FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 1027-1035 | e-ISSN: 2347-470X

Prediction and Classification of CT images for Early Detection of Lung Cancer Using Various Segmentation Models

Sneha S. Nair[1], Dr. V. N. Meena Devi[2] and Dr. Saju Bhasi[3]

[1]Department of Physics, Noorul Islam Centre for Higher Education, Thuckalay, Kumaracoil-629180, Tamil Nadu, India, n.sneha85@gmail.com
[2]Department of Physics, Noorul Islam Centre for Higher Education, Thuckalay, Kumaracoil-629180, Tamil Nadu, India, vndevi@gmail.com
[3]Department of Radiation Physics, Regional Cancer Centre, Thiruvananthapuram, Kerala, India, sajubhasi@gmail.com

***Correspondence:** Sneha S. Nair: n.sneha85@gmail.com

ABSTRACT- One of the most serious and deadly diseases in the world is lung cancer. On the other hand, prompt diagnosis, as well as care, could save lives. Probably the most capable imaging method in the medical world, computed tomography (CT) scans are challenging for clinicians to analyze as well as detect cancer. In recent years, there has been an increase in the use of image analysis techniques for the detection of CT scan images matching cancer tissues. Using a Computer-aided detection (CAD) system employing CT scans to aid inside the early lung cancer diagnosis as well as to differentiate among benign/malignant tumors is thus interesting to address. The primary objective of this study would be to assess several computer-aided approaches, analyze the right methodology already in use, and afterward propose a new approach that integrates enhancements to the best system currently in use. This research improves the performance of the existing retrieval system by combining various image feature extraction processes and modifying the internal layer section of the classifier. The segmentation method proposed here to identify cancer is Improved Random Walker segmentation along with Random Forest (RF) classifier and K-Nearest Neighbors (KNN) classifier. Here, the research is accomplished on the Lung Image database consortium (LIDC) datasets which is a collection of CT images and is utilized as the input images to verify the effectiveness of the suggested strategy. The accuracy of the proposed method for the detection of lung cancer with the aid of the RF classifier is 99.6 % as well as the KNN classifier is 96.4% accordingly.

Keywords: Computed tomography, Lung cancer, Diagnosis, Image pre-processing, Random walk, Accuracy, Cancer, Detection, Image processing, Segmentation.

ARTICLE INFORMATION

Author(s): Sneha S. Nair, Dr. V. N. Meena Devi and Dr. Saju Bhasi;
Received: 29/08/2022 **Accepted:** 24/10/2022 **Published:** 20/11/2022
E- ISSN: 2347-470X
Paper Id: IJEER 22703
Citation: 10.37391/ijeer.100445
Webpage-link:
www.ijeer.forexjournal.co.in/archive/volume-10/ijeer-100445.html

Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

The phrase "cancer" refers to illnesses wherein abnormal cells proliferate uncontrollably but can infect other regions. A significant public health issue is cancer. About one in three persons is predicted to have cancer as a result of their lifespan. This sickness will cause about one-fourth of the populace to pass away. This has been difficult to comprehend the burden of cancer on a worldwide scale. Through the lymphatic and circulatory systems, cancerous tissue to different organs and tissues. Lung cancer also referred to as lung carcinoma, seems to be a malignant tumor of the lungs marked by unconstrained cell proliferation in the lung tissues. Through the process of metastasis, this development can invade neighboring tissue and perhaps other areas of the body in addition to the lung [2]. The majority of primary lung malignancies also referred to as lung cancers, were carcinomas. Small-cell lung cancer (SCLC), as well as Non-small cell lung cancer (NSCLC), are still the two primary kinds. Chronic cough (sometimes cough with blood), losing weight, difficulty breathing, as well as chest aches have been the most usual signs.

The national institute of health (NIH) defines lung cancer as "Cancer that forms in tissues of the lung, usually in the cells lining air passages, two main types are small cell lung cancer and non-small cell lung cancer". The cells' appearance below a microscope is used to diagnose various types [5]. More people develop lung cancer than every form of cancer. Lung cancer refers to a variety of malignant conditions (mostly carcinomas) that damage the lung as well as related organs. Histological categories, which are divisions of cell types, are used to categorize malignancy [7], [8]. There are differences in each cancer's symptoms, genesis, diagnosis, as well as treatment-acceptability. Since both require various forms of therapy, picocells (oat cells), as well as non-small cell cancers, are the main difference. The most extensively used histological classification would be that suggested mostly by World Health

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 1027-1035 | e-ISSN: 2347-470X

Organization (WHO). Squamous cell (30–50%) and large cell carcinoma (5–15%), as well as the extremely frequent adenocarcinoma, seem to be the main histologic lung malignancies (10-30 percent). Investigating how well computer-based analysis methods can expose as well as collect details regarding scene composition, such as the detection of components, entities, characteristics, as well as categories, is intriguing [9].

The outcomes of the data interpretation are affected by the performance of the CT scan generated from the diagnostic methodology [10]. Whereas if a professional cannot see the details concerning the nodule, the computers would not be able to recognize it because computerized approaches try to emulate the physician. According to the basic guideline of image processing, missing data cannot be recovered, hence missing means cannot be observed. Various methods were used to assess image quality, such as:

(a) Image Quality: It is the scale that compares the size of physical tissues to the size of pixels in an image (like a CT slice). These details on the physical anatomy of the human body are provided by image scanners.

(b) Image Contrast: It is a measure that gauges how distinctive different image elements are from one another. Imaging techniques could be created to artificially improve the contrasts or zone specificity. Contrasting in image processing refers to the variation in object class representation.

(c) Image Sensitivity: The capability of the imaging technique to improve the distinction among anatomical characteristics as well as non-anatomical characteristics is referred to as image sensitivity.

(d) Image Specificity: The capacity to identify diseases in visual information. The definition of the phrases sensitivity as well as specificity was amorphous, so they could refer to either an image or the result of an algorithm.

2. REVIEW OF LITERATURE

Image pre-processing is frequently required since a specific patient's CT scan has several ambiguities or inaccuracies. Lower contrast, motion artifact, as well as erroneous noise, are three potential sources of uncertainty inside a CT scan [1]. To minimize noise from the image anatomy or nodules, image processing techniques including the scale space approaches, median filter and optimum filters are used. Basic normalization could also help to remove image noise and ensure that the following scans are properly referenced. Low processing could be required with high-definition scans, CT scanner accuracy, as well as adherence to imaging protocol. However, there are several ambiguities and severe imaging quality declines in certain clinical investigations like the Early Lung Cancer Action Program (ELCAP).

Although pulmonary nodules often have a spherical appearance, they might be confused by nearby anatomical features like arteries as well as pleural membranes. Nodules can occur in different places inside the lung cells which can have different sizes and shapes, as shown on a CT scan. Nodules are typically divided into four categories: well-circumscribed, which refers to just a nodule that is located right in the respiratory system and is therefore not linked to the vascular system; vascularized, which refers to a nodule that is conveniently located inside the respiratory system and also has important correlations to the neighboring vessels; pulmonary tail, which refers to such a nodule that is close towards the pleural skin and is linked by a likely to have reduced [3].

The process of dividing categories in an image into contiguous as well as distinct areas is known as Image segmentation, and it is a key component of computerized image analysis [4]. For instance, a CT segment from such a thoracic scan might include lung tissues, portions of the heart, and the chest wall. Isolating the lung tissue is the purpose of division within that situation. The chest wall, heart muscle segments, and lung tissues can all be seen on a CT slice out of a thoracic scanning. Isolating the lung tissues may purpose of fragmentation throughout this situation. Image segmentation is a very ridiculous subject in computer vision and image processing vision, comparable to image processing. Three techniques for image segmentation can also be categorized: statistical, variational, and geometric. Statistical techniques simulate the image data and represent the region-processing as mappings from the original images [5]. Geometric approaches use representations of item shapes to categorize the details of an image. The use of variational methods, such as level sets, results in an underlying depiction of class boundaries as just an evolving curve or surfaces that slashes the target object there at the boundary (zero-level set). Both implicit as well as explicit models, such as level sets with 15 snakes as well as gradient vector flow, are used in variational techniques [6]. These techniques deliver similar results more quickly but with less involvement from people.

Nodule modeling, a method to separate the nodules from anatomic structures inside the lung cells is part of the module detection process [13]. Although not required, nodule identification is frequently used on lung cells following the segmentation stage. This method will overlook the other parts of both the chest as well as thorax, which may also have nodules. Because lung cancer is the primary concern, the nodule identification stage is often implemented following the segmentation of a detected region. Nodule modeling is an essential part of nodule detection. The method relies on predicting the grey-level distribution of such a templates model that used an array of nodules that was assembled through specialized feature extraction. Attributing a pathology toward the foundation as well as separated nodules was part of the nodule categorization process [14]. The result of computerized nodule detection is the early diagnosis of suspect nodules. The main goal is to create as well as evaluate a classifier using a real discriminating dataset of benign from malignant nodules was critical to the successful implementation of this phase [15].

The field of computational technology that deals with nodule identification and tracking through biomedical image processing is highly extensive. This research is focused to develop a CAD arrangement for identification as well as the division of pulmonary nodules inside lung CT scans. Since accurate segmentation of pulmonary nodules is very crucial

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 1027-1035 | e-ISSN: 2347-470X

for diagnoses as well as treatment of lung cancer, in this work an advanced segmentation algorithm is developed. Further, the lung nodules segmentation needs to be classified whether normal or cancerous followed by classification by two different types of classifiers.

The implementation of the proposed research includes Improved Random Walker (RWI) segmentation along with Random Forest (RF) classifier and K-Nearest Neighbors (KNN) classifier to identify cancer. Here, this analysis was done on the Lung Image Database Consortium (LIDC) datasets individually. This technique takes advantage of patch-based image segmentation as well as the size and form features of potential nodule possibilities.

3. PROPOSED METHODOLOGY

Figure 1 shows the block diagram of the proposed lung cancer detection method. Filtering, segmenting, as well as feature extraction are the three processes that make up the recognition system in its basic form. The images of lung cancer from the database could be selected as inputs since preserving these stages improves the precision and reliability of lung cancer diagnosis. RWI segmentation is used for segmentation, assisting in identifying the precise position of the diseased area in the input images. Classifiers are used to improve the accuracy of detection. The proposed methodologies are implemented for cancer detection in MATLAB 2018a software.

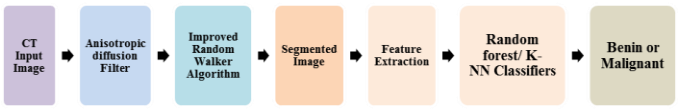

Figure 1: Block diagram of the proposed lung cancer detection method

Contextual clustering and a region-growing algorithm were first used to fragment the CT lung images [20]. The subsequent stage process of extracting characteristics from the GLCM. The third phase involves categorization using the k-NN as well as RF classifiers, which are two distinct types. The proposed model displays the implemented stages for the diagnosis of lung cancer.

It has been repeatedly demonstrated that filtration techniques can significantly influence the efficiency of the image pixels by removing unnecessary details and artifacts from the image pixel while also streamlining the presentation of complex anatomical systems. Like an optical pre-processing phase used by several discussed techniques, the idea of image filtration already was mentioned in the previous section.
The background is first smoothed out using a low pass filter convolution technique, then the patterns of concern are enhanced using a high pass filter, as well as finally the two obtained images are subtracted. The extensive variety of anatomical traits which must be taken into account was too much for this technology to handle. Many noise reduction, as well as image pre-processing algorithms, were done by developing partial differential equations (PDE).

Additionally, the amount to which PDE filters are applied to CT scans appears to be usually correlated with MRI as well as ultrasound. The classical nonlinear diffusion filter created by Rawal [11], which is centered on a PDE in a divergent format, is where anisotropic diffusion got its start.

It serves as a framework for recent advancements in multi-scale image recognition that reduce the visual appearance whilst boosting important characteristics like edges or coherence components. Whenever the amplitude of the gradients is considerable, the nonlinear anisotropic filters' primary contribution to enhancement is the softening along the isophote or across it [12].

3.1 Image Segmentation Algorithm Based on Random Walk

A specific amount of points, as well as edges, may make up the images according to the random walk (RW) theory [16]. The following equation will be generated as it was thought that images might be transferred to a weighted network,

$$T = (W, F) \qquad (1)$$

Here T represents the weighted graph in *equation (1)*, W signifies the node range, and F denotes the border range. A node, as well as an edge inside the images, were assigned to w and f, respectively, then w ∈ W & f ∈ W × W. The image's gray-scale mapping will be responsible for the edge's weight, hence, the following expression is developed using the Gaussian weight function,

$$V_{ij} = exp\,(-\,b\,(t_i - t_j)^2 \qquad (2)$$

Equation (2) with t_i signifies the pixel's luminance w_i, β indicates the weight coefficient, as well as V_{ij} corresponds to the likelihood of a stochastic process over the border of f_{it}. Due to similarities between the probabilistic solutions problem and the Dirichlet integral problem, it gets modified as,

$$G(0) = \frac{1}{2}\dot{o}.|\tilde{N}o|^2\,dW \quad (3)$$

Here *G(0)* represents the integral of *o* in *equation (3)*, ∇_o corresponds to the harmonics functional, and Ω signifies the image area. The Dirichlet integral issue could be transformed into a harmonic function to solve boundary conditions if the harmonic function could fulfill Laplace's equation. Edges linking the vertex with terminals of the image as well as edges linking the cells to the cluster centers are two different types of edges that make up an image. So, the following is one way of expressing the collection of edges,

$$F = N\,\dot{E}\,\{[q,S], [q,I]\} \qquad (4)$$

In *equation 4*, S stands for such terminal node, θ for the pixel point, as well as N for the connected neighborhood. As soon as the subset of most of the image's cut edges were identified as well as the values of such edges' total were minimized, the performance of the proposed method outcome of the image could be attained.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 1027-1035 | e-ISSN: 2347-470X

3.2 Algorithm of Improved Random Walker

Step 1: Input Image.
Step 2: Obtain seed point set (or labels), either interactively or automatically. We select 2 seed points interactively).
Step 3: Extract LBP Texture Features.
Step 4: Generate weights on the basis of image intensities and extracted LBP texture features.
Step 5: Build a Laplacian matrix. Solve random walker probabilities by solving the Dirichlet problem.
Step 6: Solve the random walker probabilities by solving the Dirichlet problem using *equation (5) or equation (6)* if one label:

$$L_U x^s = - B^T m^s \quad (5)$$

$$L_U X = - B^T M \quad (6)$$

Step 7: Assign a pixel to the label for which it has the highest probability.
Step 8: Remove small objects of fewer than 100 pixels.
Step 9: Fill holes.

Lung Segmentation is done using RWI. 13 shape features such as centroid, area, eccentricity, convex area, Euler number, Equiv diameter, extrema, extent, minor axis length, major axis length, perimeter, orientation, solidity; 7 GLCM texture geographies like entropy energy, contrast, cluster prominence, homogeneity, cluster shade as well as dissimilarity and 8 intensity features such as standard deviation, mean, variance, root mean square (RMS), kurtosis, smoothness, inverse difference moment (IDM) and skewness is computed. Intensity features are extracted from the principal component coefficients of the single-level discrete 2-D wavelet transform. Classification is done using RF and k-NN classifiers.

4. RESULTS AND DISCUSSIONS

Initially, convert all the DICOM images in the LIDC database into JPEG format and obtain a 534-training database and 150 testing databases containing both benign and malignant types of lung cancer images.
For testing purposes, select any of the testing sample images from the database and convert the image into a grayscale by eliminating the hue and saturation information while retaining the luminance. The enhancement technique is used to advance the interpretability of information in images for human viewers or to provide better input for other automated image processing techniques.

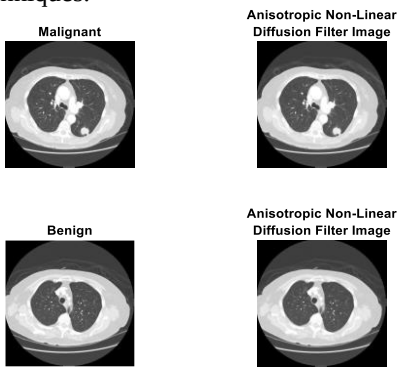

Figure 2: Anisotropic non-linear diffusion filtered image after preprocessing

Categories of image enhancement are of two types namely frequency and spatial domains. It is used in suppressing a low frequency in a frequency domain. The preprocessed images are shown in *figure 2*.

A classifier is an algorithm that effectively interacts for classification. A mathematical function carried out by a prediction model, which assigns input datasets to a category, is sometimes referred to it as a classifier. The term usage varies a lot between fields. The characteristics of observational data were also prominent as explanatory, independent variables or regressors, in statistics, where the classifier is frequently completed to logistic regression or indeed starts with a set, as well as the classifications to be anticipated were also regarded as consequences, which have been recognized to be probabilities of predictor variables. Considering machine learning, the wide varieties that can be anticipated are referred to as a class, observes were frequently referred to as the instances, as well as the independent variable were referred to as features (grouped into feature vectors). To classify data, RF, as well as k-NN classifiers, are now used.

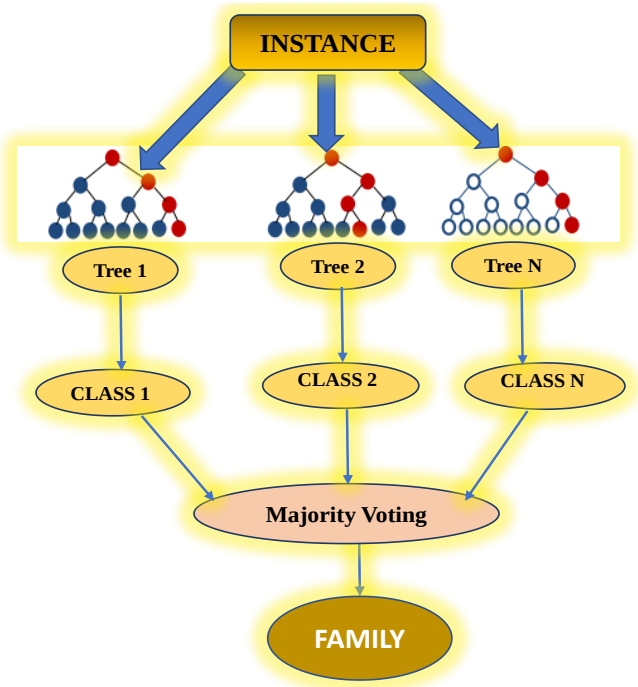

Figure 3: Random Forest Classifier

As depicted in *Figure 3*, RF algorithms are a collection of different classifiers created by mixing decision trees. Its decision tree structures are created from that level of unpredictability, which is a particularly significant feature of certain Ensembles of Classifiers [16],[17]. Depending on this concept, RF is often described as a general rule for randomized decision tree ensembles. RF's fundamental building block would be a binary tree created through recursive partitioning. The CART method, wherein binary splits iteratively divide the branch into uniform or nearly homogeneous terminal nodes [19], is frequently used to create the fundamental unit of RF trees. This technique suggests that

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 1027-1035 | e-ISSN: 2347-470X

an effective binary split should transfer information out of a parent tree node towards its two daughter nodes to increase the daughter nodes' homogenization relative to the parent node. RF is frequently a set consisting of a large number of trees, in which each tree was created via bootstrap sampling from original information [18].

Because they are produced non-deterministically but use a two-stage randomized process, RF trees vary from CART. A second randomization layer is added at the node level while building the tree, in contrast to randomization provided by bootstrap samples of the original information. RF merely chooses a random subset of the data at every node as well as utilizes these as runners to identify its optimum split for such node instead of dividing a tree node using those variables (predictors). The main goal of such two-step randomization will be to properly connect the decision trees to reduce variation in the ensembles. The non-parametric k-NN, which Cover as well as Heart developed in 1968, is being used for regression and classification [18].

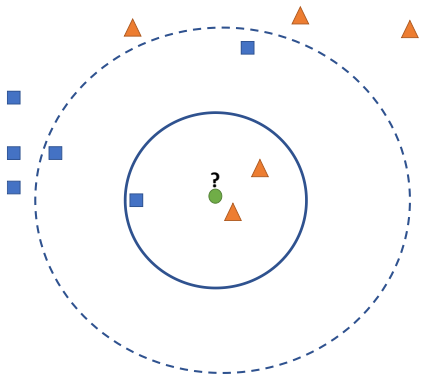

Figure 4: k-NN Classifier

Figure 4 illustrates how K-NN settles immediately from the training set. By exploring the complete dataset for such k most comparable neighbors and summing the results among those k values, a new vector can predict the future. This could be the mode class value inside a categorization situation and the mean output variable inside a regression case. A type of measure is used to identify which of the k vectors in the data were near the input.

In addition to this Manhattan, hamming, and Murkowski distances are frequently utilized for real datasets. When the input data were from the same datasets, Euclidean is utilized. If there are differences in the input datasets, Manhattan distance was applied. As the increasing size of the dataset, the k-NN calculation complexity rises.

Lung image is segmented into pieces and distributed as fundamental objects or areas. By producing more palatable results, feature extraction aims to alter how the image is represented. The practice of providing each image feature a label, which enables those pixels to have the same label that matches specific visual features, is crucial. Image segmentation was frequently used to locate objects as well as boundaries in images, such as lines and geometric.

Figure 5 shows a segmented image using RW and RWI approaches. Feature extraction is mainly used in pattern recognition and image processing. The feature is a repeated pattern of the image. The binarization method can be used in identifying the presence of lung cancer and extracting the interesting region from that image. An essential step in the process of detecting and isolating possibly different shapes as well as parts was feature extraction. Binarization is done based on the number of black-and-white pixel values. The binarization method is supported by the fact that such a count of black pixels seems to be significantly larger than the proportion of white pixels throughout usual lung images. By collecting the black pixels, usual as well as irregular pictures could be distinguished, which can serve as a useful threshold. If the count of black pixels inside the current picture is higher, it indicates that the image seems to be usual; or else, if the count of black pixels is still much lower, it indicates that such an image is unusual.

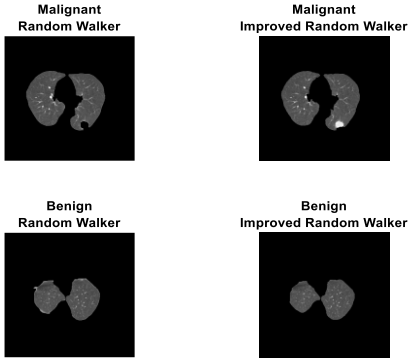

Figure 5: The output of RW and RWI for Malignant and Benign images

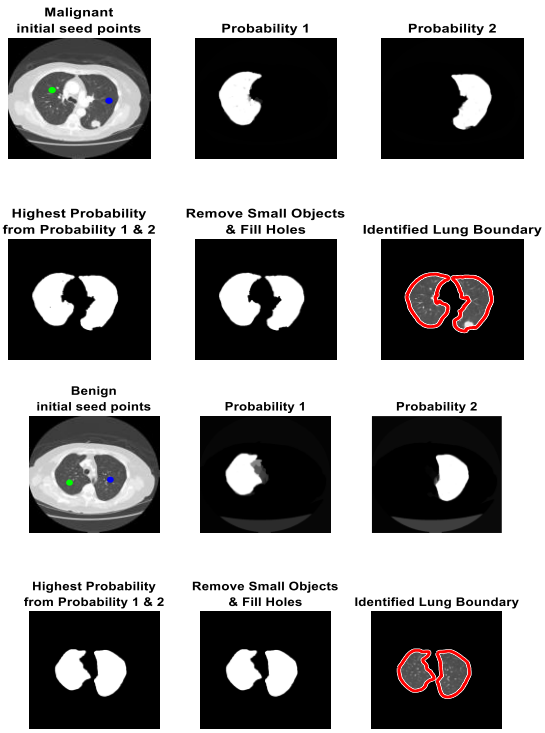

Figure 6: Binaries, cleared border, filled output of Malignant and Benign image

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 1027-1035 | e-ISSN: 2347-470X

Figure 6 shows the extracted image using the binarization method. Morphological operations are then carried out on the above binary images which erode binary images i.e., remove small objects from a binary image. Then following the dilation of the image and packed a group of surrounding pixels that could not be accessed by adding background color from the image's edge. Then, obtain the area of the segmented region and marked it using a disk shape to get a segmented binary image.

From the segmented grayscale image, compute the features such as 13 shape features, 7 GLCM texture features, and 8 intensity features. Intensity features are extracted from the principal component coefficients of the single-level discrete 2-D wavelet transform.

The 13 shape features extracted are minor axis length, eccentricity, area, Euler number, major axis length, centroid, convex area, Equiv diameter, extent, extrema, orientation, solidity, and perimeter depicted below *table 1.*

Table 1: Parameters of shape for the segmented region

S. No.	Parameters	Malignant	Benign
1	Area	21985	17756.5
2	Centroid	255.799	255.541
3	Convex Area	26610.5	19503
4	Eccentricity	0.8395	0.7624
5	Equiv Diameter	167.2869	150.2694
6	Euler Number	1	1
7	Extent	0.6548	0.7016
8	Extrema	241.9375	255.2813
9	Major Axis Length	245.2542	193.1335
10	Minor Axis Length	132.3811	123.7002
11	Orientation	-1.0283	89.1833
12	Perimeter	691.91	541.133
13	Solidity	0.8261	0.9094

The 7 GLCM features extracted are contrast, entropy energy, homogeneity, cluster prominence, cluster shade, and dissimilarity are shown in *table 2.*

Table 2: GLCM-based Texture Features

S. No.	GLCM parameters	Malignant	Benign
1	Contrast	0.0434	0.0358
2	Entropy	0.6178	0.5149
3	Energy	0.7081	0.7577
4	Homogeneity	0.9894	0.9915
5	Cluster Prominence	131.6792	48.2394
6	Cluster Shade	15.5465	8.6482
7	Dissimilarity	0.0257	0.0209

The intensity parameters like IDM, mean, RMS, standard deviation, smoothness, variance, skewness, and kurtosis, are shown in table 3. The uniformity of that image might also affect IDM. IDM may only receive a minimal advantage from heterogeneous states due to its weighting factor. As a consequence, the homogeneous image has a considerably larger IDM score than the heterogeneous one. Intensity features are extracted from the principal component coefficients of the single-level discrete 2-D wavelet transform.

Table 3: Intensity features

S. No.	Intensity features	Malignant	Benign
1	Mean	0.0019	0.0022
2	Standard Deviation	0.0597	0.0597
3	RMS	0.0598	0.0598
4	Variance	0.0036	0.0036
5	Smoothness	0.9732	0.9772
6	Kurtosis	64.4815	78.5694
7	Skewness	3.9375	4.7622
8	IDM	1.422	3.7434

From these calculated parameters, 28 features that contain 7 GLCM texture features, 8 intensity features, and 13 shape features are obtained. All 534 of the images inside the training dataset are subjected to the aforementioned procedures, but these characteristics are then utilized to develop the RF-based classifier that was created. There are 20 layers in the hidden nodes. For such investigation, the log-sigmoid transfer function is employed because it is much more appropriate than other transfer functions. The performance plots of RW and RWI with RF are shown in *Figure 5*. The designed feed-forward neural network obtained the matched input image's true label value with those existing trained image labels. Accuracy (AC), specificity (SP), and sensitivity (SE), three quality assessment parameters, can be employed to assess the classifier's performance. The capacity of the testing to properly predict a symptom is referred to as sensitivity. The capacity of problem-solving to appropriately exclude certain conditions is referred to as specificity. The percentage of correct classification determines how accurate a classifier is.

Nine performance metrics, such as false positive rate, error, accuracy, specificity, sensitivity, precision, kappa-cohen's kappa, F1 score, and Mathews correlation coefficient, are used to systematically assess the classification efficiency of the proposed technique values of RW and RWI with RF. The accuracy is the proportions of successfully predicted pixels [3]. It is expressed in the equation below,

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (7)$$

Sensitivity refers to the nodule variables percentage that is predicted, and precision is the percentage of input images that are predicted that is measured below,

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (8)$$

$$Specificity = \frac{TN}{(TN + FP)} \quad (9)$$

FPR refers to falsely described nodule pixels proportion as well as the fraction of wrongly described pixel values that are described below seems to be the false negative ratio (FNR),

$$FPR = \frac{FP}{(TP + TN)} \quad (10)$$

$$FNR = \frac{FN}{(TP + TN)} \quad (11)$$

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 1027-1035 | e-ISSN: 2347-470X

The overlapping score indicates resemblance measurements, which projects in what way the portion outcome principles suit that ground truth.

$$Overlap = \frac{TP}{(TP + FP + FN)} \quad (12)$$

TP, true positive = exactly found number nodule pixels. FP, False positive = nodule pixels incorrect found number. True positive negative = number of exact identification as background pixels. FN, False Negative = number of incorrect identification background images. Obtained outcomes for the 5 calculation measures ranging from 0 to 1. The lesser FPR, as well as the FNR segmentation performance, is excellent.

The confusion matrix is done for multiple classes from the actual Class labels and the Predict Class Labels [7]. The Two-Class of Confusion Matrix has *False Positive (*FP), *True Positive (*TP), True Negative (TN), and False Negative (FN) values. This performance evaluation of RW and RWI with RF is done by calculating the parameters of FPR-False positive rate, specificity, accuracy, error, sensitivity, Precision, F1_score, kappa-Cohen's kappa, and MCC-Matthews correlation coefficient, shown in *figure 7* and *table 4*.

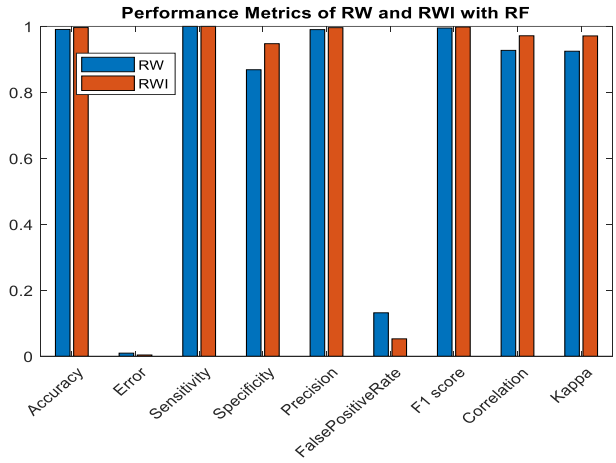

Figure 7: Performance evaluation of RW and RWI with RF

Table 4: Performance evaluation of RW and RWI with RF

S. No.	Performance analysis	RW	RWI
1	Accuracy	99.0637	99.6255
2	Error	0.9363	0.3745
3	Sensitivity	99.9999	99.9999
4	Specificity	86.8421	94.7368
5	Precision	99.002	99.5984
6	False positive rate	13.1579	5.2632
7	F1_score	99.4985	99.7988
8	Mathews Correlation Coefficient	92.7229	97.1372
9	Kappa	92.4589	97.0962

The rows of this confusion matrix represent the output class, or expected class, while the columns represent the true/actual class (Target Class). Those diagonal cells relate to accurately categorized information. The off-diagonal cells are associated with the observation that was misclassified. The fraction of correct and wrong categorization is displayed inside the confusion matrix. Upon that diagonal of the matrix, the green boxes represent correct diagnoses, while the red squares represent wrong categorizations. In each cell, the number of data points and the proportion of all observations were displayed. These percentages of most of the instances expected to belong to every class which is successfully and wrongly categorized are displayed in the columns on the utmost right of the graph. The accuracy (or positive predictive value), as well as false discovery rates, were common names for these measurements. The percentage of every instance within every category that is successfully and wrongly categorized is displayed in the rows there at bottom of such a plot. These measurements are frequently referred to as the false negative rate as well as recall, correspondingly. The accuracy level is displayed in the cells inside the plot's lower right corner.

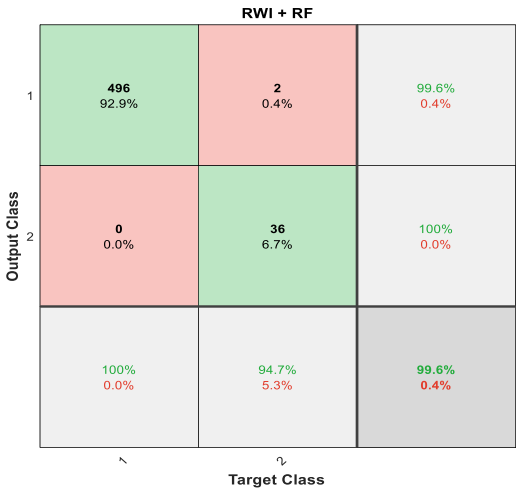

Figure 8: RWI Confusion Matrix with RF classifier

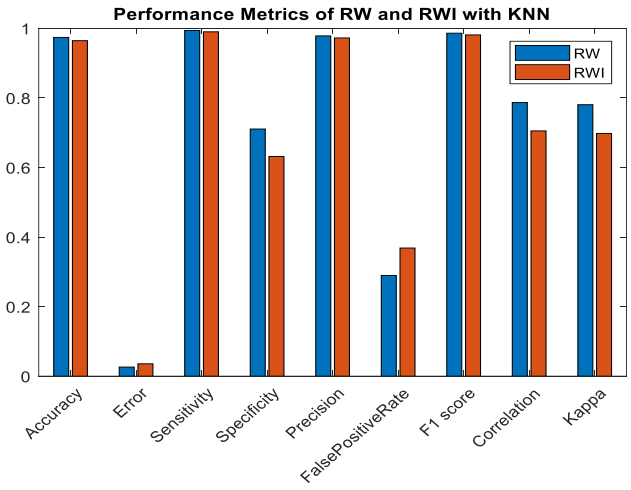

Figure 9: Performance evaluation with k-NN classifier

In *Figure 8*, the confusion matrix with the target as well as the output class is indicated. The trained network's count, as well as percentages of correctly classified instances, are displayed

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 1027-1035 | e-ISSN: 2347-470X

during the first diagonally cells. Of the 534 images, 496 are appropriately categorized under benign (the true positive). This is equivalent to 92.9% of the 534 images. Similar to this, 2 cases have been accurately identified as being malignant (true negative). This is equivalent to 0.4% of all the images. The overall accuracy for RWI with RF classifier is 99.6% with an error rate of 0.4%.

Table 5: Performance evaluation values of ADT and ADTM with SVM

S. No.	Performance analysis	RW	RWI
1	Accuracy	97.3783	96.4419
2	Error	2.6217	3.5581
3	Sensitivity	99.3952	98.9919
4	Specificity	71.0526	63.1579
5	Precision	97.8175	97.2277
6	False positive rate	28.9474	36.8421
7	F1_score	98.6	98.1019
8	Mathews Correlation Coefficient	78.6546	70.5063
9	Kappa	78.0324	69.7802

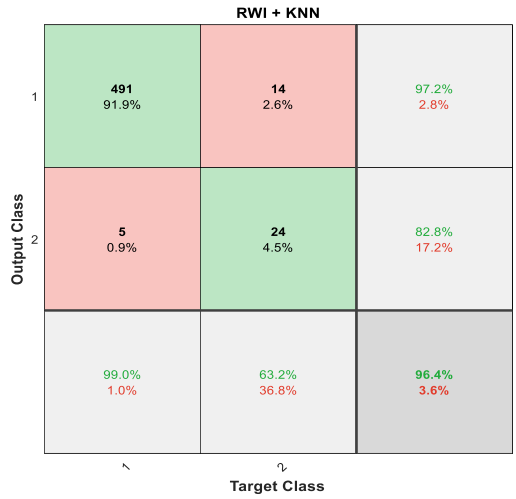

Figure 10: Confusion Matrix for RWI with k-NN

The performance evaluation with the k-NN classifier is shown in *figure 9* and *table 5.* In *Figure 10*, the confusion matrix for RWI with k-NN with target class and output class is indicated. Out of 534 images, 491 images were benign (the true positive). This equates to 91.9 percent of the 534 total images. Similar to that, 14 cases that have been accurately identified are malignant (true negative). This is equivalent to 2.6% of all images. 0.9 percent, or 5 out of the cancerous images, are mistakenly labeled as benign (false positive). In such a similar vein, 24 benign images or 4.5 percent of total data were mistakenly categorized as cancer (False negative). The overall accuracy for RWI with the k-NN classifier is 96.4% with an error rate of 3.6%.

5. CONCLUSIONS

The important components of this proposed approach are *(i)* image preprocessing, which makes-smooth the imaging and eliminates speckle noise; *(ii)* segmenting a cancer nodule from a CT scan; *(iii)* feature extraction, which acquires information from images such as area, centroid, perimeter, diameter, mean intensity, and eccentricity. The classification module determines the identified nodule as benign or malignant using the trained classification algorithm. The probability of moving out of each unmarked pixel to every labeled pixel is computed as well as vectors of possibilities are constructed for every unmarked pixel by using the enhanced random walk method, considering a dataset of user-specified (pre-labeled) pixels as labeling. The accuracy of the proposed method for lung cancer detection using the RF classifier is 99.6% and the k-NN classifier is 96.4%, which indicates that the RF classifier offers high accuracy with the RWI algorithm for lung cancer detection. Future studies will mainly focus on rating the images due to the severity of pulmonary nodule malignancy, which is crucial for the detection and management of lung cancer in clinical settings.

REFERENCES

[1] Alhaj, M. A. and Maghari, A. Y. 2017 Cancer survivability prediction using random forest and rule induction algorithms. IEEE International Conference on Information Technology (ICIT), pp. 388-391.
[2] C. Society, Cancer facts and figures 2013. American Cancer Society Atlanta, 2013.
[3] Chauhan, D. and Jaiswal, V. 2016 an efficient data mining classification approach for detecting lung cancer disease. International Conference on Communication and Electronics Systems (ICCES), pp. 1-8.
[4] Chauhan, R., Kaur, H. and Chang, V. 2017 Advancement and applicability of classifiers for variant exponential models to optimize the accuracy for deep learning. Journal of Ambient Intelligence and Humanized Computing, pp. 1-10.
[5] Hazapi, O., Lagopati, N., Pezoulas, V. C., Papayiannis, G. I., Fotiadis, D. I., Skaltsas, D. and Gorgoulis, V. G. 2022 Machine Learning: A Tool to Shape the Future of Medicine. In Handbook of Machine Learning Applications for Genomics, pp. 177-218.
[6] Ilunga–Mbuyamba, E., Avina–Cervantes, J. G., Cepeda–Negrete, J., Ibarra–Manzano, M. A. and Chalopin, C. 2017 Automatic selection of localized region-based active contour models using image content analysis applied to brain tumor segmentation. Computers in biology and medicine, 91: 69-79.
[7] Kumar, V. 2021 Evaluation of computationally intelligent techniques for breast cancer diagnosis. Neural Computing and Applications, 33(8): 3195-3208.
[8] Monirujjaman Khan, M., Islam, S., Sarkar, S., Ayaz, F. I., Ananda, M. K., Tazin, T. and Almalki, F. A. 2022 Machine Learning Based Comparative Analysis for Breast Cancer Prediction. Journal of Healthcare Engineering.
[9] Munir, K., Elahi, H., Ayub, A., Frezza, F. and Rizzi, A. 2019 Cancer diagnosis using deep learning: a bibliographic review. Cancers, 11(9): 1235.
[10] Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P. and Green, R. 2019 Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. Academic pathology, 6: 2374289519873088.
[11] Rawal, R. (2020) Breast cancer prediction using machine learning. Journal of Emerging Technologies and Innovative Research (JETIR), 13(24): 7.
[12] Rostami, M., Forouzandeh, S., Berahmand, K., Soltani, M., Shahsavari, M. and Oussalah, M. 2022 Gene selection for microarray data classification via multi-objective graph theoretic-based method. Artificial Intelligence in Medicine, 123: 102228.
[13] Roy, J., winter, C., Isik, Z. and Schroeder, M. 2014 Network information improves cancer outcome prediction. Briefings in bioinformatics, 15(4): 612-625.
[14] Shukla, A. K., Singh, P. and Vardhan, M. 2019 A new hybrid wrapper TLBO and SA with SVM approach for gene expression data. Information Sciences, 503: 238-254.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 1027-1035 | e-ISSN: 2347-470X

[15] Silva, F., Pereira, T., Neves, I., Morgado, J., Freitas, C., Malafaia, M. and Oliveira, H. P. 2022 towards Machine Learning-Aided Lung Cancer Clinical Routines: Approaches and Open Challenges. Journal of Personalized Medicine, 12(3): 480.
[16] Tie, J., Lei, X. and Pan, Y. 2021 Metabolite-disease association prediction algorithm combining DeepWalk and random forest. Tsinghua Science and Technology, 27(1): 58-67.
[17] Timilsina, M., Tandan, M. and Nováček, V. 2022 Machine learning approaches for predicting the onset time of the adverse drug events in oncology. Machine Learning with Applications, 100367.
[18] Trainor, P. J., DeFilippis, A. P. and Rai, S. N. 2017 Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. Metabolites, 7(2): 30.
[19] Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y. and Jin, Y. 2020 An improved random forest-based rule extraction method for breast cancer diagnosis. Applied Soft Computing, 86: 105941.
[20] Xu, L., Tetteh, G., Lipkova, J., Zhao, Y., Li, H., Christ, P. and Menze, B. H. 2018 Automated whole-body bone lesion detection for multiple myeloma on 68Ga-pentixafor PET/CT imaging using deep learning methods. Contrast media & molecular imaging, 2018.