ORIGINAL ARTICLE

Evaluation of Radiomics and Machine Learning for Classifying Pulmonary Nodules in CT Images

Arooj Nissar [1*], **A H Mir** [2]

[1] *Department of Information Technology, National Institute of Technology Srinagar, 190006, India*

[2] *Department of Electronics and Communicatiopy, National Institute of Technology Srinagar, 190006, India*

*Corresponding Author: Arooj Nissar
Email: arooj@nitsri.ac.in

Received: 02 September 2023 / Accepted: 15 May 2024

Abstract

Purpose: Lung cancer is a deadly disease that has high occurrence and death rates, worldwide. Clinicians are widely using computed tomography imaging for the detection of lung cancer. Radiomics extracted from medical images together with a machine learning platform has given encouraging results in lung cancer diagnosis. Therefore, this study is proposed with the aim of efficiently applying and evaluating radiomics and ML techniques to classify pulmonary nodules in CT images.

Materials and Methods: Lung Image Data Consortium is utilized in which nodules are given malignancy scores 1 through 5 i.e. benign through malignant. Three scenarios are created using these scores: G54 Vs G12, G543 Vs G12, and G54 Vs G123. Radiomics is extracted using Shape, Gray Level Co-occurrence Method, Gray Level Difference Method, and Gray Level Run Length Matrix along with Wavelet Packet Transform. To select a relevant set of features, four techniques i.e. Chi-square test, Analysis of variance, boosted ensemble classification tree and bagged ensemble classification tree are applied. The classification of nodules into benign or malignant is evaluated by using six models of support vector machine.

Results: The results, in Scenario 1, show that CGSVM+Chi-square yields the best sensitivity of 81.4%. In Scenario 2, LSVM+ANOVA yields the best sensitivity of 80.5% compared to the rest of the models, and in Scenario 3, FGSVM+BACET gives the best sensitivity of 72.3% compared to the rest of the models.

Conclusion: Overall, the study demonstrates that the radiomics and feature selection methods employed in combination with the different support vector classifiers performed significantly and achieved decent results for the classification of CT pulmonary nodules. The outcome thus can help the clinicians to diagnose, and make better decisions and treatments.

Keywords: Lung Cancer; Lung Image Data Consortium; Radiomics; Support Vector Machine; Feature Selection; Machine Learning.

PROOF

Copyright © 2025 Tehran University of Medical Sciences.
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

DOI:

Frontiers in BIOMEDICAL TECHNOLOGIES

1. Introduction

Lung Cancer (LC) is a disease impacting both male and female populations worldwide. It has occupied the second most places among all the types of cancers, having 2.21 million cases and the rate is gradually increasing. The condition in which there is an uncontrollable abnormal growth of the cells in lung tissues is referred to as nodule and slowly it spreads to other organs, too. Many factors including smoking, drug intake, and inhalation of harmful substances produced by industries and vehicles are the main cause of LC [1]. The major impact of LC is seen in people with age over 70 years while a small number of people detected with this disease age less than 45 [2]. In a report provided by the World Health Organization (WHO) [3], about 1.80 million deaths are caused just because of LC. A report on USA statistics from the period (2011-2015) revealed that 439.2 per 100,000 cases, on average, were recognized and 163.5 per 100,000 persons lost their lives each year due to LC. In the UK also, every year, approximately, 44,500 cases are diagnosed with LC [4].

For early detection of LC, Pulmonary Nodules (PNs) are primarily focused as they provide a direct picture of cancer spread. A lung nodule comprises a round lesion having a diameter of ≥ 3 cm. It can be benign which is non-cancerous or malignant which is often referred to as cancerous [2]. High mortality increases dramatically in the presence of malignant lung nodules whereas the patient's survival rate is high with benign lung nodules. The early and accurate diagnosis of LC requires proper differentiation between benign and malignant nodules [5]. One of the crucial hurdles in the detection of LC is that it doesn't show any symptoms in the early stages. Many of the cases come into knowledge or are discovered by doctors when LC reaches its advanced stage and curing the disease becomes very difficult at that time. Several clinical techniques are available to detect LC such as radiology and blood tests, endoscopy, biopsies, X-ray imaging, etc. Among these, the Computed Tomography (CT) technique is a highly adopted modality used for LC diagnosis as it provides fast results without any pain and provides in-depth details about tumor location, size, shape, etc. [4]. However, these clinical measures are effective but perform only subjective analysis and have a high risk of occurrence of human error due to manual evaluation by radiologists [6]. Hence, using the capabilities of Computer-Aided Diagnosis (CAD) is crucial in assisting medical practitioners with the detection of tumors and the proper classification of lung nodules as either benign or malignant.

The utilization of radiomics has proven its efficacy in LC diagnosis as it can extract a large number of quantitative image features [7]. Radiomics is a quantitative approach that applies data-characterization algorithms whose purpose is to improve the already available data using mathematical analysis [5, 8]. Radiomics and advanced learning approaches can be used in combination to perform an accurate diagnosis of LC. The introduction of Machine Learning (ML) in healthcare has changed the face of disease diagnosis. ML algorithms have the greater capability to deal with different types of data and produce classification output with high accuracy. Parmatasari *et al.* [9] applied a Support Vector Machine (SVM) to classify LC and yielded an accuracy of 85.63%. In another study, Abbas *et al.* [10] proposed an automated system to classify LC into benign and malignant and the implication of SVM achieved the highest accuracy.

In radiomics, we can get features from 2D Regions of Interest (ROI) and/or 3D Voxels of Interest (VOI). The proposed study aims to evaluate the performance of diagnostic systems by applying 2D radiomics and ML approaches for the diagnosis of cancer from lung nodules using CT images. The approach employs the selection of the most suitable radiomics features for classification. Various versions of SVM are evaluated through various feature selection methods under different scenarios. The performance of the model is evaluated using metrics to find the best one. The presented framework is useful and reliable in the successful classification of lung tumors as benign or malignant.

1.1. Related Work

Shakir *et al.* [8] developed radiomics-driven models to classify lung, colon, neck, and head cancer using CT images. Analytical radiomics signatures from lung nodules were extracted and derived from 105 3-D features. These signatures were incorporated into the regression model for tumor classification. Validation on 265 public datasets demonstrated high classification rates, indicating the robustness of the models. The study suggested the successful development of diagnostic mathematical functions for cancer diagnosis based on general tumor phenotype. Belfiore *et al.* [11] examined Non-Small Cell Lung Cancer (NSCLC) CT scan radiomics characteristics'

PROOF

resilience among segmentation approaches. Expert radiologists segmented three 3D-ROIs to analyze radiomics characteristics in 48 NSCLC patients. The Intra-class Correlation Coefficient (ICC) measured feature agreement and calculation parameter sensitivity. 'Shape' characteristics demonstrated good agreement (ICC>0.9) and little parameter sensitivity. A subset of 'first-order' and 'second-order' characteristics showed good agreement. The study found that certain radiomics properties can significantly improve NSCLC CT scan repeatability. Padmakumari *et al.* [12] tested CT radiomics for its ability to discriminate LC from Tuberculosis (TB) in low-income nations without lung biopsies. Radiomics characteristics were derived from 3D segmented CT images of histologically proven TB or LC patients' chests. Clinical and radiomics differences between LC and TB were significant. Radiomics may enhance resource-limited oncological patient treatment by identifying these illnesses non-invasively. However, prospective studies are needed to confirm these findings.

Radiomics [5] was used in cancer diagnosis, prognosis, and therapy response prediction by Chen *et al.* A 4-feature signature was used to classify lung nodules using radiomics and CT images. In 72 individuals with 75 PNs, benign and malignant lesions differed in 76 out of 750 imaging characteristics. The radiomics signature classified benign or malignant nodules with 84% accuracy, 92.85% sensitivity, and 72.73% specificity. The study found that radiomics can enhance lung nodule categorization non-invasively. The study in [13] developed a radiomics nomogram using wavelet characteristics to differentiate between malignant and benign early-stage lung nodules for high-risk screening purposes. The training set (N = 70) and validation set (N = 46) of 116 patients were considered with early-stage solitary PNs of size 3 cm. Standard CT pictures were used to extract each patient's radiomics characteristics. Using a multivariate logistic regression model, the researchers generated a radiomics nomogram with an Area Under the Curve (AUC) of 0.9406, accuracy of 95%, and Confidence Interval (CI) of (0.8831-0.9982) in the training set, and an AUC of 0.8454, accuracy of 95% CI: 0.7196-0.9712) in the validation set. Donga *et al.* [3] used modified gradient boosting ML to classify pulmonary nodules in CT images. They preprocess CT images, segment nodule borders, extract intensity and texture data, and train/test the modified gradient boost classifier to discriminate benign from malignant nodules. The suggested framework achieves good precision, recall, F1 score, and validation accuracy on the LIDC-IDRI dataset (0.957%, 0.91, 0.941, and 95.67%). Comparative research shows that suggested technique classifies benign or malignant lung nodules better.

The study in [14] designed a computerized system trained on samples of Colorectal Cancer (CRC) tissue to distinguish between eight distinct types of CRC. Visual descriptors such as local binary patterns, wavelet transforms, and Gabor filters were used to generate 532 pathomics characteristics incorporated into the system. Scale affects CRC tissue differentiation, as shown by a thorough analysis of wavelet families and characteristics. With tenfold cross-validation, the model outperformed previous research with an accuracy of 95.3%. Importantly, the research confirmed that classification performance was preserved when applying wavelet approximations at the first and second levels. Khehrah *et al.* [6] automate lung nodule identification using CT scans. Grayscale histograms and morphological techniques isolate lung regions and extract interior features. A threshold-based method isolates candidate nodules. Statistical and shape-based characteristics from nodule candidates produce feature vectors categorized by SVM. The method's 93.75% sensitivity on a large lung CT dataset (LIDC) outperforms comparable approaches. The framework improved lung nodule identification and diagnosis. SVM classification using GLCM and RLM features is used to identify lung cancer by Permatasari *et al.* [8]. The study classifies 500 Cancer Imaging Archive Database CT pictures into normal and LC clusters. The study investigated image preprocessing, region of interest (ROI) segmentation, and feature extraction. Default SVM classification accuracy is 85.63%.

Torres *et al.* [15] experimented Feed forward networks generalized radiomic CT scan nodule features. They suggested incorporating statistically important radiomic features for malignancy detection to improve repeatability with limited training data. The best model identified malignancies with 100% sensitivity and 83% specificity (AUC = 0.94) in an independent patient population. Alzubaidi *et al.* [4] developed a comprehensive and comparative

PROOF

PROOF

methodology for LC diagnosis utilizing CT scan images, covering global and local aspects. 1000 CT scan pictures were preprocessed by warping and cropping. Global and local features' training and testing make up the framework. Global features from ten image feature categories are extracted to provide feature vectors for six machine-learning algorithm detection models. Gabor Filter, Haar Wavelet feature, and Histogram of Oriented Gradients (HOG) outperform others, while SVM outperforms learning techniques. SVM with Haar Wavelet, HOG, and Gabor Filter features achieves 90% accuracy, 88% sensitivity, and 97% specificity, outperforming global approaches.

2. Materials and Methods

2.1. Data Set

This research work is proposed to execute the classification of lung nodules in CT images as benign or malignant using radiomics, feature selection, and SVM. The strategy comprises different stages including dataset collection, feature extraction, feature selection, classification, and performance assessment. Firstly, the dataset is acquired from an online repository of CT images, and preprocessing is done to improve the quality of the image. Then the features are extracted from images using shape and texture analysis on images directly and on multi-spectral images as well. Fourthly, filter and embedded-type feature selection methods are employed to select relevant features. At last, classifiers are used and the performance of each model is analyzed using various evaluation measures.

A dataset plays a vital role in the diagnostic system. In this work, CT images from the Lung Image Data Consortium (LIDC) database are utilized. This LIDC database has 1018 CT patient cases along with four experienced radiologists' ground truth reports. The Malignancy Score (MS), 1 through 5, of nodules ≥ 3mm and the annotations accorded by radiologists are described in detail in [16, 17, 18]. In this study, random 160 cases were used. The slice count varied in the range of 110-388. A total of 4157 DICOM slices of CT scans were hence collected and considered. The nodules in these CT slices with different MS i.e. score 1 indicating benign, score 2 likely benign, score 3 indeterminate, score 4 likely to be malignant, and score 5 highly likely to be malignant were separated. (Table 1a).

Three scenarios were created: Scenario 1 (G54 Vs G12), Scenario 2 (G543 Vs G12), and Scenario 3 (G54 Vs G123). In Scenario 1, nodules with MS 5 and 4 were taken as malignant and that of 3 and 4 were taken as benign. We discarded lung nodules with MS 3 to lessen the consequences of an indeterminate assessment of nodule malignancy. Thus there are 1703 malignant and 1265 benign nodules. Again, Scenario 2 and 3 were created so that indeterminate and uncertain nodules are grouped as malignant and benign ones, respectively, to assess the effect of nodules with malignancy suspicion on the proposed model's performance. Accordingly, 2892 malignant and 1265 benign nodules are grouped in Scenario 2, and 1703 malignant and 2454 benign nodules are categorized in Scenario 3 (Table 1b). Some of the samples from the LIDC dataset with different MS are shown in Figure 1.

Table 1a. Lung cancer score with respective meaning [18] and number of ROI's considered

Malignancy Score	Meaning	# of nodule ROIs
1	Benign	324
2	Likely Benign	941
3	Intermediate	1189
4	Likely Malignant	820
5	Malignant	883

Table 1b. Distribution of nodules with malignancy scores in 3 different Scenarios

Dataset		Malignant	Benign
Senario 1	Malinancy Score # of nodules	5 4 883 820	1 2 324 941
	Aggregate	1703	1265
Senario 2	Malinancy Score # of nodules	5 4 3 883 820 1189	1 2 324 941
	Aggregate	2892	1265
Senario 3	Malinancy Score # of nodules	5 4 883 820	1 2 3 324 941 1189
	Aggregate	1703	2454

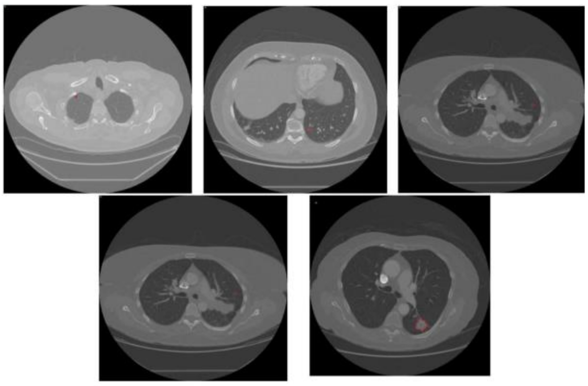

Figure 1. LIDC dataset sample images with ROI's having malignancy score: 1 to 5

In this work, median filtering is performed as the pre-processing step to remove redundant noise from the data. A median filter is a non-linear filter and is widely used to remove noise from images. The framework of the proposed methodology is illustrated in Figure 2.

2.2. Feature Extraction

Feature extraction is performed on the entire dataset. In this work, radiomics based on texture and shape features are extracted using statistical techniques. Initially, from annotations of the radiologist, the ROI of nodules is obtained. Shape features of all nodules are extracted. A sub-image of 11×11 pixels is selected around the centroid of each nodule and texture analysis is carried out. An overview of these features is briefly described as:

2.2.1. Shape Features

The classification process relies significantly on several shape factors. These characteristics are critical since they are directly related to the identification and prognosis of cancer [19]. Seven such features are extracted namely Area, Perimeter, Major-axis-Length, Minor-axis-Length, Max-Intensity, Mean-Intensity, and Min-Intensity. The list of these extracted features is given in Table 2.

2.2.2. Texture Features

Texture analysis is a method for image analysis and classification [20]. It is a way of describing the spatial distribution of intensities [21] hence enabling the description of tissue heterogeneity, a property believed to influence the outcome of cancer treatment [22]. In this work, Haralick's texture features are calculated as per the equations given in [20] from GLCM, GLDM, and GLRLM [6, 9, 10, 20, 23]. The second-order statistical method counts the relationship between two surrounding pixels in GLCM and GLDM whereas high-order features employ a run-length matrix such as GLRLM [20]. 88, 20, 44 features are extracted using GLCM, GLDM, and GLRLM, respectively, taking 4 directions (θ) and inter-pixel distance (d) of one into consideration. Using GLCM twenty-two texture features are computed viz. Autocorrelation (ACOR), Contrast (CON), Correlation 1 (COR1), Correlation 2 (COR2), Cluster

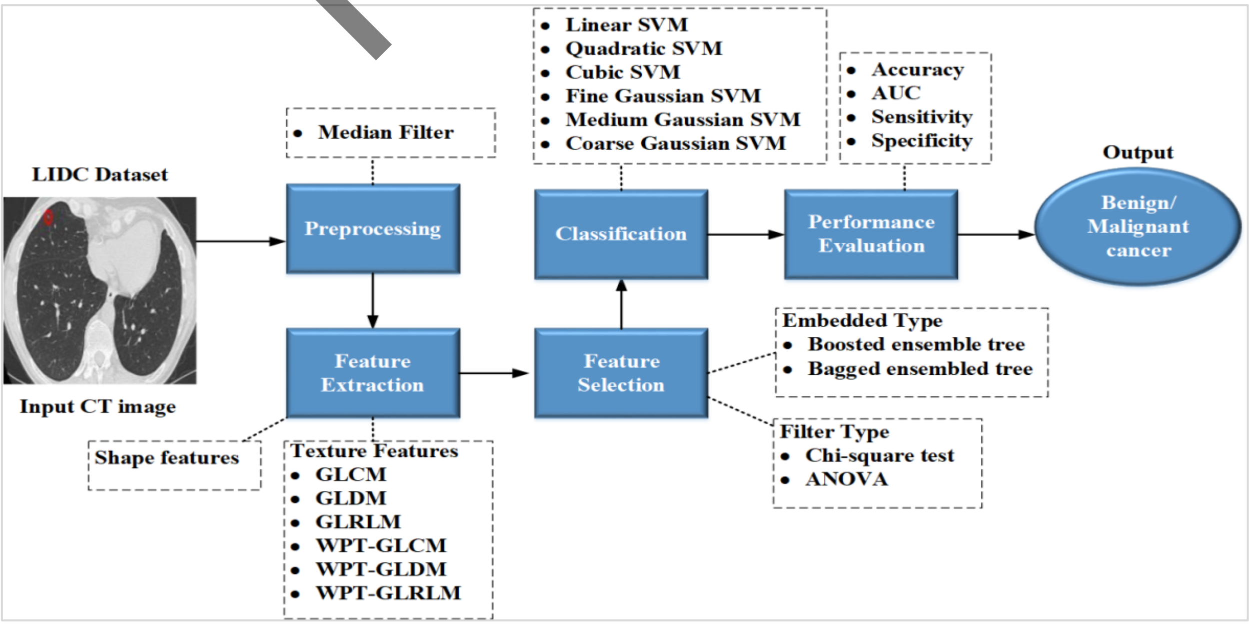

Figure 2. Proposed framework to diagnose lung cancer

PROOF

Table 2. List of features per class

GLCM	Autocorrelation (ACOR), Contrast (CON), Correlation1 (COR1), Correlation2 (COR2), Cluster Prominence (CP), Cluster Shade (CS),Dissimilarity (DS), Energy (ENR), Entropy(ENT), Homogeneity1 (HMG1), Homogeneity2 (HMG2), Maximum Probability (MP), Sum of Squares: Variance(SOS), Sum Average (SA), Sum Variance (SV), Sum Entropy (SE), Difference Variance (DV), Difference Entropy (DE), Information Measure of Correlation1 (IMC1), Information Measure of Correlation2 (IMC2), Inverse Difference Moment(IDM), Inverse Difference Moment Normalized (IDMN)
GLDM	Contrast (CON), Angular Second Moment (ASM), Entropy (ENT), Mean, Inverse Difference Moment (IDM)
GLRLM	Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray Level Non-uniformity (GLN), Run Length Non-uniformity (RLN), Run Percentage (RP), Low Gray-Level Run Emphasis (LGRE), High Gray-Level Run Emphasis (HGRE), Short Run Low Gray-Level Emphasis (SGLGE), Short Run High Gray-Level Emphasis (SRHGE), Long Run Low Gray-Level Emphasis (LRLGE), Long Run High Gray-Level Emphasis (LRHGE)
Shape Features	Area, Perimeter, MajorAxisLength, MinorAxisLength, Max_Intensity, Mean_Intensity,Min_Intensity

Prominence (CP), Cluster Shade (CS), Dissimilarity (DS), Energy (ENR), Entropy (ENT), Homogeneity 1 (HMG1), Homogeneity 2 (HMG2), Maximum Probability (MP), Sum of Squares: Variance (SOS), Sum Average (SA), Sum Variance (SV), Sum Entropy (SENT), Difference Variance (DV), Difference Entropy (DENT), Information Measure of Correlation1 (IMC1), Information Measure of Correlation 2 (IMC2), Inverse Difference Moment (IDM), Inverse Difference Moment Normalized (IDMN).

Five texture features; Contrast (CON), Angular Second Moment (ASM), Entropy (ENT), Mean (M), Inverse Difference Moment (IDM) are computed from GLDM.

Also, using GLRLM eleven features are computed namely Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray Level Non-uniformity (GLN), Run Length Non-uniformity (RLN), Run Percentage (RP), Low Gray-Level Run Emphasis (LGRE), High Gray-Level Run Emphasis (HGRE), Short Run Low Gray-Level Emphasis (SGLGE), Short Run High Gray-Level Emphasis (SRHGE), Long Run Low Gray-Level Emphasis (LRLGE), Long Run High Gray-Level Emphasis (LRHGE). The list of features extracted is given in Table 2 and the list of feature classes along with the no of features extracted in each class is provided in Table 3.

2.2.3. WPT Features

The discrete wavelet transform (DWT) is a multi-leveled sub-band framework which decomposes an image into the approximation image (LL) and details images (LH, LV, LD). The approximation sub-band, LL is then decomposed further into a second level of approximation and details, and so on. WPT is an extension of Discrete Wavelet Transform (DWT) where decomposition is carried on both approximations and details into a further level of approximations and details [24, 25]. In this proposed scheme, a two-level WPT is performed, as shown in Figure 3. There is no need to perform a deeper decomposition because, after the second level, the size of the image becomes too small, and no more valuable information is obtained [24]. The second level of decomposition provides one image of approximation and 15 images of details which are displayed in Figure 3. A comprehensive description of realization and equations used are provided in [24- 26]. In this work, Daubechies wavelet family, db1, db2, and db3, introduced by Daubechies [27], are applied to implicate WPT on each of the sub-images. As discussed, this step generates 16 sub-images whose

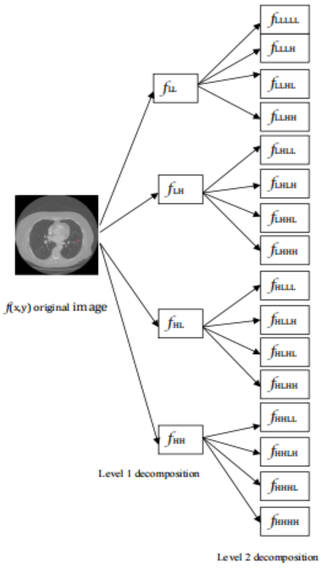

Figure 3. Block diagram of WPT

PROOF

Table 3. List of feature classes and feature count per class

Feature class			No of the features extracted			Total
Shape			7			7455
GLCM	Directions (θ)	*0^0*	22	88	152	
		45^0	22			
		90^0	22			
		135^0	22			
GLDM		*0^0*	5	20		
		45^0	5			
		90^0	5			
		135^0	5			
GLRLM		*0^0*	11	44		
		45^0	11			
		90^0	11			
		135^0	11			
WPT-GLCM	WPT family (Level=2)	*db1*	88*16	1408	4224	
		db2	88*16	1408		
		db3	88*16	1408		
WPT-GLDM		*db1*	20*16	320	960	
		db2	20*16	320		
		db3	20*16	320		
WPT-GLRLM		*db1*	44*16	704	2112	
		db2	44*16	704		
		db3	44*16	704		

PROOF

texture was re-analyzed using texture analysis techniques (3.2.2) and are denoted as WPT-GLCM, WPT-GLDM, and WPT- GLRLM. The list of features extracted is given in Table 2 and the list of feature classes along with the no of features extracted in each class is provided in Table 3.

2.3. Feature Selection

ML models benefit from Feature Selection (FS), which aims to extract only the most informative features and remove noisy non-informative irrelevant, and redundant features [28]. The FS that are routinely used are grouped into three methodological categories: Filter Type FS (FTFS), Wrapper Type FS, and Embedded Type FS (ETFS) methods. FTFS methods use feature ranking as the evaluation metric for FS. In this work, four algorithms were used for all three scenarios. Two FTFS methods, Chi-square tests and the Analysis of Variance (ANOVA), which have proven significant to the detection of lung nodules using radiomics and ML [29] are used. Also, two ETFS methods, Boosted Classification Ensemble Tree (BOCET) and Bagged Classification Ensemble Tree (BACET) are used. The ETFS entails integrating the feature selection process directly into the model training process [30].

2.3.1. Chi-Square (χ^2) Test

χ^2 tests are statistical tests used to determine if categorical variables are significantly associated. The calculated χ^2 statistic can be compared against a critical value from the chi-square distribution with degrees of freedom determined by the number of categories in the feature and target variables. If the calculated χ^2 value exceeds the critical value, it indicates a significant association between the feature and the target, suggesting that the feature is relevant for classification or prediction [31]. Features with a high χ^2 value and a low p-value are selected for further analysis because they are deemed more pertinent to the task. The mathematical formula for calculating the χ^2 statistic for a single cell is as follows (Equation 1):

$$\chi^2 = \frac{((O - E)^2)}{E} \tag{1}$$

where χ^2 is the Chi-Square statistic for a specific cell. '*O*' is the observed frequency in the cell (intersection of a feature category and a target category). '*E*' is the expected frequency in the cell under the assumption of independence. The expected frequency *E* is calculated using the following formula (Equation 2):

$$E = \frac{Row\ Total\ *\ Column\ Total}{Total\ Observations} \qquad (2)$$

Higher Chi-Square values suggest a stronger association between the feature and the target, which can indicate the relevance of the feature for classification or prediction tasks. For detecting lung nodules, this technique identifies which radiomic characteristics have a significant correlation with the presence or absence of cancer.

2.3.2. Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical technique used to examine the differences between group means in a dataset [32]. In the context of lung nodule detection using radiomics and ML, ANOVA assists in evaluating the variability of radiomics features across distinct classes or groups, such as nodules and non-nodules. Features that demonstrate significant variability between these categories are regarded essential for differentiating them and thus are selected for further analysis as shown in Table 4. The variations between the sample mean, as well as the variation within each of the samples, are computed. Higher F-statistic values indicate greater variation between groups and suggest that the feature is relevant for differentiating the groups. Thus, features with higher F-statistic values are typically selected for further analysis or model building.

Table 4. The basic mathematical equation for performing ANOVA

Calculate the group means For each group i; (where i = 1 to m), calculate the mean of the numeric feature Y	Mean(i) = Σ(X_{ij}) / n_i Y_{ij} is the value of feature Y for the j^{th} observation in group i, and n_i is the number of observations in group i
Calculate the overall mean of all values of the numeric feature Y across all groups	Overall Mean = Σ(Σ(Y_{ij})) / N Where, N is the total number of observations
Calculate the between-group sum of squares (SSB)	SSB = Σ(n_i * (Mean(i) - Overall Mean)^2)
Calculate the within-group sum of squares (SSW)	SSW(i) = Σ((Xij - Mean(i))^2)
Then, sum up the SSW(l) values for all groups	SSW = Σ(SSW(i))
Calculate the degrees of freedom (df)	Between-group df (dfB) = m-1 Within-group df (dfW) = N -m
Calculate the mean squares (MS)	Mean Square Between (MSB) = SSB / dfB Mean Square Within (MSW) = SSW / dfW
Calculate the F-statistic	F = MSB / MSW

Radiomic characteristics that demonstrate the most significant associations or variations concerning the presence or absence of LC may be systematically recognized and retained. This improves the precision as well as the efficacy of predictive ML for lung nodule detection by ensuring that only the most relevant and discriminatory features are taken into account during the model-building stage.

2.3.3. Boosted Classification Ensemble Tree (BOCET)

Boosted Classification Ensemble Tree [33] is a robust ML approach that constructs a highly accurate predictive model by aggregating predictions from numerous weak models such that decision trees as shown in Figure 4. In FS, the boosting technique entails the sequential training of decision trees on distinct subsets of the data, with a heightened emphasis on misclassified occurrences throughout each iteration. The method prioritizes characteristics that significantly contribute to proper classification, resulting in the automated inclusion of important features throughout the constructing process of the model. This guarantees that the most informative characteristics are highlighted and employed in the ultimate ensemble model.

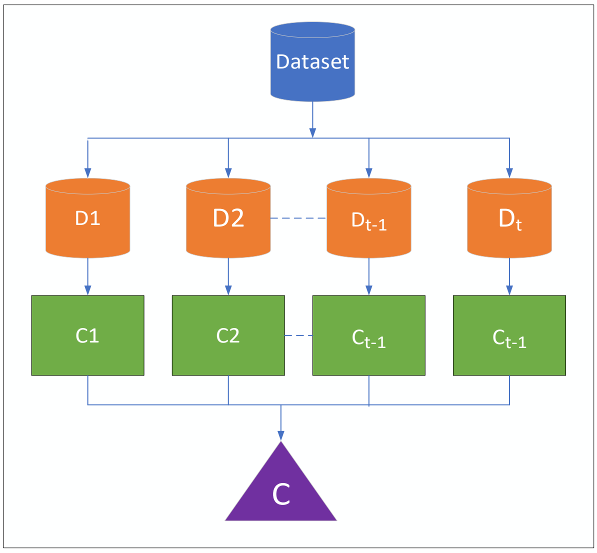

Figure 4. Boosted Classification Ensemble Tree

PROOF

PROOF

2.3.4. Bagged Classification Ensemble Tree (BACET)

The Bagged Classification Ensemble Trees, also referred to as bagging, is an ensemble learning approach that seeks to enhance the accuracy of a model by creating numerous models trained on distinct subsets of the training data [34]. Subsequently, each model is employed to generate predictions, and the outcomes derived from these models are aggregated to yield a conclusive forecast. During FS, the bagging technique entails the random picking of subsets from the dataset, followed by the training of separate decision trees on each of these subsets as depicted in Figure 5. Features that regularly manifest in the highest-performing trees are deemed significant and are preserved for subsequent study. Both approaches dynamically detect and prioritize pertinent characteristics while constructing intricate models. This phenomenon facilitates the development of more precise and robust prediction models for the identification of lung nodules. The algorithms for decision tree-based BOCET and BACET are presented in Algorithm 1 and Algorithm 2.

```
Algorithm : Boosting
Input: Training Sample , Classifier L, iterations I
Output: Result L_E

Training:
normalize the weights and make the total weight equal to m
S_i = Sample from S according to the distribution
L_i = Train a classifier on S_i via L

  $e_i = \frac{1}{m} \sum_{x_i \in S_i;\, Li(x)=y} weight(xi)$

  $\beta_i = \frac{e_i}{1-e_i}$

weight(x_i) = weight(x_i) β_i, for all x_i, where L_i(x_i)=y_i
end for
      $L_E = arg\ max \sum_{L_i(x)=y} \log(1/\beta i)$
```

```
Algorithm : Bagging
Input: Training Sample S, Classifier L, iterations I
Output: Result L_E
Training:

for i = 1 to I
S_i = bootstrap sample from S
L_i = train classifier on S, via L
end for

$L_E = arg\ max \sum_{Li(x)=y} 1$
```

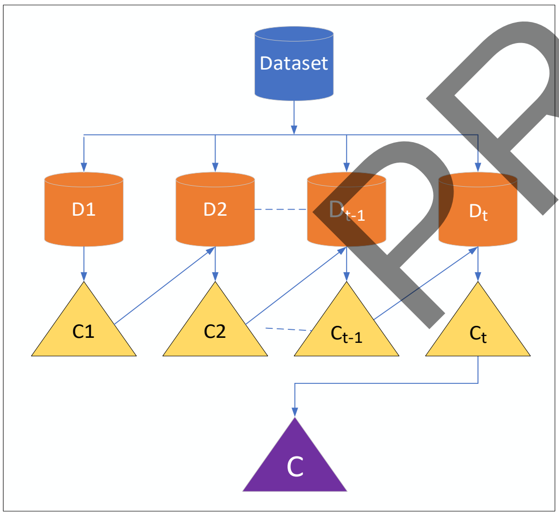

Figure 5. Bagging Classification Ensemble Tree

2.4. Classification and Performance Evaluation

Once the relevant features are selected using approaches explained in the above section, the classification of LC into 2 classes (i.e. benign and malignant) is then performed using a robust approach namely the SVM (Figure 6). SVM has emerged as a significant classifier in the domain of medical image analysis as it requires less training and is easy to implement [35]. SVM can be extended to handle nonlinearly independent data by transforming the input features into a higher-dimensional space using a kernel function. This allows for finding a non-linear decision boundary in the original feature space. The decision function with a kernel can be represented as (Equation 3):

$$f(x) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \ y_i \, \mathrm{k}\,(x_i, x) + \mathrm{b} \tag{3}$$

Where k(x_i, x) is the kernel function that computes the similarity between data points x_i and x, and α_i are the learned coefficients. k can be Linear Kernel, Polynomial Kernel, and Radial Basis Function (RBF) Kernel [35, 36] also known as Gaussian SVM is given in the following equations (Equations 4-6):

$$k(x_i, x) = \ x_i . x \tag{4}$$

$$k(x_i, x) = \ (x_i . x + c)^2 \tag{5}$$

$$k(x_i, x) = \ e^{|x_i . x + c)^2} \tag{6}$$

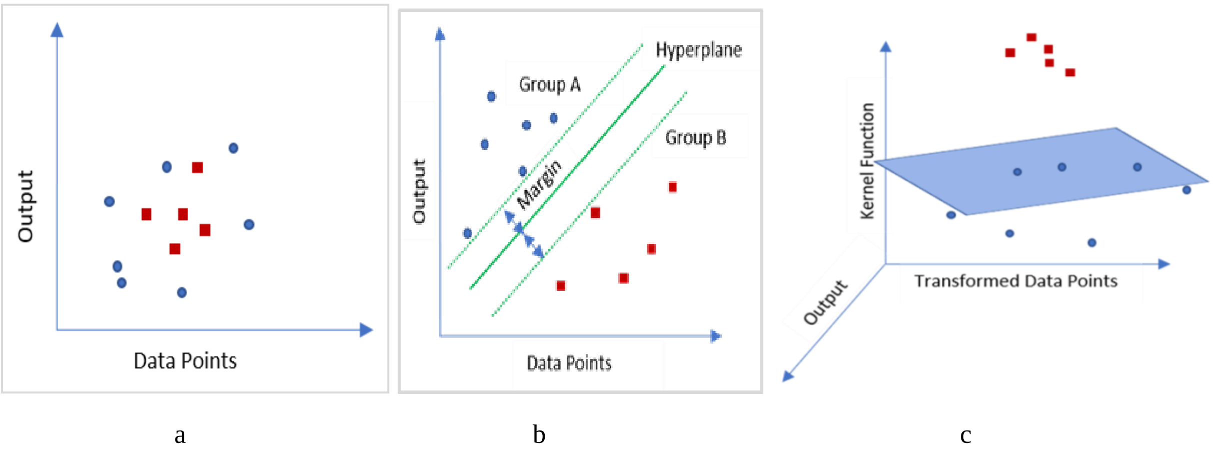

Figure 6. SVM for classification; (a) illustrates the feature space of data points and their decision boundary. (b) and (c) illustrate the non-linear data points and their transformation into higher space using a kernel function

The training of SVM involves solving a constrained optimization problem to find the optimal hyperplane or decision boundary [37]. The kernel trick is employed to transform the feature space, enabling the algorithm to capture complex decision boundaries. The training process involves solving an optimization problem with the help of Lagrange multipliers and dual problem formulation. Hence, SVM offers a versatile framework for classification tasks, with different kernel functions enabling the modeling of complex decision boundaries [38].

To evaluate a model, it is necessary to check its performance using some metrics called performance evaluation metrics like Accuracy, Sensitivity, Specificity, and Area under Curve (AUC) [37]. Accuracy is simply the ability of the model to compute many accurate predictions to the total figure of predictions. Sensitivity, also known as Recall, is used to compute the number of true positives (*tp*) and Specificity refers to the ability of the model to predict true negatives (*tn*). For all these metrics, a value close to 1 indicates a good classification result and vice-versa. The Receiver Operating Curve (ROC) tells how well a model performs. The data is divided into 'k'= 5 folds and the model is trained using 'k-1' folds. The AUC is computed and the process is repeated until all the 5 folds are utilized as test sets [37, 39]. In the end, 'k' AUC values are averaged to get cross-validated AUC. The mathematical equations used to calculate the evaluation metrics are provided as follows (Equations 7-9):

$$Accuracy = \frac{Y_{tp} + Y_{tn}}{Y_{tp} + Y_{tn} + Y_{fp} + Y_{fn}} \tag{7}$$

$$Sensitivity/Re\,c\,all = \frac{Y_{tp}}{Y_{tp} + Y_{fn}} \tag{8}$$

$$Specificity = \frac{Y_{tn}}{Y_{tn} + Y_{fp}} \tag{9}$$

Here, Ytp, Ytn, Yfp, and Yfn denote true-positive, true-negative, false-positive, and false-negative.

Within the framework of this investigation, a thorough examination was performed utilizing SVM on CT images taken from the LIDC database. We experimented with different kernels of SVM such as Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian, and Coarse Gaussian. The Linear SVM (LSVM) employs a straightforward linear kernel, ideal for data separable by a straight line. Quadratic SVM (QSVM) enhances this by introducing quadratic kernels, accommodating more intricate separations. QSVM may be preferable to linear SVM when the border does not comprise a straight line but rather a curved boundary. Cubic SVM (CSVM) goes further, leveraging cubic kernels to capture even more complex relationships. A more precise classification model may be generated using a cubic SVM when the lung cancer data reveals very complicated and curved correlations among its components. Fine Gaussian SVM (FGSVM), featuring a narrow Gaussian kernel, excels in intricate pattern recognition, while Medium Gaussian SVM (MGSVM) finds the balance between detail and generalization. It works by projecting data into an infinite-dimensional space using Gaussian functions. On the other hand, Coarse Gaussian SVM

PROOF

(CGSVM) employs a wider Gaussian kernel, focusing on broader data trends. These diverse SVM kernels empower data scientists to tailor their approach, matching the kernel's complexity to the underlying data distribution, thereby optimizing classification accuracy and robustness. All the above six SVM models are used for classification, in all the 3 scenarios, and are evaluated for different performance metrics.

3. Results

In this proposed work, the LIDC database is used to diagnose LC. The whole process of implementation has been performed using MATLAB 2017b and 2021a. A 64-bit computer system with 16 GB RAM was utilized for the purpose. A total of 4157 slices of CT images are used.

ROI of nodules for every slice was obtained using the radiologists' annotations. The shape features of every nodule are retrieved (Section 2.2.1). To compute all 152 of Haralick's texture characteristics (Section 2.2.2), a sub-image consisting of 11×11 pixels is chosen around the centroid of every nodule. Using all four spatial directions at θ = 0o, 45o, 90o, and 135o, the GLCM, GLDM, and GLRLM matrices are created, keeping inter-pixel distance ‘d’=1, which can have major implications. Moreover, WPT is applied up to level 2 for each sub-image, producing 16 multi-scaled mini-images. Daubechies wavelet family db1, db2, and db3 were used as the basis functions and the WPT texture features (Section 2.2.3) were evaluated in all 4 directions as above. A total of 4224 WPT-GLCM features, 960 WPT-GLDM features, and 2112 WPT-GLRLM features were retrieved. Hence, a cohort of 7455 features was computed (Table 3). Feature scaling (min-max normalization) is employed to normalize the feature range in preparation for further evaluation.

FS is done to obtain the most discriminative features using two FTFS techniques, Chi-square tests and ANOVA, and two ETFS techniques, BOCET and BACET. Based upon the ranking established individually by four FS techniques, four different radiomics feature sub-sets, each consisting of only eight relevant features, were selected to discriminate between benign and malignant nodules. We restricted the use of the first 8 features only for classification because the use of more than 8 features did not help the classifier to improve its accuracy any further. Subsequently, four distinct sets of relevant features are available for classification in the next phase.

In this study, six types of ML classifiers were used: LSVM, QSVM, CSVM, FGSVM, MGSVM, and CGSVM. All were evaluated to check the efficacy of four different sets of selected shapes and radiomic features in detecting lung nodules. For classification, to get cross-validated AUC for all classifiers, a fivefold cross-validation approach was used and evaluated around 50 times. All are evaluated and compared for AUC, accuracy, sensitivity, precision, and specificity. Each scenario, i.e. Scenario 1 (G54 Vs G12), Scenario 2 (G543 Vs G12), and Scenario 3 (G54 Vs G123) is evaluated one by one. A comprehensive analysis of the above metrics w.r.t. different classifiers as well as the ranking algorithms in three scenarios is presented in Table 5, Table 6, and Table 7, respectively.

PROOF

Table 5. Results for Scenario 1(G54 Vs G12)

Feature Selection method	**Classifier**	**Scenario 1 (**G54 Vs G12**)**			
		AUC	*Acc.%*	*Senst.%*	*Spec.%*
Without Feature Selection	LSVM	0.74	63.2	62.5	66.4
	QSVM	0.67	60.8	61.3	58.7
	CSVM	0.68	61.5	61.8	60.1
	FGSVM	0.65	59.8	59.4	63.7
	MGSVM	0.69	58.5	58.1	73.2
	CGSVM	0.63	58.4	58.1	73.8
ANOVA	LSVM	0.79	72.8	80.1	65.4
	QSVM	0.79	74.7	77.6	70.7
	CSVM	0.78	73.8	76.5	69.9
	FGSVM	0.74	71.9	72.8	70.2
	MGSVM	0.79	75.0	77.2	71.7
	CGSVM	0.78	73.9	80.1	67.2
Chi-Square test	LSVM	0.80	73.7	**81.4**	67.0
	QSVM	0.80	75.0	77.8	70.8
	CSVM	0.66	57.2	68.9	49.2
	FGSVM	0.78	75.2	77.2	72.2
	MGSVM	0.80	**75.3**	77.9	71.5
	CGSVM	0.79	74.4	**81.4**	67.0
BOCET	LSVM	0.79	66.7	66.6	67.3
	QSVM	**0.81**	67.6	66.5	**71.4**
	CSVM	0.77	66.2	65.7	68.1
	FGSVM	0.78	67.4	66.1	72.5
	MGSVM	0.80	67.9	66.7	72.4
	CGSVM	0.79	67.1	67.0	67.2
BACET	LSVM	0.80	62.8	62.4	65.1
	QSVM	0.79	62.8	62.1	68.0
	CSVM	0.79	63.3	62.4	70.0
	FGSVM	0.75	62.2	61.6	66.8
	MGSVM	0.80	63.3	62.4	69.1
	CGSVM	0.80	62.4	62.3	62.7

Table 6. Results for Scenario 2 (G543 Vs G12)

Feature Selection method	Classifier	**Scenario 2** (G543 Vs G12)			
		AUC	*Acc.%*	*Senst. %*	*Spec. %*
Without Feature Selection	LSVM	0.772	65.7	73.4	64.4
	QSVM	0.683	62.6	59.8	63.3
	CSVM	0.679	62.5	59.1	63.4
	FGSVM	0.646	60.8	61.0	60.8
	MGSVM	0.730	59.2	87.5	59.1
	CGSVM		Failed		
ANOVA	LSVM	0.78	72.2	**80.5**	64.4
	QSVM	0.79	74.3	77.5	70.0
	CSVM	0.72	68.8	71.4	63.8
	FGSVM	0.76	73.3	74.7	71.1
	MGSVM	0.79	74.3	76.7	70.8
	CGSVM	0.79	73.4	80.2	66.4
Chi-Square test	LSVM	0.79	72.6	71.3	73.2
	QSVM	0.79	**74.7**	70.3	77.4
	CSVM	0.68	65.7	61.4	67.4
	FGSVM	0.77	73.1	68.4	76.1
	MGSVM	0.79	73.6	69.9	75.8
	CGSVM	0.79	72.5	70.6	73.5
BOCET	LSVM	0.79	72.5	71.2	73.1
	QSVM	0.79	73.9	69.7	76.5
	CSVM	0.74	70.5	70.0	70.7
	FGSVM	0.75	72.0	67.2	75.0
	MGSVM	0.78	73.4	69.5	75.6
	CGSVM	0.79	72.4	70.8	73.1
BACET	LSVM	0.79	72.8	72.0	73.1
	QSVM	**0.80**	74.2	70.2	**76.6**
	CSVM	0.75	71.1	70.4	71.5
	FGSVM	0.75	70.7	66.6	73.0
	MGSVM	0.79	73.8	70.8	75.4
	CGSVM	0.79	72.6	71.4	73.1

Table 7. Results for Scenario 3 (G54 Vs G123)

Feature Selection method	Classifier	**Scenario 3 (G54 Vs G123)**			
		AUC	*Acc. %*	*Senst. %*	*Spec. %*
Without Feature Selection	LSVM	0.67	69.6	70.0	25.0
	QSVM			Failed	
	CSVM	0.60	68.2	71.7	44.8
	FGSVM			Failed	
	MGSVM	0.65	69.9	69.9	70.4
	CGSVM	0.68	69.6	69.6	62.01
ANOVA	LSVM	0.64	69.6	70.0	25.0
	QSVM	0.63	69.6	69.6	41.6
	CSVM	0.64	70.0	71.0	53.7
	FGSVM	0.60	68.9	71.8	47.1
	MGSVM	0.64	70.0	70.4	58.1
	CGSVM	0.62	69.6	69.6	33.5
Chi-Square test	LSVM	0.56	69.8	69.6	31.8
	QSVM	0.60	69.6	69.4	51.7
	CSVM	0.63	**70.3**	70.6	62.1
	FGSVM	0.59	69.2	71.9	48.5
	MGSVM	0.61	69.7	69.9	56.0
	CGSVM	0.59	67.6	70.6	27.6
BOCET	LSVM	0.51	69.9	69.6	42.8
	QSVM	0.63	69.9	69.6	50.1
	CSVM	0.54	64.3	40.1	33.1
	FGSVM	0.61	69.5	71.9	49.8
	MGSVM	0.63	70.1	70.2	**63.0**
	CGSVM	0.56	69.6	69.0	46.1
BACET	LSVM	0.55	68.1	69.6	55.3
	QSVM	0.64	69.6	70.6	50.0
	CSVM	0.66	70.3	72.0	54.1
	FGSVM	0.61	70.7	**72.3**	55.8
	MGSVM	0.66	70.1	70.4	60.2
	CGSVM	**0.67**	69.6	69.6	56.0

PROOF

4. Discussion

Recent studies and related literature have consistently highlighted the possible significance of shape and radiomics in the characterization of lung nodules. In our work, the shape and selected radiomics based on Daubechies db1, db2, and db3 WPT were examined with nine ML classifier models to determine the effectiveness of selected features and the model pair.

Based on the values obtained from evaluation metrics, in Scenario 1 (G54 Vs G12), it can be analyzed that the features selected using each FS method, ANOVA, Chi-square, BOCET, and BACET give good classification results when combined with the various ML models. The detailed results obtained are provided in Table 5. It is seen that Chi-Square gives overall best sensitivity /recall (81.4%) with CGSVM and LSVM. However, ANOVA gives the best values among the rest of the different classifier metrics and sensitivity is also reasonably very good with FGSVM (80.1%). The best values for AUC, accuracy, and specificity are given by QSVM + BOCET (81%), MGSVM + Chi-Square (75.3%), and QSVM + BOCET (71.4%), respectively. FTFS techniques give better performance results than ETFS. Similarly, in Scenario 2 (G543 Vs G12), where G543 is the malignant group and G12 is the benign group, all SVMs are evaluated as per the calculated performance metrics. The values obtained for evaluation parameters revealed that the overall best sensitivity (80.5%) with LSVM. The best values for AUC, accuracy and specificity are given by QSVM + BACET (80%), QSVM + Chi-Square (74.7%), and QSVM + BACET (76.6%), respectively. Finally, in the last scenario, i.e. Scenario 3 (G54 Vs G123), where G54 is the malignant group and G12 is the benign group, the evaluation metrics obtained demonstrate that the best results for sensitivity are given by FGSVM+BACET (72.3%) in comparison to other models. The best values for AUC, accuracy and specificity are given by CGSVM + BACET (67%), CSVM + Chi-Square (70.3%), and MGSVM + BOCET (63%), respectively.

If we analyze the results achieved, it is clear that the FTFS method showed the best results by yielding the best sensitivity and other parameter values. Further, it can't be denied that many among the rest of the classification models also achieved good results with comparable metrics. However, it is worth mentioning that the results attained in Scenario 1 are better than Scenario 2, and Scenario 3. The rationale for this could be the incorporation of indeterminate nodules with MS 3 in Scenario 2, and Scenario 3. Furthermore, the outcomes of Scenarios 2 and 3 reveal that classifying indeterminate lung nodules into the malignant category results in a higher classification accuracy than classifying them into the benign category, suggesting a greater degree of similarity between those indeterminate nodules and malignant nodules. Therefore, the central finding from the sum total of results shows the implications of utilizing predictive radiomics features in conjunction with SVM models that can be reliable for LC prediction

5. Conclusion

Lung cancer stands as the prevailing and most fatal form of cancer, accounting for 2.21 million fresh cases and resulting in 1.80 million fatalities. The key to fighting lung cancer is early diagnosis of pulmonary lesions and nodules. In recent years, radiomics has received considerable attention and investigation for lung nodule identification. But so far it is murky and unclear which radiomics feature(s) to use for the prediction of pulmonary nodules. In this study, an attempt has been made towards evaluation of CT radiomics extracted using shape, texture analysis, and WPT features in amalgamation with ML algorithms. The results are quite promising in the prediction of pulmonary lung nodules.

In this study, the LIDC dataset consisting of 4157 CT images is used. Shape features were extracted. A sub-image of 11 by 11 pixels, around the nodule centroid, was analyzed for its texture. Three statistical texture analysis approaches, i.e. GLCM, GLDM, and GLRLM were then employed to extract texture features. Further, Daubechies wavelet family (db1, db2, and db3) was used to apply WPT on each of the sub-images, up to decomposition level 2. The texture analysis techniques were applied again on 16 sub-images. FTFS methods, Chi-square test, ANOVA, and ETFS algorithms (BOCET and BACET) were used to determine relevant features. Finally, the classification of cancer into benign or malignant was performed in three scenarios. Pairing of nodules based upon malignancy scores, 1(benign) through 5(malignant), accordingly three scenarios were created: Scenario1 (G45 Vs G12), Scenario 2 (G453 Vs G12), and Scenario 3 (G45 Vs G123). Six different SVM models, LSVM, QSVM, CSVM, FGSVM, MGSVM, and CGSVM kernels were used for classification. The intricate framework of this approach showed how the SVM algorithm with six different kernel approaches works efficiently to extract information from CT images.

In Scenario 1, the best sensitivity of 81.4% was achieved by the MGSVM+Chi-Square model. The best sensitivity of 80.5% was achieved in Scenario 2 using the LSVM+ANOVA model. The third scenario's best sensitivity, 72.3%, was achieved by the FGSVM+BACET. Overall, the study demonstrates that the radiomics-based shape and WPT texture achieve decent results for the classification of CT pulmonary nodules. The outcome thus can help the clinicians to diagnose, and make better decisions and treatments.

In future work, the study can be extended by applying different ML algorithms, and/or Deep Learning (DL) techniques, nature-inspired optimization approaches, and considering different lung cancer datasets for better lung cancer outcomes.

References

1- Ziyad, S.R., Radha, V., Vaiyapuri, T., "Noise removal in lung LDCT images by novel discrete wavelet-based denoising with adaptive thresholding technique." *International Journal of E-Health and Medical Communications (IJEHMC)*, 12(5), pp.1-15, 2021. http://doi.org/10.4018/IJEHMC.20210901.oa1

2- Orozco, H.M., *et al.*, "Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine." *Biomedical Engineering Online*, 14(1), pp. 1-20, (2015). https://doi.org/10.1186/s12938-015-0003-y

3- Donga, H.V., Karlapati, J.S.A.N., Desineedi, H.S.S., Periasamy, P., TR, S.,"Effective Framework for Pulmonary Nodule Classification from CT Images Using the Modified Gradient Boosting Method" *Applied Sciences*, 12(16), pp.8264, (2022). https://doi.org/10.3390/app12168264

4- Alzubaidi, M. A., Mwaffaq, O., Jaradat, H., "Comprehensive and comparative global and local feature

PROOF

extraction framework for lung cancer detection using ct scan images" *IEEE Access* 9, pp. 158140-158154, (2021). Doi: 10.1109/ACCESS.2021.3129597 https://ui.adsabs.harvard.edu/abs/2021IEEEA...9o8140A

5- Chen, C.H., *et al.*, "Radiomic features analysis in computed tomography images of lung nodule classification." *PloS One*, 13(2), e0192002, (2018). doi: https://doi.org/10.1371%2Fjournal.pone.0192002

6- Khehrah, N., Farid, M. S., Bilal, S., Khan, M. H. "Lung nodule detection in CT images using statistical and shape-based features." *Journal of Imaging*, 6(2), pp. 6, (2020). https://doi.org/10.3390/jimaging6020006

7- Gillies, R. J., Paul E. K., and Hedvig H., "Radiomics: images are more than pictures, they are data." *Radiology*, 278(2), pp. 563-577, (2016). https://doi.org/10.1148/radiol.2015151169 PMID: 26579733; PMCID: PMC4734157.

8- Shakir, H., Yiming, D., Rasheed H., Khan, T.M.R., "Radiomics based likelihood functions for cancer diagnosis" *Scientific Reports*, 9(1), pp. 9501, (2019). https://doi.org/10.1038/s41598-019-45053-x

9- Permatasari, Z., Mauridhi H.P., I. K.E.P., "Lung nodule detection of CT and image-based GLCM and RLM CT scan using the support vector machine (SVM) method." *JAREE (Journal on Advanced Research in Electrical Engineering)*, 5(2), (2021). http://dx.doi.org/10.12962/jaree.v5i2.125

10- Abbas, W., Khan,.*et al.*, "Lungs nodule cancer detection using statistical techniques." In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, pp. 1-6, (2020).

11- Belfiore, M. P. *et al.*, "Robustness of Radiomics in Pre-Surgical Computer Tomography of Non-Small-Cell Lung Cancer." *Journal of Personalized Medicine*, 13(1), pp. 83, (2022). https://doi.org/10.3390/jpm13010083

12- Padmakumari, T., *et al.*, "The role of chest CT radiomics in diagnosis of lung cancer or tuberculosis: a pilot study." *Diagnostics*, 12(3), pp. 739, (2022). http://doi: 10.3390/diagnostics12030739. PMID: 35328296; PMCID: PMC8947348.

13- Jing, R., *et al.*, "A wavelet features derived radiomics nomogram for prediction of malignant and benign early-stage lung nodules." *Scientific Reports*, 11(1), pp. 22330, (2021). https://doi.org/10.1038/s41598-021-01470-5

14- Trivizakis, E., *et al.*, "A neural pathomics framework for classifying colorectal cancer histopathology images based on wavelet multi-scale texture analysis." *Scientific Reports*, 11(1), pp. 15546, (2021). https://doi.org/10.1038/s41598-021-94781-6

15- Torres, G., Baeza, S., Sanchez, C., Guasch, I., Rosell, A, Gil, D., "An intelligent radiomic approach for lung cancer screening." *Applied Sciences*, 12(3), pp.1568, (2022). https://doi.org/10.3390/app12031568

16- McNitt-Gray, F., *et al.*, "The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation." *Academic Radiology*, 14(12), pp. 1464-1474, (2007). doi: http://10.1016/j.acra.2007.07.021. PMID: 18035276; PMCID: PMC2176079.

17- Jun, T., Pu, J., Zheng, B., Wang, X., Leader, J. K., "Computerized comprehensive data analysis of lung imaging database consortium (LIDC)." *Medical Physics*, 37(7), pp. 3802-3808, (2010). https://doi.org/10.1118/1.3455701

18- Armato III, S. G., *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans." *Medical Physics*, 38(2), 915-931, (2011). https://doi.org/10.1118/1.3528204

19-Baba, T., *et al.*, “The tumour shape of lung adenocarcinoma is related to the postoperative prognosis”, *Interact Cardiovasc Thorac Surg.*, 15(1), pp. 73-6, (2012). doi: 10.1093/icvts/ivs055. Epub 2012 Apr 18. PMID: 22514255; PMCID: PMC3380970.

20- Haralick, R. M., Shanmugam, K., Dinstein, I., "Textural features for image classification." *IEEE Transactions on Systems, Man, and Cybernetics*, 6, pp. 610-621, (1973). doi: 10.1109/TSMC.1973.4309314

21- Materka, A., “Texture analysis methodologies for magnetic resonance imaging”, *Dialogues in Clinical Neuroscience*, **6**, pp. 243–250, (2004). https://doi.org/10.31887/DCNS.2004.6.2/amaterka

22- O’Connor, J. P. B. *et al.*, “Imaging intratumor heterogeneity: Role in therapy response, resistance, and clinical outcome”, *Clinical Cancer Research*, **21**, 249–257, (2015). doi:10.1158/1078-0432.CCR-14-0990

23- Mir, A.H., Hanmandlu, M., Tandon, S.N., “Texture analysis of CT images” *IEEE Engineering in Medicine and Biology Magazine* 14 (6), pp. 781-786, (1995). doi: 10.1109/51.473275.

24- Garcia, C., Zikos, G., Tziritas, G., “Wavelet packet analysis for face recognition Image and Vision Computing”, 18, pp. 289–297, (2000). https://doi.org/10.1016/S0262-8856(99)00056-6.

25- Han, J.G., Ren, W.X., Sun, Z.S.,”Wavelet packet based damage identification of beam structures”, *International Journal of Solids and Structures*, 42(26), 2005, pp. 6610-6627, (2005). https://doi.org/10.1016/j.ijsolstr.2005.04.031

26- Perlibakas, V., “Face Recognition Using Principal Component Analysis of the Wavelet Packet Decomposition”, *Science Direct Working Paper* No S1574-034X(04)70005-8, April (2004).

27- Daubechies, I., Alex, G., Yves, M., "Painless nonorthogonal expansions", *Journal of Mathematical Physics*, 27(5), pp. 1271-1283, (1986).

28- Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., O'Sullivan J.M., “A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction”, Frontiers in *Bioinformatics*, vol 2, (2022). https://doi.org/10.3389/fbinf.2022.927312

PROOF

29- Bommert, A., Welchowski, T., Schmid, M., Rahnenführer, J., "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data." *Briefings in Bioinformatics*, 23(1), bbab354, (2022). https://doi.org/10.1093/bib/bbab354

30- Zhang, Y., *et al.*, "Alzheimer’s disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion." *Information Fusion*, 66, pp. 170-183, (2021).

31- Shehu, Ina., "Testing Statistical Hypothesis on Learning Effectiviness: pre-and post-COVID 19." *South East European Journal of Sustainable Development*, 6(2), (2022).

32- Yu, Z., *et al.*, "Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research.", *Neuron*, 110(1), pp. 21-35, (2022). DOI: 10.1016/j.neuron.2021.10.030

33- Hatwell, J., Gaber, M.M., Azad, R.M.A., "gbt-hips: Explaining the classifications of gradient boosted tree ensembles.", *Applied Sciences*, 11(6). Pp. 2511, (2021). https://doi.org/10.3390/app11062511

34- Mosavi, A.,*et al.*, "Ensemble boosting and bagging based machine learning models for groundwater potential prediction.", *Water Resources Management* , 35, pp. 23-37, (2021).

35- Khan, Y. F., Kaushik, B., Chowdhary C.L., Srivastava, G, "Ensemble model for diagnostic classification of Alzheimer’s disease based on brain anatomical magnetic resonance imaging.", *Diagnostics*, 12(12), pp. 3193, (2022). https://doi.org/10.3390/diagnostics12123193

36- Goudjil, M., Koudil, M., Bedda, M. *et al.* “A Novel Active Learning Method Using SVM for Text Classification.” *Int. J. Autom. Comput.*, **15**, pp. 290–298, (2018). https://doi.org/10.1007/s11633-015-0912-z

37- Khan, Y.F., Kaushik, B., Mir, B.A., "Computational Intelligent Models for Alzheimer's Prediction Using Audio Transcript Data." *Computing and Informatics* 41(6), pp. 1589-1624.

https://doi.org/10.31577/cai_2022_6_1589

38- Vapnik, V. N., "An overview of statistical learning theory." *IEEE Transactions on Neural Networks*, 10(5), pp. 988-999, (1999).

39- Pisner, D.A., Schnyer, D.M., Support vector machine. In *Machine Learning*; Elsevier: Amsterdam, The Netherlands, pp. 101–121, (2020).

PROOF