process has made it easier to easy to figure out lesion in the proposed method. In Figure 5, six non-lesion samples were randomly selected from these consecutive slices. As depicted in the figure, it's evident that non-lesions outnumber lung lesions. This is a typical observation, as non-lesions often include representations of blood vessels or various tissues [29].

Table 3. Comparison of lesion and non-lesion in segmented images before and after filtering

Image	No. of lesion and non-lesion before filtering	No. of lesion and non-lesion after filtering
1	48	12
2	191	16
3	288	13
4	338	20
5	366	29

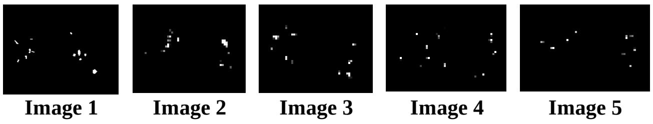


Figure 4. Segmented images of lesion and non-lesion after filtering using geometric features

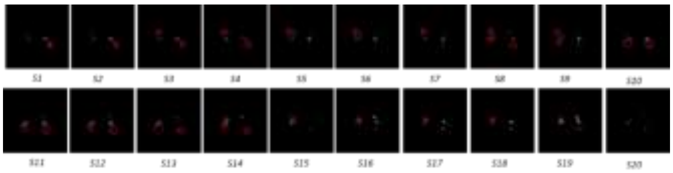


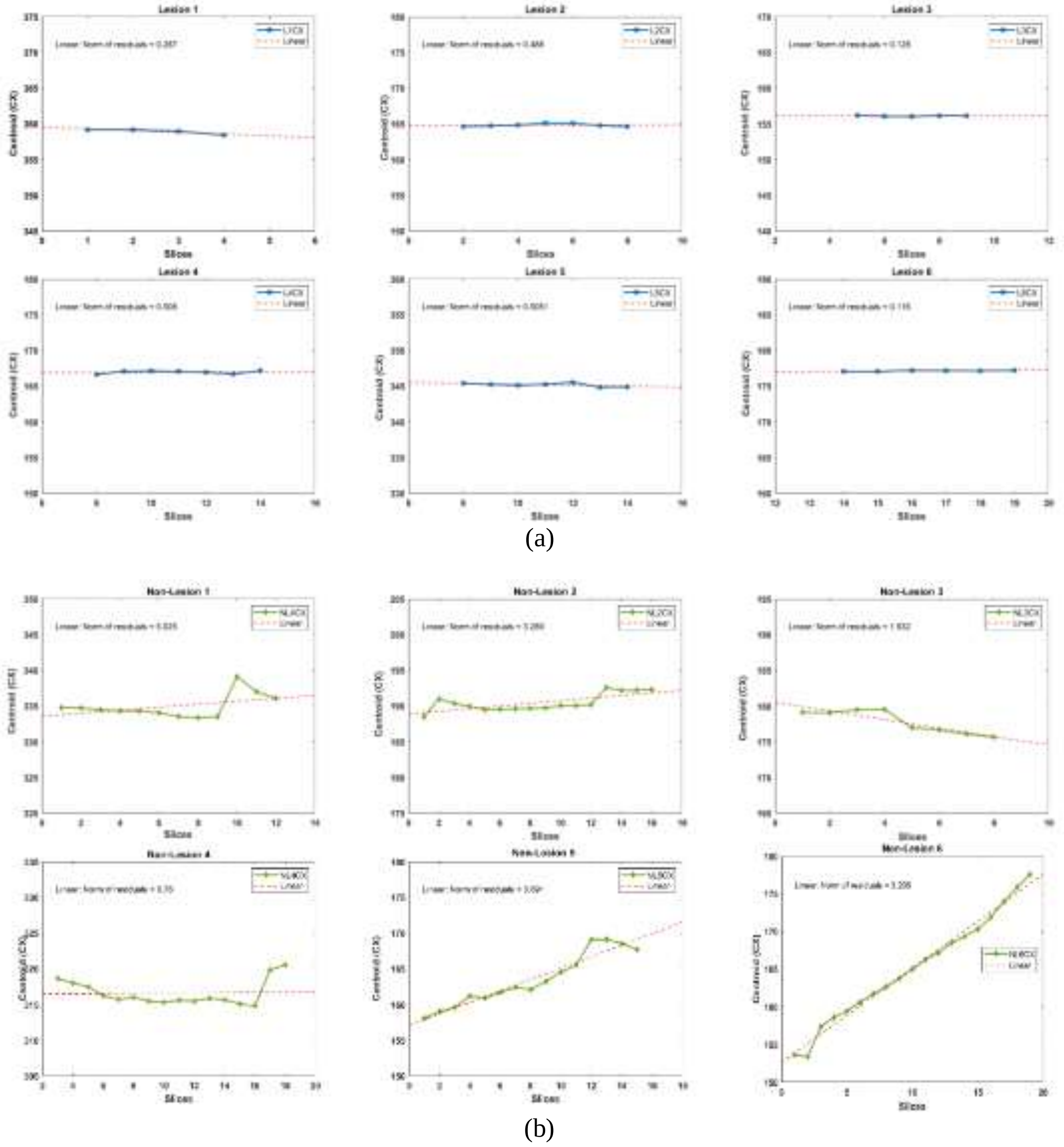
Figure 5. Represent the 20 lung CT consecutive slices

3.2. Statistical analysis of FE_CXY

In this experiment, the contributions of feature extraction approaches which are used in this work are evaluated. Choosing the right features is a crucial step in designing a pattern recognition system, as it enables the system to automatically identify the most relevant attributes within the feature set. To create a classification system that operates effectively, it is essential to select features that accurately capture the significant distinctions between the two classes under consideration, which in this case are lesions and non-lesions [30].

3.2.1. Trend and pattern analysis

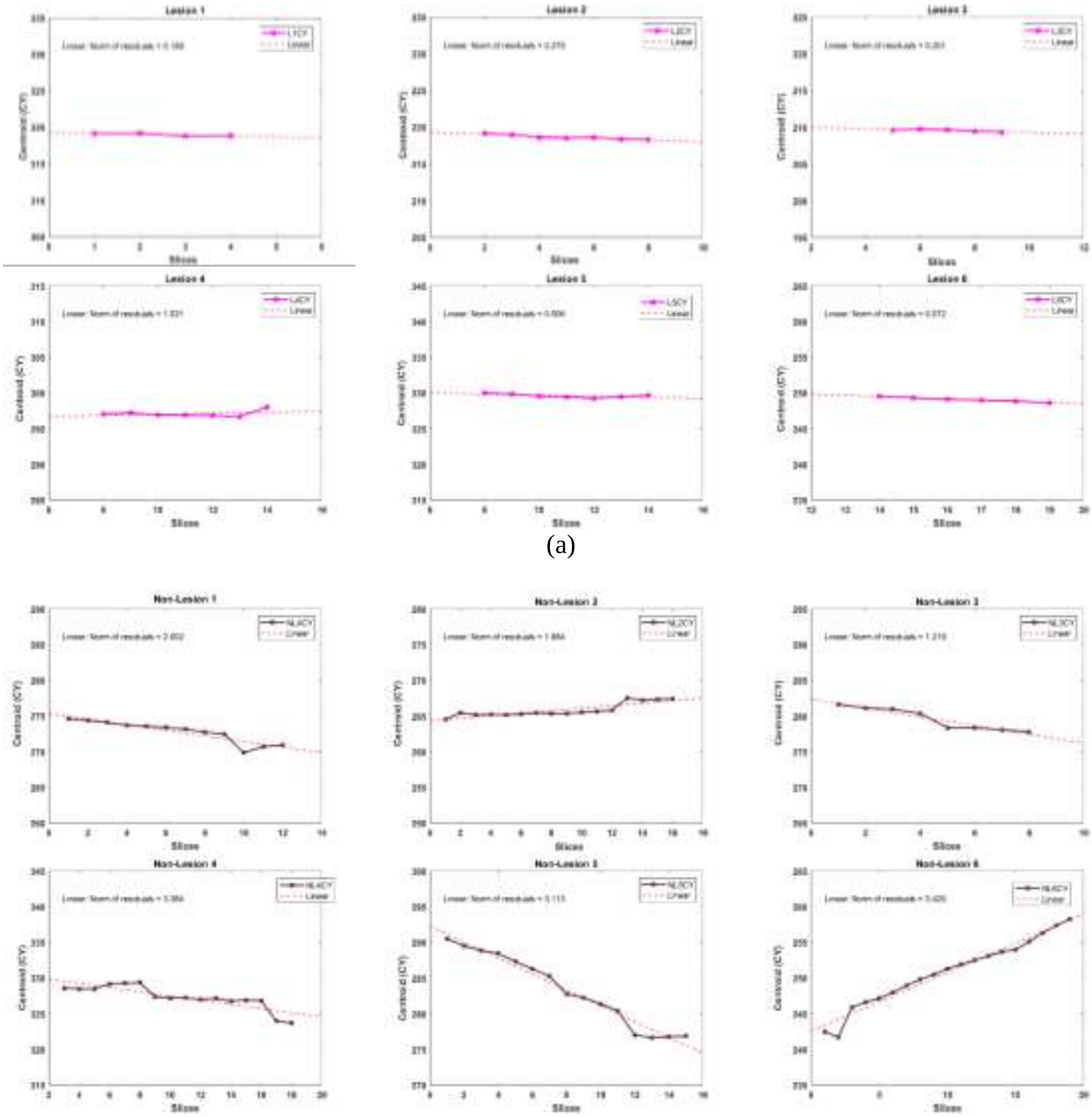
The Figures in this section serve the purpose of providing a visual representation of the data, aiding in comprehending the overall trends and variations in centroid features for both lesion and non-lesion samples. In Figures 6 and 7, we present a plot illustrating the distribution of FE_CX and FE_CY for six identified lesions (L1, L2, L3, L4, L5, and L6) and six non-lesions (NL1, NL2, NL3, NL4, NL5, NL6) within the 20 consecutive slices of lung CT. The plot in Figure 6 corresponds to the X-coordinate of the FE_CX and Y-coordinate of FE_CY in Figure 7. The x-axis denotes the CT slices, while the y-axis represents the centroid (FE_CX and FE_CY) values. On close examination, consistent trends of data dispersion are noticeable in lesions (Figure 6(a) and Figure 7 (a)). This consistent trend is characterized by a straight, horizontal line that maintains a constant value throughout the FE_CX and FE_CY values of lesions. But for FE_CX and FE_CY of non-lesions, there exhibits a discernible trend of fluctuation within the graph in Figure 6(b) and Figure 7(b). The significance of FE_CXY lies in its capacity to unveil the movement patterns of lesions across different slices. When we graph the values of FE_CXY for both lesion and non-lesion samples, a clear trend in the data emerges. Non-lesions tend to display ascending or descending patterns, while lesions exhibit a more consistent pattern. This is because of the shape of the lesion and non-lesion in the collocated of the lung CT scan. Essentially, if lesion presence in one slice, then it is expected to be appear in the preceding slices at the same location (FE_CXY) as previous slice. Conversely, the non-lesion transforms into new shapes, so if they are detected in one slice there are high chances that they will not be present in the exact location (FE_CXY) in the next slice of the series [31].



(a)

(b)

Figure 6. Distribution for FE_CX for (a) lesion and (b) non-lesion



(a)

(b)

Figure 7. Distribution for FE_CY (a) lesion and (b) non-lesion

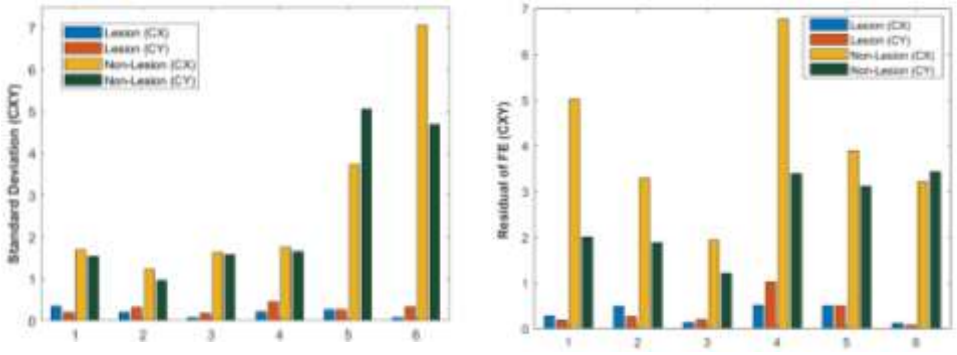
3.2.2. Further analysis

A statistical approach was applied to conduct a thorough examination and derive results to analyze the motion patterns between lesions and non-lesions, specifically focusing on the spread of X and Y centroid coordinates. In particular, standard deviation, residual, and distance metrics were computed for this experiment. These metrics furnish crucial insights into the localization of lesions.

Table 4 summarizes the standard deviation and residual values for the centroid feature, allowing for a direct comparison between lesions and non-lesions. It shows that lesions have smaller standard deviation and residual values compared to non-lesions. For lesions, the standard deviation ranges from 0.072 to 0.352 for CX and 0.165 to 0.442 for CY, while for non-lesions, it ranges from 1.223 to 7.072 for CX and 0.967 to 6.780 for CY. Similarly, the residual values for lesions are smaller, ranging from 0.116 to 0.508 for CX and 0.072 to 1.021 for CY, whereas for non-lesions, they range from 1.932 to 5.053 for CX and 1.219 to 3.429 for CY. The graphs in Figure 8 serve to visualize the differences between lesions and non-lesions more clearly based on the standard deviation and residuals of FE_CXY values, which cannot be seen in Table 4. Figure 8(a) represents the standard deviation values, and Figure 8(b) represents the residual values. The height of the bars in the graph signifies that lesions have small values in standard deviation and residual of FE_CXY, while non-lesions have large values in standard deviation and residual of FE_CXY.

Table 4. Description of standard deviation and residual for centroid

Lesion/non-lesion	Standard deviation (CX)	Residual (CX)	Standard deviation (CY)	Residual (CY)
L1	0.352	0.287	0.193	0.188
L2	0.202	0.488	0.312	0.270
L3	0.064	0.128	0.165	0.201
L4	0.212	0.508	0.442	1.021
L5	0.252	0.505	0.260	0.506
L6	0.072	0.116	0.331	0.072
NL1	1.692	5.025	1.537	2.002
NL2	1.223	3.289	0.967	1.884
NL3	1.640	1.932	1.581	1.219
NL4	1.752	6.780	1.656	3.384
NL5	3.728	3.891	5.053	3.113
NL6	7.072	3.208	4.69	3.429



(a)

(b)

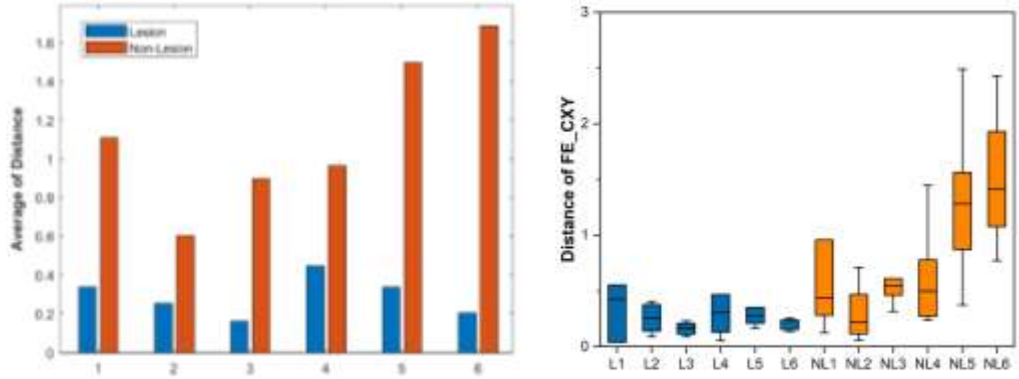
Figure 8. Graph of (a) standard deviation of centroid and (b) residual for centroid

In Table 5, the average distances for both lesion and non-lesion samples are outlined. According to the table, lesions exhibit smaller average distances, falling within the range of 0.162 to 0.446. Conversely, non-lesions display larger average distances, ranging from 0.603 to 1.686. The Figure 9 represent average and distances distribution for six lesions and six non-lesions. Figure 9(a) serves as a visual representation of the data presented in Table 5. This graph specifically illustrates the average distance of centroids. Lesions are represented by the blue bars, while non-lesions are depicted by the red bars. Upon examination, it's evident that all lesions exhibit smaller values for average distance in compared to non-lesions. The multiple box-plot diagram depicted in Figure 9(b) offers a comparative view of the distance distribution for each lesion (L1, L2, L3, L4, and L6) and non-lesion (NL1, NL2, NL3, NL4, NL5, and NL6). This graphical representation allows for a rapid assessment of how distances were distributed among lesions and non-lesions. The graph clearly illustrates that the majority of non-lesion (orange) boxplots are of larger size in comparison to the lesion (blue) boxplots. This discrepancy arises from differences in distance values and the respective counts of lesions and non-lesions.

Table 5. Description of average distance for centroid

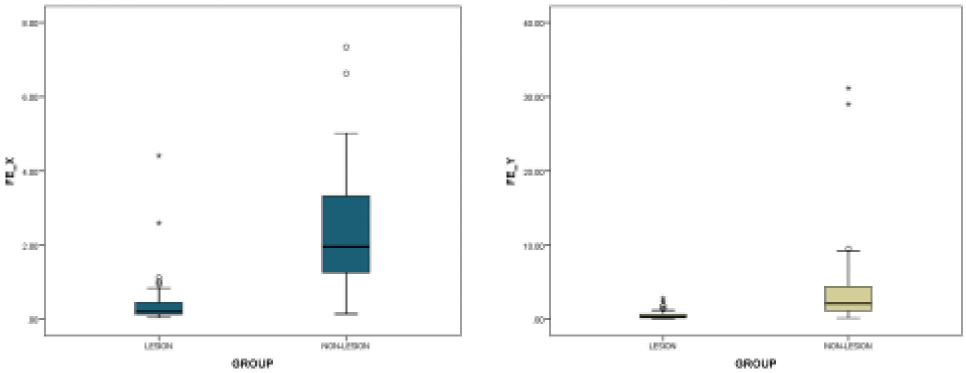
Average distance											
L1	L2	L3	L4	L5	L6	NL1	NL2	NL3	NL4	NL5	NL6
0.340	0.253	0.162	0.446	0.337	0.203	1.109	0.603	0.898	0.964	1.499	1.686

Figure 10 displays boxplots showing the standard deviation of FE_CXY for both lesions and non-lesions included in the study. Figure 10(a) shows FE_CX, while Figure 10(b) shows FE_CY. In both graphs, the boxplot for lesions appears smaller compared to non-lesions. This is because all lesions had relatively small standard deviation values. Comparison of the results from Figures 8-10 reveals no significant differences when considering only 12 lesions and non-lesions versus all lesions and non-lesions in the study. The patterns observed in Figures 8-10 may explain the varying trends between lesions and non-lesions seen in Figures 6 and 7. Specifically, the centroid data for lesions exhibited a more consistent trend compared to non-lesions, likely due to the greater dispersion of FE_CXY values for non-lesions, characterized by larger standard deviation, residual, distance, and average distance values compared to lesions, which had smaller standard deviation values.



(a) (b)

Figure 9. Graph of (a) average distance and (b) distance of FE_CXY of lesions and non-lesions



(a) (b)

Figure 10. Boxplot for (a) FE_CX and (b) FE_CY of lung lesion and non-lesion

To further investigates the performance of FE_CXY, we conducted a student’s t-test to determine if there are statistically significant differences between lesions and non-lesions. Additionally, we trained and tested two different sets of features: one without FE_CXY and one with FE_CXY, using a CNN model. The selection of features to pair with FE_CXY was based on a literature review and the opinion of radiologist regarding lung lesion diagnosis. The results revealed a p-value of 0.000 for both FE_CX and FE_CY, indicating statistically significant differences between lesions and non-lesions, with p-values below 0.05. Regarding the CNN model's performance, Table 6 illustrates an improvement of 1.33% in accuracy and 2.25% in F1-score when FE_CXY is included in the feature set, achieving 92% accuracy and 83.33% F1-score, compared to 90.67% accuracy and 81.08% F1-score without FE_CXY.

FE_CXY emerges as a potent feature, as underscored by the student's t-test results which reveal substantial distinctions in trends. By integrating such effective features, it becomes possible to construct robust models characterized by high performance and efficiency in pinpointing lung lesions. The accuracy presented in Table 7 reveals that our proposed features achieved convincing results that outperformed all compared features with 92% accuracy. While extensive research has delved into a multitude of techniques for feature extraction, incorporating an excessive number of features can overly complicate classifier models. Hence, it is imperative to selectively include only those features that contribute meaningfully. There have been studies where good classification rates were achieved using only two features, as indicates in [14], [32], [33].

Table 6. Performance evaluation between proposed features other features

Model	Accuracy	F1-score
Diameter+roundness	90.67%	81.08%
Diameter+roundness+FE_CXY	92.00%	83.33%

Table 7. Model accuracy for proposed features with similar investigations

Reference	Features	Accuracy
[20]	Super pixels features	83.40%
[15]	Geometrical features and statistical features	84.00%
[34]	Tumour shape and boundary features	89.80%
Our study	Diameter+roundness+FE_CXY	92.00%

4. CONCLUSION

In conclusion, this approach utilises the centroid feature (FE_CXY) in detecting lesion localisation, aided by statistical analysis to uncover hidden information. The statistical findings highlight the potential of FE_CXY as an alternative feature for lesion detection system. By understanding the distribution of centroid data within lung CT scan images, it becomes possible to extract powerful features that provide valuable insights of the lesion. Future research endeavours will focus on implementation of FE_CXY as effective features to simplify machine learning and deep learning models. These models can serve as valuable second opinions, assisting medical doctors and physicians in accurately interpreting medical modalities. Moreover, they have the potential to enhance screening, diagnosis, and prognosis processes by providing uninterrupted and faster analysis.

ACKNOWLEDGEMENTS

This research work is financially supported by Ministry of Higher Education Grant Scheme (FRGS) "A new motion-based optical flow lung cancer detection method for lung nodule segmentation on lung CT-scan images" (Ref: FRGS/1/2021/TK0/UITM/02/58). Ethics approval is obtained from Universiti Sains Malaysia (USM/JEPeM/21110721). The authors would like to express their gratitude to members of the Advanced Control System and Computing Research Group (ACSCRG), Advanced Rehabilitation Engineering in Diagnostic and Monitoring Research Group (AREDiM), Integrative Pharmacogenomics Institute (iPROMISE), and Centre for Electrical Engineering Studies, Universiti Teknologi MARA, Cawangan Pulau Pinang for their assistance and guidance during the fieldwork. Finally, the authors are grateful to Universiti Teknologi MARA, Cawangan Pulau Pinang for their immense administrative and financial support.

REFERENCES

[1] K. Chaitanya Thandra, A. Barsouk, K. Saginala, J. S. Aluru, and A. Barsouk, “Epidemiology of lung cancer,” *Współczesna Onkologia*, vol. 25, no. 1, pp. 45–52, 2021, doi: 10.5114/wo.2021.103829.

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, “Cancer statistics, 2022,” *CA: A Cancer Journal for Clinicians*, vol. 72, no. 1, pp. 7–33, Jan. 2022, doi: 10.3322/caac.21708.

[3] F. Siddiqui, S. Vaqar, and A. H. Siddiqui, *Lung cancer*. Treasure Island (FL): StatPearls Publishing, 2023.

[4] I. Haq *et al.*, “Lung nodules localization and report analysis from computerized tomography (CT) scan using a novel machine learning approach,” *Applied Sciences*, vol. 12, no. 24, p. 12614, Dec. 2022, doi: 10.3390/app122412614.

[5] P. Kuśnierczyk, “Genetic differences between smokers and never-smokers with lung cancer,” *Frontiers in Immunology*, vol. 14, Feb. 2023, doi: 10.3389/fimmu.2023.1063716.

[6] S. Suresh and S. Mohan, “NROI based feature learning for automated tumor stage classification of pulmonary lung nodules using deep convolutional neural networks,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1706–1717, May 2022, doi: 10.1016/j.jksuci.2019.11.013.

[7] H. Elwahsh, M. A. Tawfeek, A. A. Abd El-Aziz, M. A. Mahmood, M. Alsabaan, and E. El-shafeiy, “A new approach for cancer prediction based on deep neural learning,” *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 6, p. 101565, Jun. 2023, doi: 10.1016/j.jksuci.2023.101565.

[8] A. A. Alsheikhy, Y. Said, T. Shawly, A. K. Alzahrani, and H. Lahza, “A CAD system for lung cancer detection using hybrid deep learning techniques,” *Diagnostics*, vol. 13, no. 6, p. 1174, Mar. 2023, doi: 10.3390/diagnostics13061174.

[9] B. Ait Skourt, A. El Hassani, and A. Majda, “Lung CT image segmentation using deep neural networks,” *Procedia Computer Science*, vol. 127, pp. 109–113, 2018, doi: 10.1016/j.procs.2018.01.104.

[10] A. Patil, V. R. Udupi, C. D. Kane, A. I. Wasif, J. V. Desai and A. N. Jadhav, “Geometrical and texture features estimation of lung cancer and TB images using chest X-ray database,” *2009 International Conference on Biomedical and Pharmaceutical Engineering*, Singapore, 2009, pp. 1-7, doi: 10.1109/ICBPE.2009.5384113.

[11] F. H. Kuwil, “A new feature extraction approach of medical image based on data distribution skew,” *Neuroscience Informatics*, vol. 2, no. 3, p. 100097, Sep. 2022, doi: 10.1016/j.neuri.2022.100097.

[12] Z. Yang, L. Zhao, S. Wu, and C. Y.-C. Chen, “Lung lesion localization of COVID-19 from chest CT image: a novel weakly supervised learning method,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 1864–1872, Jun. 2021, doi: 10.1109/JBHI.2021.3067465.

[13] A. Asuntha and A. Srinivasan, “Deep learning for lung cancer detection and classification,” *Multimedia Tools and Applications*, vol. 79, no. 11–12, pp. 7731–7762, Mar. 2020, doi: 10.1007/s11042-019-08394-3.

[14] H. M. Orozco, O. O. V. Villegas, V. G. C. Sánchez, H. de J. O. Domínguez, and M. de J. N. Alfaro, “Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine,” *BioMedical Engineering OnLine*, vol. 14, no. 1, p. 9, Dec. 2015, doi: 10.1186/s12938-015-0003-y.

[15] T. Aggarwal, A. Furqan, and K. Kalra, “Feature extraction and LDA based classification of lung nodules in chest CT scan images,” in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, Aug. 2015, pp. 1189–1193. doi: 10.1109/ICACCI.2015.7275773.

[16] L. Lu *et al.*, “Identifying robust radiomics features for lung cancer by using In-Vivo and phantom lung lesions,” *Tomography*, vol. 7, no. 1, pp. 55–64, Feb. 2021, doi: 10.3390/tomography7010005.

[17] A. E. Sebastian and D. Dua, “Lung nodule detection via optimized convolutional neural network: impact of improved moth flame algorithm,” *Sensing and Imaging*, vol. 24, no. 1, p. 11, Mar. 2023, doi: 10.1007/s11220-022-00406-1.

[18] N. Maleki and S. T. A. Niaki, “An intelligent algorithm for lung cancer diagnosis using extracted features from computerized tomography images,” *Healthcare Analytics*, vol. 3, p. 100150, Nov. 2023, doi: 10.1016/j.health.2023.100150.

[19] J. P. Appadurai, S. G, B. Prabhu Kavin, K. C, and W.-C. Lai, “Multi-process remora enhanced hyperparameters of convolutional neural network for lung cancer prediction,” *Biomedicines*, vol. 11, no. 3, p. 679, Feb. 2023, doi: 10.3390/biomedicines11030679.

[20] D. D. Althubaity *et al.*, “Automated lung cancer segmentation in tissue micro array analysis histopathological images using a prototype of computer-assisted diagnosis,” *Journal of Personalized Medicine*, vol. 13, no. 3, p. 388, Feb. 2023, doi: 10.3390/jpm13030388.

[21] S. Wang *et al.*, “Artificial intelligence in lung cancer pathology image analysis,” *Cancers*, vol. 11, no. 11, p. 1673, Oct. 2019, doi: 10.3390/cancers11111673.

[22] J. John and M. G. Mini, “Multilevel thresholding based segmentation and feature extraction for pulmonary nodule detection,” *Procedia Technology*, vol. 24, pp. 957–963, 2016, doi: 10.1016/j.protcy.2016.05.209.

[23] M. F. Abdullah, S. N. Sulaiman, M. K. Osman, N. K. A. Karim, S. Setumin, and I. S. Isa, “A new procedure for lung region segmentation from computed tomography images,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, p. 4978, Oct. 2022, doi: 10.11591/ijece.v12i5.pp4978-4987.

[24] S. Aja-Fernández, A. H. Curiale, and G. Vegas-Sánchez-Ferrero, “A local fuzzy thresholding methodology for multiregion image segmentation,” *Knowledge-Based Systems*, vol. 83, pp. 1–12, Jul. 2015, doi: 10.1016/j.knosys.2015.02.029.

[25] B. Preim and C. Botha, “Image analysis for medical visualization,” in *Visual Computing for Medicine*, Elsevier, 2014, pp. 111–175. doi: 10.1016/B978-0-12-415873-3.00004-3.

[26] N. N. Jafery *et al.*, “Application of geometric features on lung lesion and non-lesion segmentation,” in *2023 IEEE 13th International Conference on Control System, Computing and Engineering (ICCSCE)*, IEEE, Aug. 2023, pp. 167–172. doi: 10.1109/ICCSCE58721.2023.10237160.

[27] N. F. Razali, I. S. Isa, S. N. Sulaiman, N. K. A. Karim, and M. K. Osman, “Improvement of breast density classifier based on cnn features extraction and SVM in mammogram images,” *Journal of Electrical & Electronic Systems Research*, vol. 21, no. OCT2022, pp. 63–72, Nov. 2022, doi: 10.24191/jeesr.v21i1.009.

[28] A. Srdanov, D. Stojiljkovic, and A. Lazic, “Euler line and the nine-point circle,” *International Journal of Mathematics Trends and Technology*, vol. 67, no. 9, pp. 72–80, Sep. 2021, doi: 10.14445/22315373/IJMTT-V67I9P508.

[29] Z. Xiao, B. Liu, L. Geng, F. Zhang, and Y. Liu, “Segmentation of lung nodules using improved 3D-UNet neural network,” *Symmetry*, vol. 12, no. 11, p. 1787, Oct. 2020, doi: 10.3390/sym12111787.

[30] D. Liu, L. Zhang, X. Lai, and H. Liu, “Image feature selection embedded distribution differences between classes for convolutional neural network,” *Applied Soft Computing*, vol. 131, p. 109715, Dec. 2022, doi: 10.1016/j.asoc.2022.109715.

[31] N. Khehrah, M. S. Farid, S. Bilal, and M. H. Khan, “Lung nodule detection in CT images using statistical and shape-based features,” *Journal of Imaging*, vol. 6, no. 2, p. 6, Feb. 2020, doi: 10.3390/jimaging6020006.

[32] F. Löw, U. Michel, S. Dech, and C. Conrad, “Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using support vector machines,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 85, pp. 102–119, Nov. 2013, doi: 10.1016/j.isprsjprs.2013.08.007.

[33] L. Boroczky, L. Zhao, and K. P. Lee, “Feature subset selection for improving the performance of false positive reduction in lung nodule CAD,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 3, 2006, doi: 10.1109/TITB.2006.872063.

[34] S. Wang *et al.*, “Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome,” *Scientific Reports*, vol. 8, no. 1, p. 10393, Jul. 2018, doi: 10.1038/s41598-018-27707-4.

BIOGRAPHIES OF AUTHORS

Nurul Najiha Jafery received the B.Sc. degree (Hons.) in Statistics from Faculty of Computer and Mathematics Sciences, Universiti Teknologi MARA (UiTM), Kelantan Campus, Malaysia in 2018 and the M.Sc. degree in Data Sciences and Analytic from School of Computer Sciences, Universiti Sains Malaysia (USM), Penang in 2021. She is currently doing her Ph.D. in Electrical Engineering, focusing application of image processing and deep learning on medical images. Her research interest includes machine learning, deep learning, and image processing. She can be contacted at email: jihajafery@gmail.com.

Siti Noraini Sulaiman received the B.Eng. degree (Hons.) in electrical and electronics and the M.Sc. degree in electrical and electronics engineering (medical imaging) from Universiti Sains Malaysia, in 2000 and 2003, respectively, and Ph.D. degree in imaging from School of Electrical and Electronics Engineering, Universiti Sains Malaysia in 2012. She is currently an Associate Professor at the Electrical Engineering Studies, and a deputy chair of Advanced Rehabilitation Engineering in Diagnostic and Monitoring (AREDiM) and the deputy chair of Advanced Control System and Computing Research Group (ACSCRG), Electrical Engineering Studies, College of Engineering, UiTM. She has published numerous research articles in international journals and conference proceedings. Her research interests include intelligent systems, image processing, neural networks for medical applications and algorithms. She can be contacted at email: sitinoraini@uitm.edu.my.

Muhammad Khusairi Osman graduated from Universiti Sains Malaysia with a B. Eng Degree in Electrical and Electronic Engineering in 2000 and M.Sc. in Electrical and Electronic Engineering in 2004. In 2014, he received his Ph.D. in medical electronic engineering from Universiti Malaysia Perlis (UniMAP), Malaysia. He is currently a senior lecturer at the Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Malaysia. Image processing, pattern recognition, and artificial intelligence are among his research interests. He can be contacted at email: khusairi@uitm.edu.my.

Noor Khairiah A. Karim received her bachelor's degree in medicine, bachelor's degree in surgery (MBBS), and Master of Radiology (MRad) from the University of Malaya, Malaysia. She then obtained her Fellowship in Cardiac Imaging from the National Heart Center Singapore. She is currently a Senior Medical Lecturer of the Regenerative Medicine Cluster, and a Consultant Radiologist at the Advanced Medical and Dental Institute, Universiti Sains Malaysia. Her current research areas include medical image processing and analysis with a special interest in cardiac, breast and brain imaging. She can be contacted at email: drkhairiah@usm.my.

Mohd Firdaus Abdullah received his MSc Degree in Science (Electrical Engineering) from the Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia in 2012. In 2023, he finishes his PhD at Universiti Teknologi MARA, Malaysia. His areas of research focus include the image processing of medical imaging, specifically in analyzing CT scan image. He can be contacted at email: f.abdullah@uitm.edu.my.

Iza Sazanita Isa received her bachelor's in electrical engineering from the Universiti Teknologi MARA, Malaysia in 2004 and the M.Sc. degree from the Universiti Sains Malaysia, Malaysia in 2008. Since 2009, she joined Universiti Teknologi MARA, Penang Campus, Malaysia as young lecturer and promoted as senior lecturer at the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA (UiTM), Penang branch, in 2013. She pursues her PhD in Electrical Engineering under the SLAB/SLAI scholarship and graduated in 2018 and she finished her postdoctoral fellowship under School of Computer Sciences, Universiti Sains Malaysia (USM). She is attached to the department of Control System Engineering at the faculty, a member of AREDiM research group, the head of research group RIDyLT and actively involved in teaching. Her research interest includes the model development using image processing and artificial intelligence. She can be contacted at email: izasazanita@uitm.edu.my.

Zainal Hisham Che Soh graduated from Leed University United Kingdom with a B. Eng. (Honors) Degree in Electronic Engineering in 1997. He received an M.Sc. in Computer Science in 2004 at Universiti Teknologi Malaysia (UTM) Kuala Lumpur. In 2013, he received his Ph.D. in Electrical and Electronic Engineering from Universiti Sains Malaysia (USM), Malaysia. He is currently an associate profesor at the Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Malaysia. Image processing, ARM embedded system, intenet of things (IoT), data science and big data analytics are among his research interests. He can be contacted at email: zainal872@uitm.edu.my.