IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 24, 2021, accepted November 4, 2021, date of publication November 19, 2021,
date of current version December 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129597

Comprehensive and Comparative Global and Local Feature Extraction Framework for Lung Cancer Detection Using CT Scan Images

MOHAMMAD A. ALZUBAIDI, MWAFFAQ OTOOM, (Senior Member, IEEE), AND HAMZA JARADAT
Department of Computer Engineering, Yarmouk University, Irbid 21163, Jordan
Corresponding author: Mohammad A. Alzubaidi (maalzubaidi@yu.edu.jo)

ABSTRACT Lung cancer is reported to be the second most common cancer disease. This paper proposes a comprehensive and comparative global and local feature extraction framework for lung cancer detection using CT scan images. This framework consists of three main phases: data collection, global training and testing, and local training and testing. A set of 1000 CT scan images is used in this study. During the global training and testing phase, the collected images are preprocessed through image warping and cropping. Global features are then extracted from images to represent each image with feature vectors, using ten different image feature types. The feature vectors are then used to build detection models with six different machine learning algorithms. In the local training and testing phase, each image is divided into a set of local blocks. Those feature types that performed well in the global phase are then extracted from each of these blocks, to represent each block with feature vectors. These feature vectors are then used to build detection models for all of the image blocks, using the learning algorithms that performed well in the global phase. The results show that the Gabor Filter, the Histogram of Oriented Gradients (HOG), and the Haar Wavelet feature types outperformed the other seven feature types. The results also show that Support Vector Machine (SVM) outperforms the other five learning algorithms. Of most importance, the proposed local feature extraction approach outperforms the traditional global one. In the local phase, using SVM with Haar Wavelet features achieved 90% accuracy, 88% sensitivity, and 91% specificity. Using SVM with HOG features achieved 88% accuracy, 85% sensitivity, and 89% specificity. Finally, using SVM with Gabor Filter features achieved the best accuracy, sensitivity, and specificity rates of 97%, 96%, and 97%, respectively.

INDEX TERMS CT scan, lung cancer, global feature extraction, local feature extraction, SVM, Gabor filter.

I. INTRODUCTION

Cancer is one of the most common and dangerous human diseases. It involves the growth and the spread of abnormal cells within the human body. It can be successfully treated if discovered and diagnosed in its early stages [1]. In the normal biological process, older cells that become damaged are replaced by new cells. Cancer occurs when this process breaks down, and damaged cells are not replaced. These abnormal cells can metastasize into other organs of the body [2].

Lung cancer is one type of cancer that starts in the lung tissue [5]. It is reported to be the second most common cancer [3]. Statistics for the 2011-2015 period showed that, on average, 439.2 per 100,000 people were diagnosed with cancer each year in the USA, and 163.5 per 100,000 people died each year from that cancer [3]. In 2021, about 235,760 new cases and 131,880 deaths are expected in USA. In the UK, about 44,500 people are diagnosed with lung cancer every year [7]. In 2008, 1.37 million deaths were caused by lung cancer [8], and in 2012, 1.6 million deaths [9], with an increased rate of about 17% between 2008 and 2012.

Unfortunately, lung cancer often does not show any symptoms in its early stages. Symptoms appear in the later stages of the disease [5], [7]. In the USA, only 17.4% of people survive for 5 years after diagnosis without treatment, and the percentage is lower in developing countries [3]. However, early detection and diagnosis of lung cancer leads to

The associate editor coordinating the review of this manuscript and approving it for publication was Yin Zhang.

quicker recovery, makes the treatment less complex and less expensive and (most importantly) increases the recovery rate from the disease. In the USA, early diagnosis and treatment can increase a patient's 5-year survival time from 15% to 65-80% [6].

Several methods are available to diagnose and detect lung cancer, including blood tests, radiology tests, endoscopy procedures and biopsies. Each type of test has some advantages, disadvantages, and some special applications. CT (Computed Tomography) scanning can provide a fast test result without pain, and it provides information about the tumor shape, size, and location [4]. A CT scan is a 3-D image of the inside of the body, produced by an x-ray machine that takes multiple images of the same anatomical location from different angles [10]. In addition, a CT scan helps to evaluate intrathoracic pathological conditions [11]. To detect lung cancer, specialists typically perform a CT scan with a contrast enhancing medium injected into the blood. This shows the details in the lung more clearly [7]. Such a CT scan provides detailed images of the patient's chest, to allow for better detection of lung cancer. [12].

Computer Aided Diagnosis (CAD) systems help a radiologist detect cancer in the images. These systems use image processing and machine learning techniques to detect suspicious regions in radiograph images. Detection of these regions helps doctors and specialists, who provide the final interpretation of the images [13].

Researchers have proposed many different CAD systems to detect and classify lung cancer within CT scan images [61]. A survey of these systems is provided in the next section. These systems use several types of image features and machine learning algorithms in an attempt to provide accurate detection and classification. This raises the question of what combination of image feature types and machine learning methods works best.

This paper proposes a comprehensive and comparative framework for evaluating and comparing computerized global and local feature extraction methods for lung cancer detection in CT scan images. This framework compares the effectiveness of ten different types of extracted image features. It also compares six well-known machine learning algorithms: Support Vector Machines (SVM), Neural Networks, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Naïve Bayes. The purpose of building such a framework is to determine the type of feature extractor, and the learning method that provide the most accurate detection system.

The rest of this paper is organized as follows. Section II provides some background information and surveys the literature and related work. A theoretical framework for the proposed research is presented in Section III. Section IV presents the methodology of the proposed framework, providing a step-by-step procedure that highlights the main contributions of this work. Section V presents and discusses the results. Finally, Section VI concludes the paper and presents some future work directions.

II. BACKGROUND AND LITERATURE REVIEW

In this section, some background and information about lung cancer is presented. Then, some related research studies in lung cancer detection are briefly reviewed.

A. LUNG CANCER

Metabolic processes in the body use oxygen, which is acquired through gas exchange between the air inside the lungs and the blood. The lungs are spongy organs that take in oxygen from outside. Because they are the first line of defense, they are vulnerable to diseases, including lung cancer [5].

Lung cancer starts when cells in the lungs grow abnormally. It can start anywhere in the lungs, and it can spread to other parts of the body. Smoking is one the main causes of lung cancer, and it is the leading cause of cancer deaths [5].

Unfortunately, lung cancer often does not show any symptoms in the early stages. Symptoms appear in the late stages of the disease. These symptoms include coughing (especially continuous coughing, and coughing with blood), shortness of breath, chest pain, bone pain, and headache [5], [7].

There are two main types of lung cancers: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC) [5].

SCLC is less common and appears among heavy smokers. It is also known as *oat cell cancer*. This type spread faster than NSCLC, and thus, it is treated using chemotherapy and radiation therapy [5].

NSCLC is more common and includes several types of lung cancer such as squamous cell carcinoma, adenocarcinoma, and large cell carcinoma. These subtypes are grouped together because they are usually use the same treatment procedure [5].

Doctors can distinguish these various types of lung cancer through a microscope. Based on that diagnosis, they can select an appropriate treatment [5].

Annual lung cancer screening (using low-dose CT scans) is highly recommended for those with higher risk, such as smokers and elderly people [5]. Additional tests are warranted if the patient has symptoms. These tests include a breathing test using spirometer device, a blood test, and a chest X-ray. These tests help determine if the symptoms are from other possible diseases, such as a chest infection. A chest X-ray can also detect the cancer as a white-grey mass [7].

Once the patient is diagnosed with cancer, a CT (Computed Tomography) scan is employed to get a clear image of the inside of the body, and the location and size of the tumor. This CT scan is a 3-D image of the inside of the body, taken with an X-ray machine that takes multiple images, from different angles, of the same anatomical location [10]. In addition, a CT scan can reveal other intrathoracic pathological conditions [11]. To detect lung cancer, a specialist uses a CT scan with a contrast enhancing medium injected into the blood, which shows the details of the lung more clearly [7]. The

CT scan images the patient's chest, allowing for detection of cancerous tissues [12].

More background information about lung cancer can be found in [5] and [7].

To sum up, early detection and diagnosis of lung cancer leads to quicker recovery, makes the treatment less complex and less expensive and (most importantly) increases the recovery rate from the disease. Thus, the focus of this work is to use CT scan images, along with image processing and machine learning techniques, to provide earlier detection of lung cancer.

B. LUNG CANCER DETECTION USING CT SCAN IMAGES

To detect lung cancer, researchers have implemented many different techniques to process the images, extract features and apply machine learning algorithms.

For example, Jin *et al*. [37] proposed a CAD system that used image segmentation to extract the regions of interest from CT scan images, and then used a convolution neural network as a classifier, to detect pulmonary nodules. Their system achieved an accuracy of 85%.

In [54], the authors developed a CAD system to detect lung nodules using a genetic algorithm. First, they segmented the CT images to detect the regions of interest (ROIs) based on the density values of pixels within the image. Then they employed various thresholds to scan the pixels in all directions. After reducing the number of ROIs (based on upper and lower slices) they used a genetic algorithm to classify each nodule. A sensitivity of 93.4% on 276 CT scan images was achieved.

Shafiei and Fekri-Ershad [55] used textural features and morphological operations to detect lung cancer in CT scan images. They used the super pixel algorithm to cluster the images, followed by morphological operations. Then, they used the active contour algorithm to identify the tumors in the images. Their system achieved a Dice Similarity of 84.88%.

In [39], the authors proposed a new methodology to detect nodules in lungs. They relied on 128 features, based on intensity, shape, texture, and context features. They achieved a sensitivity of only 80%. They compared the performance of SVM against the K-Nearest Neighbor classifier, and the Nearest Mean classifier. Based on their work, SVM outperformed the other methods.

Gonzalez and Ponomaryvo [40] proposed a CAD system to classify lung cancers into benign or malignant. The proposed system included four steps: (1) preprocessing, (2) lung segmentation, (3) nodule detection, and (4) classification. In the preprocessing step, they calculated several masks, using thresholding techniques and morphological operations. To determine the Region of Interest (ROI), they used priori information and the Hounsfield Unit (HU) scale, which uses area, eccentricity, circularity, and fractal dimension as features. For classification, the system uses the SVM algorithm. They reported an accuracy of 78.08%.

In [41], the authors developed a new CAD system. They randomly selected 420 cases from LIDC-IDRI database. The system used the Watershed technique to detect possible nodules and to distinguish them from other structures. The Histogram of Oriented Gradients (HOG) technique was used for feature extraction. To reduce false positives, it used a rule-based classifier and SVM. Using 10-fold cross validation, they achieved a 93.9% sensitivity.

Silva *et al*. [22] developed a CAD system to detect lung nodules. SVM was used as a classification method, which was applied to 33 exams. Their proposed methodology achieved an accuracy of 95.21%.

The authors in [42] used the Random Subspace Method (RSM) to build a CAD system. They used a two-step supervised learning system, which employed RSM to detect pulmonary nodules in lungs. From a database of 125 samples, they extracted 216 features, and built a classifier based on RSM, and genetic-algorithm-based feature selection. Their proposed system achieved an accuracy of 88.9%.

Mean clustering was used in [43] to detect lung cancer. The authors used EK-Mean clustering to detect and classify lung tumors. First, they removed the noise using a median filter. Then, they used a K-means algorithm for clustering and segmentation. They then used a gray-level co-occurrence matrix (GLCM) method to extract features, such entropy, correlation, homogeneity, PSNR, and SSIM. Their system achieved an accuracy of 90.7%.

Aggarwal *et al*. [38] used Linear Discriminate Analysis (LDA) to classify nodules in lungs, and to distinguish them from normal anatomy. They used thresholding and gray-level characteristics for segmentation. An accuracy of only 84% was obtained using this system.

Marker-controlled watershed segmentation was used in [44] to detect lung cancer from CT scan images. To enhance the image quality, it used Gabor filters as a preprocessing step. A 90% accuracy was achieved in their system.

A thresholding algorithm was used as a segmentation method in [45]. The authors used a 3-step process to detect lung nodules. First, they used a thresholding algorithm to segment the lung region in CT data. Second, they removed the lung vessels by using an active contour model (ACM). Then, they detected the nodules using a selective shape filter. Finally, they used a classifier to distinguish true or false positive nodules, depending on features. This system had an 85% detection rate.

Liu *et al*. [56] used a thresholding and region growing algorithm. They used pulmonary parenchyma for segmentation and a circle shape descriptor for ROI extraction. The system had 85.6% sensitivity and a 13.4% false positive rate.

Wook-Jin Choi and Tae-Sun Choi [47] developed a CAD system to detect solitary pulmonary nodules. They used optimal thresholding and neighborhood for segmentation. To detect the nodule, they used multi-scale dot enhancement filtering, and angular histograms. SVM was used for classification. The system produced a 97.5% sensitivity. However, no other performance measures were reported.

A deep learning approach was used in [48] to develop a CAD system. The authors evaluated a deep belief network (DBN), a convolutional neural network (CNN) and a scale invariant feature transform (SIFT). The specificity of the three methods were 82.2%, 78.7% and 66.8% respectively.

Shenglin Ma MD *et al*. [49] used data from 844 lung cancer patients in their work. Four serum proteins were found with high concentrations in the patients, compared to the normal health conditions. This work obtained a 98.25% specificity.

C. OTHER CANCER DETECTION USING CT SCAN IMAGES

In [46], the authors proposed a novel Multiple Instance Learning (MIL) algorithm to detect gastric cancer with their CAD system. Bag-level features (which look at characteristics of the whole image) and instance-level features (which look at the intensity and texture characteristics within the gastric wall) were extracted as two-level features. The accuracy in the system was 76.9%.

In [50], the authors used a decision tree technique to detect breast cancer with their CAD system. In their system, they used 24 features to discriminate malignant breast-cancers. Their accuracy was 95.50%.

D. SUMMARY

Several CAD systems have been proposed in the literature for cancer detection and classification within CT scan images. Many different image feature types, as well as learning algorithms, have been used. This raises the question of what combination of feature types and learning methods works best.

Moreover, CT scan images (like other medical images) are special types of images, which are interpreted and diagnosed by radiologists based on anatomical regions within the images. Normal content in some anatomical regions might be abnormal in other regions, and vice versa. This suggests that building local models for each anatomical region within the images might be useful. However, there has been no focus within the literature on anatomically-based local feature extraction and model learning.

This leads to our research question:

What image features and learning methods would be most useful for building localized learning models for lung cancer detection within CT scan images?

III. THEORITICAL FRAMEWORK

In this section, the key concepts and terms of this work are identified and explained.

A. CT SCAN IMAGES

There are several types of medical images. CT scan images are commonly used in lung cancer detection. To build a successful computer-aided cancer detection system, a balanced data set that contains both normal and abnormal images is needed. This allows the machine to learn the borders between normal images and abnormal ones.

B. IMAGE WARPING

Each CT scan image is captured by a particular type of CT scanner. This produced a variety of different sizes and orientations - each referenced to its own scanner's x-y coordinate system. For consistency, there is a need to align the results of all the CT scans with a common reference anatomical coordinate system, which replaced the original x-y coordinate system unique to each CT scanner. The purpose of image warping is to align the anatomical features in each image with a common reference anatomical coordinate system.

C. IMAGE CROPPING

Since each CT scan image is captured differently, each image could have some unwanted regions, such as regions that fall outside the body. The purpose of image cropping is to remove all such unwanted regions in images.

D. FEATURE EXTRACTORS

Several feature extraction methods are proposed in the literature [57], [58]. Candidates include intensity histogram, histogram of oriented gradients, Gabor filter, entropy filter, grayscale contrast, grayscale correlation grayscale energy, grayscale homogeneity, standard deviation, and Haar wavelet.

1) INTENSITY HISTOGRAM

This feature shows how often each pixel intensity value is repeated within the image [30]. Thus, this feature depends on the number of pixel intensity repetitions, without taking into account the pixel locations within the image.

2) HISTOGRAM OF ORIENTED GRADIENTS (HOG)

This feature is computed as follows. The image is divided into cells. Then the orientation of the gradient within each cell is computed. Finally, a count is taken of each gradient orientation, and a histogram is created from those counts. [31].

3) GABOR FILTER

The image's local spatial frequency content (i.e. texture) is analyzed using a Gaussian window with a Gabor filter. [32].

4) ENTROPY FILTER

This feature is a statistical measure of the randomness of the pixel values within an image. It is given by Equation (1) [33].

$$Entropy = -\sum \left(p\log_2(p)\right) \tag{1}$$

where:

p is the normalized histogram counts for the image.

5) GRAYSCALE FEATURES

Each of these features describe the image with a single number. These parameters can be calculated using Equations (2)-(5).

$$Contrast = \sum_i^{N_g} \sum_j^{N_g} (i-j)^2 p(i,j) \tag{2}$$

gives the intensity difference between a pixel and its neighborhood. [33]

$$Correlation = \sum\sum \frac{p\,(i,j) - \mu_{row}\mu_{col}}{\mu_{row}\mu_{col}} \tag{3}$$

measures how a pixel is correlated to its neighborhood pixels. [33]

$$Energy = \sum_{i,j} p\,(i,j)^2 \tag{4}$$

provides a measure of the uniformity of pixels. [33]

$$Homogeneity = \sum_{i,j} \frac{p\,(i,j)}{1 + |i+j|} \tag{5}$$

measures the closeness of the elements within the gray level co-occurrence matrix to its diagonal. [33]

where:
$p\,(i,j)$ is the value of the pixel in row i and column j.
μ_{row} is the mean of data across a row.
μ_{col} is the mean of data across a column.
N is the total number of the pixels.

6) STANDARD DEVIATION

This feature is used in statistics to represent the variation or dispersion of the data [34]. It is given by Equation (6).

$$Standard\ Deviation = \sqrt{\frac{\sum_{i,j} (p\,(i,j) - \mu)^2}{N-1}} \tag{6}$$

where:
$p\,(i,j)$ is the value of the pixel in row i and column j.
μ is the mean of a set of data.
N is the total number of the pixels.

7) HAAR WAVELET

The 1D discrete signal of Haar Wavelet is defined using Equation (7) below [35]:

$$\psi\,(x) = \begin{cases} 1, & for\ 0 \le x < \frac{1}{2} \\ -1, & for\ \frac{1}{2} \le x < 1 \\ 0, & otherwise \end{cases} \tag{7}$$

When applying the transform to 2D images, Equation (8) is used [36].

$$\boldsymbol{\psi}^{\boldsymbol{i}}_{\boldsymbol{j},\boldsymbol{m},\boldsymbol{n}}\,(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{2}^{\frac{\boldsymbol{j}}{\boldsymbol{2}}}\boldsymbol{\psi}^{\boldsymbol{i}}\left(\boldsymbol{2}^{\boldsymbol{j}}_{\boldsymbol{x}-\boldsymbol{m}}, \boldsymbol{2}^{\boldsymbol{j}}_{\boldsymbol{y}-\boldsymbol{n}}\right), \quad \boldsymbol{i} = \{\boldsymbol{H}, \boldsymbol{V}, \boldsymbol{D}\} \tag{8}$$

where **H** is the Horizontal components, **V** is the vertical components and **D** is the Diagonal components.

E. LEARNING METHODS

Once image features are extracted, some method must be found to learn the normal and abnormal regions in the images. Several machine learning algorithms are available in the literature. Well-known methods include Support Vector Machines (SVM), Neural Networks, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Naïve Bayes.

1) SUPPORT VECTOR MACHINE (SVM)

The Support Vector Machine (SVM) method [14] classifies our feature vectors into two separate groups (in this case abnormal and normal) [15]. Classification is learnt by analyzing a set of labeled vectors, and the resulting learning is then used to label unlabeled vectors, as being in one of two classes [16].

Given training data (i.e. feature vectors for each image) (x_i, y_i), where $x_i \in R^n$, and $y_i \in \{1, -1\}$ [17], and the two classes to be separated using SVM, this classifier finds the best hyper-plane that partitions the vector space into two regions. This hyperplane provides the largest possible margin between the hyperplane and the nearest feature vectors in each partition [18].

The hyper-plane can be defined using Equation (9) [18]:

$$f\,(x) = x'\beta + b = 0 \tag{9}$$

where:
$\beta \in R^n$
b is a real number

This hyper-plane separates the classes, and is called the decision boundary. Any data point above this boundary is considered of class 1 ($x_i\beta + b > 0$ then $y_i = 1$), and any data point below this boundary is considered of class -1 ($x_i \beta + b < 0$ then $y_i = 1$). The method computes the values of β and b, with largest possible margin. This type of classification is considered linear SVM [19].

2) NEURAL NETWORKS

Our brains are networks of neurons, interconnected with effectors (axons) and receptors (dendrites) [20]. The dendrites are inputs and axon terminals are the outputs. Each neuron includes a cell body that contains a nucleus and associated dendrites. The axon is a long fiber. Complex processes happen in the axon, using the inputs to get the outputs [21]. At the end of the axon are branches called axon terminals that send the output signal to the dendrites on other neurons [21].

Artificial neural network is inspired by the human nervous system. Consider a set of inputs x_1 and x_2, and an output y. The inputs and the output are connected with weighted interconnection links w_1 and w_2. The output can be calculated using Equation (10).

$$y = f\,(w_1x_1 + w_2x_2) \tag{10}$$

The weights refer to the strength of the connection between the neurons [20].

A single-layer artificial neural network has a single output layer [22]. Consider a network with n inputs and m outputs, this leads to $n \times m$ weights, starting with w_{11} and ending with w_{mn}.

The output vector will be $\bar{y} = (y_1, y_2, \ldots, y_m)$ and the input vector will be $\bar{x} = (x_1, x_2, \ldots, x_n)$. The single output is then computed using Equation (11) [23]:

$$y_m = f\left(\sum_{i=1}^{n} w_{mi}x_i\right) \tag{11}$$

A multi-layer artificial neural network has one or more *hidden* layer(s) between the input and output layers [22]. These hidden layers allow for non-linear relationship between the input and the output of the network.

In this work, a multi-layer artificial neural network is used with three hidden layers.

3) K-NEAREST NEIGHBORS

The k-nearest neighbors (KNN) classifier works as follows. Each feature vector is assigned a label/class based on the majority of the k-nearest neighbor vectors, using a distance metric to compute the nearest neighbors [24].

A feature vector is compared to a set of labeled feature vectors previously used as training data (prior data [25]). The label for the feature vector is determined based on the majority of nearby labeled feature vectors [26]. An odd number for K is used [25].

The distance used in this work is Euclidean, which is defined in Equation (12) [26]:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{K} (x_i - y_i)^2} \tag{12}$$

In our work, different values for K were used, in an attempt to find the optimal value [27].

4) DECISION TREE

A decision tree has three components: Root, Branch, and Leaf. Each step of the decision process is based on one element of the feature vector. The process starts at the Root, proceeds through the Branches, and ends at the Leaf, which represents the label [28]. Ideally, the Root node is based on the most informative element of the vector [29].

5) RANDOM FOREST

The Random Forest learning method computes and builds a set of multiple decision tree models to distinguish between positive and negative instances. Each decision tree is built based on a randomly selected instances of a given data set [59].

During prediction, a feature vector is compared to the set of decision tree models previously constructed. Prediction results are collected from all decision trees. The label for the feature vector is determined based on the majority votes of all decision tree prediction results [59].

6) NAÏVE BAYES

The Naïve Bayes is a probability based supervised learning method. It uses Bayes theorem to compute conditional probabilities of any given instance to belong to each class label. The instance is then assigned to the label with the highest conditional probability [60].

Given a set of training instances represented as feature vectors for each instance, and the set of possible classes $(C_1, C_2 \ldots C_n)$, a feature vector $\mathbf{x}$ for any given instance to be classified is assigned conditional probabilities for each possible class using Equation (13) [60].

$$p(C_i|\mathbf{x}) = \frac{p(C_i) \times p(\mathbf{x}|C_i)}{p(\mathbf{x})} \tag{13}$$

An output label that is most probable is then assigned to $\mathbf{x}$ based on the *maximum a posteriori* rule, using Equation (14) [60].

$$y = argmax\, p(C_i) \prod p(\mathbf{x}|C_i) \tag{14}$$

F. PERFORMANCE MEASURES

It is possible to evaluate and compare the performance of different image feature extraction methods and different machine learning algorithms using the well-known 10-fold cross validation method. In this method, each labeled feature vector set is divided into ten partitions, where nine partitions are used for training, and one partition is used for testing. The process is repeated ten times such that each partition is used for testing once [53]. This results in ten values for each performance measure, which are then averaged.

Several performance measures can be used. Some researchers in the field use an accuracy metric, while others use sensitivity and/or specificity metrics.

Accuracy shows the ability of the system to distinguish between abnormal and normal cases. Sensitivity shows the ability of the system to identify the abnormal cases correctly, and specificity shows the ability of the system to identify the normal cases correctly [52]. These measures are computed as follow.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{17}$$

where:

- True positive (TP): number of the abnormal cases that are correctly diagnosed as abnormal.
- False positive (FP): number of the normal cases that are incorrectly diagnosed as abnormal.
- True negative (TN): number of the normal cases that are correctly diagnosed as normal.
- False negative (FN): number of the abnormal cases that are incorrectly diagnosed as normal.

G. GLOBAL MODELS

To build global models, image features are extracted from the entire image. Each of these extracted feature types is used to generate a feature vector to represent each image. These feature vectors are then used to compute and build the global detection models.

H. LOCAL MODELS

To be able to localize the image regions with suspicious content, images are warped, cropped, and then divided into

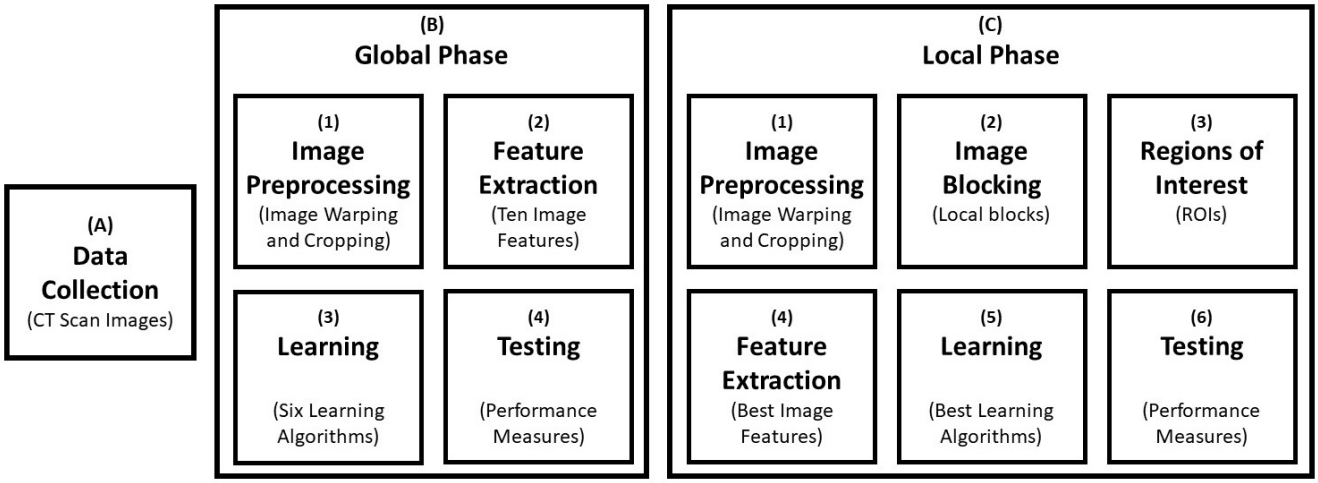

FIGURE 1. Overall methodology.

a number of local blocks. Within each local block, image features are extracted from the local block. Each of these extracted feature types is used to generate a feature vector to represent each local block. These feature vectors are then used to compute and build the local detection models.

IV. METHODOLOGY

In this work, we propose a comprehensive and comparative global and local feature extraction framework to build a detection system for lung cancer from CT scan images. It applies multiple classification techniques to features extracted from those images.

Figure 1 shows the overall methodology used in this work. This methodology consists of three major sequential phases: the Data Collection Phase, the Global Phase, and the Local Phase. In the Data Collection phase, a set of CT scan images is collected.

During the Global Phase, the collected images are preprocessed (1) using image warping and cropping. Then ten different types of global features are extracted from each preprocessed image (2). This produces ten feature vectors for each image. These feature vectors are then used to build detection models using six different machine learning algorithms (3). The result is 60 different detection models. The performance of each detection model is then measured, and compared with the other 59 models (4).

During the Local Phase, the images are preprocessed (1) and then each image is subdivided into an array of local blocks (2) which are used to define ROIs (3). Using the types of image features that performed well during the Global Phase, features are then extracted from each image block to produce feature vectors (4). These feature vectors are then used to build detection models for the image blocks, using the learning algorithms that performed well in the Global Phase (5). In this third phase, we experiment with different numbers of blocks per image (including a single block, i.e. global approach) to determine the optimal number of blocks. The performance of each detection model is then measured, and compared with the other models (6).

A. DATA COLLECTION

A set of 1000 CT images were used in this work: 500 abnormal cases and 500 normal cases. The images were selected randomly from thousands of images from TCIA database [51].

These CT scan images were taken after the patients were diagnosed, but before any treatment or surgery. Each image represents a single slice from the stack of slices produced by a CT scan. The thickness of each slice is between 3 and 6 mm, depending on the CT scanner. This TCIA database contains more than 4000 images from different patients.

B. GLOBAL PHASE

In the global phase, we train our system to detect lung cancer from the CT scan images. This is done in three stages (as shown in Figure 1): preprocessing, feature extraction, and learning. The resulting learnt model is then used in the testing phase for performance evaluation.

Several feature extraction methods and learning models are used. This work compares the performance of each feature extraction method and each learning method for building the final lung cancer detection model.

1) IMAGE PREPROCESSING

In this stage, two image preprocessing steps are applied to each image in our data set: image warping and image cropping.

a: IMAGE WARPING

To align the anatomical features in each image with a common reference anatomical coordinate system, we first selected one image, and manually marked 19 anatomically-based control points within the image, as shown in Figure 2.

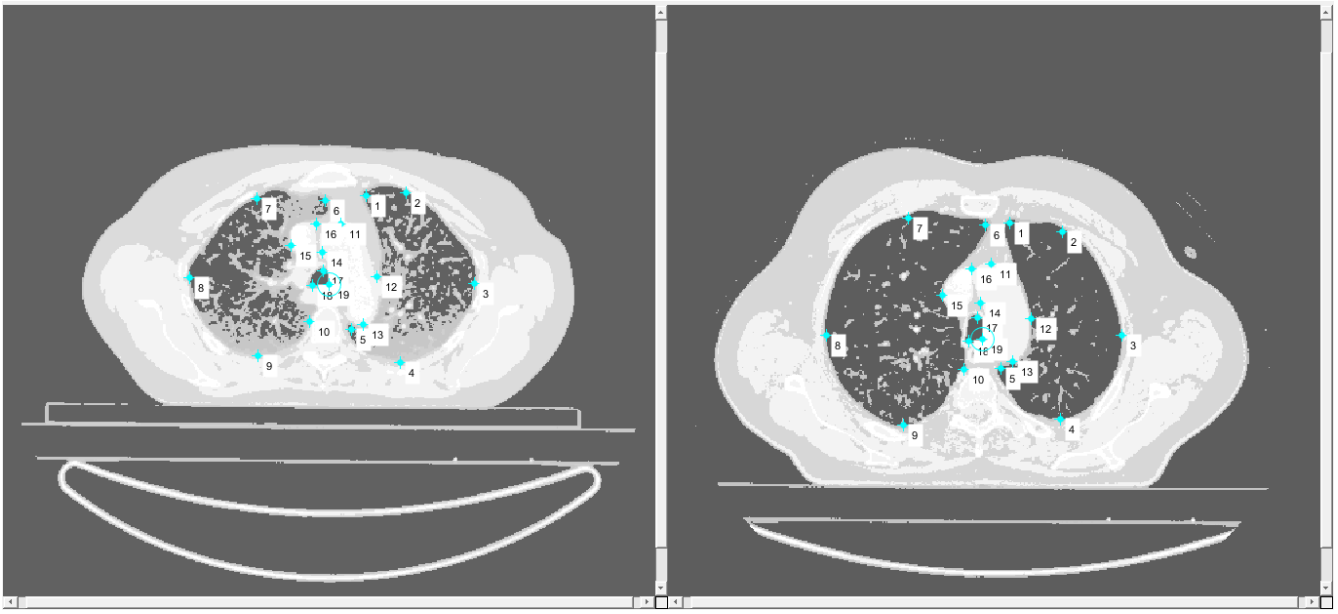

FIGURE 2. Image warping.

(These control points were corner points that surrounded the lungs, and corner points that surround the heart.) We then used these 19 control points to warp the images, so that they aligned with the corresponding control points in our reference anatomical coordinate system, as shown in Figure 2.

A MATLAB software tool was used to mark the control points within the reference anatomical coordinate system image (right side of Figure 2) and the image to be warped (left side of Figure 2).

We then computed a *geometric transformation* that aligns the control points within the chosen image to those within our reference anatomical coordinate system. Given the (x,y) pixel coordinates for the 19 control points in the reference image, and the (x,y) coordinates of the same 19 points in the chosen image, a non-linear transformation was computed that warps the chosen image to align its control points with those in the reference image.

We computed and applied a geometric transformation to each of the images in our 1000-image data set, to generate warped versions of all the images. The resulting images contain the same textural details as the original images, but with their control points aligned to the ones in the reference image.

b: IMAGE CROPPING

The warped image from the previous step is cropped. Figure 3 shows an example of a cropped image. The cropping points were the far-right and the far-left points, and the highest and the lowest points.

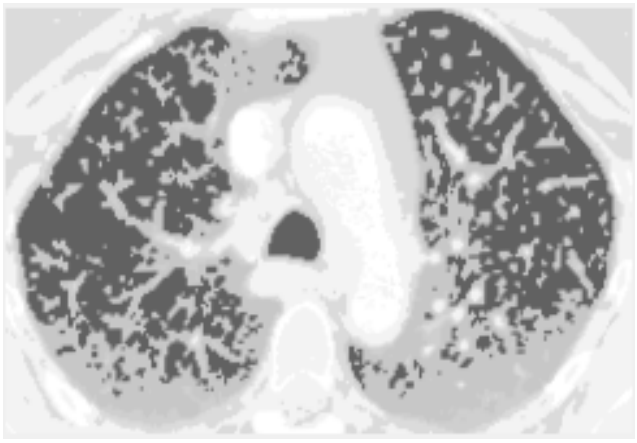

FIGURE 3. Cropped image.

2) FEATURE EXTRACTION

Extracted image features play an important role in our proposed framework. In this paper, we compare the performance of ten feature extraction methods. The features used in this work are shown in Table 1.

For global feature extraction, we extract these ten feature types from the entire image. This produces 10 feature vectors for each image. These feature vectors are then used in the learning phase.

3) LEARNING

Each of these ten extracted feature types is used to generate a feature vector to represent each image. These ten feature vectors are then used in the learning process to compute and build the final detection model. To do so, several learning algorithms are used. This work compares the performance of six well-known machine learning algorithms: Support Vector Machines (SVM), Neural Networks, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Naïve Bayes.

TABLE 1. Image features used.

No.	Feature
1	Intensity Histogram
2	Histogram of Oriented Gradients (HOG)
3	Gabor Filter
4	Entropy Filter
5	Grayscale Contrast
6	Grayscale Correlation
7	Grayscale Energy
8	Grayscale Homogeneity
9	Standard Deviation
10	Haar Wavelet

4) TESTING

For each CT scan image, the same preprocessing and feature extraction processes were applied. The learnt model is then used to label each image as abnormal or normal

As stated earlier, in this work, we compare the performance of ten different image feature extraction methods, and six different machine learning algorithms.

To evaluate the performance of each of the 60 learnt models, a 10-fold cross validation method is used to compute three performance measures: accuracy, sensitivity, and specificity.

C. LOCAL PHASE

In the Global Training and Testing phase, we compare the performance of the ten types of extracted image features and the six machine learning algorithms, and identify those features and the learning methods that outperformed the other features and methods. The Local Training and Testing phase then uses those better-performing features and methods.

1) IMAGE PREPROCESSING

Image warping and cropping operations that were used in the global phase are also used in the local phase.

2) IMAGE BLOCKING

In this stage, we divide each warped image into a number of blocks, as shown in Figure 4. Within each block, we then use the types of feature extraction that performed well during the Global Phase. Each feature type produces b feature vectors for each image, where b is the total number of blocks in the image. Each of these b feature vectors can then be used to label one block.

3) REGIONS OF INTEREST

To reduce the learning and training time, we go through all of the 500 abnormal images to determine which blocks contain suspicious content. Many of the blocks in abnormal images do not contain any suspicious content, and thus, they are not included in the feature extraction nor in the training and learning processes. In other words, we only train models with the blocks that contain suspicious content in the abnormal images – those are our Regions of Interest (ROIs).

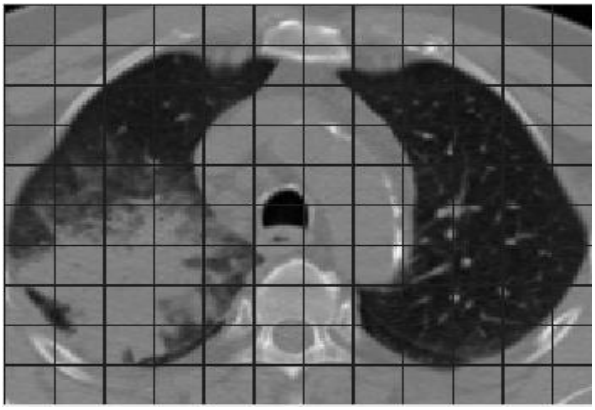

FIGURE 4. Local blocks.

4) FEATURE EXTRACTION

In this stage, we apply the best feature extraction methods (identified in the global phase) to each ROI.

5) LEARNING

In this stage we apply the best machine learning algorithms (identified in the global phase) to each ROI.

6) TESTING

In the testing process, we use these models to label each ROI (i.e. each block) in any given test image as normal or abnormal. If any ROI within the test image is labeled as abnormal, then that entire test image is labeled as abnormal. Only if the none of the ROIs in the test image is labeled as abnormal is that test image labeled as normal.

For training purposes, we would ideally like to have training sets that contain a balance in the number of abnormal cases and normal cases. That is the reason that we chose to use 500 abnormal CT scans, and 500 normal CT scans. However, when CT scan images are subdivided into multiple blocks, there are many more normal blocks than abnormal blocks. In addition, for any given block, there might be only 10 abnormal blocks across the 500 abnormal images. This leads to a large imbalance in the count of normal versus abnormal blocks in the training set for that block.

This work avoids this imbalance problem in two ways:

(1) We chose our ROIs to be only the blocks with suspicious content in at least 250 abnormal images (out of the 500 abnormal images).
(2) We experiment with different block sizes, to find the optimal number of blocks to define our ROIs for local feature extraction. This allows merging neighboring blocks, which in turn increases the number of suspicious blocks in a local ROI across the images.

Table 2 presents the algorithm for our proposed local feature extraction approach.

V. RESULTS AND DISCUSSION

In this section, the detailed results (in terms of accuracy, sensitivity, and specificity) are presented and discussed for

TABLE 2. Proposed local feature extraction algorithm.

I is the set of input images

$I_w = warp(I)$
$I_c = crop(I_w)$
for $b = b_1, b_2 \ldots b_n$
  $B = blocks(I_c, b)$
  $ROIs = RegionsOfInterest(B)$

  for each ROI
    $FV = FeatureExtraction(ROI)$
    $Model = Learning(FV)$
  End *for*

  $P = Performance(Model, ROIs)$
End *for*

Output b_i with highest P

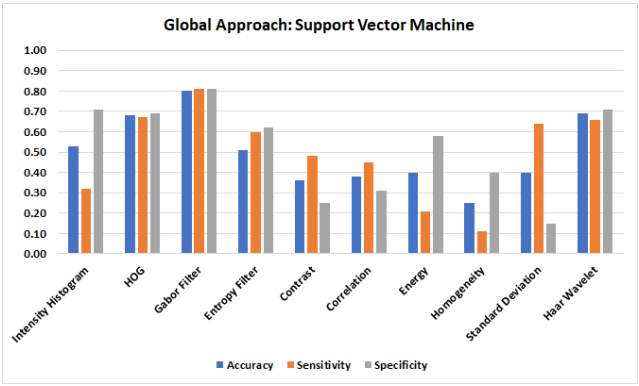

FIGURE 5. SVM performance for global feature extraction.

the six learning methods, across the ten types of extracted features using the global feature extraction approach and the local feature extraction approach.

For the local feature extraction approach, the accuracy, sensitivity, and specificity rates are presented and discussed for different ROI sizes, ranging from a 1×1 block (i.e. the global approach) to 20×20 blocks, using the image features and the learning methods that outperformed others in the global approach.

A. GLOBAL FEATURE EXTRACTION RESULTS

1) SVM LEARNING METHOD

Figure 5 shows the accuracy, sensitivity, and specificity performance measures using SVM and global feature extraction for the ten image features.

The results presented in Figure 5 suggest that the Gabor Filter, the Haar Wavelet, and the HOG feature types outperform the rest of the feature types, when using SVM with global feature extraction. Overall, the Gabor Filter achieved the best performance measures, with accuracy, sensitivity, and specificity rates of 80%, 81%, and 81%, respectively.

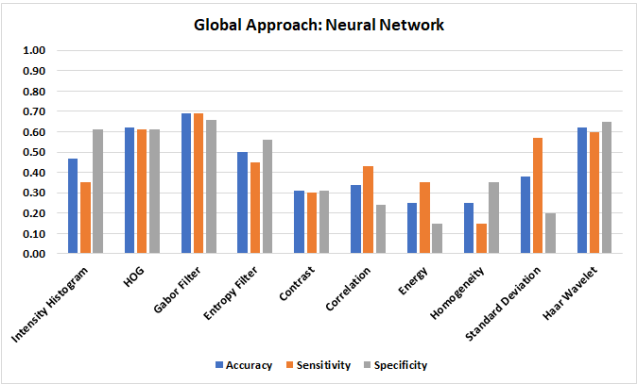

FIGURE 6. Neural Network performance for global feature extraction.

TABLE 3. The best k values in global feature extraction.

Image Feature	Best k	Best Accuracy
Intensity Histogram	15	0.51
HOG	1	0.60
Gabor Filter	1	0.65
Entropy Filter	7	0.56
Grayscale Contrast	1	0.45
Grayscale Correlation	9	0.46
Grayscale Energy	13	0.46
Grayscale Homogeneity	9	0.37
Standard Deviation	1	0.26
Haar Wavelet	15	0.59

2) NEURAL NETWORK LEARNING METHOD

Figure 6 shows the accuracy, sensitivity, and specificity performance measures using the Neural Network with global feature extraction for the ten image features.

The results presented in Figure 6 suggest that the Gabor Filter, the Haar Wavelet, and the HOG feature types outperform the rest of the feature types, when using the Neural Network with global feature extraction. Overall, the Gabor Filter achieved the best performance measures with accuracy, sensitivity, and specificity rates of 69%, 69%, and 66%, respectively.

3) K-NEAREST NEIGHBORS LEARNING METHOD

This section presents the results of using the k-nearest neighbors (KNN) learning method. However, we first conducted several experiments to find the values of k that achieved the best accuracy values for each type of image feature with the global feature extraction approach.

Table 3 shows the values of k that achieved best accuracy rates for each image feature with the global feature extraction approach.

Figure 7 shows the accuracy, sensitivity, and specificity performance measures using KNN with global feature extraction for the ten image features (using their best k values).

The results presented in Figure 7 suggest that the Gabor Filter, the HOG, and the Haar Wavelet feature types

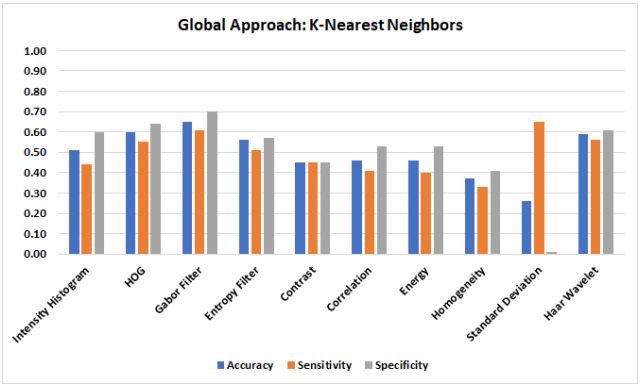

FIGURE 7. KNN performance for global feature extraction.

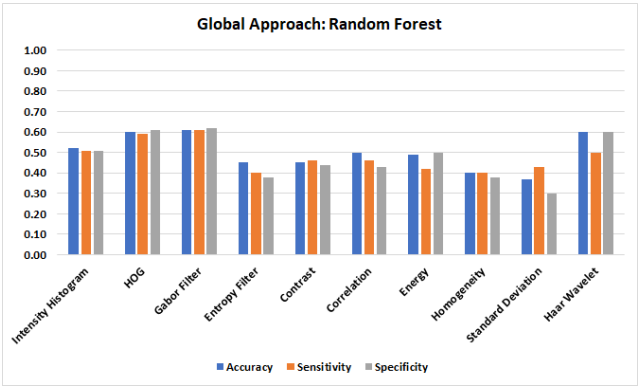

FIGURE 9. Random Forest performance for global feature extraction.

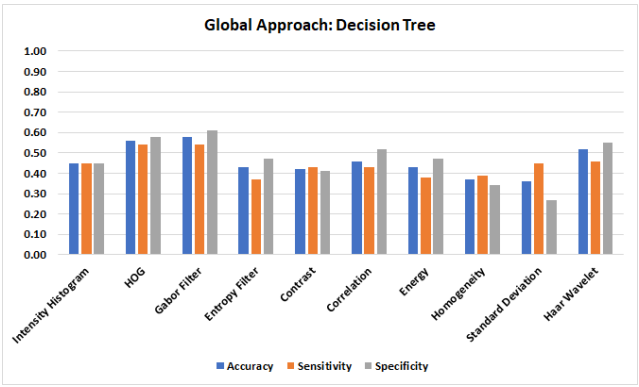

FIGURE 8. Decision Tree performance for global feature extraction.

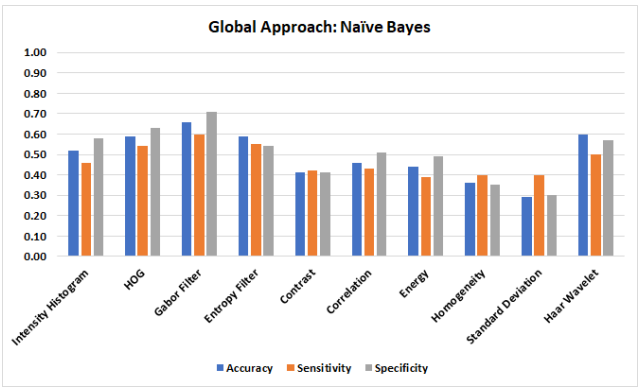

FIGURE 10. Naïve Bayes performance for global feature extraction.

outperform the rest of the feature types, when using KNN with global feature extraction. Overall, the Gabor Filter achieved the best performance measures with accuracy, sensitivity, and specificity rates of 65%, 61%, and 70%, respectively with a k value of 1.

4) DECISION TREE LEARNING METHOD

Figure 8 shows the accuracy, sensitivity, and specificity performance measures using the Decision Tree with global feature extraction for the ten types of image features.

The results presented in Figure 8 suggest that the Gabor Filter, the HOG, and the Haar Wavelet feature types outperform the rest of the feature types, when using a Decision Tree with global feature extraction. Overall, the Gabor Filter achieved the best performance measures with accuracy, sensitivity, and specificity rates of 58%, 54%, and 61%, respectively.

5) RANDOM FOREST LEARNING METHOD

Figure 9 shows the accuracy, sensitivity, and specificity performance measures using the Random Forest with global feature extraction for the ten types of image features.

The results presented in Figure 9 suggest that the Gabor Filter, the HOG, and the Haar Wavelet feature types outperform the rest of the feature types, when using a Random Forest model with global feature extraction. Overall, the Gabor Filter achieved the best performance measures with accuracy, sensitivity, and specificity rates of 61%, 61%, and 62%, respectively.

6) NAÏVE BAYES LEARNING METHOD

Figure 10 shows the accuracy, sensitivity, and specificity performance measures using the Naïve Bayes with global feature extraction for the ten types of image features.

The results presented in Figure 10 suggest that the Gabor Filter, the HOG, and the Haar Wavelet feature types outperform the rest of the feature types, when using a Naïve Bayes model with global feature extraction. Overall, the Gabor Filter achieved the best performance measures with accuracy, sensitivity, and specificity rates of 66%, 60%, and 71%, respectively.

7) THE BEST IMAGE FEATURE(S)

In order to identify the best image feature types, we averaged the performance measures (accuracy, sensitivity, and specificity) for each of the ten image feature types, across all six learning methods. Figure 11 shows the average accuracy, sensitivity, and specificity performance measures for the ten types of image features, when using SVM, Neural Network, KNN, Decision Tree, Random Forest, and Naïve Bayes with the global feature extraction approach.

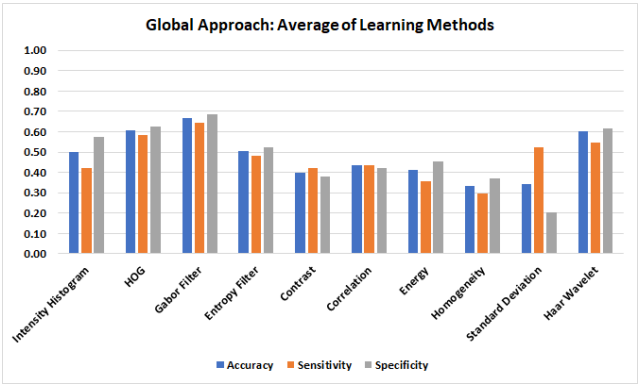

FIGURE 11. Average of learning methods for global feature extraction.

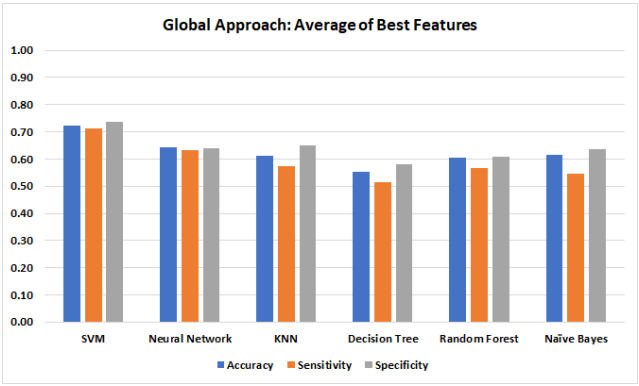

FIGURE 12. Average of best features for global feature extraction.

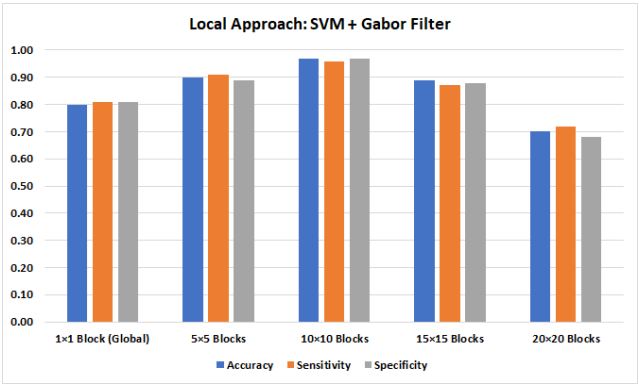

FIGURE 13. SVM and Gabor Filter performance for local feature extraction.

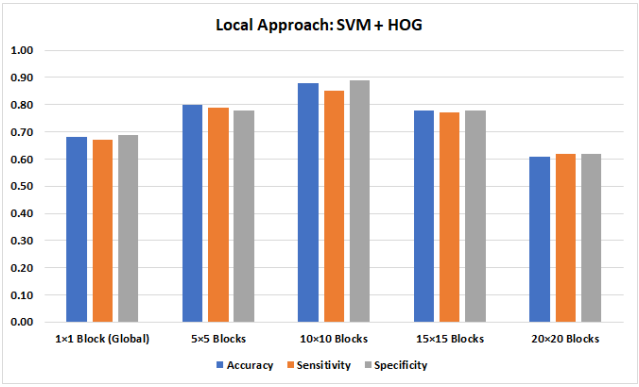

FIGURE 14. SVM and HOG performance for local feature extraction.

The results presented in Figure 11 suggest that the Gabor Filter, the HOG, and the Haar Wavelet feature types outperform the rest of the feature types, with the global feature extraction approach. Moreover, the Gabor Filter achieved the best performance measures with average accuracy, sensitivity, and specificity rates of 67%, 64%, and 69%, respectively.

8) THE BEST LEARNING METHOD(S)

To identify the best learning method(s), we averaged the performance measures (accuracy, sensitivity, and specificity) for each of the six learning methods across the best three types of image features identified in the previous section. Figure 12 shows the average accuracy, sensitivity, and specificity performance measures for the six different learning methods, when using the Gabor Filter, the HOG, and the Haar Wavelet with the global feature extraction approach.

The results presented in Figure 12 suggest that SVM outperforms the other five learning methods, with the global feature extraction approach. It achieved the best performance measures, with average accuracy, sensitivity, and specificity rates of 72%, 71%, and 74%, respectively.

Taken together, the results presented in figures 11 and 12 suggest that the Gabor Filter, the HOG, and the Haar Wavelet feature types used with SVM, outperformed the rest of image feature types and learning methods, with the global feature extraction approach. Thus, they were used in presenting the local feature extraction results in the following section.

B. LOCAL FEATURE EXTRACTION RESULTS

1) SVM WITH GABOR FILTER

Figure 13 shows the accuracy, sensitivity, and specificity performance measures using SVM and the Gabor Filter with local feature extraction, using ROI arrays sizes of 1×1, 5×5, 10×10, 15×15, and 20×20 blocks. Note that an ROI array size of 1×1 is equivalent to the global feature extraction approach for SVM and the Gabor Filter.

The results presented in Figure 13 suggest that an ROI array size of 10×10 blocks outperforms all the other array sizes, when using SVM and the Gabor Filter with the local feature extraction approach. It achieved the best performance measures, with accuracy, sensitivity, and specificity rates of 97%, 96%, and 97%, respectively.

2) SVM WITH HOG

Figure 14 shows the accuracy, sensitivity, and specificity performance measures using SVM and HOG with local feature extraction using ROI array sizes of 1×1, 5×5, 10×10, 15×15, and 20×20 blocks. Note that the ROI array size of 1×1 represents the global feature extraction approach for SVM and HOG.

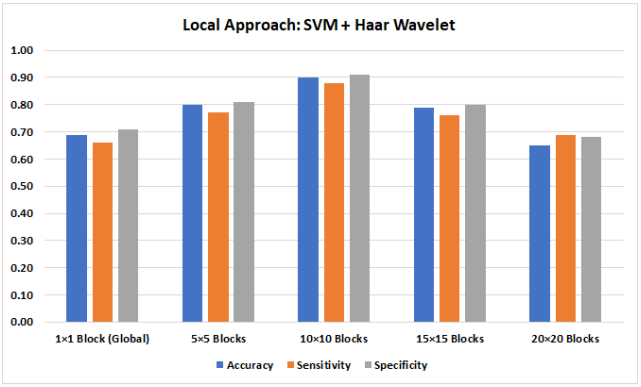

FIGURE 15. SVM and Haar Wavelet performance for local feature extraction.

The results presented in Figure 14 suggest that an ROI array size of 10 × 10 blocks outperforms all the other sizes, when using SVM and HOG with the local feature extraction approach. It achieved the best performance measures with accuracy, sensitivity, and specificity rates of 88%, 85%, and 89%, respectively.

3) SVM WITH HAAR WAVELET

Figure 15 shows the accuracy, sensitivity, and specificity performance measures using SVM and the Haar Wavelet with local feature extraction using ROI array sizes of 1 × 1, 5 × 5, 10 × 10, 15 × 15, and 20 × 20 blocks. Note that the ROI array size of 1 × 1 block represents the global feature extraction approach for SVM and the Haar Wavelet.

The results presented in Figure 15 suggest that the ROI array size of 10 × 10 blocks outperforms the other sizes, when using SVM and the Haar Wavelet with the local feature extraction approach. It achieved the best performance measures with accuracy, sensitivity, and specificity rates of 90%, 88%, and 91%, respectively.

C. OVERALL DISCUSSION

Taken together, the results presented in figures 13, 14, and 15 suggest that the local feature extraction approach with an ROI array size of 10 × 10 blocks, when used with a Gabor Filter, a HOG, or a Haar Wavelet feature extractor and with SVM outperformed the rest of ROI array sizes.

The results also suggest that with local feature extraction, an ROI array size of 10 × 10 blocks and SVM, the Gabor Filter outperformed the other feature extractors.

Overall, the results presented in this work, which was performed with a large data set of 1000 CT scan images, suggests that the local feature extraction approach outperforms the global one. It also suggests that the proposed method performs better than the other methods cited in the literature review section.

Specifically, the proposed research in this work is better in comparison towards other methods in the literature for several reasons. First, it uses larger data set than those in the literature. Our work uses a large data set of 1000 CT scan images, while the cited work in the literature used less than 500 images. For example, the work in [41], [42], [54], and [55] used 420, 125, 276, and 400 images, respectively.

Second, our work achieved better performance when compared with the cited methods in the literature. It achieved accuracy, sensitivity, and specificity rates of 97%, 96%, and 97%, respectively. Although the work of [41], [47], and [54] achieved a sensitivity rate of 93.4%, 93.9%, and 97.5%, respectively, however, accuracy and specificity rates were not reported.

Finally, our work builds localized learning models for lung cancer detection, while the cited work in the literature used global detection methods using different machine learning algorithms [22], [39]–[42], [55]. Building such localized learning models is very important in CT scan images, because these images are interpreted and diagnosed by radiologists based on anatomical regions within the images. Normal content in some anatomical regions might be abnormal in other regions, and vice versa. In addition, our work compares the effectiveness of ten different types of extracted image features and six well-known machine learning algorithms. This explains the importance of our proposed local feature extraction and learning method when compared with traditional methods.

VI. CONCLUSION AND FUTURE WORK

In this work, we proposed to answer the following research question:

Q: *What image features and learning methods would be most useful for building localized learning models for lung cancer detection within CT scan images?*

This work proposed a comprehensive and comparative global and local feature extraction framework for lung cancer detection using CT scan images. It compared between six well-known machine learning algorithms, and ten image feature extraction methods, using global and local feature extraction approaches. Image warping was performed to allow for anatomically-based local feature extraction and model learning.

The results presented in this work showed that the Gabor Filter, the Histogram of Oriented Gradients (HOG), and the Haar Wavelet feature extraction methods outperformed seven other feature extraction methods, and that a Support Vector Machine (SVM) outperformed five other types of learning algorithms.

The results also showed that the proposed local feature extraction approach outperformed the traditional global approach. SVM with Haar Wavelet feature extraction achieved 90% accuracy, 88% sensitivity, and 91% specificity. SVM with HOG feature extraction achieved 88% accuracy, 85% sensitivity, and 89% specificity. SVM with Gabor Filter feature extraction achieved the *best* accuracy, sensitivity, and specificity rates of 97%, 96%, and 97%, respectively.

These results show that the proposed method performs better than other cited methods within the literature not only in achieving better accuracy, sensitivity, and specificity rates but

also in terms of using a large data set of 1000 CT scan images and building localized learning models for lung cancer detection. This suggests that using SVM with Gabor Filter feature extraction could be useful for detecting suspicious regions within CT scan images, to assist radiologists in detecting lung cancer.

As for future work, more comparison algorithms will be considered, and more data sets will be included.

REFERENCES

[1] *Cancer*. Accessed: Apr. 30, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Cancer

[2] *What is Cancer?*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[3] *Cancer Statistics*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/statistics

[4] *CT Scan for Cancer*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.cancer.org/treatment/understanding-your-diagnosis/tests/ct-scan-for-cancer.html

[5] *Lung Cancer*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptoms-causes/syc-20374620

[6] J. John and M. G. Mini, "Multilevel thresholding based segmentation and feature extraction for pulmonary nodule detection," *Proc. Technol.*, vol. 24, pp. 957–963, Jan. 2016.

[7] *Lung Cancer*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.nhs.U.K./conditions/lung-cancer

[8] S. M. B. Netto, A. C. Silva, R. A. Nunes, and M. Gattass, "Automatic segmentation of lung nodules with growing neural gas and support vector machine," *Comput. Biol. Med.*, vol. 42, no. 11, pp. 1110–1121, Nov. 2012.

[9] C. P. Wild and B. W. Stewart, Eds., *World Cancer Report 2014*, World Health Organization, Geneva, Switzerland, 2014, pp. 482–494.

[10] *What is a CT Scan?*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.ucdmc.ucdavis.edu/radiology/UCDHS_CT_FAQ_v1.pdf

[11] P. Pelosi and M. G. D. Abreu, "Lung CT scan," *Open Nucl. Med. J.*, vol. 2, no. 1, pp. 86–98, 2010.

[12] S. J. Swensen, J. R. Jett, J. A. Sloan, D. E. Midthun, T. E. Hartman, A. M. Sykes, G. L. Aughenbaugh, F. E. Zink, S. L. Hillman, G. R. Noetzel, and R. S. Marks, "Screening for lung cancer with low-dose spiral computed tomography," *Amer. J. Respiratory Crit. Care Med.*, vol. 165, no. 4, pp. 508–513, 2002.

[13] *Computer-Aided Diagnosis: The Tipping Point for Digital Pathology*. Accessed: Apr. 30, 2021. [Online]. Available: https://digitalpathologyassociation.org/blog/computer-aided-diagnosis-the-tipping-point-for-digital-pathology

[14] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *J. Mach. Learn. Res.*, vol. 2, pp. 125–137, Mar. 2002.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[16] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory (COLT)*, 1992, pp. 144–152.

[17] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Nat. Taiwan Univ., Taipei, Taiwan, Tech. Rep., 2003, pp. 1396–1400.

[18] *Support Vector Machines for Binary Classification*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html

[19] *SVM Tutorial*. Accessed: Apr. 30, 2021. [Online]. Available: http://web.mit.edu/zoya/www/SVM.pdf

[20] A. D. Dongare, R. R. Kharde, and A. D. Kachare, "Introduction to artificial neural network," *Int. J. Eng. Innov. Technol.*, vol. 2, no. 1, pp. 189–194, 2012.

[21] S. K. Pratap, "Artificial neural network (ANN) inspired from biological nervous system," *Int. J. Appl. Innov. Eng. Manage. (IJAIEM)*, vol. 2, no. 1, pp. 227–231, 2013.

[22] I. Nunes and H. S. da Silva, *Artificial Neural Networks: A Practical Course*. Springer, 2018.

[23] J. Kacprzyk and W. Pedrycz, Eds., *Springer Handbook of Computational Intelligence*. Springer, 2015.

[24] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[25] *K-Nearest Neighbours*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.geeksforgeeks.org/k-nearest-neighbours

[26] *K Nearest Neighbors—Classification*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.saedsayad.com/k_nearest_neighbors.htm

[27] *Introduction to K-Nearest Neighbors: A Powerful Machine Learning Algorithm (With Implementation in Python & R)*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering

[28] *Chapter 4: Decision Trees Algorithms*. Accessed: Apr. 30, 2021. [Online]. Available: https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1

[29] *Decision Tree: An Algorithm That Works Like the Human Brain*. Accessed: Apr. 30, 2021. [Online]. Available: https://towardsdatascience.com/decision-tree-an-algorithm-that-works-like-the-human-brain-8bc0652f1fc6

[30] *Intensity Histogram*. Accessed: Apr. 30, 2021. [Online]. Available: https://homepages.inf.ed.ac.U.K./rbf/HIPR2/histgram.htm

[31] *Histogram of Oriented Gradients (HOG) for Object Detection*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.vocal.com/video/histogram-of-oriented-gradients-hog-for-object-detection

[32] X. Song, F. Liu, Z. Zhang, C. Yang, X. Luo, and L. Chen, "2D Gabor filters-based steganalysis of content-adaptive JPEG steganography," *Multimedia Tools Appl.*, vol. 76, no. 24, pp. 26391–26419, 2017.

[33] J. Kaur, N. Garg, and D. Kaur, "Segmentation and feature extraction of lung region for the early detection of lung tumor," *Int. J. Sci. Res. (IJSR)*, vol. 3, no. 6, pp. 2327–2330, 2014.

[34] J. M. Bland and D. G. Altman, "Statistics notes: Measurement error," *BMJ*, vol. 312, no. 7047, p. 1654, 1996.

[35] N. Jacob and A. Martin, "Image denoising in the wavelet domain using Wiener filtering," Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. 17, 2004, pp. 1–21.

[36] *Short Intro. to Haar Wavelet Transform*. Accessed: Apr. 30, 2021. [Online]. Available: https://chengtsolin.wordpress.com/2015/04/15/real-time-2d-discrete-wavelet-transform-using-opengl-compute-shader

[37] X.-Y. Jin, Y.-C. Zhang, and Q.-L. Jin, "Pulmonary nodule detection based on CT images using convolution neural network," in *Proc. 9th Int. Symp. Comput. Intell. Design (ISCID)*, Dec. 2016, pp. 202–204.

[38] T. Aggarwal, A. Furqan, and K. Kalra, "Feature extraction and LDA based classification of lung nodules in chest CT scan images," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Aug. 2015, pp. 1189–1193.

[39] C. Jacobs, E. M. van Rikxoort, T. Twellmann, E. T. Scholten, P. A. de Jong, J.-M. Kuhnigk, M. Oudkerk, H. J. de Koning, M. Prokop, C. Schaefer-Prokop, and B. van Ginneken, "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images," *Med. Image Anal.*, vol. 18, pp. 374–384, Feb. 2014.

[40] E. Rendon-Gonzalez and V. Ponomaryov, "Automatic lung nodule segmentation and classification in CT images based on SVM," in *Proc. 9th Int. Kharkiv Symp. Phys. Eng. Microw., Millim. Submillimeter Waves (MSMW)*, Jun. 2016, pp. 1–4.

[41] M. Firmino, G. Angelo, H. Morais, M. R. Dantas, and R. Valentim, "Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy," *BioMed. Eng. OnLine*, vol. 15, no. 1, pp. 1–17, Jan. 2016.

[42] M. C. Lee, L. Boroczky, K. Sungur-Stasik, A. D. Cann, S. M. Kawut, and C. A. Powell, "Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction," *Artif. Intell. Med.*, vol. 50, no. 1, pp. 43–53, 2010.

[43] P. B. Sangamithraa and S. Govindaraju, "Lung tumour detection and classification using EK-mean clustering," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2016, pp. 2201–2206.

[44] S. Ignatious and R. Joseph, "Computer aided lung cancer detection system," in *Proc. Global Conf. Commun. Technol. (GCCT)*, Apr. 2015, pp. 555–558.

[45] J. Tong, W. Ying, and W. C. Dong, "A lung cancer lesions dectection scheme based on CT image," in *Proc. 2nd Int. Conf. Signal Process. Syst.*, vol. 1, Jul. 2010, pp. V1-557–V1-560.

[46] C. Li, C. Shi, H. Zhang, Y. Chen, and S. Zhang, "Multiple instance learning for computer aided detection and diagnosis of gastric cancer with dual-energy CT imaging," *J. Biomed. Inform.*, vol. 57, pp. 358–368, Oct. 2015.

[47] W.-J. Choi and T.-S. Choi, "Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 37–54, 2014.
[48] K. L. Hua, C. H. Hsu, S. C. Hidayati, W. H. Cheng, and Y. J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," *OncoTargets Therapy*, vol. 8, p. 2022, Aug. 2015.
[49] S. Ma, W. Wang, B. Xia, S. Zhang, H. Yuan, H. Jiang, W. Meng, X. Zheng, and X. Wang, "Multiplexed serum biomarkers for the detection of lung cancer," *EBioMedicine*, vol. 11, pp. 210–218, Sep. 2016.
[50] W.-J. Kuo, R.-F. Chang, D.-R. Chen, and C. C. Lee, "Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images," *Breast Cancer Res. Treatment*, vol. 66, no. 1, pp. 51–57, Mar. 2001.
[51] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, 2013.
[52] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal, "Part 1: Simple definition and calculation of accuracy, sensitivity and specificity," *Arch. Acad. Emergency Med.*, vol. 3, no. 2, pp. 48–49, 2015.
[53] *A Gentle Introduction to K-Fold Cross-Validation*. Accessed: Apr. 30, 2021. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation
[54] S. Özekes, "Rule-based Lung region segmentation and nodule detection via genetic algorithm trained template matching," *Istanbul Commerce Univ. J. Sci.*, vol. 6, no. 11, pp. 17–30, 2007.
[55] F. Shafiei and S. Fekri-Ershad, "Detection of lung cancer tumor in CT scan images using novel combination of super pixel and active contour algorithms," *Traitement du Signal*, vol. 37, no. 6, pp. 1029–1035, 2020.
[56] Y. Liu, Z. Xing, C. Deng, P. Li, and M. Guo, "Automatically detecting lung nodules based on shape descriptor and semi-supervised learning," in *Proc. Int. Conf. Comput. Appl. Syst. Modeling (ICCASM)*, Oct. 2010.
[57] F. Tajeri Pour, M. Rezaei, M. Saberi, and S. F. Ershad, "Texture classification approach based on combination of random threshold vector technique and co-occurrence matrixes," in *Proc. Int. Conf. Comput. Sci. Netw. Technol.*, vol. 4, Dec. 2011, pp. 2303–2306.
[58] C. L. Chowdhary and D. P. Acharjya, "Segmentation and feature extraction in medical imaging: A systematic review," *Proc. Comput. Sci.*, vol. 167, pp. 26–36, Jan. 2020.
[59] C. Zhang and Y. Ma, Eds., *Ensemble Machine Learning: Methods and Applications*. Springer, 2012.
[60] M. N. Murty and V. S. Devi, *Pattern Recognition: An Algorithmic Approach*. Springer, 2011.
[61] D. M. Abdullah and N. S. Ahmed, "A review of most recent lung cancer detection techniques using machine learning," *Int. J. Sci. Bus.*, vol. 5, no. 3, pp. 159–173, 2012.

MOHAMMAD A. ALZUBAIDI received the Ph.D. degree in computer science and engineering from the Ira A. Fulton Schools of Engineering, Arizona State University, in 2012.

He is currently an Associate Professor of computer engineering at Yarmouk University, Jordan. His research interests include medical imaging perception and understanding, computer vision and pattern recognition, assistive technology, and machine learning.

MWAFFAQ OTOOM (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Virginia Tech, in 2012.

He is currently a Full Professor of computer engineering at Yarmouk University, Jordan. His research interests include novel modeling techniques for embedded systems, assistive technology, and machine learning.

HAMZA JARADAT received the M.S. degree in computer engineering from Yarmouk University, in 2019.

His research interests include medical imaging perception and understanding, computer vision and pattern recognition, and machine learning.