Improving Lungs Cancer Detection Based on Hybrid Features and Employing Machine Learning Techniques

Jing Yang
Universiti Malaya
Por Lip Yee
Universiti Malaya
Abdullah Ayub Khan
Benazir Bhutto Shaheed University Lyari
Mohammad Shahbaz Khan
Children’s National Hospital
Hanen Karamti
Princess Nourah bint Abdulrahman University
Amjad Aldweesh
Shaqra University
Lal Hussain (✉ lall_hussain2008@live.com)
University of Azad Jammu and Kashmir
Abdulfattah Omar
Prince Sattam Bin Abdulaziz University

Research Article

Keywords: Non-small lung cancer (NSCLC), Small cell lung cancer (SCLC), Support vector Machine (SVM), Classification, autoencoder, Gray-level co-occurrence (GLCM), and Haralick texture features

Posted Date: May 24th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2929395/v1

Additional Declarations: No competing interests reported.

Improving Lungs Cancer Detection Based on Hybrid Features and Employing Machine Learning Techniques

Jing Yang[1,*], Por Lip Yee[2], Abdullah Ayub Khan[3], Mohammad Shahbaz Khan[4], Hanen Karamti[5], , Amjad Aldweesh[6,*], Lal Hussain[7,8, *], Abdulfattah Omar[9]

[1]Department of Computer System and Technology, Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia (Email: Yj741655109@163.com)

[2]Faculty of Computer Science & Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia. (Email:porlip@um.edu.my)

[3]Department of Computer Science and Information Technology, Benazir Bhutto Shaheed University Lyari, Karachi 75660, Pakistan (Email: abdullah.ayub@bbsul.edu.pk)

[4]Children's National Hospital, 111 Michigan AVE NW, Washington, DC, 20854, USA

[5]Department of computer sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

[6]College of Computer science and information technology, Shaqra University, Saudi Arabia

[7]Department of Computer Science and Information Technology, King Abdullah Campus Chatter Kalas, University of Azad Jammu and Kashmir, Muzaffarabad, 13100, Azad Kashmir, Pakistan

[8]Department of Computer Science and Information Technology, Neelum Campus, University of Azad Jammu and Kashmir, Athmuqam, 13230, Azad Kashmir, Pakistan.

[9]Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University, Saudi Arabia,

Corresponding Authors: Lal Hussain (lall_hussain2008@live.com), Amjad Aldweesh and Jing Yang

Abstract: Lung cancer detection using machine learning involves training a model on a dataset of medical images, such as CT scans, to identify patterns and features associated with lung cancer. Past researchers developed different computer aided diagnostic (CAD) systems for early prediction of lung cancer. The researchers extracted single features such as texture, morphology etc.; however, by combining the features, accuracy can be improved. In this study, we extracted Gray-level co-occurrence (GLCM), autoencoder and Haralick texture features. We combined these features and computed the performance using robust machine algorithms including Decision tree (DT), Naïve Bayes (NB) and support vector machine (SVM) with different kernel functions. The performance was evaluated using standard performance measures. The hybrid methods such as GLCM + Autoencoder, and Haralick + Autoencoder yielded highest detection performance using SVM Gaussian and radial base function (RBF) with sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) with accuracy of 100% and AUC 1.00 followed by SVM polynomial yielded an accuracy of 99.89% and AUC of 1.00; GLCM + Haralick using SVM Gaussian yielded accuracy (99.56%), SVM RBF yielded accuracy (99.35%). The results reveal that the proposed feature extraction methodology can be usefully used to predict the lung cancer for further diagnosis at early stage.

Keywords: Non-small lung cancer (NSCLC), Small cell lung cancer (SCLC), Support vector Machine (SVM), Classification, autoencoder, Gray-level co-occurrence (GLCM), and Haralick texture features

MSC: Artificial Intelligence, Machine Learning, Lung Cancer, cross validation

1. Introduction

According to the recent statistics of lung cancer in 2022 [1], there were about 2.36 million new cases of lung cancer expected for diagnosis and out of which 85% belong to non-small cell lung cancer. The non-small cell lung cancer (NSCLC) is diagnosed using radiofrequency (RF) and stereotactic body radiotherapy (SBRT). The other type of lung cancer is small cell lung carcinoma (SCLC). Both types have different methods for treatment and spreading. NSCLC is different from SCLC and slowly grows. While SCLC is growing rapidly related to smoking, spread in whole body quickly and forms tumor. The lung cancer deaths are related to the number of cigarette smoked [2]. NSCLC is so named because the cancer cells in this type of lung cancer do not look small and uniform under a microscope, as they do in small cell lung cancer [3]. The further subtypes of NSCLC include squamous cell carcinoma, large cell carcinoma, and adenocarcinoma. NSCLC is commonly caused by smoking, exposure to radon and air pollution, but also can occur in people who never smoked. The symptoms of NSCLC can include a persistent cough, shortness of breath, chest pain, and coughing up blood.

The SCLC, which is a more aggressive and strong-growing lung cancer type, accounts for about 10-15% of all lung cancer cases [4,5]. It is so named because the cancer cells in this type of lung cancer look small and uniform under a microscope. SCLC is commonly caused by smoking but can also occur in people who never smoked. SCLC often spreads (metastasize) to other parts of body early in the course of the disease, so it is frequently advanced at the time of diagnosis. The diagnosis is typically made with imaging tests, such as a computer tomography (CT) scan or chest X-ray, and confirmed through a biopsy. Treatment options for SCLC typically involve a combination of chemotherapy and radiation therapy [6]. If the cancer is limited to one area of the chest, surgery may be used as well. Prognosis for SCLC is generally poor, with a median survival time of about a year from the time of diagnosis. Due to the aggressive nature of this disease, early diagnosis and treatment are important for improving outcomes.

The SCLC is directly linked with cigarette smoking and aggressive types of lung cancer. Therefore, SCLC have different methods for treatment and diagnosis than NSCLC. The NSCLC early detection can be very helpful with survival rate of 35% to 85% depending on the stage and tumor type. Usually, most of the tumor are late detected so overall 5-year survival rate for NSCLC remains 16% only. Chemotherapy is utilized for SCLC which provokes 60% of response for NSCLC patients. The excessive tobacco uses, and smoking causes the lung cancer around 90% cases. Other factors that may lead to lung cancer include air pollution exposures, radon gas, asbestos and chronic infections. In addition, many hereditary and there have been suggested both inherited and acquired mechanisms of lung cancer susceptibility. Radiation therapy, surgery, targeted therapy and chemotherapy are also choices for lung cancer treatment [7].

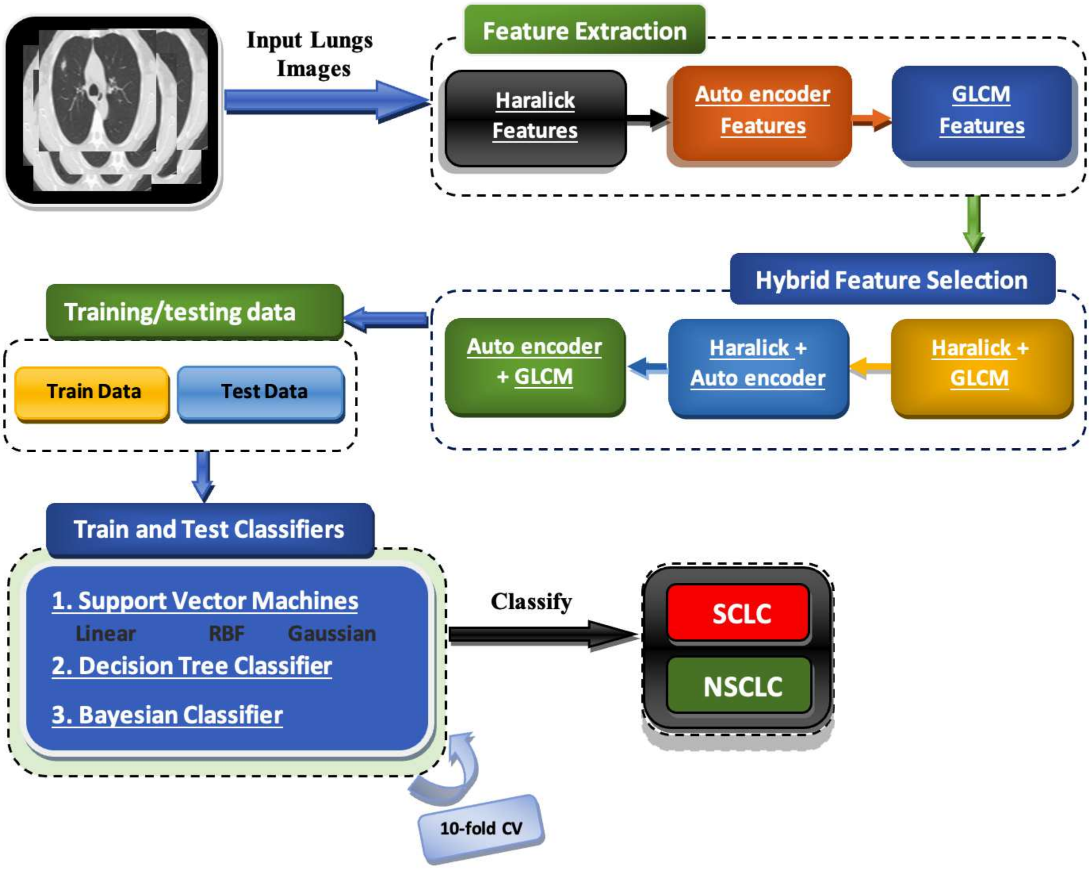
As radiation and x-rays were discovered at the end of the 19th century, physicians used these results to examine the human body and approaches to non-surgical cancer treatment came along. Hospital radiologists and surgeons started working together and with the use of computers, significant cancer data began to accumulate in 1968. For the past 50 years, considerable effort has been made in this field. Tests or imagining modalities typically conducted to evaluate the stage of lung cancer some of them are Computed Tomography (CT), this is the process that includes the detailed pictures of the anatomy and lung tumor are precarious for treatment planning. For cancer staging, CT scans of the chest are essential and the abdominal CT scan is used for locating secondaries and metastases [8]. Positron emission tomography (PET) scan utilize radioactive sugar as cancer cells rapidly uses sugar and is essential for the identification of spread to lymph nodes or other organs [9]. One of the best currently available scans is magnetic resonance imaging (MRI) scan that is used for the scanning of brain. Scanning of brain may be necessary to decide the propagation of tumor in brain [10].

Artificial intelligence (AI) can be used in the detection and diagnosis of lung cancer [11]. One approach is using AI algorithms to analyze medical imaging, such as CT scans, to identify signs of lung cancer. This can help radiologists make more accurate and efficient diagnoses [12]. Additionally, AI can also be used to predict lung cancer progression and the effectiveness of treatment plans using hand-crafted features and dynamic features extraction approaches [13–17]. However, it is important to note that the use of AI in the medical field is still in early stages and further research is required to fully understand its potential and limitations in the detection and treatment of lung cancer.

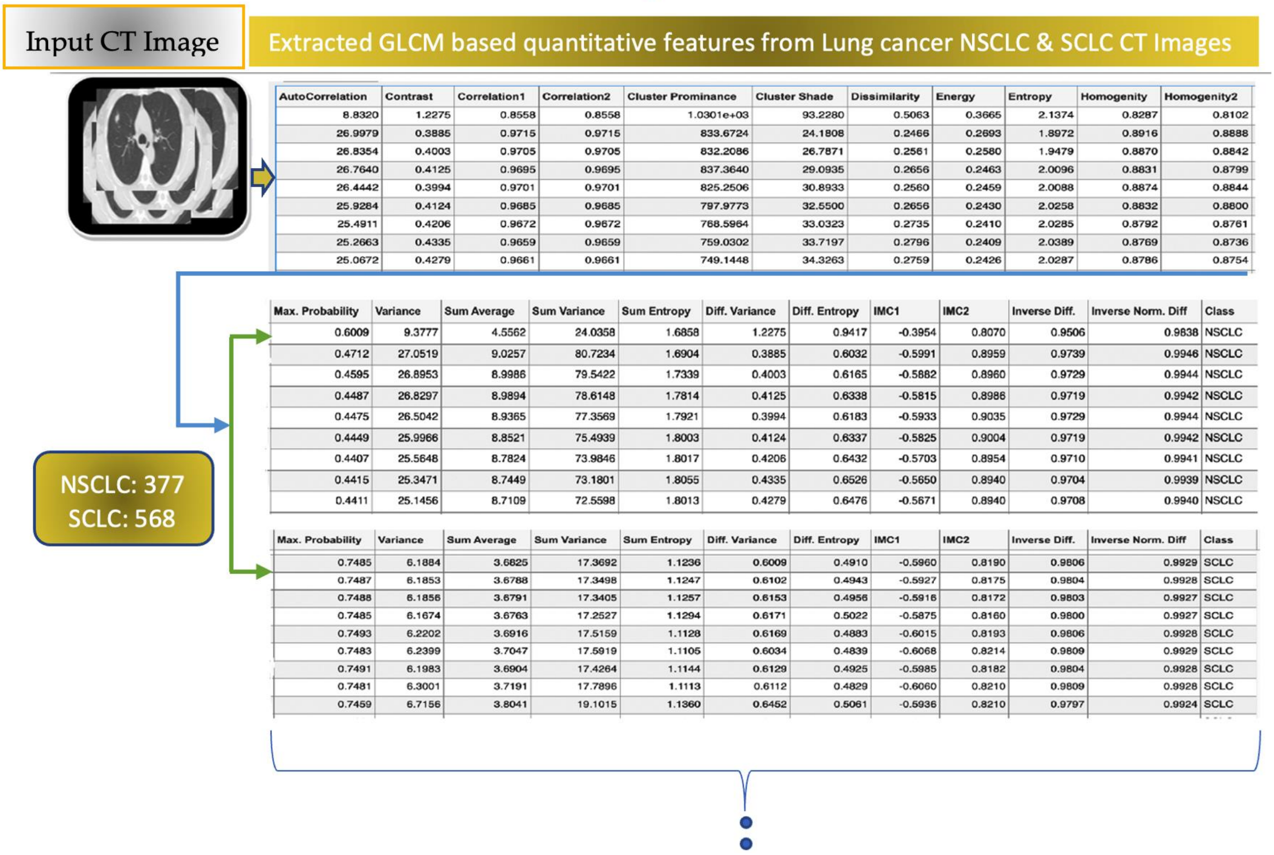
A study proved low-dose spiral CT to be more effective than conventional chest radiography in detecting lung cancer at early stages [18]. Radiological features of CT lung cancer are often Solitary Pulmonary nodules. However, most of the Pulmonary Nodules (PN) in lung cancer have a similar appearance to benign ailments such as tuberculosis, inflammatory pseudo tumor, cardiac tames, and aspergillosis [19]. In 1991, helical CT was implemented in chest imagery and the state of CT images of thoracic structures was dramatically improved [20]. Various rows (4, 8, 16, 32, 64 and 128) of the Analyzer are used for these new CT scanners. With the introduction of Multislice CT scan, high resolution images can be obtained quickly which provide a greater volume of information and a more accurate detection of lung pathologies. A conventional CAD system involves processing multiple images, performing different tasks, and then classifying these into tumors or benign lesions [21]. The CAD device is used specifically to detect lung cancer. This method addresses the issue of designing a computer-based system to obtain the highest features from the differentiated unusual region of the lung CT images, and those features could be utilized to specifically identify lung tumors from the CT as favorable or destructive. A recent study achieved sensitivity of 80% for detecting nodules with malignant potential and resulted in 0.85 false positive readings per section. In short, computer aided diagnosis of lung nodules is likely to have an important role in CT based screening tests in the near future [22].

The previous researchers utilized the single feature extracting approach, limited pre-processing steps and default parameters for machine learning algorithms. The pre-processing steps play a vital role for providing a better and accurate analysis. As this dataset was previously investigated by [23] comprised of CT images using entropy based complexity techniques to investigate the nonlinear hidden dynamics with limited pre-processing steps and single features extracting strategy. The dataset was imbalanced and small, so to avoid overfitting, we utilized the 10-fold cross validation and data augmentation based on random cropping, random flipping, color shifting, Gaussian noise, image scaling. Moreover, the feature extraction strategy also plays a vital role to improve the prediction performance. Researchers are devising tools to improve the feature engineering approach. Apart, single and hybrid features can also matter to improve the diagnostic capability. The hybrid feature approaches are often used for classification tasks, as the relationships between the features and the class labels are often complex and cannot be modeled using a single feature. The main contributions of this study are:

- Hybrid approach: We utilized the hybrid features approach. We extracted different features comprising of GLCM, Haralick texture, and autoencoder. We combined these features by concatenating which combined the contributions of different features combination. To the best of our knowledge, the hybrid features extracting strategy along with diverse pre-processing steps and parametric optimization approach is utilized which further improved the prediction performance.
- Parametric optimization: Likewise, the machine learning classification algorithms performance can be further improved by optimizing the hyperparameters. We utilized grid search method to obtain the optimal features of machine learning algorithms. We then fed these features as single, and hybrid features by concatenating different features to different robust machine learning (ML) classification algorithms after optimizing their parameters as reflected in Figure 1. The proposed approach yielded the highest detection performance.



a)


Input CT Image
Extracted GLCM based quantitative features from Lung cancer NSCLC & SCLC CT Images
NSCLC: 377
SCLC: 568

AutoCorrelation	Contrast	Correlation1	Correlation2	Cluster Prominance	Cluster Shade	Dissimilarity	Energy	Entropy	Homogenity	Homogenity2
8.8320	1.2275	0.8558	0.8558	1.0301e+03	93.2280	0.5063	0.3665	2.1374	0.8287	0.8102
26.9979	0.3885	0.9715	0.9715	833.6724	24.1808	0.2466	0.2693	1.8972	0.8916	0.8888
26.8354	0.4003	0.9705	0.9705	832.2086	26.7871	0.2561	0.2580	1.9479	0.8870	0.8842
26.7640	0.4125	0.9695	0.9695	837.3640	29.0935	0.2656	0.2463	2.0096	0.8831	0.8799
26.4442	0.3994	0.9701	0.9701	825.2506	30.8933	0.2560	0.2459	2.0088	0.8874	0.8844
25.9284	0.4124	0.9685	0.9685	797.9773	32.5500	0.2656	0.2430	2.0258	0.8832	0.8800
25.4911	0.4206	0.9672	0.9672	768.5964	33.0323	0.2735	0.2410	2.0285	0.8792	0.8761
25.2663	0.4335	0.9659	0.9659	759.0302	33.7197	0.2796	0.2409	2.0389	0.8769	0.8736
25.0672	0.4279	0.9661	0.9661	749.1448	34.3263	0.2759	0.2426	2.0287	0.8786	0.8754

Max. Probability	Variance	Sum Average	Sum Variance	Sum Entropy	Diff. Variance	Diff. Entropy	IMC1	IMC2	Inverse Diff.	Inverse Norm. Diff	Class
0.6009	9.3777	4.5562	24.0358	1.6858	1.2275	0.9417	-0.3954	0.8070	0.9506	0.9838	NSCLC
0.4712	27.0519	9.0257	80.7234	1.6904	0.3885	0.6032	-0.5991	0.8959	0.9739	0.9946	NSCLC
0.4595	26.8953	8.9986	79.5422	1.7339	0.4003	0.6165	-0.5882	0.8960	0.9729	0.9944	NSCLC
0.4487	26.8297	8.9894	78.6148	1.7814	0.4125	0.6338	-0.5815	0.8986	0.9719	0.9942	NSCLC
0.4475	26.5042	8.9365	77.3569	1.7921	0.3994	0.6183	-0.5933	0.9035	0.9729	0.9944	NSCLC
0.4449	25.9966	8.8521	75.4939	1.8003	0.4124	0.6337	-0.5825	0.9004	0.9719	0.9942	NSCLC
0.4407	25.5648	8.7824	73.9846	1.8017	0.4206	0.6432	-0.5703	0.8954	0.9710	0.9941	NSCLC
0.4415	25.3471	8.7449	73.1801	1.8055	0.4335	0.6526	-0.5650	0.8940	0.9704	0.9939	NSCLC
0.4411	25.1456	8.7109	72.5598	1.8013	0.4279	0.6476	-0.5671	0.8940	0.9708	0.9940	NSCLC

Max. Probability	Variance	Sum Average	Sum Variance	Sum Entropy	Diff. Variance	Diff. Entropy	IMC1	IMC2	Inverse Diff.	Inverse Norm. Diff	Class
0.7485	6.1884	3.6825	17.3692	1.1236	0.6009	0.4910	-0.5960	0.8190	0.9806	0.9929	SCLC
0.7487	6.1853	3.6788	17.3498	1.1247	0.6102	0.4943	-0.5927	0.8175	0.9804	0.9928	SCLC
0.7488	6.1856	3.6791	17.3405	1.1257	0.6153	0.4956	-0.5916	0.8172	0.9803	0.9927	SCLC
0.7485	6.1674	3.6763	17.2527	1.1294	0.6171	0.5022	-0.5875	0.8160	0.9800	0.9927	SCLC
0.7493	6.2202	3.6916	17.5159	1.1128	0.6169	0.4883	-0.6015	0.8193	0.9806	0.9928	SCLC
0.7483	6.2399	3.7047	17.5919	1.1105	0.6034	0.4839	-0.6068	0.8214	0.9809	0.9929	SCLC
0.7491	6.1983	3.6904	17.4264	1.1144	0.6129	0.4925	-0.5985	0.8182	0.9804	0.9928	SCLC
0.7481	6.3001	3.7191	17.7896	1.1113	0.6112	0.4829	-0.6060	0.8210	0.9809	0.9928	SCLC
0.7459	6.7156	3.8041	19.1015	1.1360	0.6452	0.5061	-0.5936	0.8210	0.9797	0.9924	SCLC

b)

Fig 1: Schematic Diagram with hybrid features extraction approach to detect lung cancer a) Over schematic diagram, b) Extraction of GLCM quantitative features from lung cancer CT images

Figure 1 a) depicts the schematic flow of our work. In the first step, the lung cancer images are read as input. We then extracted Autoencoder, Haralick and GLCM features from these images. Figure 1 b) reflects few examples of the GLCM features extracted from NSCLC and SCLC subjects. Similarly other features were extracted from lung cancer types. We then utilized the single and hybrid features extracting approach. We combined the features with hybrid approach such as GLCM + Autoencoder, Haralick + Autoencoder, GLCM + Haralick. These hybrid features are then fed as input to machine learning classifiers such as SVM with different kernels, NB and DT by optimizing the parameters of these algorithms.

2. Materials and Methods

2.1. Datasets

The dataset utilized in this study, provided by Lung Cancer Alliance (LCA) can be obtained at request on their official website (https://www.prnewswire.com/news-releases/lung-cancer-alliance-launches-first-open-access-patient-driven-website-for-ct-scans-and-clinical-data-95842964.html). This dataset is utilized and detailed previously by [23] and similar other studies. The database images are in the Digital Imaging and Communications in Medicin (DICOM) format with total 76 patients with a total of 945 images including 377 images of NSCLC and 568 of SCLC subjects.

2.2.Pre-Processing

Image pre-processing refers to the techniques used to prepare an image for analysis or processing in computer vision applications [24,25]. The purpose of image pre-processing is to enhance the quality of an image and make it easier to analyze, segment, and extract features from image.

Image Resize

Image resizing refers to the process of changing the size of an image. This can be done for various reasons, such as to fit the image into a specific space, to reduce the file size, or to increase the resolution. There are two common methods for resizing images: interpolation and resampling [26]. We used interpolation which involves estimating the value of pixels in an enlarged image based on the values of surrounding pixels. When resizing an image, it is important to consider the aspect ratio of the image, which is the proportion of the width to the height. If the aspect ratio is not preserved during resizing, the image may become distorted. To preserve the aspect ratio, the image can be resized proportionally [27], either by specifying only one dimension and letting the other dimension be calculated automatically, or by using an aspect ratio constraint.

Data Augmentation

Data augmentation is a technique used in machine learning to artificially increase the size of a dataset by generating modified versions of the original data [28]. The goal of data augmentation is to reduce overfitting, which occurs when a machine learning model performs well on the training data but poorly on new, unseen data. By increasing the size and diversity of the training dataset, data augmentation can help to prevent overfitting and improve the generalization performance of a machine learning model.

We utilized the following data augmentation techniques:

1. Random cropping: Randomly cropping a portion of the original image can increase the size of the dataset and provide new perspectives on the objects in the image [29].
2. Random flipping: Randomly flipping [30] the image horizontally or vertically can provide new views of the objects in the image and help the model to learn more robust features.
3. Random rotation: Randomly rotating [31] the image can help the model to learn features that are invariant to orientation.
4. Color shifting: Changing the brightness, saturation, or hue of the image can help the model to learn features that are invariant to color [32].
5. Gaussian noise: Adding Gaussian noise to the image can help the model to be more robust to noise in the input data.
6. Image scaling: Scaling the image up or down can help the model to learn features that are invariant to scale [33].

2.3.Hyperparameters optimization

Hyperparameter optimization is the process of selecting the best hyperparameters for a machine learning algorithm. Hyperparameters are parameters that are not learned from the training data, but rather set prior to training [34]. They control the learning process of the algorithm and can have a significant impact on its performance.

We utilized the following hyperparameters for Support Vector Machines (SVM), Naïve Bayes, and Decision Trees:

Support Vector Machines (SVM:

1. C: The C hyperparameter controls the trade-off between achieving a low training error and a low testing error. A smaller C value will result in a wider margin and a lower training error, while a larger C value will result in a narrower margin and a higher training error.
2. Gamma: The gamma hyperparameter determines the shape of the radial basis function that is used to map the input data to a higher-dimensional space. A smaller gamma value will result in a more complex model, while a larger gamma value will result in a simpler model

Naïve Bayes Smoothing parameter: The smoothing parameter, also known as Laplace smoothing, controls the strength of the smoothing applied to the probabilities in the Naïve Bayes model [35–37]. A larger smoothing parameter will result in a smoother probability distribution, while a smaller smoothing parameter will result in a more discrete distribution.

Decision Trees

For decision tree, the parameters utilized were [38–41]:

1. Maximum depth: The maximum depth hyperparameter controls the maximum depth of the decision tree. A smaller maximum depth will result in a simpler model, while a larger maximum depth will result in a more complex model.
2. Minimum samples per leaf: The minimum samples per leaf hyperparameter controls the minimum number of samples that must be present in a leaf node in order for it to split. A smaller minimum samples per leaf will result in a more complex model, while a larger minimum samples per leaf will result in a simpler model.

   Minimum samples per split: The minimum samples per split hyperparameter controls the minimum number of samples that must be present in a split in order for it to occur. A smaller minimum samples per split will result in a more complex model, while a larger minimum samples per split will result in a simpler model.

2.4.Grid Search Method

To determine the optimal hyperparameters for a machine learning algorithm, one commonly used method is grid search, which involves training the model with different combinations of hyperparameters and selecting the combination that results in the best performance on a validation dataset [42–45]. Another method is random search, which involves randomly sampling hyperparameter combinations and selecting the combination that results in the best performance on a validation dataset. Grid search is a technique for hyperparameter optimization in machine learning. It involves systematically searching over a predefined set of hyperparameter values, training the model with each combination of hyperparameter values, and selecting the combination of values that results in the best performance on a validation set.

The following procedure was utilized:

1. Define a set of hyperparameters for the model, along with a range of possible values for each hyperparameter.
2. Create a grid of all possible combinations of hyperparameter values.
3. Train the model with each combination of hyperparameter values, using a validation set to evaluate the performance of the model for each combination.
4. Select the combination of hyperparameter values that result in the best performance on the validation set.
5. Train the final model using the selected hyperparameter values and the entire training dataset.

Grid search is a simple and effective method for hyperparameter optimization [42–45], but it can be computationally expensive and time-consuming, especially for models with many hyperparameters or a large range of possible values for each hyperparameter. For this reason, more efficient methods for hyperparameter optimization, such as randomized search and Bayesian optimization, have been developed. These methods are able to explore the hyperparameter space more efficiently and often converge to the optimal hyperparameters more quickly than grid search.

2.5.Training/ Testing Data Validation

The 10-fold cross validation is a technique used to evaluate the performance of a machine learning model [46]. It is a resampling method that involves dividing the dataset into 10 equal-sized subsets, or "folds", and then training and evaluating the model on 9 of the folds while using the remaining fold as the validation set [47]. It involves dividing the original training dataset into K folds (where K is a positive integer), and then training the model K times, with each time using a different fold as the validation set and the remaining K-1 folds as the training set.

2.6.Feature extraction

Feature extraction is a process of identifying and extracting relevant information from datasets in order to represent them in a more compact and informative way. In the context of machine learning, feature extraction is often used to pre-process data before they are fed into a model. The goal is to extract features that are most relevant to the task at hand and that capture the underlying patterns in the data. The machine learning algorithms requires the most specific feature extracting approach. Recently, the researchers computed hybrid features [48–50] ,and different features extracting approaches [50–54] to improve various imaging pathologies detection. In this study, we have applied feature extraction strategies which are GLCM, Haralick, and Autoencoder.

2.6.1. *Haralick Texture Features*

The images texture properties are computed using Haralick Features. The Haralick features were utilized in previous studies to solve many classification problems [55–57], specifically for colon biopsy classification [58,59]. In this effort, we suggest the Haralick texture feature are extracted from lung cancer Computer Tomography (CT) images.

1.Mean:

It is instantly associated with the heterogeneity of the image spectrum. The mean for an image is measured utilizing below equation.

$$M = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} p(i,j) \tag{1}$$

2. Variance:

The variance measures the spread of the distribution about the mean value in the image. It describes the intensity variation around the mean.

$$V_{ar} \ = \sum_{i=1}^{N-1}\sum_{J=1}^{N-1}(i-\mu)^2 p(i,j) \tag{2}$$

3. Standard Deviation:

The standard deviation has been the second pivotal moment that measured contrast population. It is likely to be distributed and therefore can be represented as a measure of high and low contrast homogeneity.

$$S_{dev} = \left[\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(p(i,j)-\mu)\right]^{1/2} \tag{3}$$

4. Contrast:

The contrast is defined as a measurement of the intensity between the pixels and the surrounding images. Mathematically:

$$C_{ont} = \sum_{i=1}^{N}\sum_{j=1}^{N}(i-j)^2\, p_{ij} \tag{4}$$

5. Entropy:

Entropy is the randomness measurement used to distinguish the input image texture. The following expression describes it as:

$$E_{tro} = \sum_{i}\sum_{j} p(i,j)\log p(i,j) \tag{5}$$

6. Correlation:

Correlation is widely used in statistics, data analysis and machine learning. It is used to identify the relationship between different features and target variable in a dataset to select the most relevant features and understand the underlying patterns in the data. Mathematically.

$$C_{orr} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{(i-m_i)(j-m_j)}{\sigma_i\sigma_j} \quad (6)$$

7. Energy:

Energy can be described as measuring the magnitude of repetitions of pixel pairs. It estimates the image's regularity. The energy preference will be high if the pixels are quite similar.

$$E_{ner} = \sum_{i=1}^{N}\sum_{j=1}^{N} p_{(ij)^2} \quad (7)$$

8. Homogeneity:

Homogeneity is an important consideration when working with data, as it can affect the performance of a model. In general, a more homogeneous dataset will lead to better model performance, while a more heterogeneous dataset may require more advanced techniques to handle the variability in the data. Mathematically,

$$H_{gen} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{p_{ij}}{1+|i-j|} \quad (8)$$

9. Kurtosis:

In digital image processing, kurtosis values are computed through noise and resolutions measurements.

$$K_{urt} = \frac{1}{MN\sigma^4}\sum_{i=1}^{M}\sum_{j=1}^{N}(p(i.j)-\mu)^4 \quad (9)$$

10. Skewness:

This feature is based on geometrical moments of patches of images. Skewness is an asymmetry measure or absence of symmetry.

$$\boldsymbol{S_{kew} = \frac{1}{MN\sigma^3}\sum_{i=1}^{M}\sum_{j=1}^{N}(p(i,j)-\mu)^3} \quad (\mathbf{10})$$

2.6.2. *Grey Level Co-occurrence Matrix (GLCM)*

GLCM was one of the most popular techniques of consistency studies. As a unique and popular texture analysis tool, it considers the properties of images associated with second-order figures.

The performance is equated with a microprocessor solution. GLCM, first presented by Haralick, is a forceful method for calculating consistency features. Suppose a picture to be investigated is quadrangular and has N_x column and N_y row. Supposing that the gray level looking to pixels is quantize to N_g Ng levels. Let $L_x = 1,2,3,4,5,6 \ldots\ldots\ldots. N_x$ be the column, $L_y = 1,2,3,4,5,6 \ldots\ldots\ldots. N_y$ be the row, and $H = 1,2,3,4,5,6 \ldots\ldots\ldots. N_g$ be the set of N_g quantized grey level. The texture framework data is specified by the matrix of comparative frequencies $Q_{u,v}$ with two adjacent pixels parted by shift c and angle θ.1 The GLCM is calculated with the subsequent equation:

Supposing a picture to be investigated is quadrangular and has N_x column and N_y row. Supposing that the gray level looking to pixels is quantize to N_g Ng levels. Let $L_x = 1,2,3,4,5,6 \ldots\ldots\ldots. N_x$ be the column, $L_y = 1,2,3,4,5,6 \ldots\ldots\ldots. N_y$ be the row, and $H = 1,2,3,4,5,6 \ldots\ldots\ldots. N_g$ be the set of N_g quantized grey level. The texture framework data is specified by the matrix of comparative frequencies $Q_{u,v}$ with two adjacent pixels parted by shift c and angle θ.1 The GLCM is calculated with the subsequent equation:

$$Q(u,v,d,0) = \{(x_1,y_1),(x_2,y_2)\, f(x_1,y_1) = i$$
$$H(\, x_2,y_2) = j,(x_1,y_1)-(x_2,y_2) = c \quad (11)$$
$$K = (x_1,y_1),(x_2,y_2)$$

Where x, y is the amount of incidences within the windows magnitudes, where the strength rank of a pixel two of kind deviations as of v to u the position of the 1st pixel is (x_1, y_1) and that of the 2nd pixel is $(x_2, y_2), c$ is the distance be-tween the pixel couple, θ is the point of view in the two pixels. The synchronize matrix define is not the symmetric. If the GLCM is considered with a symmetry, after that one viewpoint up to 180o want to be measured. A symmetric synchronize matrix can be figured by an appearance.

$$Q(u, v, d, 0)\ k = (Q(u, v, d, 0) + Q(u, v, d, 0), T)\ divid\ by\ 2 \quad (12)$$

Where $Q(u, v, d, 0), T$ is a transpose of $Q(u, v, d, 0)$. Possibility approximations are gained by dividing to each entry in $Q(u, v, d, 0)$ by the sum of entirely probable intensity variations (K_x, K_y) with the space d and track θ i.e.

$$(K_x, K_y), d, 0$$

Thus, a normal form of GLCM is gained as seen in the following equation.

$$Q(u, v, d, 0)^{\ X} = \frac{P(i, j, d, \theta) + P^T(i, j, d\theta)}{2 * \sum \theta L_x L_y, d, \theta} \quad (13)$$

Where, $Q(u, v, d, 0)^{\ X}$ is standardized GLCM. The expression:

$$2 \times (K_x, K_y), c, 0$$

Equation (3.13) is constant, and Equation (3.12) can be reworded as in Equation 14 to be appropriate and to evade separation on FPGA while addressing the separation needs of the additional hardware properties that can reduce the performance of FPGA strategy.

$$Q(u, v, d, 0)^{\ N} = \frac{P(i, j, d, \theta) + P^T(i, j, d\theta)}{2 * \sum \theta L_x L_y, d, \theta} \quad (14)$$

The above mathematical framework of GLCM is a square matrix with components that corresponds to the relative frequency of occurrence of gray-level pairs of pixels separated in a specified direction by a certain range. For a displacement vector $d(dx, dy)$, the elements of a $G \times G$ gray level co-occurrence matrix are identified as $Pd = (i, j) = \left|\{((r, s)), (t, v)) : I(r, s) = i, I(t, v) = j\}\right|$ where I indicates the GLCM image value and $(r, s), (t, v)$ and (dx, dy) is the cardinality set. Let $Q\ (u, v, d, 0)$ is a (normalized) occurrence of rate of grey level pair (u, v) at space c and viewpoint θ and N_g be the quantity of gray levels.

$$ASM = \sum_{i=1}^{N_R} \sum_{j=1}^{N_R} P(i, j)^2 \quad (15)$$

$$SM = \sum_{i=1}^{N_R} \sum_{j=1}^{N_R} P(i - j)^2 P(i, j) \quad (16)$$

$$SM = \sum_{i=1}^{N_R} \sum_{j=1}^{N_R} \frac{(i - \mu_x)(j - \mu_y)(P(i, j))}{\sigma_x \sigma_y} \quad (17)$$

Where μ_x, μ_y, σ_x, σ_y are the variances and the mean of the rows and columns sums correspondingly, they are defined as follows:

$$\mu_x = \sum_{i=1}^{N_R} \sum_{j=1}^{N_R} i\ P(i, j)^2 \quad (18)$$

$$\mu_y = \sum_{i=1}^{N_R} \sum_{j=1}^{N_R} jP(i, j)^2 \quad (19)$$

$$\sigma_x = \sqrt{\sum_{i=1}^{N_R}\sum_{j=1}^{N_R}(i - \mu_x)^2 P(i,j)^2} \qquad (20)$$

$$\sigma_y = \sqrt{\sum_{i=1}^{N_R}\sum_{j=1}^{N_R}(i - \mu_y)^2 P(i,j)^2} \qquad (21$$

For the symmetric matrix μ_x= $\mu_y$and the de-nominator for the Equation. (2.18) decreases to the variance (sigma squared) are as stated in Equation (2.20). Inequality is utilized to calculate the disparity in the two gray levels of i and j. Entropy is the second calculation at an angle to the minute. This informal journalist has little value in a series of irregular events.

According to Brynolfsson et al., [60] the analysis of the medical imaging plot in studies to diagnose, classify and evaluate cancer is very popular. Despite many requests for oncology and medical imaging in general, there is no agreement on the workbench of the plots or on the generation of reports of defined parameters to replicate the results.

2.6.3. *Autoencoder*

Autoencoder is a kind of artificial neural network (ANN) that is used to learn to encode data efficiently. Autoencoder picks up to compress the input phrase data into a function code, which is then compressed to get something that exactly matches the basic data. This allows the car to reduce size by learning, for example, to ignore noise. We start by evoking the outmoded autoencoder model, for instance the one castoff in (Bengio & Lecun, 2007) to make profound systems. An autoencoder takes in an effort trajectory $i = [0,1], d$ and first plots it to an unseen picture $j = [0,1], d$ over a regulate plotting $j = f_0(x) = k(wx + b)$, parameter by $0 = \{w, b\}$. W is a $d \times d$ bulk matrix and b is a prejudice vector. The resultant dormant symbol y is then plotted support to a reconstructed vector a $[v, d]$in input space $a = g_0(y) = s(wy + b)$ with the $0 = (wb)$.

The Weight matrix w of the opposite plotting might predictably be forced by $w = (wt)$, in which event the autoencoder is supposed to consume secured weights. Each exercise x_i is thus plotted to an equivalent y_i and a rebuilding z_i. The parameters of this prototype are enhanced. To abate the standard rebuilding error:

$$\theta^*, \theta'^* = arg_{\theta,\theta'} \frac{1}{n} \sum_{i=0}^{n} L\left(x^{(i)}, z^{(i)}\right) \qquad (22)$$

$$arg_{\theta,\underline{\theta}} \frac{1}{n} \sum_{i=0}^{n} L\left(x^{(i)}, g_\theta\left(f_\theta\left(x^{(i)}\right)\right)\right) \qquad (23)$$

Where 0l is a cost function for instance the basis formed error L (x, z) =||x − ||2, $L = (x, z) = \left||x - 2|\right|$ another defeat, recommended by the explanation of z and x as moreover bit routes or vectors of spot likelihoods (Bernoulli's) is the makeover cross-entropy:

$$L_H(x, z) = H(B_x \,||\, B_z) = -\sum_{k=1}^{d} x\,(x_k log z_k + (1 - x_k)\text{Log}\,(1 - z_k))\ (24)$$

Record that if $x$is a binary vector, (x,z) is a negative log-likelihood. For the example x, set the Bernoulli parameters z. Equation can be written as:

$$\theta^*, \theta'^* = arg_{\theta,\theta'} E_{q^o}(x) \left[L_H\left(X, g_\theta\left(f_\theta(X)\right)\right)\right] \qquad (25)$$

Where $q_0 = (x)$ symbolizes the empirical delivery related to n working out effort. This optimization will classically be approved out by stochastic rise origin. We will use the automatic encoder in this survey. An automatic encoder is a type of artificial neural network that can be used to learn how to effectively encode data unattended. The purpose of the automatic encoder is to represent (encode) a data set to learn how to generally reduce dimensionality and to train the network to ignore the "noise" of the signal.

Hybrid features

Hybrid features in machine learning refer to the combination of multiple features or feature sets in order to improve the performance of a machine learning model [61]. This can be done by combining features from different sources, such as text, images, or audio, or by combining different types of features, such as low-level and high-level features [62].

There are several ways to generate hybrid features as detailed in [63,64]. We utilized the hybrid features by concatenating different extracted features. Hybrid features can often improve the performance of a machine learning model by providing additional information or context that can help the model make more accurate predictions. However, it's important to note that the use of hybrid features may also increase the complexity of the model and require more computational resources.

A good feature extraction technique and feature descriptor should be capable of extracting the required features for lung cancer nodule recognition[65]. We had a bag of features available to represent an image. The basic primitive features are based on Haralick, GLCM, and Autoencoder. Any one of these features is not sufficient to get high performance results [66]. Few studies [61-64] have been conducted on Hybrid features systems to solve problems and got significant performance. In this paper, hybrid features are introduced with the combination of Haralick + GLCM, GLCM + Autoencoder, and Autoencoder + Haralick.

The Individual features are combined by concatenating single features to make a hybrid feature vector as utilized by [68,69] with following procedure.

a) Single Features

GLCM (22)	Haralick (14)	Autoencoder (50)
Autocorrelation Contrast Correlation1 Correlation2 Cluster Prominance Cluster shade Dissimilarity Energy Entropy Homogenity1 Homogenity2 Max. Probability Sum of Sqr. Var. Sum avg Sum Var. Sum entropy Diff. Var. Diff. Ent. Inf. measure of Corr1 Info. measure of Corr2 Inverse Diff. Inverse Diff. Normalized Inverse Diff. Movement normalized	Contrast Correlation Variance Inverse Diff. Movement Sum Avg. Sum Var. Sum Ent. Entropy Diff. Var. Diff. Entropy Inf. measure of Corr1 Info. measure of Corr2 Maximal Correlation Coefficient	Latent variables: Encoder features: Decoder features: Bottleneck features: Reconstruction features: Other features including edges, textures, shapes, or colors

b) Hybrid Features

GLCM + Haralick (36)	GLCM + Autoencoder (72)	Haralick + Autoencoder (64)
(1-22) + (23-36)	(1-22) + (23-72)	(1-14) +(15-64)

GLCM + Haralick

1. Load images and apply any necessary preprocessing steps
2. For each preprocessed image
2.1. Compute the gray-level co-occurrence matrix (GLCM) with specified parameters
2.2. Compute Haralick texture features from the GLCM
3. Combine the Haralick features into a feature vector for each image
4. Train a machine learning model on the feature vector

GLCM + Autoencoder

1. Load images and apply any necessary preprocessing steps
2. For each preprocessed image
    2.1. Compute the gray-level co-occurrence matrix (GLCM) with specified parameters.

2.2. Flatten the image and use the encoder part of a pre-trained autoencoder to extract features.

3. Combine the autoencoder features with the GLCM features into a feature vector for each image.
4. Train a machine learning model on the feature vector

Haralick + Autoencoder

1. Load images and apply any necessary preprocessing steps
2. For each preprocessed image

    2.1. Compute Haralick texture features using the GLCM with specified parameters.

    2.2. Flatten the image and use the encoder part of a pre-trained autoencoder to extract features.
3. Combine the autoencoder features with the Haralick features into a feature vector for each image.
4. Train a machine learning model on the feature vectors

2.7.Classification

The final phase of the proposed method is classification, in which three classifiers comprised of SVM, NB and DT were utilized. The classification is a type of supervised learning method [70–73]. Recently, there are many applications of machine learning classification algorithms detailed in [69–75]. It is a categorization method that distinguishes, predicts, and understands objects and ideas. Using this model, the accuracy and many other parameters of the performance assessment are calculated based on extracted features. The identified test sample label matches the model's classified results. Training, testing, and validation were used for 10-fold cross-validation.

2.7.Area under the receiver operating characterstics (ROC) curve (AUC-ROC)

The AUC-ROC is a measure of the performance of a binary classification model [81]. It is a way to evaluate how well a model is able to distinguish between two classes, such as positive and negative, by plotting TPR against the at various threshold settings. The AUC-ROC is then calculated as the area under this curve. A value of 1.0 represents a perfect model, and a value of 0.5 represents a model that is no better than random guessing. The AUC-ROC is a useful metric because it is independent of the threshold settings, and it can be used to compare the performance of different models [82]. It can also be used when the dataset is imbalanced, as it does not rely on a predefined threshold for classifying instances.

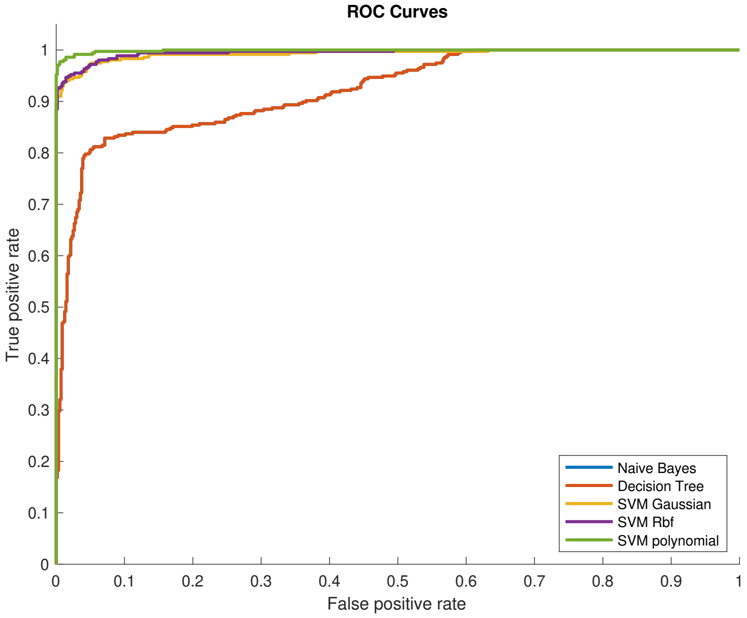
3. Results and Discussions

In this study, we extracted different features such as GLCM, autoencoder, and Haralick features from NSCLC and SCLC subjects and employed machine learning techniques and evaluated performance using standard performance metrics such as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), false positive rate (FPR), accuracy and area under the curve (AUC). We applied single and hybrid feature extracting approach. The aim of this study was to employ the hybrid features for improving the detection performance.

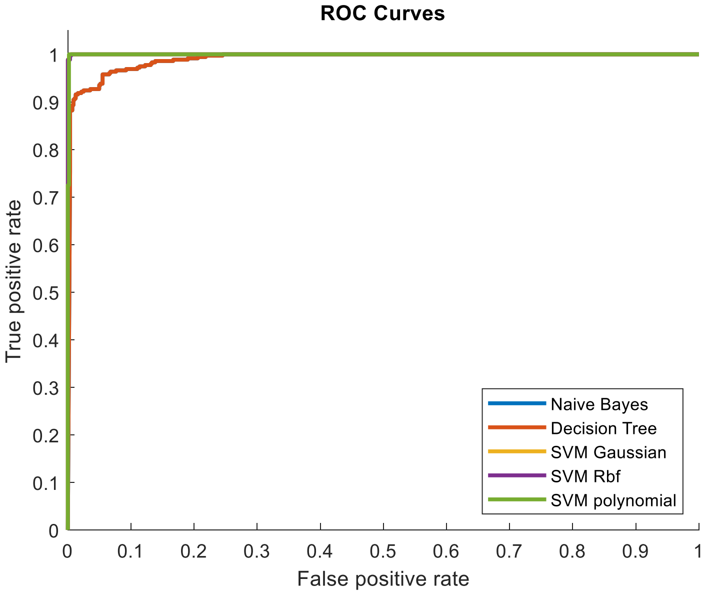
Table 1 presents the lung cancer detection results using single feature extracting strategy by computing Haralick, GLCM and SIFT features. The Haralick texture features yielded they highest accuracy (99.89%), sensitivity (100%), specificity (99.72%), PPV (99.82%), NPV (100%), FPR (0.002809) and AUC (0.9984) using SVM polynomial followed by SVM RBF with accuracy (99.24%), SVM Gaussian with accuracy (98.91%), Decision tree with accuracy (97.78%) and Naïve Bayes with accuracy (93.47%). By extracting the GLCM features, the SVM polynomial yielded the highest performance with accuracy (98.69%) followed by SVM RBF with accuracy (96.84%) and so on. The SIFT features yielded highest accuracy (98.31%) using SVM Gaussian followed by Naïve Bayes with accuracy (96.58%).

Table 1: Detection performance with single features based extracting strategy and employing robust machine learning techniques to distinguish NSCLC from SCLC

Method	Sensitivity	Specificity	PPV	NPV	Accuracy	FPR	AUC
Haralick							
Naïve Bayes	0.9485	0.9129	0.9451	0.9181	0.9347	0.08708	0.9837
Decision Tree	0.9858	0.9916	0.9946	0.9778	0.988	0.008427	0.9837
SVM Gaussian	0.9964	0.9775	0.9859	0.9943	0.9891	0.02247	0.9999
SVM RBF	0.9982	0.9831	0.9894	0.9972	0.9924	0.01685	0.9999
SVM poly.	1	0.9972	0.9982	1	0.9989	0.002809	0.9984
GLCM							
Naïve Bayes	0.7869	0.8567	0.8968	0.7176	0.8139	0.1433	0.8524
Decision Tree	0.9574	0.9579	0.9729	0.9342	0.9576	0.04213	0.9224
SVM Gaussian	0.9929	0.9129	0.9475	0.9878	0.9619	0.08708	0.9928
SVM RBF	0.9964	0.9242	0.9541	0.994	0.9684	0.07584	0.9948
SVM poly.	0.9982	0.9691	0.9808	0.9971	0.9869	0.0309	0.9989
Scale Invariant Feature transform (SIFT)							
Naïve Bayes	0.9658	0.9672	0.9658	0.9637	0.9658	0.00481	0.9797
Decision Tree	0.9474	0.9415	0.9413	0.9451	0.9475	0.0621	0.9631
SVM Gaussian	0.983	0.9816	0.983	0.9816	0.9831	0.01842	0.995
SVM RBF	0.9605	0.9619	0.9605	0.9675	0.9605	0.00812	0.9691
SVM poly.	0.9405	0.9419	0.9405	0.9475	0.9405	0.00112	0.9697



a)



b)

Fig. 2 Area under the receiver operating curve to distinguish NSCLC from SCLC by extracting hybrid feature a) GLCM b) Haralick texture features.

Figure 2 shows the area under the receiver operating characteristic curve (AUC) with single feature extracting approach by computing a) GLCM and b) Haralick texture features. Based on GLCM features, the highest separation was obtained using SVM polynomial with AUC (0.9989) followed SVM RBF with AUC (0.9948), SVM Gaussian with AUC (0.9928), Decision tree with AUC (0.9837) and Naïve Bayes with AUC (0.8524). Based on Haralick texture features, the highest separation was obtained using SVM Gaussian and RBF with AUC (0.9999) followed SVM polynomial with AUC (0.9984), Naïve Bayes and Decision tree with AUC (0.9837)

Table 2: Detection performance with Hybrid GLCM + Haralick Featuers and employing robust machine learning techniques to distinguish NSCLC from SCLC

Methods	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Accuracy**	**FPR**	**AUC**
Naïve Bayes	0.9449	0.9326	0.9568	0.9146	0.9402	0.06742	0.9895
Decision Tree	0.9858	0.9775	0.9858	0.9775	0.9826	0.02247	0.9895
SVM Gaussian	1	0.9888	0.9929	1	0.9956	0.01124	1
SVM RBF	1	0.9831	0.9895	1	0.9935	0.01685	1
SVM Polynomial	0.9982	0.9972	0.9982	0.9972	0.9978	0.002809	0.9995

Table 2 presents the detection results with Hybrid (GLCM + Haralick) features and employing machine learning techniques. The highest detection performance was obtained utilizing SVM with polynomial kernel produced specificity (99.72%), sensitivity (99.82%), NPV (99.72%), PPV (99.82%), accuracy (99.78%), FPR (0.00280) and AUC (0.9995) followed by SVM Gaussian with sensitivity (100%), specificity (9888%), PPV (99.29%), NPV (100%), accuracy (99.56%), FPR (0.01124) and AUC (1.00). The SVM RBF yields accuracy (99.35%), AUC (1.00), decision tree yields accuracy (98.26%), AUC (0.9895) and Naïve Bayes provided accuracy (94.02%) and AUC (0.9895).

Table 3: Detection performance with Hybrid GLCM + Autoencoder featuers and employing robust machine learning techniques to distinguish NSCLC from SCLC

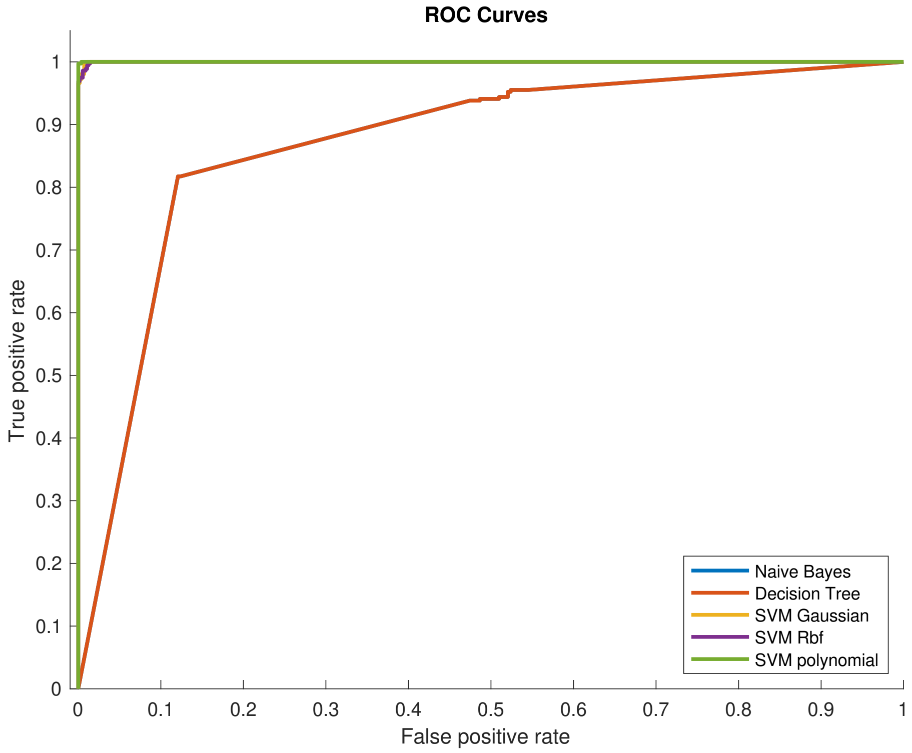
Methods	Sensitivity	Specificity	PPV	NPV	Accuracy	FPR	AUC
Naïve Bayes	0.9627	0.8736	0.9233	0.9367	0.9282	0.12640	0.9358
Decision Tree	0.9929	0.9803	0.9876	0.9887	0.9820	0.01966	0.9358
SVM Gaussian	1	1	1	1	1	0	1
SVM RBF	1	1	1	1	1	0	1
SVM Polynomial	1	0.9972	0.9982	1	0.9989	0.002809	0.9999

Table 3 depicts the detection performance of hybrid (GLCM + Autoencoder) features extracting methodology and employing the robust machine learning techniques. The highest detection performance was obtained by employing SVM Gaussian and RBF kernels yielded 100% all performance metrics followed by SVM polynomial kernel with sensitivity (100%), specificity (99.72%), PPV (99.82%), NPV (100%), accuracy (99.89%), FPR (0.00280) and AUC (0.9999). The Decision Tree and Naïve Bayes classifiers yield accuracy (98.80%) and AUC (0.9358).

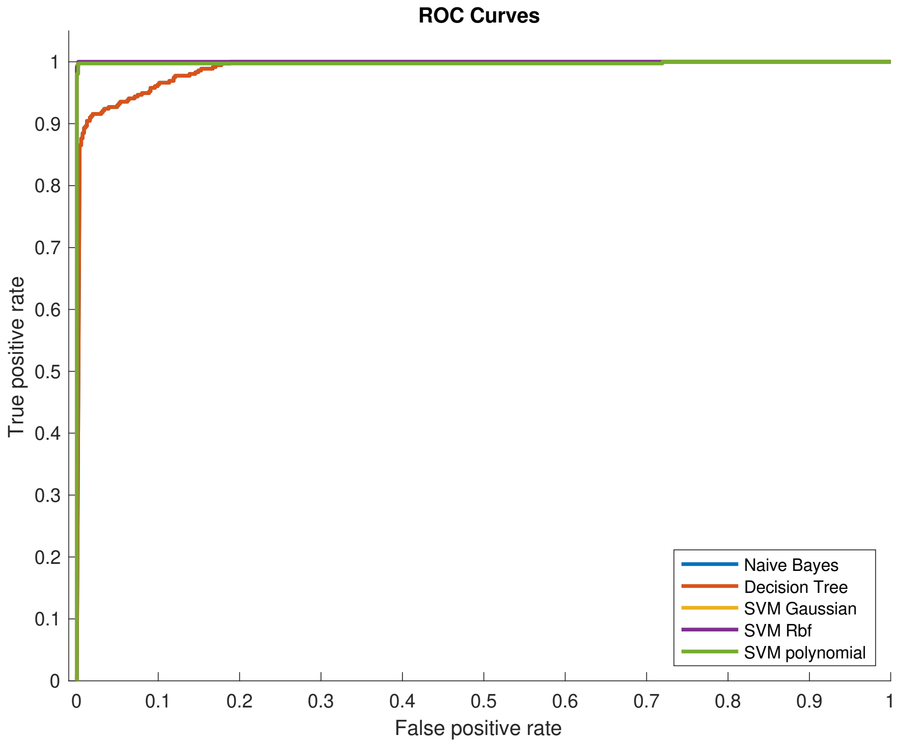
Table 4: Detection performance with Hybrid Haralick + Autoencoder featuers and employing robust machine learning techniques to distinguish NSCLC from SCLC

Methods	Sensitivity	Specificity	PPV	NPV	Accuracy	FPR	AUC
Naïve Bayes	0.9556	0.8567	0.9134	0.9242	0.9173	0.14330	0.9276
Decision Tree	0.9929	0.9888	0.9929	0.9888	0.9913	0.01124	0.9276
SVM Gaussian	1	1	1	1	1	0	1
SVM RBF	1	1	1	1	1	0	1
SVM Polynomial	1	0.9972	0.9982	1	0.9989	0.002809	1

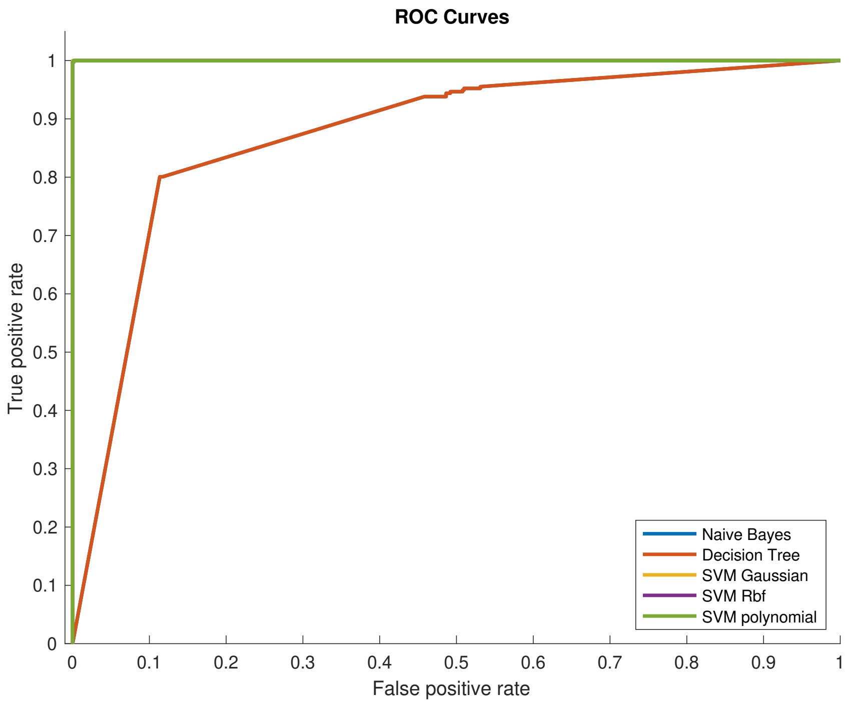
Table 4 depicts the detection performance of hybrid (Haralick + Autoencoder) features extracting methodology and employing the robust machine learning techniques. The highest detection performance was obtained using SVM Gaussian and RBF with 100% performance followed by SVM polynomial with sensitivity and NPV (100%), specificity (99.72%), PPV (99.82%) and AUC (1.00), Decision tree with accuracy (99.13%), AUC (0.9276) and Naïve Bayes yields accuracy (91.73%) and AUC (0.9276).


ROC Curves
True positive rate
False positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Naive Bayes
Decision Tree
SVM Gaussian
SVM Rbf
SVM polynomial

(a)


ROC Curves
True positive rate
False positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Naive Bayes
Decision Tree
SVM Gaussian
SVM Rbf
SVM polynomial

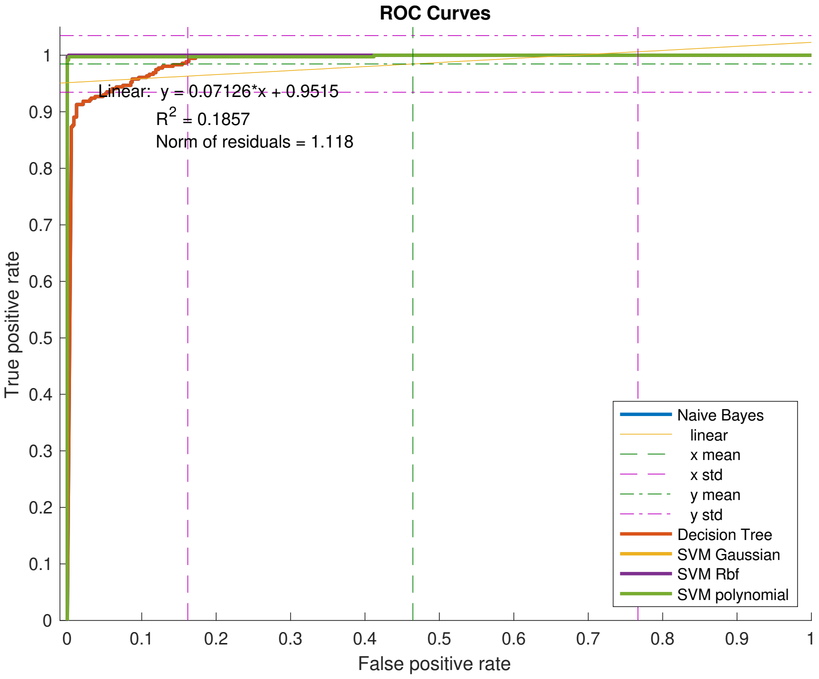
(b)



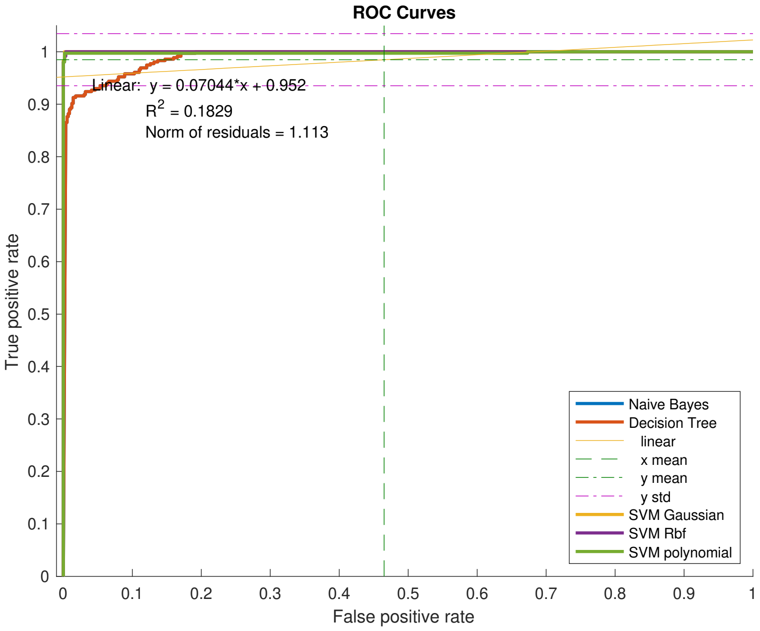
(c)

Fig. 3 Area under the receiver operating curve to distinguish NSCLC from SCLC by extracting hybrid feature a) Haralick + Autoencoder, b) Haralick + GLCM, c) Autoencoder + GLCM

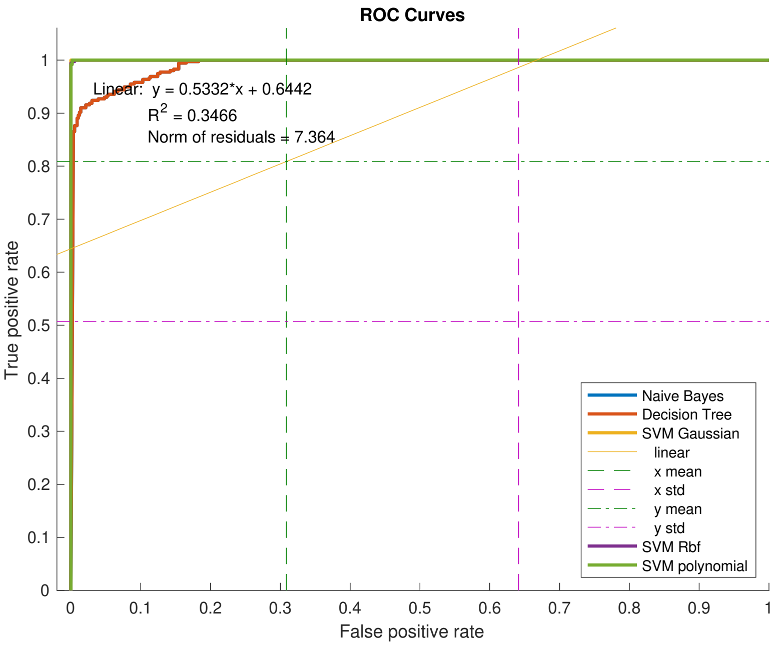
Figure 3 (a-c) reflects AUC-ROC to distinguish the NSCLC from SCLC subjects by extracting hybrid features and employing robust machine learning techniques. The highest Separation was obtained to distinguish NSCLC from SCLC by extracting hybrid features Haralick + autoencoder with AUC (1.00) using SVM Gaussian, RBF and polynomial followed by Naïve Bayes and Decision tree with AUC (0.9276) as reflected in Figure 4 (a). The highest separation to distinguish NSCLC from SCLC by extracting hybrid Haralick + GLCM features was obtained with AUC (1.00) using SVM Gaussian and RBF followed by SVM polynomial with AUC (0.9995), Naïve Bayes and Decision tree with AUC (0.9895) as reflected in Figure 4 (b). To distinguish the NSCLC from SCLC, the highest separation by extracting hybrid GLCM + Autoencoder features was obtained using SVM Gaussian and RBF with AUC (1.00) followed by SVM polynomial with AUC (0.9999), Naïve Bayes and Decision tree with AUC (0.9358) as depicted in Figure 4 (c). The Table 4 reflects the main findings and comparison of results with previous studies.


ROC Curves
Linear: y = 0.07126*x + 0.9515
R^2 = 0.1857
Norm of residuals = 1.118
True positive rate
False positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Naive Bayes
linear
x mean
x std
y mean
y std
Decision Tree
SVM Gaussian
SVM Rbf
SVM polynomial

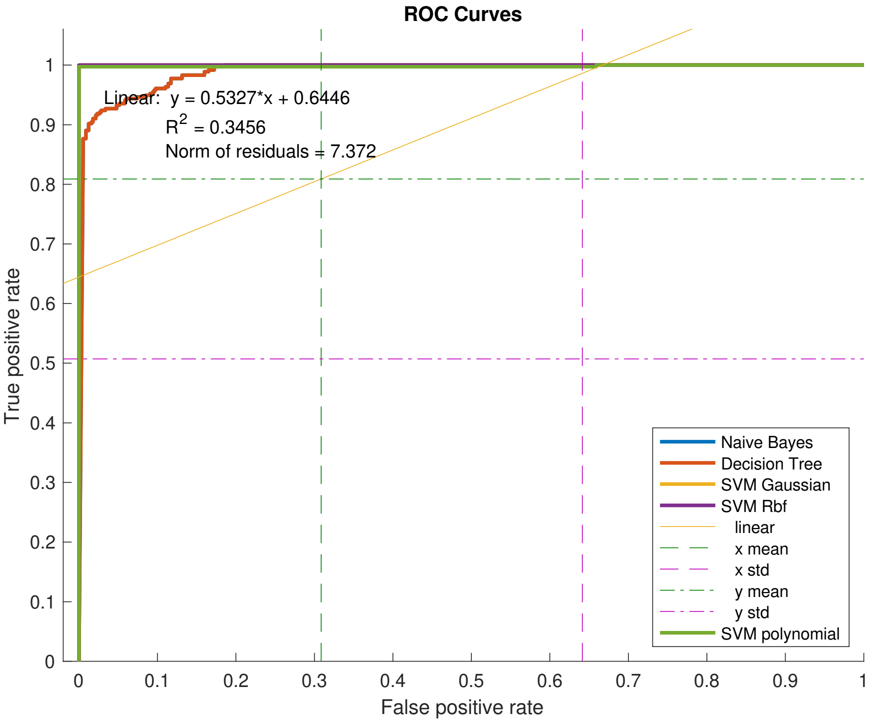
a)


ROC Curves
Linear: y = 0.07044*x + 0.952
R^2 = 0.1829
Norm of residuals = 1.113
True positive rate
False positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Naive Bayes
Decision Tree
linear
x mean
y mean
y std
SVM Gaussian
SVM Rbf
SVM polynomial

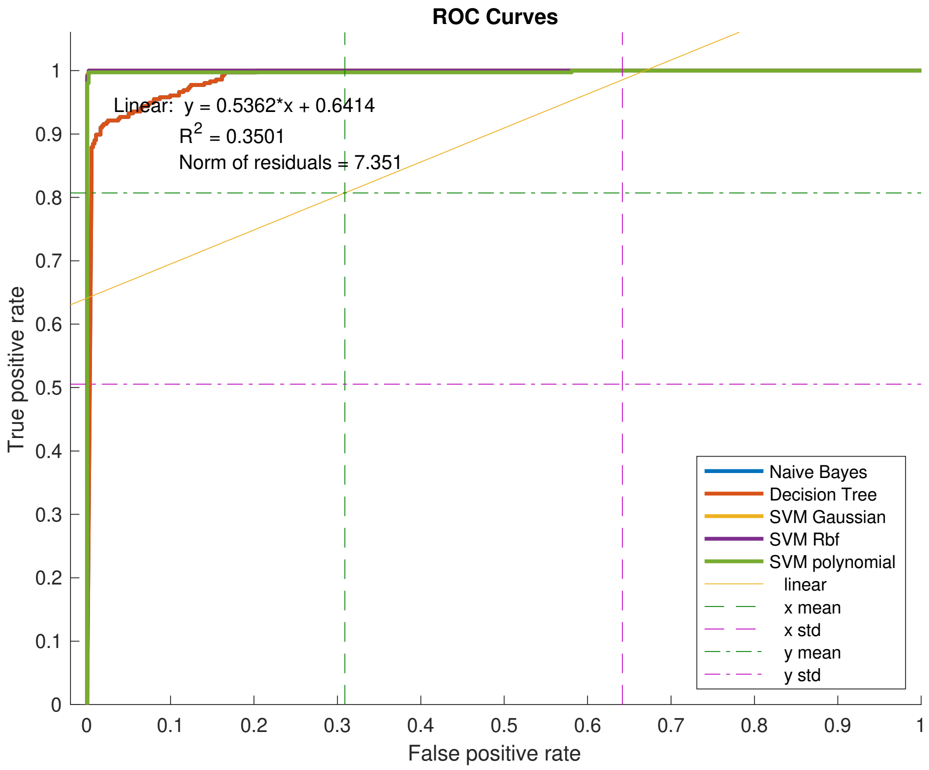
b)


ROC Curves
Linear: y = 0.5332*x + 0.6442
R^2 = 0.3466
Norm of residuals = 7.364
True positive rate
False positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Naive Bayes
Decision Tree
SVM Gaussian
linear
x mean
x std
y mean
y std
SVM Rbf
SVM polynomial

c)


ROC Curves
Linear: y = 0.5327*x + 0.6446
R^2 = 0.3456
Norm of residuals = 7.372
True positive rate
False positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Naive Bayes
Decision Tree
SVM Gaussian
SVM Rbf
linear
x mean
x std
y mean
y std
SVM polynomial

d)



e)

Fig. 4 Area under the receiver operating curve to distinguish NSCLC from SCLC by extracting hybrid Haralick + GLCM features by fitting linear curve on AUC and computing mean and standard deviation a) Naïve Bayes, b) Decision tree, c) SVM Gaussian, d) SVM RBF, e) SVM Polynomial

Figure 4 presents the AUC separation by extracting hybrid GLCM + Haralick features to distinguish the SCLC from NSCLC and computing the mean values and standard deviations of x-values (FPR) and y-values (TPR) and fitting linear curves. The Naïve Bayes and decision tree yielded the mean FPR (0.4652) and TPR (0.9848) and std FPR (0.3021) and TPR (0.049). The SVM Gaussian yielded mean FPR (0.3085) and TPR (0.8087) and std FPR (0.3329) and TPR (0.3015). The SVM RBF produced mean FPR (0.3085) and TPR (0.8089) and std FPR (0.3329) and TPR (0.3016). Similarly the SVM polynomial yielded mean FPR (0.3088) and TPR (0.8070) and std FPR (0.3328) and TPR (0.3016). The corresponding linear fits in Figure 4 (a-e) are reflected accordingly.

Table 5: Comparison of findings with previous studies

Author	Features Used	Performance
Guo et al. [83]	1. Texture 2. Shape	Sensitivity = 94%,

Sousa et al. [84]	1. Gradient
	2. Histogram
	3. Spatial
		Sensitivity = 84%,
		Specificity = 96%
		Accuracy=95%

Orozco et al. [85]	1. Texture	Sensitivity = 84%,
Messay et al. [86]	1. Gradient 2. Shape 3. Intensity	Sensitivity = 82%,

Retico et al. [87]	1. Morphology 2. Texture	Sensitivity = 72%,
Teramoto et al. [88]	1. Shape 2. Intensity	Sensitivity = 83%,
Hussain et al. [23]	Lung cancer detection based on Multimodal features such as texture, morphological, and EFDs Texture features using MFE with standard deviation, Morphological features using RCMFE with mean EFDs features using MFE	P-value (1.95E-50) P-value (3.01E-14) P-value (1.04E-13)
Hussain et al. [89]	RICA features and SVM	Accuracy = 99.77%

Dandil et al. [90]	1. GLCM
	2. Shape
	3. Statistical
	4. Energy
		Sensitivity = 97%,
		Specificity = 94%
		Accuracy =95%

This study	Single features
	
	Haralick using SVM polynomial
	GLCM using SVM polynomial
	SIFT using SVM Gaussian
	
	Hybrid features approach
	
	(GLCM + autoencoder, GLCM + Haralick, Haralick + Autoencoder) features
		Single Features
		Accuracy= 99.89%
		Accuracy= 98.69%
		Accuracy = 98.39%
		
		Hybrid features approach
		
		Sensitivity = 100%,
		Specificity = 100%
		AUC = 1.00
		Accuracy = 100%

Table 5 presents the findings of the current study and compared the results with previous studies. This study was trifold to improve the lung cancer detection performance i.e. i) improving the preprocessing steps, ii) improving the feature extracting strategy, iii) and optimizing the hyperparameters of machine learning algorithms. Image preprocessing is the process of preparing an image for further analysis and processing. It involves a series of steps that are applied to an image to enhance its quality, remove any noise or artifacts, and ensure that the image is in the proper format for further analysis. Image feature extraction is an important step in many machine learning applications that deal with image data. The goal of feature extraction is to extract meaningful and relevant information from an image and represent it in a compact and numerical form that can be used for further analysis and processing.: Images often have a large number of pixels, which results in a high-dimensional feature space. Feature extraction helps to reduce the dimensionality of the data, making it easier to process and analyze. By extracting relevant features, the machine learning model can focus on the most important information in the image and make predictions with higher accuracy. Moreover, by focusing on the most important features, the model is able to generalize better to new, unseen data. Grid search is a commonly used technique for optimizing the hyperparameters of a machine learning model. The idea behind grid search is to specify a range of possible values for each hyperparameter and exhaustively search through all possible combinations of these values to find the best set of hyperparameters for a given machine learning problem. The first step is to define the hyperparameters that need to be optimized. For example, the hyperparameters of a support vector machine (SVM) could include the regularization parameter and the kernel type. For each hyperparameter, a range of possible values is specified. This could be a discrete set of values or a continuous range. Extracting the most relevant feature is a tedious task on which the machine learning algorithms are to be trained. Previously, researchers computed few traditional features extracting methods which are not much helpful in extracting the most relevant information. Guo et al. [83] extracted texture and shape based features and obtained a sensitivity of 94.0%. Sousa et al. [84] computed different features such as gradient, histogram and spatial and obtained an accuracy of 95.0%. Messay et al. [86] extracted gradient, shape and intensity features and yielded a sensitivity of 82.0%. Moreover, Dandil et al. [90] extracted different features such as GLCM, shape, statistical and energy and obtained accuracy of 95.0%, sensitivity of 97.0% and specificity of 94.0%. In this study, we first extracted different features such as GLCM, Haralick, autoencoder, and SIFT features. The single feature extracting strategy using Haralick yielded highest accuracy of 99.89% with SVM polynomial, GLCM yielded the highest accuracy of 98.69% using SVM polynomial and SIFT features yielded highest accuracy of 98.31% using SVM Gaussian. With hybrid feature extracting approach (i.e. GLCM + Autoencoder, GLCM + Haralick, Haralick + Autoencoder) yielded the accuracy of 100%, AUC of 1.00. The current approach improved the lung cancer detection which can be utilized as a better tool for improving healthcare systems.

4. **Conclusion and Future Directions**

In the recent study, we improved the lung cancer detection by applying and optimizing the pre-processing steps and optimizing the feature extraction strategies along with hyper-parameters optimization of machine learning algorithms. Based on signel features extraction approach, the a highest accuracy of 99.89% was obtained with Haralick features using SVM polynomial, an accuracy of 98.89% with GLCM features using SVM polynomial. The SVM RBF with hybrid features GLCM + Autoencoder and Haralick + Autoencoder yielded the highest detection performance out of all methods yielded a 100% sensitivity, specificity, PPV, NPV, accuracy and AUC. SVM polynomial and GLCM + Haralick using SVM Gaussian yielded the second highest detection accuracy of 99.56%. The hybrid features extraction methods. By utilizing the hybrid features can capture more comprehensive and diverse information about the patient, which may help the model to better distinguish between cancerous and non-cancerous cases. Hybrid feature extraction can also help to reduce noise in the data by filtering out irrelevant or redundant features. This can improve the performance of the machine learning model and make it more robust to variations in the input data. hybrid feature extraction is a promising approach for detecting lung cancer that offers several potential advantages over using either type of data alone. Based on these results, the proposed methodology can be very helpful in the early detection and treatment of lung cancer, with the potential to decrease mortality rate and increase survival rate. The dataset utilized in this study was small and unbalanced. We utilized k-fold cross validation and data augmentation techniques to avoid overfitting. Though many recent artificial intelligence-based methods were utilized. However, we further improved the prediction performance by improving the preprocessing and feature extraction methods. There is still a room for further methodological improvement utilizing hybrid deep learning methods and optimizing the parameters with different methods. We will also compute the performance with other metrics and visualization methods. We will also test these methodologies on larger datasets with more diverse lung cancer types. Moreover, clinical information was not available, we will improve the future prediction by incorporating the clinical information along with the imaging features for better diagnostic and improving the disease recurrence, survival and severity.

Declarations

Funding

NIL

Conflicts of interest/Competing interests

The authors declares that they do not have any conflict of interest

Data availability

Data is publicaly available and is used in studies

Code availability

Will be provided on request

Authors' contributions

J.Y. edited the manuscript and wrote the manuscript

P.L.Y. Edited, supervised

A.A.K. Edited, supervised

M.S.K. Edited

H.K. Edited

A.A. Edited

L.H. edited the manuscript and wrote the manuscript, results, implementation, graphs

A.O. Edited, supervised

References

[1] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer statistics, 2022, CA. Cancer J. Clin. 72 (2022) 7–33. https://doi.org/10.3322/caac.21708.

[2] D. Moldovanu, H.J. de Koning, C.M. van der Aalst, Lung cancer screening and smoking cessation efforts, Transl. Lung Cancer Res. 10 (2021) 1099–1109. https://doi.org/10.21037/tlcr-20-899.

[3] T. Funakoshi, I. Tachibana, Y. Hoshida, H. Kimura, Y. Takeda, T. Kijima, K. Nishino, H. Goto, T. Yoneda, T. Kumagai, T. Osaki, S. Hayashi, K. Aozasa, I. Kawase, Expression of tetraspanins in human lung cancer cells: frequent downregulation of CD9 and its contribution to cell motility in small cell lung cancer, Oncogene. 22 (2003) 674–687. https://doi.org/10.1038/sj.onc.1206106.

[4] J.E. Walter, M.A. Heuvelmans, P.A. de Jong, R. Vliegenthart, P.M.A. van Ooijen, R.B. Peters, K. ten Haaf, U. Yousaf-Khan, C.M. van der Aalst, G.H. de Bock, W. Mali, H.J.M. Groen, H.J. de Koning, M. Oudkerk, Occurrence and lung cancer probability of new solid nodules at incidence screening with low-dose CT: analysis of data from the randomised, controlled NELSON trial, Lancet Oncol. 17 (2016) 907–916. https://doi.org/10.1016/S1470-2045(16)30069-9.

[5] P. Correale, R. Giannicola, R.E. Saladino, V. Nardone, L. Pirtoli, P. Tassone, A. Luce, S. Cappabianca, M. Scrima, P. Tagliaferri, M. Caraglia, On the way of the new strategies aimed to improve the efficacy of PD-1/PD-L1 immune checkpoint blocking mAbs in small cell lung cancer, Transl. Lung Cancer Res. 9 (2020) 1712–1719. https://doi.org/10.21037/tlcr-20-536.

[6] K.E. Rosenzweig, S.E. Sim, B. Mychalczak, L.E. Braban, R. Schindelheim, S.A. Leibel, Elective nodal irradiation in the treatment of non–small-cell lung cancer with three-dimensional conformal radiation therapy, Int. J. Radiat. Oncol. 50 (2001) 681–685. https://doi.org/10.1016/S0360-3016(01)01482-1.

[7] J. Zang, H. Horinouchi, J. Hanaoka, K. Funai, N. Sakakura, H. Liao, The role of salvage surgery in the treatment of a

gefitinib-resistant non-small cell lung cancer patient: a case report, J. Thorac. Dis. 13 (2021) 4554–4559. https://doi.org/10.21037/jtd-21-171.

[8] P.G. Kemps, M. Bol, E.J.A. Steller, L.M.H. de Pont, C. Holterhues, L. van Gerven, W. Kolkman, Colon carcinoma presenting as ovarian metastasis, Radiol. Case Reports. 16 (2021) 2799–2803. https://doi.org/10.1016/j.radcr.2021.06.072.

[9] Z. Zhang, S. Zhao, K. Wang, M. Shang, Z. Chen, H. Yang, Y. Chen, B. Chen, Identification of biomarkers associated with cervical lymph node metastasis in papillary thyroid carcinoma: Evidence from an integrated bioinformatic analysis, Clin. Hemorheol. Microcirc. 78 (2021) 117–126. https://doi.org/10.3233/CH-201074.

[10] T. Hamdeni, F. Fnaiech, S. Gasmi, J.M. Ginoux, R. Naeck, M. Bouchouicha, A. Ben Khedher Zidi, F. Tshibasu, Overview and definitions on lung cancer diagnosis, Middle East Conf. Biomed. Eng. MECBME. 2018-March (2018) 165–170. https://doi.org/10.1109/MECBME.2018.8402427.

[11] Q. Pei, Y. Luo, Y. Chen, J. Li, D. Xie, T. Ye, Artificial intelligence in clinical applications for lung cancer: diagnosis, treatment and prognosis, Clin. Chem. Lab. Med. 60 (2022) 1974–1983. https://doi.org/10.1515/cclm-2022-0291.

[12] Q. Ni, Z.Y. Sun, L. Qi, W. Chen, Y. Yang, L. Wang, X. Zhang, L. Yang, Y. Fang, Z. Xing, Z. Zhou, Y. Yu, G.M. Lu, L.J. Zhang, A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images, Eur. Radiol. 30 (2020) 6517–6527. https://doi.org/10.1007/s00330-020-07044-9.

[13] Z. Raizah, U.K. Kodipalya Nanjappa, H.U. Ajjipura Shankar, U. Khan, S.M. Eldin, R. Kumar, A.M. Galal, Windmill Global Sourcing in an Initiative Using a Spherical Fuzzy Multiple-Criteria Decision Prototype, Energies. 15 (2022) 8000. https://doi.org/10.3390/en15218000.

[14] A.A. Mir, L. Hussain, M.H. Waseem, A. Aldweesh, S. Rasheed, E.S. Yousef, M.S.A. Nadeem, E.T. Eldin, Analysis of Proposed and Traditional Boosting Algorithm with Standalone Classification Methods for Classifying Gene Expresssion Microarray Data Using a Reject Option, Appl. Artif. Intell. 36 (2022). https://doi.org/10.1080/08839514.2022.2151171.

[15] L. Hussain, S.A. Qureshi, A. Aldweesh, J. ur R. Pirzada, F.M. Butt, E.T. Eldin, M. Ali, A. Algarni, M.A. Nadim, Automated breast cancer detection by reconstruction independent component analysis (RICA) based hybrid features using machine learning paradigms, Conn. Sci. 34 (2022) 2784–2806. https://doi.org/10.1080/09540091.2022.2151566.

[16] F. Althoey, M.N. Akhter, Z.S. Nagra, H.H. Awan, F. Alanazi, M.A. Khan, M.F. Javed, S.M. Eldin, Y.O. Özkılıç, Prediction models for marshall mix parameters using bio-inspired genetic programming and deep machine learning approaches: A comparative study, Case Stud. Constr. Mater. 18 (2023) e01774. https://doi.org/10.1016/j.cscm.2022.e01774.

[17] M.M.A. Lashin, M.I. Khan, N. Ben Khedher, S.M. Eldin, Optimization of Display Window Design for Females' Clothes for Fashion Stores through Artificial Intelligence and Fuzzy System, Appl. Sci. 12 (2022) 11594. https://doi.org/10.3390/app122211594.

[18] C.I. Henschke, D.I. McCauley, D.F. Yankelevitz, D.P. Naidich, G. McGuinness, O.S. Miettinen, D.M. Libby, M.W. Pasmantier, J. Koizumi, N.K. Altorki, J.P. Smith, Early Lung Cancer Action Project: overall design and findings from baseline screening, Lancet. 354 (1999) 99–105. https://doi.org/10.1016/S0140-6736(99)06093-6.

[19] T. Sun, J. Wang, X. Li, P. Lv, F. Liu, Y. Luo, Q. Gao, H. Zhu, X. Guo, Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set, Comput. Methods Programs Biomed. 111 (2013) 519–524. https://doi.org/10.1016/j.cmpb.2013.04.016.

[20] W. de Wever, J. Coolen, J.A. Verschakelen, Imaging techniques in lung cancer, Breathe. 7 (2011) 338–346. https://doi.org/10.1183/20734735.022110.

[21] Y.-J. Yu-Jen Chen, K.-L. Hua, C.-H. Hsu, W.-H. Cheng, S.C. Hidayati, Computer-aided classification of lung nodules on computed tomography images via deep learning technique, Onco. Targets. Ther. (2015) 2015. https://doi.org/10.2147/OTT.S80733.

[22] S.G. Armato, A.S. Roy, H. MacMahon, F. Li, K. Doi, S. Sone, M.B. Altman, Evaluation of automated lung nodule

detection on low-dose computed tomography scans from a lung cancer screening program1, Acad. Radiol. 12 (2005) 337–346. https://doi.org/10.1016/j.acra.2004.10.061.

[23] L. Hussain, W. Aziz, A.A.A. Alshdadi, M.S. Ahmed Nadeem, I.R. Khan, Q.-U.-A. Chaudhry, Analyzing the Dynamics of Lung Cancer Imaging Data Using Refined Fuzzy Entropy Methods by Extracting Different Features, IEEE Access. 7 (2019) 64704–64721. https://doi.org/10.1109/ACCESS.2019.2917303.

[24] R. Ramani, N.S. Vanitha, S. Valarmathy, The Pre-Processing Techniques for Breast Cancer Detection in Mammography Images, Int. J. Image, Graph. Signal Process. 5 (2013) 47–54. https://doi.org/10.5815/ijigsp.2013.05.06.

[25] H. Golnabi, A. Asadpour, Design and application of industrial machine vision systems, Robot. Comput. Integr. Manuf. 23 (2007) 630–637. https://doi.org/10.1016/j.rcim.2007.02.005.

[26] T. Fu, K. Zhang, L. Zhang, S. Wang, S. Ma, An Efficient Framework of Reference Picture Resampling (RPR) for Video Coding, IEEE Trans. Circuits Syst. Video Technol. 32 (2022) 7107–7119. https://doi.org/10.1109/TCSVT.2022.3176934.

[27] Z. Tang, J. Yao, Q. Zhang, Multi-operator image retargeting in compressed domain by preserving aspect ratio of important contents, Multimed. Tools Appl. 81 (2022) 1501–1522. https://doi.org/10.1007/s11042-021-11376-z.

[28] A. Mikolajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 Int. Interdiscip. PhD Work., IEEE, 2018: pp. 117–122. https://doi.org/10.1109/IIPHDW.2018.8388338.

[29] R. Takahashi, T. Matsubara, K. Uehara, Data Augmentation Using Random Image Cropping and Patching for Deep CNNs, IEEE Trans. Circuits Syst. Video Technol. 30 (2020) 2917–2931. https://doi.org/10.1109/TCSVT.2019.2935128.

[30] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random Erasing Data Augmentation, Proc. AAAI Conf. Artif. Intell. 34 (2020) 13001–13008. https://doi.org/10.1609/aaai.v34i07.7000.

[31] E. Okafor, L. Schomaker, M.A. Wiering, An analysis of rotation matrix and colour constancy data augmentation in classifying images of animals, J. Inf. Telecommun. 2 (2018) 465–491. https://doi.org/10.1080/24751839.2018.1479932.

[32] E. Salvador, A. Cavallaro, T. Ebrahimi, Cast shadow segmentation using invariant color features, Comput. Vis. Image Underst. 95 (2004) 238–259. https://doi.org/10.1016/j.cviu.2004.03.008.

[33] D.L. Ruderman, W. Bialek, Statistics of natural images: Scaling in the woods, Phys. Rev. Lett. 73 (1994) 814–817. https://doi.org/10.1103/PhysRevLett.73.814.

[34] X. Dong, J. Shen, W. Wang, L. Shao, H. Ling, F. Porikli, Dynamical Hyperparameter Optimization via Deep Reinforcement Learning in Tracking, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 1515–1529. https://doi.org/10.1109/TPAMI.2019.2956703.

[35] S.N. Wood, N. Pya, B. Säfken, Smoothing Parameter and Model Selection for General Smooth Models, J. Am. Stat. Assoc. 111 (2016) 1548–1563. https://doi.org/10.1080/01621459.2016.1180986.

[36] Zhenmei Gu, N. Cercone, Naive Bayes Modeling with Proper Smoothing for Information Extraction, in: 2006 IEEE Int. Conf. Fuzzy Syst., IEEE, 2006: pp. 393–400. https://doi.org/10.1109/FUZZY.2006.1681742.

[37] A.Y. Liu, C.E. Martin, Smoothing Multinomial Naïve Bayes in the Presence of Imbalance, in: 2011: pp. 46–59. https://doi.org/10.1007/978-3-642-23199-5_4.

[38] V. Nourani, Z. Razzaghzadeh, A.H. Baghanam, A. Molajou, ANN-based statistical downscaling of climatic parameters using decision tree predictor screening method, Theor. Appl. Climatol. 137 (2019) 1729–1746. https://doi.org/10.1007/s00704-018-2686-z.

[39] M. Tayefi, H. Esmaeili, M. Saberi Karimian, A. Amirabadi Zadeh, M. Ebrahimi, M. Safarian, M. Nematy, S.M.R. Parizadeh, G.A. Ferns, M. Ghayour-Mobarhan, The application of a decision tree to establish the parameters associated with hypertension, Comput. Methods Programs Biomed. 139 (2017) 83–91. https://doi.org/10.1016/j.cmpb.2016.10.020.

[40] R.G. Mantovani, T. Horvath, R. Cerri, J. Vanschoren, A.C.P.L.F. de Carvalho, Hyper-Parameter Tuning of a Decision Tree Induction Algorithm, in: 2016 5th Brazilian Conf. Intell. Syst., IEEE, 2016: pp. 37–42. https://doi.org/10.1109/BRACIS.2016.018.

[41] S. Hussain, Relationships Among Various Parameters for Decision Tree Optimization, in: 2014: pp. 393–410. https://doi.org/10.1007/978-3-319-01866-9_13.

[42] F.J. Pontes, G.F. Amorim, P.P. Balestrassi, A.P. Paiva, J.R. Ferreira, Design of experiments and focused grid search for neural network parameter optimization, Neurocomputing. 186 (2016) 22–34. https://doi.org/10.1016/j.neucom.2015.12.061.

[43] Y. Sun, S. Ding, Z. Zhang, W. Jia, An improved grid search algorithm to optimize SVR for prediction, Soft Comput. 25 (2021) 5633–5644. https://doi.org/10.1007/s00500-020-05560-w.

[44] I. Syarif, A. Prugel-Bennett, G. Wills, SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance, TELKOMNIKA (Telecommunication Comput. Electron. Control. 14 (2016) 1502. https://doi.org/10.12928/telkomnika.v14i4.3956.

[45] Qiujun Huang, Jingli Mao, Yong Liu, An improved grid search algorithm of SVR parameters optimization, in: 2012 IEEE 14th Int. Conf. Commun. Technol., IEEE, 2012: pp. 1022–1026. https://doi.org/10.1109/ICCT.2012.6511415.

[46] Z. Nematzadeh, R. Ibrahim, A. Selamat, Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques, in: 2015 10th Asian Control Conf., IEEE, 2015: pp. 1–6. https://doi.org/10.1109/ASCC.2015.7244654.

[47] I. Tsamardinos, E. Greasidou, G. Borboudakis, Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation, Mach. Learn. 107 (2018) 1895–1922. https://doi.org/10.1007/s10994-018-5714-4.

[48] S. Rathore, M. Hussain, M. Aksam Iftikhar, A. Jalil, Ensemble classification of colon biopsy images based on information rich hybrid features, Comput. Biol. Med. 47 (2014) 76–92. https://doi.org/10.1016/j.compbiomed.2013.12.010.

[49] S. Rathore, A. Iftikhar, A. Ali, M. Hussain, A. Jalil, Capture largest included circles: An approach for counting red blood cells, Commun. Comput. Inf. Sci. 281 CCIS (2012) 373–384. https://doi.org/10.1007/978-3-642-28962-0_36.

[50] Automated colon cancer detection using hybrid of novel geometric features and some traditional features, (2016). https://doi.org/10.1016/j.compbiomed.2015.03.004.

[51] L. Hussain, A. Ahmed, S. Saeed, S. Rathore, I.A. Awan, S.A. Shah, A. Majid, A. Idris, A.A. Awan, Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies, Cancer Biomarkers. 21 (2018) 393–413. https://doi.org/10.3233/CBM-170643.

[52] L. Hussain, W. Aziz, S. Saeed, S. Rathore, M. Rafique, Automated Breast Cancer Detection Using Machine Learning Techniques by Extracting Different Feature Extracting Strategies, in: 2018 17th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. 12th IEEE Int. Conf. Big Data Sci. Eng., IEEE, 2018: pp. 327–331. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00057.

[53] L. Hussain, S. Saeed, I.A. Awan, A. Idris, M.S.A.A.A.A.A. Nadeem, Q.-A. Chaudhry, Q.-A. Chaudhary, Q.-A. Chaudhry, Q.-A. Chaudhary, Detecting Brain Tumor Using Machine Learning Techniques Based on Different Features Extracting Strategies, Curr. Med. Imaging Former. Curr. Med. Imaging Rev. 14 (2019) 595–606. https://doi.org/10.2174/1573405614666180718123533.

[54] L. Hussain, S. Rathore, A.A. Abbasi, S. Saeed, Automated lung cancer detection based on multimodal features extracting strategy using machine learning techniques, in: H. Bosmans, G.-H. Chen, T. Gilat Schmidt (Eds.), Med. Imaging 2019 Phys. Med. Imaging, SPIE, 2019: p. 134. https://doi.org/10.1117/12.2512059.

[55] D.S. Guru, Y.H. Sharath, S. Manjunath, Texture Features and KNN in Classification of Flower Images, Int. J. Comput. Appl. (2010) 21–29.

[56] S.G. Mougiakakou, I. Valavanis, K.S. Nikita, a. Nikita, D. Kelekis, Characterization of CT liver lesions based on texture features and a multiple neural network classification scheme, Proc. 25th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (2003) 1287–1290. https://doi.org/10.1109/IEMBS.2003.1279504.

[57] M.E. Mavroforakis, H. V. Georgiou, D. Cavouras, N. Dimitropoulos, S. Theodoridis, Mammographic mass classification

using textural features and descriptive diagnostic data, Int. Conf. Digit. Signal Process. DSP. 1 (2002) 461–464. https://doi.org/10.1109/ICDSP.2002.1027918.

[58] A.N. Esgiar, R.N. Naguib, B.S. Sharif, M.K. Bennett, A. Murray, Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa., IEEE Trans. Inf. Technol. Biomed. 2 (1998) 197–203. https://doi.org/10.1109/4233.735785.

[59] A.N. Esgiar, R.N.G. Naguib, B.S. Sharif, M.K. Bennett, A. Murray, Fractal analysis in the detection of colonic cancer images., IEEE Trans. Inf. Technol. Biomed. 6 (2002) 54–8. https://doi.org/10.1109/4233.992163.

[60] P. Brynolfsson, D. Nilsson, T. Torheim, T. Asklund, C.T. Karlsson, J. Trygg, T. Nyholm, A. Garpebring, Haralick texture features from apparent diffusion coefficient (ADC) MRI images depend on imaging and pre-processing parameters, Sci. Rep. 7 (2017). https://doi.org/10.1038/s41598-017-04151-4.

[61] A. Ali, S. Qadri, W.K. Mashwani, S. Brahim Belhaouari, S. Naeem, S. Rafique, F. Jamal, C. Chesneau, S. Anam, Machine learning approach for the classification of corn seed using hybrid features, Int. J. Food Prop. 23 (2020) 1110–1124. https://doi.org/10.1080/10942912.2020.1778724.

[62] Shih-Fu Chang, R. Manmatha, Tat-Seng Chua, Combining Text and Audio-Visual Features in Video Indexing, in: Proceedings. (ICASSP '05). IEEE Int. Conf. Acoust. Speech, Signal Process. 2005., IEEE, n.d.: pp. 1005–1008. https://doi.org/10.1109/ICASSP.2005.1416476.

[63] D. Hazarika, S. Gorantla, S. Poria, R. Zimmermann, Self-Attentive Feature-Level Fusion for Multimodal Emotion Detection, in: 2018 IEEE Conf. Multimed. Inf. Process. Retr., IEEE, 2018: pp. 196–201. https://doi.org/10.1109/MIPR.2018.00043.

[64] A. Razdan, M. Bae, A hybrid approach to feature segmentation of triangle meshes, Comput. Des. 35 (2003) 783–789. https://doi.org/10.1016/S0010-4485(02)00101-X.

[65] S. Piramu Kailasam, M. Mohamed Sathik, A novel hybrid feature extraction model for classification on pulmonary nodules, Asian Pacific J. Cancer Prev. 20 (2019) 457–468. https://doi.org/10.31557/APJCP.2019.20.2.457.

[66] M. Madhubala, M. Seetha, Hybrid Feature Extraction and Selection Using Bayesian Classifier, Natl. Conf. Adv. Era Multi Discip. Syst. AEMDS,2013, Technol. Educ. Res. Integr. Institutions, Kurukshetra, Haryana, India. (2013) 449–453.

[67] B. Sanae, A.K. Mounir, F. Youssef, A hybrid feature extraction scheme based on DWT and uniform LBP for digital mammograms classification, Int. Rev. Comput. Softw. 10 (2015) 102–110. https://doi.org/10.15866/irecos.v10i1.5052.

[68] Y. Eroğlu, M. Yildirim, A. Çinar, Convolutional Neural Networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mRMR, Comput. Biol. Med. 133 (2021) 104407. https://doi.org/10.1016/j.compbiomed.2021.104407.

[69] S. Rathore, M. Hussain, A. Khan, Automated colon cancer detection using hybrid of novel geometric features and some traditional features, Comput. Biol. Med. 65 (2015) 279–296. https://doi.org/10.1016/j.compbiomed.2015.03.004.

[70] M. Amrane, S. Oukid, I. Gagaoua, T. Ensari, Breast cancer classification using machine learning, in: 2018 Electr. Electron. Comput. Sci. Biomed. Eng. Meet., IEEE, 2018: pp. 1–4. https://doi.org/10.1109/EBBT.2018.8391453.

[71] L. Hussain, W. Aziz, S.A. Nadeem, A.Q. Abbasi, Classification of Normal and Pathological Heart Signal Variability Using Machine Learning Techniques Classification of Normal and Pathological Heart Signal Variability Using Machine Learning Techniques, (2015).

[72] L. Hussain, W. Aziz, A.S. Khan, A.Q. Abbasi, S.Z. Hassan, Classification of Electroencephlography (EEG) Alcoholic and Control Subjects using Machine Learning Ensemble Methods, J. Multidiscip. Eng. Sci. Technol. 2 (2015) 126–131.

[73] L. Hussain, W. Aziz, S.A. Nadeem, A.Q. Abbasi, Classification of Normal and Pathological Heart Signal Variability Using Machine Learning Techniques Classification of Normal and Pathological Heart Signal Variability Using Machine Learning Techniques, Int. J. Darshan Inst. Eng. Res. Emerg. Technol. 3 (2015) 13–19.

[74] V.A. Memos, G. Minopoulos, K.D. Stergiou, K.E. Psannis, Internet-of-Things-Enabled Infrastructure Against Infectious

Diseases, IEEE Internet Things Mag. 4 (2021) 20–25. https://doi.org/10.1109/IOTM.0001.2100023.

[75] G.M. Minopoulos, V.A. Memos, C.L. Stergiou, K.D. Stergiou, A.P. Plageras, M.P. Koidou, K.E. Psannis, Exploitation of Emerging Technologies and Advanced Networks for a Smart Healthcare System, Appl. Sci. 12 (2022) 5859. https://doi.org/10.3390/app12125859.

[76] K.D. Stergiou, G.M. Minopoulos, V.A. Memos, C.L. Stergiou, M.P. Koidou, K.E. Psannis, A Machine Learning-Based Model for Epidemic Forecasting and Faster Drug Discovery, Appl. Sci. 12 (2022) 10766. https://doi.org/10.3390/app122110766.

[77] H. Alabduljabbar, M.N. Amin, S.M. Eldin, M.F. Javed, R. Alyousef, A.M. Mohamed, Forecasting compressive strength and electrical resistivity of graphite based nano-composites using novel artificial intelligence techniques, Case Stud. Constr. Mater. (2023) e01848. https://doi.org/10.1016/j.cscm.2023.e01848.

[78] Y. Zhou, Z. Ahmad, Z. Almaspoor, F. Khan, E. Tag-Eldin, Z. Iqbal, M. El-Morshedy, On the implementation of a new version of the Weibull distribution and machine learning approach to model the COVID-19 data, Math. Biosci. Eng. 20 (2022) 337–364. https://doi.org/10.3934/mbe.2023016.

[79] S. Ullah, S. Li, K. Khan, S. Khan, I. Khan, S.M. Eldin, An Investigation of Exhaust Gas Temperature of Aircraft Engine Using LSTM, IEEE Access. 11 (2023) 5168–5177. https://doi.org/10.1109/ACCESS.2023.3235619.

[80] H. Alabduljabbar, K. Khan, H.H. Awan, R. Alyousef, A.M. Mohamed, S.M. Eldin, Modeling the capacity of engineered cementitious composites for self-healing using AI-based ensemble techniques, Case Stud. Constr. Mater. (2022) e01805. https://doi.org/10.1016/j.cscm.2022.e01805.

[81] E. Seli, C. Bruce, L. Botros, M. Henson, P. Roos, K. Judge, T. Hardarson, A. Ahlström, P. Harrison, M. Henman, K. Go, N. Acevedo, J. Siques, M. Tucker, D. Sakkas, Receiver operating characteristic (ROC) analysis of day 5 morphology grading and metabolomic Viability Score on predicting implantation outcome, J. Assist. Reprod. Genet. 28 (2011) 137–144. https://doi.org/10.1007/s10815-010-9501-9.

[82] R.F. Fernandes, D. Scherrer, A. Guisan, Effects of simulated observation errors on the performance of species distribution models, Divers. Distrib. 25 (2019) 400–413. https://doi.org/10.1111/ddi.12868.

[83] Wei Guo, Ying Wei, Hanxun Zhou, DingYe Xue, An adaptive lung nodule detection algorithm, in: 2009 Chinese Control Decis. Conf., IEEE, 2009: pp. 2361–2365. https://doi.org/10.1109/CCDC.2009.5192686.

[84] J.R.F. da Silva Sousa, A.C. Silva, A.C. de Paiva, R.A. Nunes, Methodology for automatic detection of lung nodules in computerized tomography images, Comput. Methods Programs Biomed. 98 (2010) 1–14. https://doi.org/10.1016/j.cmpb.2009.07.006.

[85] H.M. Orozco, O.O.V. Villegas, H. de J.O. Dominguez, V.G.C. Sanchez, Lung Nodule Classification in CT Thorax Images Using Support Vector Machines, in: 2013 12th Mex. Int. Conf. Artif. Intell., IEEE, 2013: pp. 277–283. https://doi.org/10.1109/MICAI.2013.38.

[86] T. Messay, R.C. Hardie, S.K. Rogers, A new computationally efficient CAD system for pulmonary nodule detection in CT imagery, Med. Image Anal. 14 (2010) 390–406. https://doi.org/10.1016/j.media.2010.02.004.

[87] A. Retico, M.E. Fantacci, I. Gori, P. Kasae, B. Golosio, A. Piccioli, P. Cerello, G. De Nunzio, S. Tangaro, Pleural nodule identification in low-dose and thin-slice lung computed tomography, Comput. Biol. Med. 39 (2009) 1137–1144. https://doi.org/10.1016/j.compbiomed.2009.10.005.

[88] A. Teramoto, H. Fujita, K. Takahashi, O. Yamamuro, T. Tamaki, M. Nishio, T. Kobayashi, Hybrid method for the detection of pulmonary nodules using positron emission tomography/computed tomography: a preliminary study, Int. J. Comput. Assist. Radiol. Surg. 9 (2014) 59–69. https://doi.org/10.1007/s11548-013-0910-y.

[89] L. Hussain, M.S. Almaraashi, W. Aziz, N. Habib, S.-U.-R. Saif Abbasi, Machine learning-based lungs cancer detection using reconstruction independent component analysis and sparse filter features, Waves in Random and Complex Media. (2021) 1–26. https://doi.org/10.1080/17455030.2021.1905912.

[90] E. Dandıl, A Computer-Aided Pipeline for Automatic Lung Cancer Classification on Computed Tomography Scans, J. Healthc. Eng. 2018 (2018) 1–12. https://doi.org/10.1155/2018/9409267.