

Proficiency evaluation of shape and WPT radiomics based on machine learning for CT lung cancer prognosis



Arooj Nissar^{1*} and A. H. Mir²

Abstract

Background Lung cancer is a fatal disease which has high occurrence and mortality rates, worldwide. Computed tomography imaging is being widely used by clinicians for detection of lung cancer. Radiomics extracted from medical images together with machine learning platform has enabled automated lung cancer diagnosis. Therefore, this study is proposed with the aim to efficiently apply radiomics and ML techniques to classify pulmonary nodules in CT images. Lung Image Data Consortium is utilized which contains 1018 CT lung cancer cases.

Results Radiomics are 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 two techniques, Analysis of variance and Chi-square test, are applied. The classification of nodule into benign or malignant is evaluated by using state-of-art models: Support vector machine, Decision Trees, Ensemble Trees (BOCET, BACET, RUSBOCET), Ensemble Subspace KNN and Ensemble Subspace Discriminant. The results show that, BACET gives best AUROC (92.9%), MGSVM gives best accuracy (90.4%), FGSVM yields the best sensitivity (97.8%), MGSVM gives best precision (94.1%) and RUSBOCET gives best specificity (84%).

Conclusions The results show that the proposed methodology can be successfully used for the classification of pulmonary nodules based on CT images. The outcome thus can help clinicians to reach better decision, treatments and early diagnosis.

Keywords Lung cancer, LIDC, Radiomics, WPT, SVM, Feature selection, Machine learning, Diagnosis

Background

Lung cancer (LC) is a dreaded 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. Many factors including smoking, drug intake, and inhalation of harmful substances produced

² Department of Electronics and Communication, National Institute

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 World Health Organization (WHO) [3], about 1.80 million deaths are caused just because of LC. A report on USA statistics in 2020 revealed that 197,453 new lung cancers were reported and 136,084 people died from LC. In the UK, every year, approximately, 44,500 cases are diagnosed with LC [4].

For early detection of LC, pulmonary nodules 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 cancerous [2]. Mortality



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

^{*}Correspondence:

Arooj Nissar

arooj@nitsri.ac.in

¹ Department of Information Technology, National Institute of Technology Srinagar, Srinagar, Jammu and Kashmir 190006, India

of Technology Srinagar, Srinagar, Jammu and Kashmir 190006, India

increases dramatically with the presence of malignant lung nodules, whereas the patient's survival rate increases if the nodules are benign. Hence, 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 does not 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. 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].

The conversion of digital medical images into mineable high-dimensional data, a process that is known as radiomics, is motivated by the concept that biomedical images contain information that reflects underlying pathophysiology and that these relationships can be revealed via quantitative image analyses. Radiomics is designed to be used in decision support of precision medicine [7]. Radiomics is a quantitative approach that applies data-characterization algorithms whose purpose is to improve the already available data using mathematical analysis [5]. 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. Noor Khehrah et al. [6] proposed using statistical and shape-based features and SVM for LC detection. They achieved excellent results with sensitivity of 93.75%. In another study, Parmatasari et al. [8] applied SVM to classify lung cancer and yielded an accuracy of 85.63%.

In radiomics, features from 2D Region of Interests (ROI's) and/or 3D Voxels of Interests (VOI's) are extracted. The proposed study aims to evaluate the proficiency 2D CT radiomics, shape and ML approaches for the diagnosis of cancer in lung nodules. The approach employs selection of most suitable features for classification. Various state-of-art classifiers, in addition to SVM, are evaluated using metrics to find the best outcome/ model. The proposed framework is useful and reliable in the successful classification of the lung nodule as benign or malignant.

Existing literature

Donga et al. [3] investigated 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 between benign and 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 the suggested technique classifies benign or malignant lung nodules better. Alzubaidi et al. [4] developed a comprehensive and comparative methodology for lung cancer diagnosis utilizing CT scan images, covering global and local aspects. Thousand CT scans 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 support vector machine (SVM) outperforms learning techniques. SVM with Haar Wavelet, HOG, and Gabor Filter features achieves 90% accuracy, 88% sensitivity, and 97% specificity, outperforming global approaches. Radiomics was used in cancer diagnosis, prognosis, and therapy response prediction by Chen et al. [5]. A 4-feature signature was used to classify lung nodules using radiomics and CT images. In 72 individuals with 75 pulmonary nodules, benign and malignant lesions differed in 76 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. Khehrah et al. [6] automates lung nodule identification using CT scans. Statistical and shape-based characteristics from nodule candidates produce feature vectors categorized by support vector machines. 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. SVM classification accuracy obtained is 85.63%.

Shakir et al. [9] developed radiomics-driven models to classify lung, colon, and 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 regression

model for tumor classification. Validation on public dataset of 265 images 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. Palumbo et al. [10] evaluated role of shape and texture features from ¹⁸F-FDG PET/CT to classify between benign and malignant lung nodule. 18 3D features from PET and CT were used in prediction models. The outcomes of the models clearly show that an added accuracy gain is acquired when a combination of PET and CT features are used. Belfiore et al. [11] examined Non-Small Cell Lung Cancer (NSCLC) CT scan radiomics characteristics' resilience with respect to three segmentation approaches. Expert radiologists segmented three 3D-ROIs to analyze radiomics characteristics in 48 NSCLC patients. The intra-class correlation coefficient (ICC) among features was evaluated. 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 reproducibility. Padmakumari et al. [12] tested CT radiomics for its ability to discriminate lung cancer (LC) from tuberculosis (TB) in low-income nations without lung biopsies. Radiomics 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. The study showed that the radiomics may enhance resource-limited oncological patient treatment by identifying these illnesses non-invasively. However, prospective studies are needed to confirm these findings. 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. Training set (N=70) and validation set (N=46) of 116 patients were considered with early-stage solitary pulmonary nodules (SPNs) 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.

Torres et al. [14] experimented with hybrid approach of using feedforward networks and nodule radiomics from CT. 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. In another study, Balci et al. [15], proposed a new hybrid method that performs classifications using both medical image analysis and radial scanning series features. According to the results, an accuracy of 92.84%, recall of 92.41 and precision of 92.63 was obtained.

From the above studies and many more it is a deeprooted fact that not all the extracted features are equally good predictors for classification of lung nodules. Hence it is interesting and important to pick out the most discriminative feature and/or a subset of such features from a pool of extracted features. In this study, an attempt is made to investigate and analyze the discriminative power of various shape features and statistical texture features. The texture features are proposed to be extracted before and after applying WPT. Most significant and discriminative features out of an extensive pool of 7455 extracted features are selected using feature selection methods. State-of-art classification techniques will hence be applied to the selected features to evaluate the performance of each diagnostic model.

The rest of the paper proceeds as follows: Materials and methods are presented in section "Methods". The experimental results and analysis are reported in section "Results". Section "Discussion" presents the conclusion and future work.

Methods

This research work is proposed to execute the classification of lung nodules in CT images as benign or malignant by investigating the proficiency of shape and 2D radiomics, feature selection methods and ML algorithms. The whole study framework comprises of different stages, viz. dataset collection, feature-extraction, feature selection, classification, and performance assessment of various classifiers.

Dataset

A dataset plays a vital role in any diagnostic system. In this work, CT images from Lung Image Data Consortium (LIDC) is utilized which contains 1018 CT lung cancer patient scans. This LIDC database has CT images along with four experienced radiologists' ground truth reports. The presence of malignancy in nodules ≥ 3 mm and the annotations accorded by radiologists are described in detail in [16–18]. The slice count in each scan, used in this study, varied in the range of 110–388. A total of 1207 slices of CT scans were considered; 883 malignant and 324 benign. Samples from the LIDC dataset with malignant ROI's are shown in Fig. 1. The framework of the proposed methodology is shown in Fig. 2.



Fig. 1 LIDC dataset sample images with a malignant ROI's



Fig. 2 Proposed framework to classify lung nodule as malignant or benign

Feature extraction

For feature extraction the above dataset was put to use. Shape features and radiomics (texture, wavelet) are extracted using statistical techniques. A synopsis of these features is as follows:

Shape features

Different shape features play a vital role in the classification process. These features are crucial as they are closely associated with the detection 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 features is also given in Table 1.

Texture features

Texture analysis is a way of describing the spatial distribution of intensities [20] hence enabling description of tissue heterogeneity, a property believed to influence the outcome of cancer treatment [21]. In this work, to analyze Haralick's texture features, three techniques, i.e. Gray Level Co-occurrence Method (GLCM), Gray Level Difference Method (GLDM), and Gray Level Run Length Matrix (GLRLM) [6, 9, 10, 22, 23], for some value of inter pixel distance 'd' and angle ' θ ', are adopted to extract second and high-order statistics. Using GLCM twentytwo texture features are computed, viz. 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 (SENT), Difference Variance (DV), Difference Entropy (DENT), Information Measure of Correlation1 (IMC1), Information Measure of Correlation2 (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). Refer Table 1.

WPT texture features

2-level Wavelet Packet Transform (WPT) [24, 25] is used to generate multi-scale representations of the original image. While there are number of well performing wavelets available, the choice of the wavelet used depends on the application. This study focused on orthogonal wavelets of compact support, introduced by Daubechies [26]. The Daubechies wavelets can have appreciable influence into the success of texture classification because the filter affects positively the quality of the descriptors [2, 27]. Daubechies wavelet family db1, db2, and db3 were applied to implicate WPT. Given an image, 2-level WPT generates 16 multi-scaled images. Above set of features are computed, using same three Haralick's texture techniques, on wavelet multi-scaled images. Accordingly, these classes are denoted as WPT-GLCM, WPT-GLDM and WPT-GLRLM. The list of feature classes' along with the no of features extracted in each class is provided in Table 2 and the list of features extracted is given in Table 1.

Feature selection

Feature selection (FS) aims to draw out only the most informative features and remove noisy, non-informative, irrelevant and redundant features so as to benefit

GLCM	Autocorrelation (ACOR), Contrast (CON), Correlation1 (COR1), Correlation2 (COR2), Cluster Prominence (CP), Cluster Shade (CS),Dissimilarity (DS), Energy (ENR), Entropy(ENT), Homogeneity1 (HMG1), Homo- geneity2 (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

Table 1 List of Shape and Texture Features extracted using GLCM, GLDM, and GLRLM [22]

Table 2 List of features per class

Feature class		No of the f	Total			
Shape		7				
GLCM	Directions (0)	00	22	88	152	7455
		45 ⁰	22			
		<i>90</i> ⁰	22			
		135 ⁰	22			
GLDM		O^O	5	20		
		45 ⁰	5			
		<i>90⁰</i>	5			
		135 ⁰	5			
GLRLM		O^O	11	44		
		45 ⁰	11			
		<i>90⁰</i>	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		

ML models [28]. The FS methods that are routinely used are grouped into three methodological categories: Filter Type FS methods (FTFS), Wrapper Type FS methods and Embedded Type FS methods. In this work two FTFS methods were used, viz. Chi-square tests and the Analysis of variance (ANOVA). These methods use feature ranking as the evaluation metric and have proven significant to the detection of LC using radiomics and ML [29]. Chi-Square (χ^2) tests are statistical tests used to determine if categorical variables are significantly associated. 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 [30]. 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. Analysis of Variance (ANOVA) is also a statistical technique used to examine the differences between group means in a dataset [31]. Features that demonstrate significant variability between two categories are regarded essential for differentiating them and thus are selected for further analysis. Hence, features with higher F-statistic values are typically selected and retained for further analysis or model building.

Classification and performance evaluation Classification

Once the discriminative radiomics are available, using FTFS algorithms, various state-of-art classifiers; support vector machine (SVM) [32–34], decision trees [35], ensemble trees [36, 37], and ensemble subspace [38, 39] may be used to classify the lung nodule into 2 classes (benign and malignant).

SVM

Research reveals that SVM has emerged as a popular and powerful approach of ML in the domain of medical image analysis because of its relative simplicity and flexibility in implementation [33]. SVM is able to handle nonlinearly independent data by transforming the input features into a higher-dimensional space using a kernel function. Different types of kernel functions used are Linear, Polynomial, Radial Basis Function (RBF, also known as Gaussian) and sigmoid.

Decision trees

The decision tree algorithm in ML employs a tree structure for the classification process. With this approach, we decompose the dataset, root node, into nodes where each internal node denotes the feature, branches denote the rules and the leaf nodes denote decision and categorization. This method may be used with either numerical or qualitative information.

Ensemble tree

A tree ensemble is a ML technique for supervised learning that consists of a set of individually trained decision trees defined as base learners that may not perform well singly. The aggregation of the base learners produce a new strong model, which is often more accurate than the former ones. The three types of ensemble learning methods, bagging, boosting and RUSboosting are used. Bagging generates a set of bootstrapped versions of data (bags) from the original training dataset. Each bag consists of N observations, drawn from the original dataset at random but with replacement. Thus, a bag consists of approx. 63% of distinct samples and the remaining ones are the duplicates [40]. A decision tree is then trained with each bag, and combined by majority voting. Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers in series. A model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added. RUSboost is a boosting ML technique designed to improve the performance of models trained on skewed data. It applies random undersampling (RUS), a technique which randomly removes examples from the majority class until a desired class distribution is achieved.

Ensemble random subspace

The random subspace (RS) ensemble classifier achieves the benefits by applying a random subset of features over the combined set of base classifiers (KNN and Discriminant). Randomly selected subset features from the actual data set space are utilized to train the set of N number of base/weak classifiers. A majority voting combination rule is implemented over the output predictions of weak classifiers to obtain target class labels at final stage. K-Nearest Neighbor algorithm (KNN) is a nonparametric classification technique. It is a type of instance-based learning. The input consists of the *K* closest training examples in the feature space and the output is a class membership. Classification is done by a majority vote of neighbors. If K=1, then the class is single nearest neighbor [41]. LDA is a supervised learning algorithm that means it uses class labels and is suitable for class separation. It uses withinclass and between-class scatter matrices. If there are two classes, then the LDA draws one hyperplane and projects the data onto this hyperplane in such a way as to maximize the separation of the two categories. This hyperplane is created by maximizing the distance between the means of two classes and minimizing the variation between each category. It provides accepted accuracy and is widely used in medical-computer interfaces [42].

Performance evaluation

The confusion matrix obtained from prediction models is used to visualize their performance. The set of performance evaluation metrics used in this work are Accuracy, Area under Curve (AUC), Sensitivity, Precision and Specificity. Accuracy is the ability of the model to compute accurate predictions to the total figure of predictions. Sensitivity, also known as Recall, is used to compute the number of true positives (tp). Precision refers to the ability of the model to predict the quality of positive prediction and Specificity refers to the ability of the model in predicting 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 mathematical equations used to calculate the evaluation metrics are provided as under:

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

Recall = $\frac{tp}{tp + fn}$
Precision = $\frac{tp}{tp + fp}$
Specificity = $\frac{tn}{tn + fp}$

Here, *tp*, *tn*, *fp*, *and fn* denote true-positive, true-negative, false-positive, and false-negative.

Results

The implementation of the proposed strategy was performed using MATLAB 2017a and 2021b. A 64-bit computer system with 16 GB RAM was utilized for the purpose. The experimentation of this study is done on the LIDC dataset. A total of 1207 slices of CT scans were considered; 883 malignant and 324 benign.

Using the annotations of the radiologists, ROI of nodules for all slices were obtained. Shape features (section "Shape features") of all nodules are extracted. Further, a sub-image of 11×11 pixels is selected around the centroid of each nodule and all 152 Haralick's texture features (section "Texture features") are computed. The GLCM, GLDM and GLRLM matrices are formulated using all four spatial directions at $\theta = 0^{\circ}$, 45°, 90°, 135° keeping inter pixel distance 'd' = 1 which can have far reaching implications. Moreover, for each sub-image WPT is applied upto level-2 which yielded 16 multiscaled minuscule-images. Daubechies wavelet family db1, db2, and db3 were used as the basis functions and the WPT texture features (section 3.2.3) were evaluated in all the 4 directions as above. A total of 4224 WPT-GLCM features, 960 WPT-GLDM features and 2112 WPT-GLRLM features were available. Hence, a cohort of 7455 features was available for further analysis. (Table 2).

To obtain the most discriminative features from this cohort of 7455 features two FTFS techniques, ANOVA and Chi-Square, were applied. Based upon the ranking specified individually by both techniques first eight features were selected to discriminate between a benign and malignant nodule. We restricted to the use of only first 8 features for classification for the reason that use of more than 8 features was not helping the classifiers to improve their metrics any further. The selected features for ANOVA are: Area, Perimeter, MajorAxisLength, MinorAxisLength, $db3_LL_1HH_2_$ WPT_GLCM_CP_45⁰, db1_LL₁LL₂_WPT_GLDM_Contrast_45°, db3_LH1LL2_WPT_GLDM_Contrast_0° and db3_LH1LH2_ WPT_GLCM_IDMN_1350 and that of Chi-Square are: Area, Perimeter, MajorAxisLength, db1_LL1LL2_WPT_GLDM_Con-MinorAxisLength, trast_0⁰, db2_LL₁LL₂_WPT_GLDM_Contrast_90⁰, db3_ HL₁HH₂_WPT_GLRLM_GLN_0⁰ and db1_LH₁HL₂_ WPT GLRLM LRHGE 0⁰.

Subsequently, two sets of discriminative features were at hand for classification in next phase. (Table 3).

In this proposed study, several state-of-art classifiers; FGSVM, MGSVM, CGSVM, Decision Trees, BOCET, BACET, Ensemble Subspace Discriminant, Ensemble Subspace KNN, (section "Feature selection") were evaluated to check the efficacy of two different sets of selected shape and radiomic features in detecting lung nodules. For classification, to get cross-validated AUC for all classifiers, fivefold cross-validation approach was used and evaluated around 50 times. All are evaluated and compared for AUC, accuracy, sensitivity, precision and specificity. A comprehensive analysis of the above metrics w.r.t. different classifiers as well as the ranking algorithms is provided in Table 4. The summarized results are given in Figs. 3, 4.

Discussion

In latest research the related literature has continually heightened the potential role of shape and radiomics in the characterization of lung nodules. The important point is to assess the boons that radiomic features can furnish beyond usual imaging parameters alone. In our experiments the shape and selected radiomics based on Daubechies db1, db2 and db3 WPT were checked with nine models of ML classifiers to determine the proficiencies of selected features and the model duo.

It is imperative to mention here that out of 7455 features 4-shape-based and 4-WPT-based radiomic features, derived using db1, db2 and db3 kernels, are selected. This shows that selected wavelet-transformed texture features have more discriminative power than classical texture features derived using GLCM, GLDM and GLRLM. Hence our proposed hypothesis is consistent with the preliminary results.

The features thus selected using each FS method, ANOVA and Chi-square, give good classification results when combined with the various ML models. The detailed results obtained are provided in Table 4. Based on the metrics obtained from different classifiers, it is analyzed that

I. ANOVA gives overall best sensitivity /recall (97.8%) with FGSVM. However, Chi-Square gives best values for rest of the different classifier metrics and sensitivity is also reasonably very good with FGSVM (95.7%)

Rank	ANOVA		Chi-square test	
	Feature index	Feature name	Feature index	Feature name
1	1	Area	1	Area
2	4	Perimeter	4	Perimeter
3	2	Major axis length	2	Major axis length
4	3	Minor axis length	3	Minor axis length
5	5497	db3_LL ₁ HH ₂ _WPT_GLCM_CP_45 ⁰	248	db1_LL ₁ LL ₂ _WPT_GLDM_Contrast_0 ⁰
6	249	db1_LL ₁ LL ₂ _WPT_GLDM_Contrast_45 ⁰	2682	db2_LL ₁ LL ₂ _WPT_GLDM_Contrast_90 ⁰
7	6328	db3_LH ₁ LL ₂ _WPT_GLDM_Contrast_0 ⁰	6198	db3_HL1HH2_WPT_GLRLM_GLN_0 ⁰
8	6631	db3_LH ₁ LH ₂ _WPT_GLCM_IDMN_135 ⁰	1646	db1_LH1HL2_WPT_GLRLM_LRHGE_0

 Table 3
 Listicle of topmost 8 significant features selected by two FTFS techniques

Table 4 Classification results attained for nine different State-of-art classifiers using two FTFS techniques

Feature selection	Classifiers	(Malignant vs Benign)					
algorithm		AUROC	Accuracy	Sensitivity/ Recall/TPR	Precision/PPV	Specificity/TNR	
ANOVA	FGSVM	0.883	0.837	0.978	0.83	0.453	
	MGSVM	0.904	0.864	0.923	0.895	0.706	
	CGSVM	0.895	0.827	0.917	0.856	0.579	
	Decision Trees	0.878	0.856	0.91	0.895	0.709	
	Ensemble Boosted Trees (BOCET)	0.905	0.866	0.917	0.902	0.727	
	Ensemble Bagged Trees (BACET)	0.910	0.866	0.905	0.912	0.761	
	Ensemble RUSBoosted Trees (RUSBOCET)	0.914	0.852	0.861	0.932	0.829	
	Ensemble Subspace Discriminant	0.902	0.843	0.948	0.854	0.559	
	Ensemble Subspace KNN	0.909	0.869	0.929	0.896	0.706	
Chi-Square test	FGSVM	0.911	0.865	0.957	0.872	0.615	
	MGSVM	0.910	0.904	0.93	0.941	0.824	
	CGSVM	0.910	0.869	0.922	0.901	0.723	
	Decision Tree	0.883	0.857	0.926	0.885	0.67	
	Ensemble Boosted Trees (BOCET)	0.923	0.877	0.923	0.91	0.75	
	Ensemble Bagged Trees (BACET)	0.929	0.891	0.935	0.918	0.772	
	Ensemble RUSBoosted Trees (RUSBOCET)	0.925	0.861	0.869	0.937	0.84	
	Ensemble Subspace Discriminant	0.911	0.86	0.948	0.872	0.621	
	Ensemble Subspace KNN	0.896	0.865	0.93	0.891	0.689	

Overall highest values are marked in bold



Fig. 3 Comparison of performance metrics for 9 prediction models using ANOVA



Fig. 4 Comparison of performance metrics for 9 prediction models using Chi-square

- II. The best values for AUC, accuracy, precision and specificity are given by Ensemble Bagged Trees (92.9%), MGSVM (90.4%), MGSVM (94.1%) and Ensemble RUSBoosted Trees (84%) respectively using Chi-Square test.
- III. Chi-square FS technique gives better performance results than ANOVA.
- IV. Analyzing sensitivity/recall metric it can be spotted that all classifiers are furnishing promising results but Ensemble RUSBoosted Trees.
- V. Also, MGSVM (Chi-Square) gives comparatively better outcomes for different evaluation metrics. (AUROC=91%0, Accuracy=90.4%, Sensitivity=93%, Specificity=82.4%, Precision=94.1%)

The outcomes show that the methodology, specifically CT Daubechies WPT transformed texture features, can be successfully used for the classification of pulmonary nodules into benign or malignant. As a result, the approach described in this work may offer a viable standin for the precise prediction of LC, leading to early and more efficient treatment.

Comparison with previous work

Comparison of above proposed models, using two FS methods and nine cutting edge ML methods, with the previous work presented in Tables 5 and 6. The comparison has been made using AUC, accuracy, sensitivity, precision and specificity. But sensitivity and precision are the metrics that are more significant for the clinicians in assessing the predictive power of the model in cancer prediction. Considering the three parameters namely accuracy, sensitivity and precision it is clear that the best values are attained by proposed model 9 (Radiomics + Ensemble Subspace KNN, 86.9%), proposed model 1 (Radiomics + FGSVM, 97.8%) and proposed model 7 (Radiomics+Ensemble RUSBoosted Trees, 93.2%), respectively, using ANOVA as the FS method in comparison to models proposed in [43, 44]. Similarly, the best values for AUROC, accuracy, sensitivity and precision are attained by proposed model 15 (Radiomics+Ensemble Ensemble Bagged Trees, 92.9%), proposed model 11 (Radiomics + MGSVM, 90.4%), proposed model 10 (Radiomics + FGSVM, 95.7%) and proposed model 11 (Radiomics + Ensemble MGSVM, 94.1%), respectively, using Chi-square as the FS method when compared to methodologies given in [43, 44]. Although, methodologies employed in [43, 44] are based on deep learning, which are mostly thought of giving more promising results, our above proposed models are giving comparatively better results. This may be attributed to the use of FS methods and the ML classification models employed. Hence, we are confident that in a few years, radiomics and ML will be successfully integrated to come up with an efficient and effective automated assessment aid for radiologists.

Table 5 Performance metrics of proposed models compared with earlier research using ANOVA FS

	Methods	Database	Results (%)				
			AUROC	Accuracy	Sensitivity	Precision	Specificity
Cai, J et al. [43]	Deep Learning	LIDC	88.1	84.6	83.7	_	85.2
Wang et al. [44]	Deep Learning	LIDC	92.75	85.23	92.79	84.56	72.89
Proposed model 1	Radiomics + FGSVM	LIDC	88.3	83.7	97.8	83	45.3
Proposed model 2	Radiomics + MGSVM	LIDC	90.4	86.4	92.3	89.5	70.6
Proposed model 3	Radiomics + CGSVM	LIDC	89.5	82.7	91.7	85.6	57.9
Proposed model 4	Radiomics + Decision Tree	LIDC	87.8	85.6	91.0	89.5	70.9
Proposed model 5	Radiomics + Ensemble Boosted Trees	LIDC	90.5	86.6	91.7	90.2	72.7
Proposed Model 6	Radiomics + Ensemble Bagged Trees	LIDC	91.0	86.6	90.5	91.2	76.1
Proposed model 7	Radiomics + Ensemble RUSBoosted Trees	LIDC	91.4	85.2	86.1	93.2	82.9
Proposed model 8	Radiomics + Ensemble Subspace Discriminant	LIDC	90.2	84.3	94.8	85.4	55.9
Proposed model 9	Radiomics + Ensemble Subspace KNN	LIDC	90.9	86.9	92.9	89.6	70.6

Bold values give the highest values attained in that column

Table 6 Performance metrics of proposed models compared with earlier research using Chi-square FS

	Methods	Database	Results (%)				
			AUROC	Accuracy	Sensitivity	Precision	Specificity
Cai, J et al. [43]	Deep Learning	LIDC	88.1	84.6	83.7	_	85.2
Wang et al. [44]	Deep Learning	LIDC	92.75	85.23	92.79	84.56	72.89
Proposed model 10	Radiomics + FGSVM	LIDC	91.1	86.5	95.7	87.2	61.5
Proposed model 11	Radiomics + MGSVM	LIDC	91.0	90.4	93.0	94.1	82.4
Proposed model 12	Radiomics + CGSVM	LIDC	91.0	86.9	92.2	90.1	72.3
Proposed model 13	Radiomics + Decision Tree	LIDC	88.3	85.7	92.6	88.5	67.0
Proposed model 14	Radiomics + Ensemble Boosted Trees	LIDC	92.3	87.7	92.3	91.0	75.0
Proposed Model 15	Radiomics + Ensemble Bagged Trees	LIDC	92.9	89.1	93.5	91.8	77.2
Proposed model 16	Radiomics + Ensemble RUSBoosted Trees	LIDC	92.5	86.1	86.9	93.7	84.0
Proposed model 17	Radiomics + Ensemble Subspace Discriminant	LIDC	91.1	86.0	94.8	87.2	62.1
Proposed model 18	Radiomics + Ensemble Subspace KNN	LIDC	89.6	86.5	93.0	89.1	68.9

Bold values give the highest values attained in that column

Conclusions

LC 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. CT imaging provides unparalleled insight into the intricate landscape of lung structures. In recent decades, CAD systems for lung nodule identification have received considerable attention and investigation. Investigating shape and radiomics based on texture extracted using Daubechies WPT, a high-throughput computer approach for quantitative CT image analysis was the aim of this study.

In this study, LIDC dataset consisting of 1018 CT images is utilized for experimentation. From the

sub-images of 11×11 pixels around the nodule centroid, shape features, features using three statistical texture analysis approaches, i.e. GLCM, GLDM, and GLRLM, and Daubechies WPT texture features are extracted. Filter Type feature selection algorithms were used to determine relevant features; Chi-square and ANOVA. After extensive attribute selection eight features were selected as the most significant ones. Finally, classification of cancer into benign or malignant was performed using different cutting-edge ML classifiers. The results show that, BACET gives best AUROC (92.9%), MGSVM gives best accuracy (90.4%), FGSVM yields the best sensitivity (97.8%), MGSVM gives best precision (94.1%) and RUSBOCET gives best specificity (84%). The outcome is better than many cutting-edge methodologies. Therefore, the proposed methodologies can be successfully used for the classification of pulmonary nodules based on CT images and thus can help clinicians to reach better decisions and treatments.

In future work, the study can be extended to the use of other FS methods, applying ML and deep learning techniques experimented with nature-inspired optimization approaches, considering different lung cancer datasets for earlier diagnostics, better decisions, and better lung cancer outcomes.

Abbreviations

LC	Lung Cancer
СТ	Computed Tomography
ML	Machine Learning
LIDC	Lung Image Data Consortium
GLCM	Grav Level Co-occurrence Method
GLDM	Grav Level Difference Method
GLRI M	Grav Level Run Length Matrix
WPT	Wavelet Packet Transform
ANOVA	Analysis of Variance
SVM	Support Vector Machine
EGSVM	Fine Gaussian Support Vector Machine
MEGSVM	Medium Gaussian Support Vector Machine
CEGSVM	Coarse Gaussian Support Vector Machine
DT	Decision Trees
BOCET	Ensemble Boosted Trees
BACET	Ensemble Bagged Trees
	Ensemble BLISBoosted Trees
FS	Ensemble hospoosted nees
WHO	World Health Organization
POL	Pagion of Interest
NSCLC	Non Small Coll Lung Cancor
TR	Tuborculosis
SDNc	Solitary Pulmonary Nodulos
SEINS	Area Linder the Curve
AUC	Filter Turpe FS methods
	K Nearest Neighbor algorithm
	R-Nedlest Neighbor algorithm
	Linear Discriminant Analysis
LDA	
RUC	Autocorrelation
ACOR	Autocorrelation
CON	Contrast
CORT	Correlation 1
COR2	Correlation2
CP	Cluster Prominence
CS	Cluster Shade
DS	Dissimilarity
ENR	Energy
ENI	Entropy
HMGT	Homogeneity
HMG2	Homogeneity2
MP	Maximum Probability
SUS	Sum of Squares: Variance
SA	Sum Average
SV	Sum variance
SEINT	Sum Entropy
DV	Difference variance
DENI	Difference Entropy
IMCI	Information Measure of Correlation I
IMC2	Information Measure of Correlation2
IDM	Inverse Difference Moment
IDMN	Inverse Difference Moment Normalized
ASM	Angular Second Moment
M	Mean
SRF	Short Kun Emphasis

LKE	Long Run Emphasis
GLN	Gray Level Non-uniformity
RLN	Run Length Non-uniformity
RP	Run Percentage
LGRE	Low Gray-Level Run Emphasis
HGRE	High Gray-Level Run Emphasis
SGLGE	Short Run Low Gray-Level Emphasis
SRHGE	Short Run High Gray-Level Emphasis
LRLGE	Long Run Low Gray-Level Emphasis
LRHGE	Long Run High Grav-Level Emphasis

Acknowledgements

Not applicable.

Author contributions

AN contributed to the study concepts, design, literature research, experimental studies/data analysis/statistical analysis and manuscript preparation. AHM is the guarantor of integrity of the entire study. Both authors have read and approved the final manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Availability of data and materials

Available.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 20 December 2023 Accepted: 23 February 2024 Published online: 05 March 2024

References

- Ziyad SR, Radha V, Vaiyapuri T (2021) Noise removal in lung LDCT images by novel discrete wavelet-based denoising with adaptive thresholding technique. Int J E-Health Med Commun 12(5):1–15
- Madero Orozco H, Vergara Villegas OO, Cruz Sánchez VG, Ochoa Domínguez HD, Nandayapa Alfaro MD (2015) Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. Biomed Eng 14(1): 1–20
- Donga HV, Karlapati JS, Desineedi HS, Periasamy P, Sureshkumar TR (2022) Effective Framework for Pulmonary Nodule Classification from CT Images Using the Modified Gradient Boosting Method. Appl Sci 12(16):8264
- Alzubaidi MA, Otoom M, Jaradat H (2021) Comprehensive and comparative global and local feature extraction framework for lung cancer detection using ct scan images. IEEE Access 9:158140–158154
- Chen C-H, Chang C-K, Chih-Yen Tu, Liao W-C, Bing-Ru Wu, Chou K-T, Chiou Y-R, Yang S-N, Zhang G, Huang T-C (2018) Radiomic features analysis in computed tomography images of lung nodule classification. PLoS ONE 13(2):e0192002
- Khehrah N, Farid MS, Bilal S, Khan MH (2020) Lung nodule detection in CT images using statistical and shape-based features. J Imaging 6(2):6
- Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. Radiology 278(2):563–577
- Permatasari Z, Purnomo MH, Purnama IK (2021) Lung nodule detection of CT and image-based GLCM and RLM CT scan using the support vector machine (SVM) method. JAREE 5(2)

- 9. Shakir H, Deng Y, Rasheed H, Khan TM (2019) Radiomics based likelihood functions for cancer diagnosis. Sci Rep 9(1):9501
- Palumbo B, Bianconi F, Palumbo I, Fravolini ML, Minestrini M, Nuvoli S, Stazza ML, Rondini M, Spanu A (2020) Value of shape and texture features from ¹⁸F-FDG PET/CT to discriminate between benign and malignant solitary pulmonary nodules: an experimental evaluation. Diagnostics 10(9):696. https://doi.org/10.3390/diagnostics10090696
- Belfiore MP, Sansone M, Monti R, Marrone S, Fusco R, Nardone V, Grassi R, Reginelli A (2023) Robustness of radiomics in pre-surgical computer tomography of non-small-cell lung cancer. J Personal Med 13(1):83. https://doi.org/10.3390/jpm13010083
- 12. Thattaamuriyil Padmakumari L, Guido G, Caruso D, Nacci I, Del Gaudio A, Zerunian M, Polici M, Gopalakrishnan R, Sayed Mohamed AK, De Santis D, Laghi A et al (2022) The role of chest CT radiomics in diagnosis of lung cancer or tuberculosis: a pilot study.". Diagnostics 12(3):739
- Jing R, Wang J, Li J, Wang X, Li B, Xue F, Shao G, Xue H (2021) A wavelet features derived radiomics nomogram for prediction of malignant and benign early-stage lung nodules. Sci Rep 11(1):22330
- 14. Torres G, Baeza S, Sanchez C, Guasch I, Rosell A, Gil D (2022) An intelligent radiomic approach for lung cancer screening. Appl Sci 12(3):1568
- Balci MA, Batrancea LM, Akgüller Ö, Nichita A (2023) A series-based deep learning approach to lung nodule image classification. Cancers 15(3):843. https://doi.org/10.3390/cancers15030843
- McNitt-Gray MF, Armato SG III, Meyer CR, Reeves AP, McLennan G, Pais RC, Freymann J, Brown MS, Engelmann RM, Bland PH, Laderach GE (2007) The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation. Acad Radiol 14(12):1464–1474
- Tan J, Pu J, Zheng B, Wang X, Leader JK (2010) Computerized comprehensive data analysis of lung imaging database consortium (LIDC). Med Phys 37(7Part1):3802–3808
- Armato SG III, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38(2):915–931
- Baba T, Uramoto H, Takenaka M, Oka S, Shigematsu Y, Shimokawa H, Hanagiri T, Tanaka F (2012) The tumour shape of lung adenocarcinoma is related to the postoperative prognosis. Interact Cardiovasc Thorac Surg 15(1):73–76. https://doi.org/10.1093/icvts/ivs055. Epub 2012 Apr 18.
- Materka A (2004) Texture analysis methodologies for magnetic resonance imaging. Dialogues Clin Neurosci 6:243–250
- O'Connor JPB et al (2015) Imaging intratumor heterogeneity: Role in therapy response, resistance, and clinical outcome. Clin Cancer Res 21:249–257. https://doi.org/10.1158/1078-0432.CCR-14-0990
- Haralick RM, Shanmugam K, Dinstein IH (1973) Textural features for image classification. IEEE Trans Syst Man Cybern 6:610–621
- Mir AH, Hanmandlu M, Tandon SN. Texture analysis of CT images. IEEE Eng Med Biol Mag 14(6):781–786
- Garcia C et al (2000) Wavelet packet analysis for face recognition. Image Vis Comput 18:289–297
- Perlibakas V (2004) Face recognition using principal component analysis of the wavelet packet decomposition. Science Direct Working Paper No S1574-034X(04)70005-8
- Daubechies I, Grossmann A, Meyer Y (1986) Painless nonorthogonal expansions. J Math Phys 27(5):1271–1283
- Rahouma KH, Mabrouk SM, Aouf M (2021) Lung cancer diagnosis based on Chan-Vese active contour and polynomial neural network. Procedia Comput Sci 194:22–31, ISSN 1877-0509, https://doi.org/10.1016/j.procs. 2021.10.056
- Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM (2022) A review of feature selection methods for machine learning-based disease risk prediction. Front Bioinform 2
- Bommert A, Welchowski T, Schmid M, Rahnenführer J (2022) Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. Brief Bioinform 23(1): bbab354
- 30. Shehu I (2022) Testing statistical hypothesis on learning effectiveness: pre-and post-COVID 19. South East Eur J Sustain Dev 6(2)
- Yu Z, Guindani M, Grieco SF, Chen L, Holmes TC, Xiangmin Xu (2022) Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. Neuron 110(1):21–35

- Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10(5):988–999
- Pisner DA, Schnyer DM (2020) Support vector machine. In: Machine learning; Elsevier: Amsterdam, The Netherlands, pp 101–121
- Kecman V (2005) Support vector machines: an introduction. In: Wang L (ed) Support vector machines: theory and applications. studies in fuzziness and soft computing, vol 177. Springer, Berlin
- 35. Rokach L, Maimon O (2005) Decision trees. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer, Boston
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
- 37. Zhou ZH (2021). Ensemble learning. In: Machine learning. Springer, Singapore
- Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques, 3rd edn. Morgan Kaufmann Publishers, Burlington
- Ahn H, Moon H, Fazzari MJ, Lim N, Chen JJ, Kodell RL (2007) Classification by ensembles from random partitions of high-dimensional data. Comput Stat Data Anal 51(12): 6166–6179. ISSN 0167-9473
- 40. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123-140
- Sammut C, Webb GI (eds) (2011) Encyclopedia of machine learning. Springer
- Wang XG, Tang XO (2004) Experimental study on multiple LDA classifier combination for high dimensional data classification. Multiple Classifier Systems. In: 5th International workshop on multiple classifier systems, vol 3077, pp 344–353
- Cai J et al (2023) Impact of localized fine tuning in the performance of segmentation and classification of lung nodules from computed tomography scans using deep learning. Front Oncol 13:1140635. https://doi. org/10.3389/fonc.2023.1140635
- Wang H, Zhu H, Ding L et al (2023) A diagnostic classification of lung nodules using multiple-scale residual network. Sci Rep 13:11322. https:// doi.org/10.1038/s41598-023-38350-z

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.