



Research article



Enhancing lung cancer detection through hybrid features and machine learning hyperparameters optimization techniques

Liangyu Li^{a,b}, Jing Yang^{c,***}, Lip Yee Por^c, Mohammad Shahbaz Khan^d, Rim Hamdaoui^{e,**}, Lal Hussain^{f,g}, Zahoor Iqbal^{h,*}, Ionela Magdalena Rotaruⁱ, Dan Dobrotă^j, Moutaz Aldrery^k, Abdulfattah Omar^l

^a Center for Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

^b Health Informatics Laboratory, Cancer Research Institute, Chifeng Cancer Hospital (Second Affiliated Hospital of Chifeng University), Medical Department, Chifeng University, Chifeng City, Inner Mongolia Autonomous Region, 024000, China

^c Department of Computer System and Technology, Faculty of Computer Science and Information Technology, Universiti Malaya, 50603, Kuala Lumpur, Malaysia

^d Children's National Hospital, 111 Michigan Ave NW, Washington, DC, 20010, United States

^e Department of Computer Science, College of Science and Human Studies Dawadmi, Shaqra University, Shaqra, Riyadh, Saudi Arabia

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

^g Department of Computer Science and Information Technology, Neelum Campus, University of Azad Jammu and Kashmir, Athmuqam, 13230, Azad Kashmir, Pakistan

^h School of Computer Science and Technology, Zhejiang Normal University, Jinhua, 321004, China

ⁱ Department of Industrial Engineering and Management, Lucian Blaga University of Sibiu, Bulevardul Victoriei 10, Sibiu, 550024, Romania

^j Faculty of Engineering, Lucian Blaga University of Sibiu, Bulevardul Victoriei 10, Sibiu, 550024, Romania

^k Department of Chemical Engineering, College of Engineering, King Khalid University, Abha, 61411, Saudi Arabia

^l Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University, Saudi Arabia

ARTICLE INFO

ABSTRACT

Keywords:

Classification

Haralick texture features

Lung cancer types

Autoencoder and gray-level co-occurrence (GLCM)

Machine learning offers significant potential for lung cancer detection, enabling early diagnosis and potentially improving patient outcomes. Feature extraction remains a crucial challenge in this domain. Combining the most relevant features can further enhance detection accuracy. This study employed a hybrid feature extraction approach, which integrates both Gray-level co-occurrence matrix (GLCM) with Haralick and autoencoder features with an autoencoder. These features were subsequently fed into supervised machine learning methods. Support Vector Machine (SVM) Radial Base Function (RBF) and SVM Gaussian achieved perfect performance measures, while SVM polynomial produced an accuracy of 99.89% when utilizing GLCM with an autoencoder, Haralick, and autoencoder features. SVM Gaussian achieved an accuracy of 99.56%, while SVM RBF achieved an accuracy of 99.35% when utilizing GLCM with Haralick features. These results demonstrate the potential of the proposed approach for developing improved diagnostic and prognostic lung cancer treatment planning and decision-making systems.

* Corresponding author.

** Corresponding author.

*** Corresponding author.

E-mail addresses: s2147529@siswa.um.edu.my (J. Yang), r.hamdaoui@su.edu.sa (R. Hamdaoui), izahoor@zjnu.edu.cn (Z. Iqbal).

<https://doi.org/10.1016/j.heliyon.2024.e26192>

Received 31 July 2023; Received in revised form 30 January 2024; Accepted 8 February 2024

Available online 14 February 2024

2405-8440/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Contemporary statistics on lung cancer from 2022 [1] indicate that 2.36 million new cases are expected to be diagnosed, with 85% classified as non-small cell lung cancer (NSCLC). Small cell lung cancer (SCLC) is another type of lung cancer, and both exhibit distinct treatment plans and dissemination patterns. NSCLC, characterized by slow growth, differs from SCLC, which grows rapidly and forms tumors. Lung cancer deaths are predominantly linked to cigarette smoking [2]. The term "non-small cell lung cancer" arises from the microscopic appearance of NSCLC cells, contrasting with the small, uniform cells of SCLC [3]. While smoking, radon exposure, and air pollution are common causes, NSCLC can also develop in individuals who have never smoked. Individuals experiencing any of these symptoms, particularly if persistent or worsening, should seek medical attention promptly for evaluation for NSCLC.

Small cell lung cancer (SCLC) – a rapidly growing and aggressive form – accounts for approximately 10–15% of all lung cancer diagnoses [4,5]. This distinct type, named for the microscopic appearance of its uniform, diminutive cells, primarily affects smokers but can also occur in never-smokers. SCLC's tendency to spread early to other organs often leads to advanced-stage diagnoses. Imaging tests like CT scans or chest X-rays assist in initial detection, but confirmation usually requires a biopsy. Treatment for SCLC typically combines chemotherapy and radiation therapy [6]. Surgery might be an option for isolated cases confined to specific chest areas.

Due to its aggressiveness, early diagnosis and intervention are crucial in improving SCLC outcomes. The median survival time after diagnosis, unfortunately, stands at around one year, highlighting the urgency of swift action.

Artificial intelligence (AI) is poised to dramatically reshape lung cancer detection and diagnosis [6]. One powerful approach uses AI algorithms to scan medical images like CT scans for lung cancer signatures, aiding radiologists in faster and more accurate diagnoses [7]. Beyond detection, AI can also predict disease progression and treatment response through sophisticated feature extraction techniques [8–12]. This nascent technology holds immense promise for significantly improving lung cancer care for patients. However, further research is crucial to fully unlock AI's potential and manage its limitations in this critical field.

Researchers have employed a variety of machine learning and deep learning techniques for classification and prediction tasks across diverse imaging and signal modalities. Machine learning algorithms are increasingly prominent in medical diagnostic systems, particularly for lung disease prediction. Researchers have developed standardized toolkits for feature importance calculation [13], as well as machine learning algorithms for prioritizing CT exams in emergency departments [14], and electronic noses capable of differentiating COPD patients from healthy individuals [15,16]. Liquid biopsy techniques have demonstrated promising results in lung cancer diagnosis and prognosis [17]. Additionally, blood cadmium levels have been investigated as a potential lung cancer biomarker, particularly in former smokers [18].

These advancements hold great potential for improving lung disease diagnosis and treatment. Abbasi, S. F et al. [19] used machine learning ensemble methods to accurately predict the EEG-Based Neonatal Sleep Stage. Abbasi, S. F. et al. [20] also utilized deep learning for real-time neonatal sleep stage classification. Aamir, S. et al. [21] utilized machine learning techniques for breast cancer prediction.

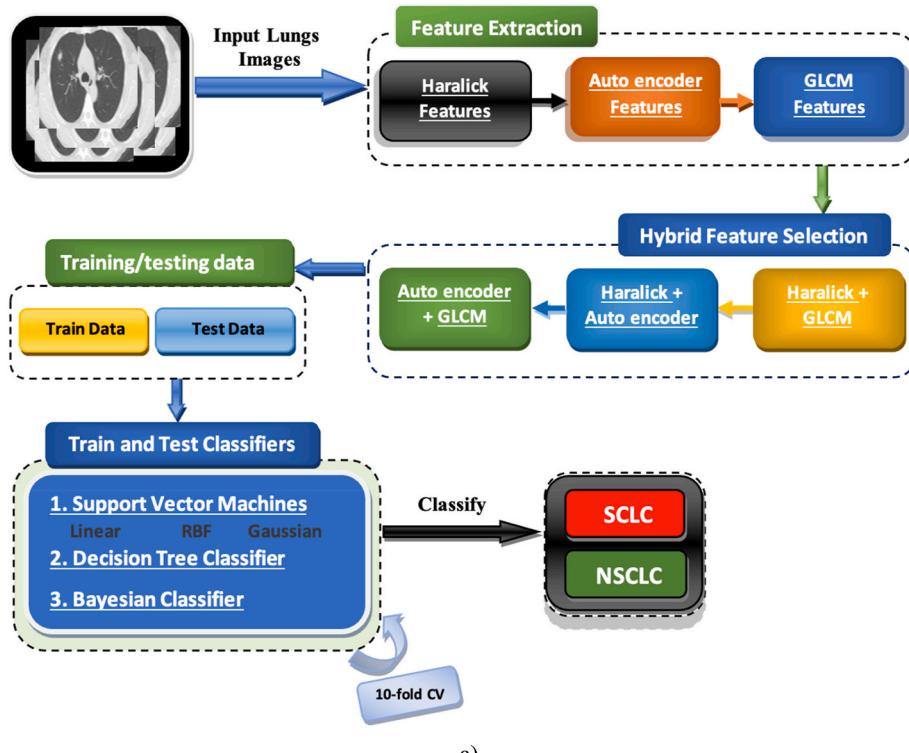
In machine learning, the most important and tedious task is to extract the most relevant features. In the past, researchers utilized various feature extraction approaches. Almutairi, S. et al. [22] used the gorilla troops optimization (GTO) algorithm for feature selection to detect breast cancer from a deep learning feature space. Park, H. J. et al. [23] computed the hybrid feature approach for object detection. Leng, L. et al. [24] used low correlation features for palmprint coding and verification. Leng, L. et al. [25] also utilized high discrimination-based features to detect the palmprint. Recent research [26–34] delves into a diverse toolbox of feature extraction methods, from single approaches to hybrid combinations, to analyze key cancer indicators. These tools, including texture analysis, GLCM (Gray Level Co-occurrence Matrix), and Haralick textures, along with morphological features, are then paired with powerful supervised machine learning algorithms like support vector machines, decision trees, and Bayesian classifiers. This potent combination has yielded promise.

A crucial study [35] demonstrated that low-dose spiral CT scans significantly outperform conventional chest X-rays in detecting early-stage lung cancer. While lung cancer often manifests as solitary pulmonary nodules (SPNs) on CT scans, differentiating them from benign lesions like tuberculosis or fungal infections can be challenging due to their similar appearance [36]. Enter helical CT, a game-changer introduced in 1991 that revolutionized chest imaging quality [37]. By employing multiple rows of detectors (ranging from 4 to a staggering 128), these cutting-edge scanners deliver rapid, high-resolution images of the thorax. This translates to a richer volume of information for clinicians, enabling more accurate detection and diagnosis of lung abnormalities.

Previous studies in this area often relied on basic data preparation, a single method for extracting features from images, and standard settings for their machine learning algorithms. These limitations can affect the accuracy of the analysis. While [38], explored this dataset using entropy-based complexity techniques on CT scans, their approach involved minimal preprocessing and only one feature extraction strategy. To address the dataset's imbalanced nature and small size, we implemented 10-fold cross-validation along with augmentation techniques like random cropping, flipping, and noise injection. This helps prevent overfitting, a common issue with limited data. Feature extraction plays a vital role in prediction accuracy, and researchers continuously develop new tools to improve this process. Choosing the right mix of features, single or hybrid, can significantly impact diagnostic outcomes. Think of complex diagnoses as locked doors. Different features are like keys, and hybrid approaches are like master keys that open the door more reliably than any single key alone.

The main contributions of this study can be summarized as follows:

Combining Feature Power: Instead of relying on a single feature type, we embraced a hybrid approach, extracting diverse features like GLCM, Haralick textures, and even those generated by an autoencoder neural network. We then "mixed" these features through concatenation, ensuring each type contributed to the analysis. As far as we know, this novel combination of hybrid features, diverse



a)

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

The table displays extracted GLCM based quantitative features from Lung cancer CT images, categorized by class:

	AutoCorrelation	Contrast	Correlation1	Correlation2	Cluster Prominence	Cluster Shade	Dissimilarity	Energy	Entropy	Homogeneity	Homogeneity2
NSCLC: 377	8.8320	1.2275	0.8568	0.8558	1.0301e+03	93.2280	0.5063	0.3865	2.1374	0.8287	0.8102
	26.9979	0.3898	0.9715	0.9715	838.6724	24.1808	0.2468	0.2693	1.8972	0.8916	0.8886
	28.8354	0.4003	0.9705	0.9705	832.2096	25.7871	0.2561	0.2586	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	768.5864	33.0323	0.2735	0.2410	2.0285	0.8792	0.8701
	25.2663	0.4335	0.9659	0.9659	759.0302	33.7197	0.2798	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
NSCLC: 377	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	0.8860	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	25.9968	8.8521	75.4939	1.8003	0.4124	0.6337	-0.5825	0.9004	0.9719	0.9942	NSCLC
	0.4407	25.5646	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	-0.5571	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
SCLC: 568	0.7485	6.1884	3.6825	17.3692	1.1206	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	1.1287	0.6153	0.4956	-0.5918	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	0.4983	-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. Hybrid features based a) Schematic diagram b) GLCM quantitative features from CT images.

preprocessing, and optimized parameters is a first, and it significantly boosted our prediction accuracy.

Tuning the Algorithm: Just like fine-tuning a musical instrument, optimizing the settings (hyperparameters) of machine learning algorithms can greatly improve their performance. We used a systematic grid search technique to find the ideal settings for each algorithm. Then, we fed our single and hybrid features (again, through concatenation) into these optimized algorithms, yielding the impressive results showcased in Fig. 1. As you can see, our proposed approach takes the top spot in terms of detection performance!

Fig. 1a): Workflow Overview.

1. Image Input: The process begins with loading lung cancer images as input.
2. Feature Extraction: From these images, we extract crucial features of two types: those generated by an autoencoder neural network and traditional texture features like Haralick and GLCM (Fig. 1b showcases examples of GLCM features for different cancer types.).
3. Single and Hybrid Approaches: We utilize both single-feature and hybrid-feature extraction strategies. In the hybrid approach, features are combined strategically, such as GLCM + Autoencoder, Haralick + Autoencoder, or GLCM + Haralick.
4. Machine Learning Analysis: These single and hybrid features are then fed into robust machine learning classifiers like SVM with various kernels, Naive Bayes, and Decision Trees. To maximize their performance, we optimize their internal parameters.

Section one provides an introduction of the research problem background and flow of work. Section two provides Materials and methods including dataset sources, pre-processing methods, parameter optimization methods, training and testing data validation, feature extraction methods, machine learning classification methods. Section three detailed the results and discussions. Section four depicts the conclusion with summary and main findings of the study.

2. Materials and methods

2.1. Datasets

Our data comes courtesy of the Lung Cancer Alliance (LCA), who generously make it available on their website (<https://wgntv.com/news/medical-watch/early-lung-cancer-detection-the-lifesaving-scan-many-smokers-skip/>). This valuable resource has been used in other studies, such as [38], and comprises DICOM-format images from 76 patients with lung cancer. In total, there are 945 images, with 377 showcasing non-small cell lung cancer (NSCLC) and 568 depicting small cell lung cancer (SCLC).

2.2. Pre-processing

Before diving into analysis, computer vision tasks often involve image pre-processing [39,40]. Think of it as tidying up a room before welcoming guests: it enhances the image's quality, making it easier to segment, analyze, and extract key features.

2.2.1. Image resize

Shrinking or stretching an image to fit a frame, save space, or sharpen details – that's image resizing [41]. We used a clever technique called interpolation, where new pixels are "guestimated" based on their neighbors, helping us resize without warping the image. But it's not just about shrinking or stretching! Keeping the "shape" of the image matters too. That's where aspect ratio, the width-to-height proportion, comes in. Ignoring it can stretch or squish the image unnaturally [42]. To avoid this, we resized proportionally, like adjusting both height and width equally, or used a built-in "keep aspect ratio" feature.

2.2.2. Data augmentation

Data augmentation is a technique that artificially adds spice to your dataset by creating new versions of your existing data [43]. This helps combat overfitting, a problem where the model becomes too focused on the specific training examples and can't perform well on unseen, new data. By boosting the size and variety of the training data, data augmentation helps the model generalize better and become more reliable overall.

2.3. Hyperparameters optimization

Think of a machine learning algorithm like a baker preparing a cake. While the ingredients (training data) are crucial, the recipe's instructions (hyperparameters) can drastically affect the outcome [44]. These predefined settings, set before baking (training), guide the algorithm's learning process and determine its effectiveness.

2.4. Grid search method

Finding the sweet spot for your machine learning model is like mixing a perfect cocktail. Just as the right ratios of ingredients make a tasty drink, choosing the ideal settings (hyperparameters) can make your model perform at its best. Grid search, like a meticulous bartender trying every possible combination, systematically tests predefined hyperparameter values, picking the winner based on validation data performance [45–48]. Random search, on the other hand, is a bit more spontaneous, throwing darts of random hyperparameter combinations and keeping the one that hits the validation bullseye. Grid search, while reliable, can become a slow and thirsty process, especially with complex models or many knobs to tweak [45–48]. This is where swifter approaches like randomized

search and Bayesian optimization come in. These techniques explore the hyperparameter space more nimbly, often reaching the peak performance point much faster than grid search.

2.5. Training/testing data validation

This popular technique splits the data into ten equal parts, called "folds," then trains the model on nine of them while testing it on the remaining one [49]. This process is repeated ten times, with each fold taking turns as the test set. This rigorous approach helps assess the model's general performance, not just how well it fits the specific training data [50].

2.6. Feature extraction

This technique identifies and extracts key features from datasets, reducing redundancy and capturing the most relevant information for analysis. Before computers can learn from data, we need to clean it up and extract the key ingredients: features! This step, called feature extraction, comes before feeding the data to the model. The objective is to identify features that hold the greatest significance for the specific task and uncover the pattern in the data. The specific feature extraction is the preliminary step in Machine learning techniques. Researchers have recently developed hybrid features [51–53] and explored various feature extraction methods [28,34, 53–55] to improve the detection of different imaging pathologies. This study employs feature extraction strategies based on the GLCM, Haralick, and Autoencoder. Before feeding data to a machine learning model, we often need to "mine" for the most valuable nuggets of information.

Previously researchers [51–53] developed different feature extraction methods [28,34,53–55] for improving different pathologies in medical imaging problems. This study utilized the single and hybrid of following feature extracting approach.

2.6.1. Haralick texture features

To capture the intricate textures within the images, we relied on Haralick features. These handy tools have proven their worth in classifying different types of data, like colon biopsies [56,57]. In our case, we're pioneering the use of Haralick features to analyze lung cancer patterns in CT scans.

2.6.2. Gray level Co-occurrence matrix (GLCM)

GLCM is a classic texture analysis technique for a reason. This powerful tool dives deeper than just looking at pixels, exploring how they relate to each other. Think of it as analyzing not just the individual brushstrokes in a painting, but also how they come together to create the bigger picture.

The GLCM features were first proposed by Haralick. Consider an image which have N_x column and N_y row. Consider, gray level to quantize to N_g levels. Let the column represented by $L_x = 1, 2, 3, 4, 5, 6, \dots, N_x$ and row represented by $L_y = 1, 2, 3, 4, 5, 6, \dots, N_y$ and set of N_g quantized gray level denoted by $H = 1, 2, 3, 4, 5, 6, \dots, N_g$. The texture data framework is denoted by matrix of comparative frequencies $Q_{u,v}$ having two adjacent pixels parted by shift c and angle θ . The GLCM features are computed using equation (1)):

$$\begin{aligned} Q(u, v, d, 0) &= \{(x_1, y_1), (x_2, y_2) | f(x_1, y_1) = i \\ H(x_2, y_2) &= j, (x_1, y_1) - (x_2, y_2) = c \\ K &= (x_1, y_1), (x_2, y_2) \end{aligned} \quad (1)$$

Here x, y denote the number of occurrences within windows magnitudes, the 1st pixel is (x_1, y_1) and 2nd pixel (x_2, y_2) denote the strength rank of derivatives from v to u . The distance between the pixel is denoted by c , the point of view between th pixel is denoted by θ . Synchronization matrix is inherently not symmetric. For GLCM to be symmetric, one of the viewpoints up to 180° required to be measured. equation (2)): is computed to formulate the symmetric synchronization matrix.

$$Q(u, v, d, 0) k = (Q(u, v, d, 0) + Q(u, v, d, 0)^T, T) \text{ divid by } 2 \quad (2)$$

Where $Q(u, v, d, 0)$, T is a transpose of $Q(u, v, d, 0)$. Possibility approximations are obtained to divide 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 is computed using equation (3)).

$$Q(u, v, d, 0)^X = \frac{P(i, j, d, \theta) + P^T(i, j, d\theta)}{2 * \sum \theta L_x L_y, d, \theta} \quad (3)$$

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

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

Equation (3) is constant, and equation (2) can be modified as in equation (4)) to be appropriate and to evade separation on the FPGA represented by equation (4)).

$$Q(u, v, d, 0)^N = \frac{P(i, j, d, \theta) + P^T(i, j, d\theta)}{2 * \sum \theta L_x L_y, d, \theta} \quad (4)$$

For a displacement vector $d(dx, dy)$, the elements of a $G \times G$ GLCM are identified as $Pd = (i, j) = |\{(r, s), (t, v) : I(r, s) = i, I(t, v) = j\}|$, where I indicate the GLCM image value and $(r, s), (t, v)$ and (dx, dy) is the cardinality set. Let $Q(u, v, d, 0)$ is occurrence of rate. Medical imaging plots are like visual fingerprints for cancer diagnoses, helping classify and assess tumors [58].

2.6.3. Autoencoder

Imagine trying to pack your entire wardrobe into a tiny suitcase without sacrificing your favourite outfits! Autoencoders, a type of artificial brain, can do something similar with data. They squeeze it down into a smaller, neater version, called a "latent space," without losing the important bits. This compressed form helps remove noise and clutter, making it easier to analyze and extract crucial features, like packing only the clothes you'll actually wear.

Autoencoder takes 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 Weight matrix w i.e. $w = (wt)$, pre-trained weights are often beneficial when your task aligns closely with the task the weights were initially trained on. This allows for knowledge transfer and faster convergence. To abate the standard rebuilding error as reflected in equations (5) and (6)):

$$\theta^*, \theta'^* = \arg_{\theta, \theta'} \frac{1}{n} \sum_{i=0}^n L(x^{(i)}, z^{(i)}) \quad (5)$$

$$\arg_{\theta, g} \frac{1}{n} \sum_{i=0}^n L(x^{(i)}, g_\theta(f_\theta(x^{(i)}))) \quad (6)$$

Corresponding likelihoods (Bernoulli's) is represented by equation 7):

$$L_H(x, z) = H(B_x \parallel B_z) = - \sum_{k=1}^d x_k \log z_k + (1 - x_k) \log (1 - z_k) \quad (7)$$

If x is a binary vector, the negative log-likelihood (x, z) can be expressed. For the given example x , the Bernoulli parameters are set as z . The equation can be written as depicted in equation (8)):

$$\theta^*, \theta'^* = \arg_{\theta, \theta'} E_{q^*}(x)[L_H(X, g_\theta(f_\theta(X)))] \quad (8)$$

2.6.3.1. Hybrid features. Hybrid features combine multiple types of information, like image textures and extracted data patterns, to boost a model's performance [56]. We used them by simply stitching together [57] (concatenating) different features we extracted from lung cancer images [59,60].

While hybrids can pack a punch, they can also make the model a bit more complex and demanding to run [61,62]. But the benefits

Table 1
Single Features extraction strategy.

GLCM (22)	Haralick (14)	Autoencoder (50)
Autocorrelation	Contrast	Latent variables
Contrast	Variance	Encoder features
Correlation1	Sum Avg.	Bottleneck features
Correlation2	Sum Ent.	Reconstruction features
Cluster shade	Correlation	Decoder features
Energy	Maximal Correlation Coefficient	Other features including textures, edges, colors or shapes
Cluster Prominance	Diff. Entropy	
Homogeneity2	Inf. measure of Corr1	
Sum avg	Inverse Diff. Movement	
Sum entropy	Sum Var. Entropy	
Dissimilarity	Diff. Var.	
Entropy	Info. measure of Corr2	
Inf. measure of Corr1		
Inverse Diff. Movement normalized.		
Homogeneity1		
Info. measure of Corr2		
Diff. Ent.		
Inverse Diff.		
Max. Probability		
Sum of Sqr. Var.		
Sum Var.		
Diff. Var.		
Inverse Diff. Normalized		

often outweigh the challenges: studies show hybrid features can significantly improve accuracy in various tasks [61,63–65]. In this paper, we explore three hybrid combinations: Haralick + GLCM, GLCM + Autoencoder, and Autoencoder + Haralick. Each adds a unique perspective to the puzzle, helping our model identify lung cancer nodules more effectively. We chose these specific features because no single one tells the whole story [66]. Haralick and GLCM capture image textures, while Autoencoder uncovers hidden patterns. By putting them together, we get a richer picture of the data, leading to better results. The procedure is outlined in [Tables 1](#) and [2](#) regarding single and hybrid feature strategy.

2.6.3.2. Algorithmic steps. Step 1. Load and preprocess the images.

- Load the images into a numpy array.
- Convert the images to grayscale, if necessary.
- Normalize the pixel values of the images to the range [0, 1].

Step 2. Compute the GLCM

- Choose the offset distance and angle for the GLCM.
- Compute the GLCM for each image.

Step 3. Compute the Haralick texture features

- Compute the Haralick texture features from the GLCM. Some common features include contrast, energy, homogeneity, and correlation.

Step 4. Compute the Autoencoder features

1. Flatten each image.
2. Extract features from the flattened image using the autoencoder.

Step 5. Combine the Haralick texture features into a feature vector

- Concatenate the GLCM, Haralick texture and Autoencoder features for each image to create a feature vector.

Step 5. Machine learning model training on the feature vector

- Train the machine learning model on the feature vector to perform the desired task, such as image classification or segmentation.

Table 2
Hybrid features approach.

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

Table 3

Hybrid (Haralick + (GLCM) Features based detection performance and utilizing ML Algorithms to differentiate NSCLC from SCLC.

Methods	Sensitivity	Specificity	PPV	NPV	Accuracy	FPR	AUC
Naïve Bayes (NB)	0.9449	0.9326	0.9568	0.9146	0.9402	0.06742	0.9895
SVM RBF	1	0.9831	0.9895	1	0.9935	0.01685	1
Decision Tree (DT)	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 Polynomial	0.9982	0.9972	0.9982	0.9972	0.9978	0.002809	0.9995

2.7. Classification

Classification, a widely used technique [31,67–69]. Let's label and understand data based on predefined categories. In our case, these classifiers analyze extracted features from lung cancer images [59,60,67–71] to identify potential tumors. To get a well-rounded picture, we used a 10-fold cross-validation approach, essentially repeating the sorting process ten times with different training and testing sets.

2.8. Area under the receiver operating characteristics (ROC) curve (AUC-ROC)

AUC-ROC, a popular tool for evaluating binary classification models [72], assesses how well they can distinguish between two classes, like apples and oranges. It does this by looking at a special curve called a ROC curve.

2.9. Cohens kappa

Cohen's kappa is a statistical coefficient assessing the extent of agreement between multiple raters classifying categorical data. It corrects for agreement attributable to chance alone, providing a more reliable measure of true consensus. The formula for Cohen's kappa is as reflected in equation (9):

$$Kappa = \frac{(po - pe)}{1 - pe} \quad (9)$$

where:

- po is the observed agreement between the two raters.
- pe is the expected agreement between the two raters, given that they are rating independently.
- 1 - pe is the maximum possible agreement between the two raters, given their marginal distributions.

Cohen's kappa values range from 0 to 1, with 0 indicating no agreement and 1 indicating perfect agreement. Kappa scores in the ranges 0.01–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1.00 represent slight, fair, moderate, substantial, and almost perfect agreement, respectively.

Quadratic weighted Cohen's kappa

Quadratic weighted Cohen's kappa (QWKO) is a variant of Cohen's kappa that considers the severity of disagreements. It does so by assigning greater weight to disagreements between more distant categories. The formula for QWKO is as shown in equation (10):

$$QWKO = \frac{(1 - po)}{(1 - pe(w))} \quad (10)$$

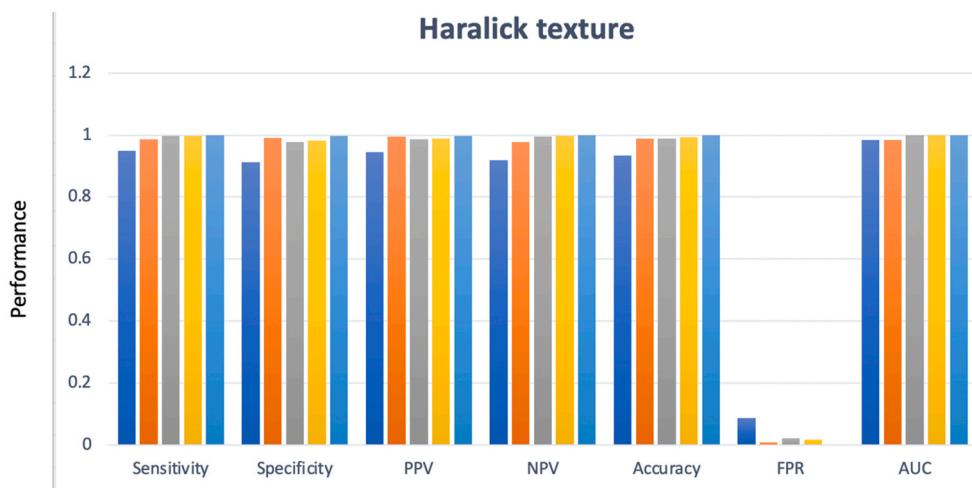
where:

- po is the observed agreement between the two raters.
- pe(w) is the expected agreement between the two raters, given that they are rating independently and that disagreements are weighted according to a weighting matrix w.

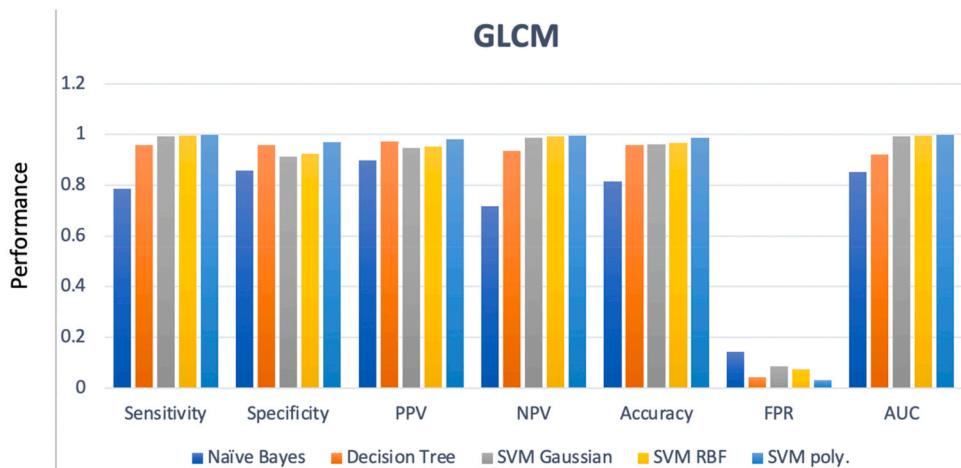
The weighting matrix w determines the severity of disagreements between categories. For instance, a matrix with off-diagonal elements of 1 indicates equal weighting for all disagreements, while one with zeros indicates only perfect agreement counts. QWKO's sensitivity to disagreements between distant categories makes it suitable for ordered categories, like Likert scale ratings.

3. Results and discussions

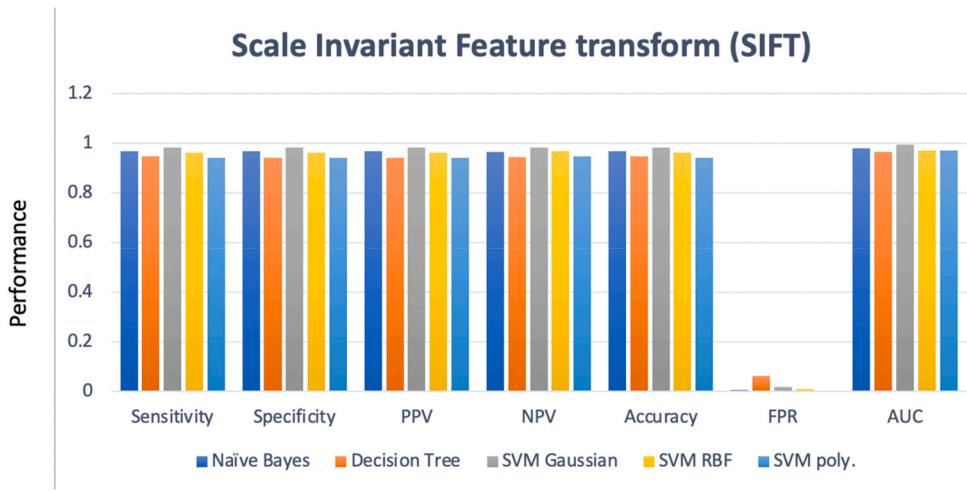
Our goal is to boost lung cancer detection. To achieve this, we explored a treasure trove of features from lung cancer images, like textures analyzed by GLCM and Haralick, and hidden patterns uncovered by autoencoders. We then used these features, both individually and in clever combinations, as clues for machine learning models to identify both non-small cell lung carcinoma (NSCLC) and



a)



b)



c)

(caption on next page)

Fig. 2. Single features based detection performance a) Haralick texture b) GLCM c) Scale Invariant Feature Transform (SIFT) to differentiate NSCLC from SCLC using machine learning techniques.

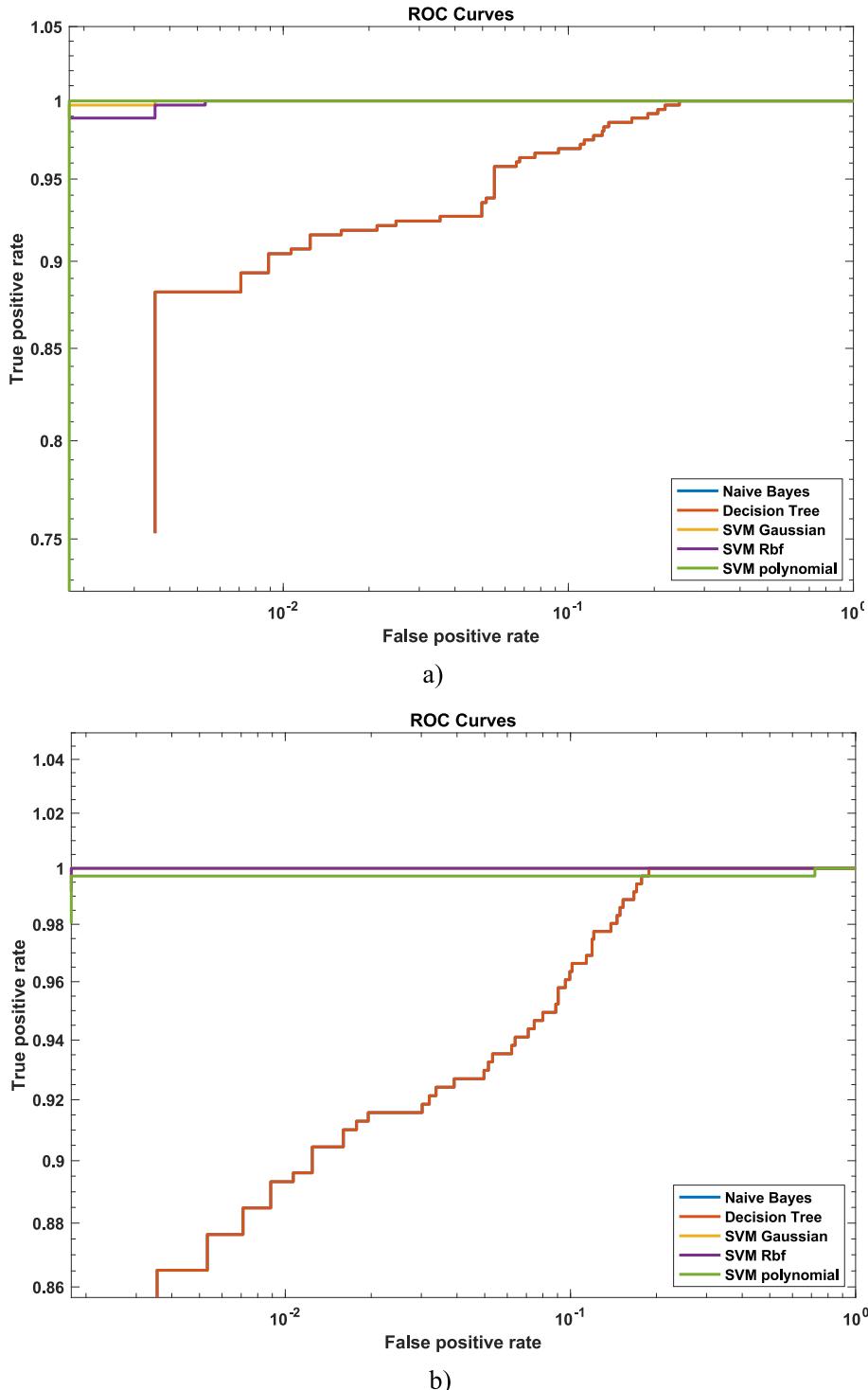


Fig. 3. AUC to differentiate NSCLC from SCLC by extracting hybrid feature a) GLCM b) GLCM + Haralick.

small cell lung carcinoma (SCLC). Standard metrics helped us compare their performance and pick the winning combinations.

Fig. 2(a–c) presents the lung cancer detection results using single feature extraction strategies. Haralick texture features **Fig. 2a**) exhibited the highest performance, achieving an accuracy of 99.89% and an AUC of 0.9984. GLCM features **Fig. 2b**) yielded an accuracy of 98.69% with SVM polynomial, while SIFT features **Fig. 2c**) attained an accuracy of 98.31% using SVM Gaussian.

Fig. 3(a–b) depicts the AUC) for single feature extraction approach by computing **Fig. 3a**) GLCM and **Fig. 3b**) Haralick texture features.

Using GLCM and Haralick features in combination with machine learning techniques are reflected in **Table 3**. The SVM polynomial achieved the highest performance, reaching an accuracy of 99.78% and AUC of 0.9995.

Table 4 presents the detection performance achieved using the hybrid (GLCM + Autoencoder) feature extraction methodology and robust machine learning techniques. A 100% highest detection performance was yielded using hybrid features (Haralick + autoencoder, GLCM + autoencoder) utilizing SVM RBF and Gaussian, as shown in **Tables 4 and 5**.

Table 6 reflects Cohen's Kappa statistics using SVM Gaussian and SVM polynomial based on the extracted GLCM features. Using SVM Gaussian, the observed agreement (po) is 0.9989, the random agreement (pe) is 0.5256, and the agreement due to true concordance (po-pe) is 0.4733. This indicates that there is a very high level of agreement between the two raters, and that this agreement is not simply due to chance. Cohen's kappa is 0.9977, which indicates almost perfect agreement. The kappa error is 0.0023, and the kappa C.I. (alpha = 0.0500) is 0.9932–1.0022. This indicates that the observed kappa is statistically significant. The maximum possible kappa, given the observed marginal frequencies, is 0.9977. This means that the observed kappa is very close to the maximum possible kappa, which further supports the conclusion that there is a very high level of agreement between the two raters. The ratio of the observed kappa to the maximum possible kappa is 1.0000. This indicates that the observed kappa is perfect agreement. The z-score for the observed kappa is 435.1988, and the p-value is 0.0000. This indicates that the observed kappa is statistically significant at the alpha = 0.05 level.

The AUC-ROC curves for distinguishing SCLC from NSCLC using hybrid GLCM + Haralick features are presented in **Fig. 4(a–e)**. The mean and standard deviation values of false positive rates (FPRs) and true positive rates (TPRs) were calculated for each machine learning algorithm, and linear curves were fitted to the data. The Naïve Bayes **Fig. 4a**) and decision tree **Fig. 4b**) classifiers yielded mean FPRs of 0.4652 and mean TPRs of 0.9848, with standard deviations of 0.3021 for FPR and 0.049 for TPR. The SVM Gaussian **Fig. 4c**), SVM RBF **Fig. 4d**), and SVM polynomial **Fig. 4e**) classifiers achieved mean FPRs of 0.3085, 0.3085, and 0.3088, respectively, and mean TPRs of 0.8087, 0.8089, and 0.8070, with standard deviations of 0.3329 for FPR and 0.3015, 0.3329 for FPR and 0.3016 for TPR, and 0.3328 for FPR and 0.3016 for TPR, respectively.

Fig. 5(a–b) presents parallel coordinate plots for visualizing high-dimensional data. These plots represent each observation in a dataset as a line connecting its values for each variable, enabling the visualization of patterns and relationships within the data. The plots utilize GLCM features extracted from MRI images of lung cancer types. The dataset includes 356 NSCLC MRI subjects and 563 SCLC MRI subjects. Blue lines represent NSCLC cases, while red lines represent SCLC cases. Solid lines indicate correctly predicted cases, while asterisks with lines indicate incorrectly detected cases. In **Fig. 5a**), the validation accuracy reaches 99.9%, with true positives (TP) of 355, false positives (FP) of 1, true negatives (TN) of 563, and false negatives (FN) of 0. **Fig. 5b**) shows the predictions using SVM Gaussian, with true positives (TP) of 347, false positives (FP) of 9, true negatives (TN) of 561, and false negatives (FN) of 2.

Table 7 summarizes the findings of the current study and compares them to those of previous studies. This study aimed to enhance lung cancer detection performance in three ways: 1) refining preprocessing steps; 2) improving feature extraction strategies; 3) parameters optimization of ML algorithms.

Image preprocessing is like that, but for digital pictures. It's the crucial first step where we prepare images for further analysis by removing noise, adjusting brightness and contrast, and sometimes even straightening them out. Previous approaches to lung cancer detection relied on traditional feature extraction methods, which often missed crucial information. For example, Guo et al. [75] focused on texture and shape features, reaching a sensitivity of 94.0%. Sousa et al. [80] explored a broader range with gradients, histograms, and spatial features, reaching 95.0% accuracy. Messay et al. [76] combined gradients, shape, and intensity, but only achieved 82.0% sensitivity. Dandil et al. [79] took a multi-pronged approach with GLCM, shape, statistical, and energy features, reaching their highest results: 95.0% accuracy, 97.0% sensitivity, and 94.0% specificity.

We computed single and hybrid features. The Haralick yielded an accuracy of 99.89% using SVM polynomial followed by GLCM with accuracy of 98.69% and SIFT with 98.31% accuracy using single feature extracting approach. The hybrid features yielded 100% accuracy and 1.00 AUC.

Table 4

Hybrid (Autoencoder + GLCM) features based performance detection and employing machine learning techniques to differentiate SCLC from NSCLC.

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 Polynomial	1	0.9972	0.9982	1	0.9989	0.002809	0.9999
SVM RBF	1	1	1	1	1	0	1
SVM Gaussian	1	1	1	1	1	0	1

Table 5

Hybrid (Haralick + Autoencoder) features based performance detection and utilizing robust machine learning techniques.

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 Polynomial	1	0.9972	0.9982	1	0.9989	0.002809	1
SVM RBF	1	1	1	1	1	0	1
SVM Gaussian	1	1	1	1	1	0	1

Table 6

Quadratic weighted Cohen's Kappa statistics.

Kappa statistics	SVM Gaussian	SVM Polynomial
Observed agreement (po)	0.9989	0.9880
Random agreement (pe)	0.5256	0.5271
Agreement due to true concordance (po-pe)	0.4733	0.4609
Maximum possible kappa, given the observed marginal frequencies	0.9977	0.9839
Cohen's kappa	0.9977	0.9747
Residual not random agreement (1-pe)	0.4744	0.4729
kappa C.I. (alpha = 0.0500)	0.9932	1.0022
z (k/kappa error)	435.1988	p = 0.0000
kappa error	0.0023	0.0076
k observed as proportion of maximum possible	1.0000	0.9906

4. Conclusion

Our recent study aimed to enhance lung cancer detection through refined preprocessing steps, optimized feature extraction strategies, and fine-tuned machine learning hyperparameters. Employing Haralick features and SVM polynomial, we achieved the highest accuracy of 99.89% among single feature extraction approaches, followed by 98.69% with GLCM features and SVM polynomial. However, the most impressive detection performance was achieved using SVM RBF with hybrid features GLCM + Autoencoder and Haralick + Autoencoder, resulting in perfect 100% sensitivity, specificity, PPV, NPV, accuracy, and AUC. The second-highest detection accuracy of 99.56% was obtained using SVM polynomial and GLCM + Haralick features, with SVM Gaussian achieving a detection accuracy of 99.35%.

Harnessing the power of diverse data: We propose a novel approach to lung cancer detection using hybrid features, combining multiple data types like scans and tissue textures. This strategy offers several advantages over single data types:

- Enhanced Patient Representation: By capturing a wider range of information,

Leveraging diverse feature sets allows the model to capture a more multifaceted portrait of patients, thereby enhancing its ability to differentiate between cancer and non-cancer. Imagine it like having a detective with multiple lines of evidence instead of just one blurry photo.

- Improved Data Quality: Hybrid features act as built-in data filters, automatically removing irrelevant or redundant information. This reduces noise and strengthens the signal, leading to a more robust and accurate machine learning model. Think of it like cleaning up a messy crime scene before starting the investigation.
- Promising Clinical Impact: The potential benefits extend beyond detection. This approach can inform decision-making, diagnosis, and even prognosis, ultimately improving patient outcomes.

Addressing Data Challenges: We were aware of the limitations of our study. The dataset was relatively small and imbalanced, which could lead to overfitting. However, we employed several strategies to mitigate this risk:

- K-fold Cross-Validation: This technique splits the data into multiple subsets, training the model on different combinations and preventing overconfidence on limited information. Imagine testing your detective skills on various cases instead of just one.
- Data Augmentation: We artificially expanded the dataset by creating variations of existing data points, providing the model with a richer training experience. Think of it like giving your detective additional evidence created from existing clues.
- Refined Preprocessing and Feature Extraction: We meticulously optimized every step of the pipeline, from data cleaning to feature selection, ensuring the model receives the most relevant and effective information. Imagine the detective honing their observation skills and focusing on the crucial details.
- Cutting-edge AI Techniques: We further enhanced prediction performance by incorporating recent advancements in artificial intelligence. Think of it as equipping your detective with the latest forensic tools.

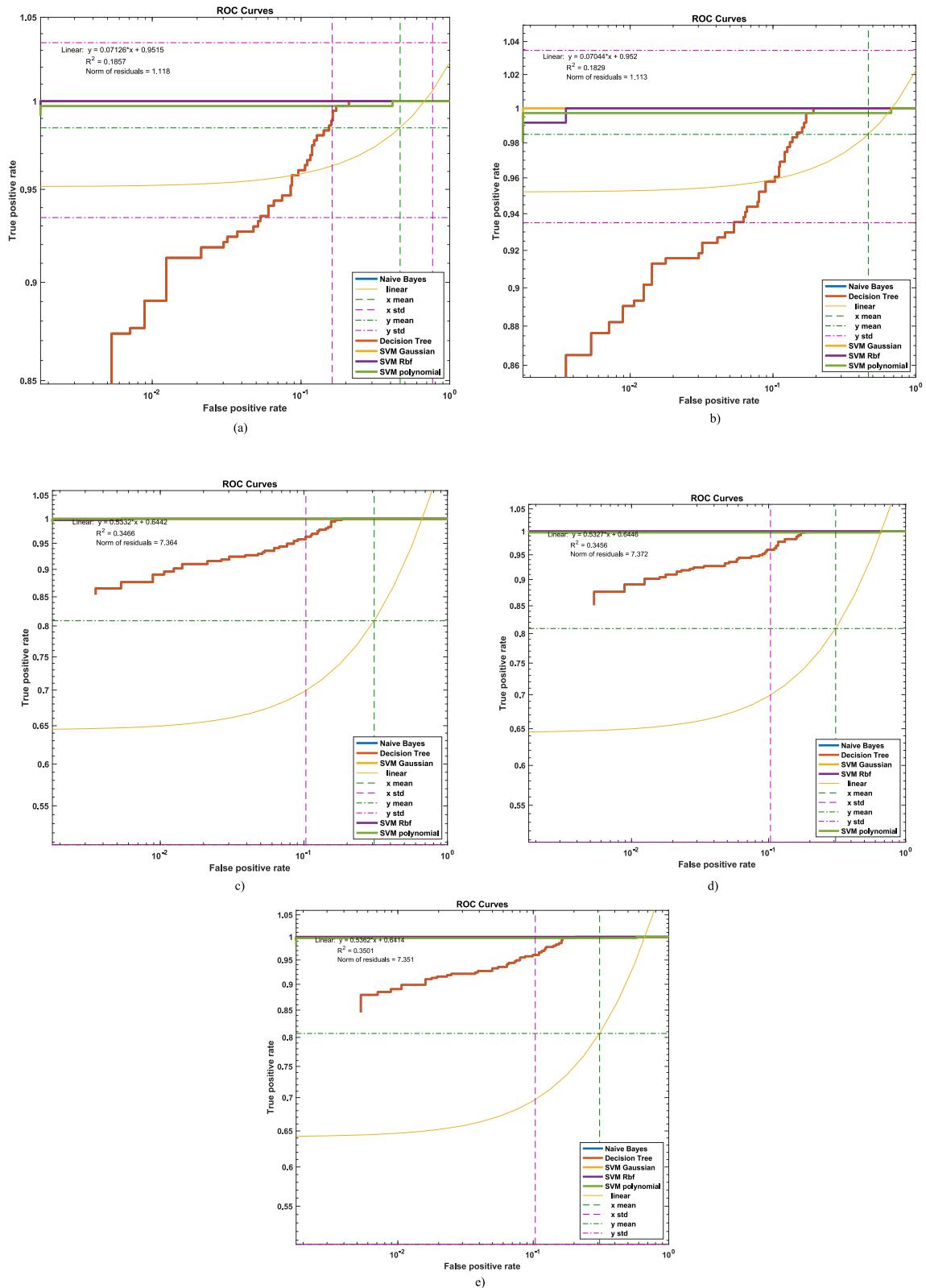


Fig. 4. Hybrid (Haralick + GLCM) features based AUC to differentiate NSCLC from SCLC by computing by fitting a linear curve on the AUC along with mean and standard deviation using logarithmic scales a) Naïve Bayes, b) Decision Tree, c) SVM Gaussian, d) SVM RBF, e) SVM Polynomial.

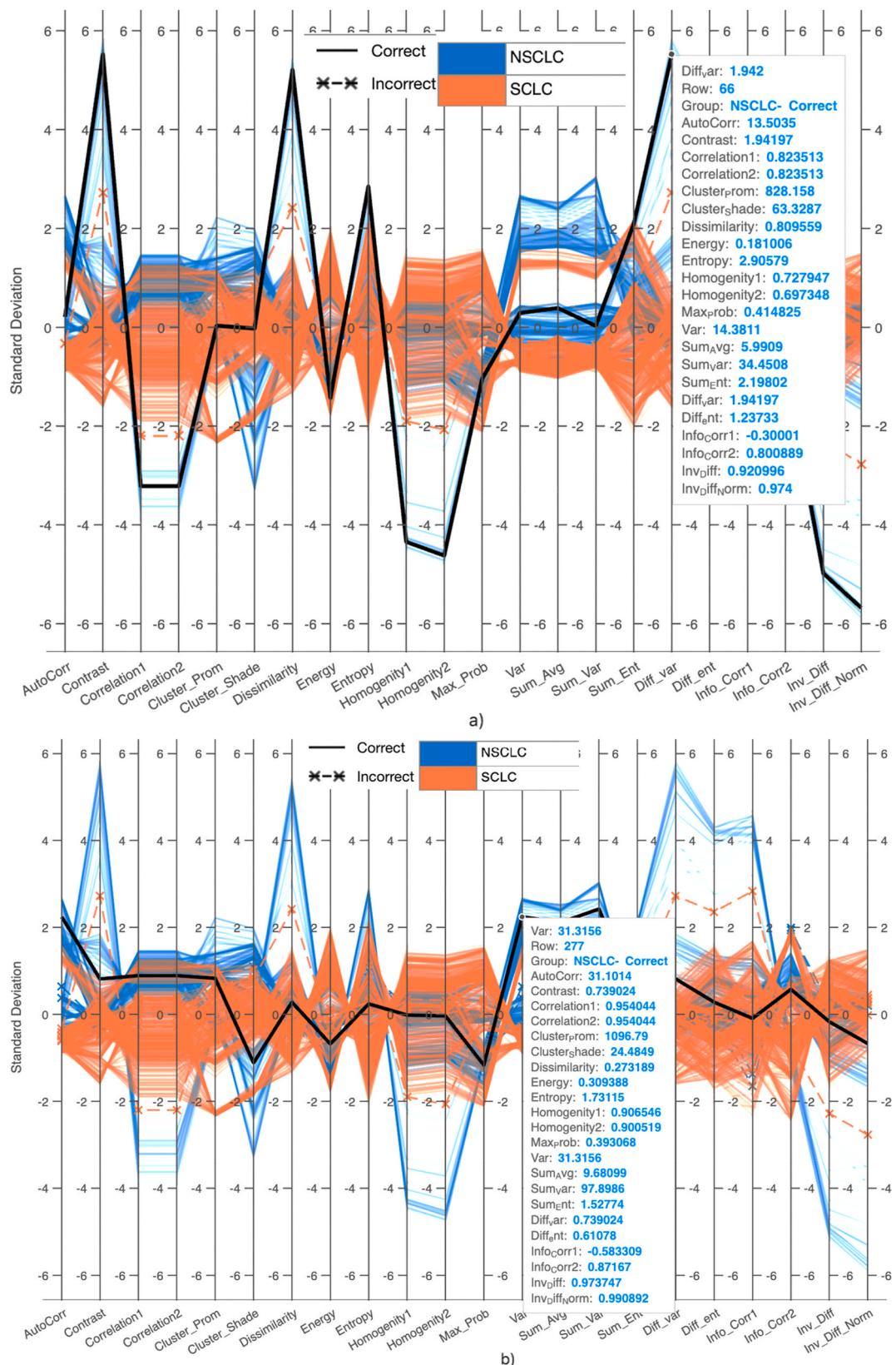


Fig. 5. Parallel Coordinate graph by extracting GLCM features from lung cancer sub-types MRI images using a) SVM Gaussian, b) SVM Polynomial.

Table 7

Comparison of results with previous studies.

Author	Features Used	Performance
Teramoto et al. [73]	1 Shape 2 Intensity	Sensitivity = 83.00%,
Orozco et al. [74]	1 Texture	Sensitivity = 84.00%,
Guo et al. [75]	1 Texture 2 Shape	Sensitivity = 94.00%,
Messay et al. [76]	1 Shape 2 Gradient 3 Intensity	Sensitivity = 82.00%,
Retico et al. [77]	1 Texture 2 Morphology	Sensitivity = 72.00%,
Hussain et al. [78]	RICA features and SVM	Accuracy = 99.77%
Dandil et al. [79]	1 Statistical 2 Shape 3 GLCM 4 Energy	Sensitivity = 97.00%, Specificity = 94.00% Accuracy = 95.00%
This study	Single features a) Haralick (SVM polynomial) b) GLCM (SVM polynomial) c) SIFT (SVM Gaussian) <u>Hybrid features approach</u> (GLCM + autoencoder, Haralick + Autoencoder, GLCM + Haralick) features	Single Features a) Accuracy = 99.89% b) Accuracy = 98.69% c) Accuracy = 98.39% <u>Hybrid features approach</u> Specificity = 100% Sensitivity = 100%, AUC = 1.00 Accuracy = 100%

5. Limitations and future directions

Further methodological improvements can be achieved through hybrid deep learning methods and optimized parameter tuning. We will also evaluate the performance using additional metrics and visualization techniques. Additionally, we will test these methodologies on larger and more diverse lung cancer datasets. Furthermore, we will incorporate clinical information alongside imaging features to enhance diagnostic accuracy and improve disease recurrence, survival, and severity predictions. The current study employed a small and imbalanced dataset and evaluated the hybrid feature extraction approach with a limited number of machine learning algorithms. Other machine learning algorithms may potentially yield even better performance. Additionally, the clinical significance of the hybrid feature extraction approach requires further evaluation. Larger datasets and comparisons with existing clinical tools for lung cancer detection are necessary. We will address these limitations by evaluating the hybrid feature extraction approach on larger and more diverse lung cancer datasets. We will also incorporate clinical information, such as patient demographics, medical history, and symptoms, to improve the performance and clinical significance of the approach. Additionally, we will compare the hybrid feature extraction approach to other state-of-the-art lung cancer detection methods. Furthermore, we will develop deep learning models capable of automatically extracting hybrid features from medical images and test the hybrid feature extraction approach for other medical imaging tasks, such as the detection of other types of cancer, neurodegenerative diseases, and cardiovascular diseases.

Finally, we will develop a hybrid feature extraction approach tailored to specific lung cancer subtypes, such as adenocarcinoma or squamous cell carcinoma, and utilize the proposed approach as a predictive model for lung cancer recurrence or survival.

CRediT authorship contribution statement

Liangyu Li: Methodology, Funding acquisition, Formal analysis, Conceptualization. **Jing Yang:** Investigation. **Lip Yee Por:** Investigation, Formal analysis, Data curation, Conceptualization. **Mohammad Shahbaz Khan:** Investigation, Formal analysis, Conceptualization. **Rim Hamdaoui:** Methodology, Funding acquisition, Data curation, Conceptualization. **Lal Hussain:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zahoor Iqbal:** Project administration, Methodology, Investigation, Conceptualization. **Ionela Magdalena Rotaru:** Writing – review & editing, Supervision, Formal analysis. **Dan Dobrotă:** Writing – review & editing, Funding acquisition, Formal analysis. **Moutaz Aldrdery:** Writing – review & editing, Investigation, Formal analysis. **Abdulfattah Omar:** Software, Methodology, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by King Khalid University, Abha, Saudi Arabia. The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number (R.G.P. 2/57/44). The author would like to thank the Deanship of Scientific Research at Shaqra University for supporting this work. This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2023/R/1445).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.heliyon.2024.e26192>.

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/tlc-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 tetraspanin 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](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. Scrimali, 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/tlc-20-536>.
- [6] 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>.
- [7] 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>.
- [8] Z. Raizah, U.K. Kodipalya Nanjappa, H.U. Ajipura 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>.
- [9] 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 expression microarray data using a reject option, *Appl. Artif. Intell.* 36 (2022), <https://doi.org/10.1080/08839514.2022.2151171>.
- [10] 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>.
- [11] F. Althoeij, M.N. Akhter, Z.S. Nagra, H.H. Awan, F. Alanaizi, M.A. Khan, M.F. Javed, S.M. Eldin, Y.O. Özkilic, 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>.
- [12] 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>.
- [13] P. Liang, H. Wang, Y. Liang, J. Zhou, H. Li, Y. Zuo, Feature-scML: an open-source Python package for the feature importance visualization of single-cell omics with machine learning, *Curr. Bioinform.* 17 (2022) 578–585, <https://doi.org/10.2174/1574893617666220608123804>.
- [14] A. Shahbandegan, V. Mago, A. Alaref, C.B. van der Pol, D.W. Savage, Developing a machine learning model to predict patient need for computed tomography imaging in the emergency department, *PLoS One* 17 (2022) e0278229, <https://doi.org/10.1371/journal.pone.0278229>.
- [15] B. V A. M. Subramonian, L. Mathew, Noninvasive detection of COPD and Lung Cancer through breath analysis using MOS Sensor array based e-nose, *Expert Rev. Mol. Diagn.* 21 (2021) 1223–1233, <https://doi.org/10.1080/14737159.2021.1971079>.
- [16] V.A. Binson, M. Subramonian, L. Mathew, Discrimination of COPD and lung cancer from controls through breath analysis using a self-developed e-nose, *J. Breath Res.* 15 (2021) 046003, <https://doi.org/10.1088/1752-7166/ac1326>.
- [17] C. Freitas, C. Sousa, F. Machado, M. Serino, V. Santos, N. Cruz-Martins, A. Teixeira, A. Cunha, T. Pereira, H.P. Oliveira, J.L. Costa, V. Hespanhol, The role of liquid biopsy in early diagnosis of lung cancer, *Front. Oncol.* 11 (2021), <https://doi.org/10.3389/fonc.2021.634316>.
- [18] M.R. Lener, E. Reszka, W. Marciniak, M. Lesicka, P. Baszuk, E. Jabłońska, K. Białkowska, M. Muszyńska, S. Pietrzak, R. Derkacz, T. Grodzki, J. Wójcik, M. Wojtyś, T. Dębniak, C. Cybulski, J. Gronwald, B. Kubisa, J. Pieróg, P. Waloszczyk, R.J. Scott, A. Jakubowska, S.A. Narod, J. Lubinski, Blood cadmium levels as a marker for early lung cancer detection, *J. Trace Elem. Med. Biol.* 64 (2021) 126682, <https://doi.org/10.1016/j.jtemb.2020.126682>.
- [19] S. Farooq Abbasi, H. Jamil, W. Chen, EEG-based neonatal sleep stage classification using ensemble learning, *Comput. Mater. Contin.* 70 (2022) 4619–4633, <https://doi.org/10.32604/cmc.2022.020318>.
- [20] S.F. Abbasi, Q.H. Abbasi, F. Saeed, N.S. Alghamdi, A convolutional neural network-based decision support system for neonatal quiet sleep detection, *Math. Biosci. Eng.* 20 (2023) 17018–17036, <https://doi.org/10.3934/mbe.2023759>.
- [21] S. Aamir, A. Rahim, Z. Aamir, S.F. Abbasi, M.S. Khan, M. Alhaisoni, M.A. Khan, K. Khan, J. Ahmad, Predicting breast cancer leveraging supervised machine learning techniques, *Comput. Math. Methods Med.* 2022 (2022) 1–13, <https://doi.org/10.1155/2022/5869529>.
- [22] S. Almutairi, M. S, B.-G. Kim, M.M. Aborokbah, N. C, Breast cancer classification using Deep Q Learning (DQL) and gorilla troops optimization (GTO), *Appl. Soft Comput.* 142 (2023) 110292, <https://doi.org/10.1016/j.asoc.2023.110292>.
- [23] H.-J. Park, J.-W. Kang, B.-G. Kim, ssFPN: scale sequence (S2) feature-based feature pyramid network for object detection, *Sensors* 23 (2023) 4432, <https://doi.org/10.3390/s23094432>.
- [24] L. Leng, A.B. Jin Teoh, M. Li, M.K. Khan, Analysis of correlation of 2DPalmHash Code and orientation range suitable for transposition, *Neurocomputing* 131 (2014) 377–387, <https://doi.org/10.1016/j.neucom.2013.10.005>.
- [25] L. Leng, M. Li, C. Kim, X. Bi, Dual-source discrimination power analysis for multi-instance contactless palmprint recognition, *Multimed. Tools Appl.* 76 (2017) 333–354, <https://doi.org/10.1007/s11042-015-3058-7>.
- [26] L. Hussain, T. Nguyen, H. Li, A.A. Abbasi, K.J. Lone, Z. Zhao, M. Zaib, A. Chen, T.Q. Duong, Machine-learning classification of texture features of portable chest X-ray accurately classifies COVID-19 lung infection, *Biomed. Eng. Online* 19 (2020) 88, <https://doi.org/10.1186/s12938-020-00831-x>.

- [27] K. Shaheed, P. Szczuko, Q. Abbas, A. Hussain, M. Albathan, Computer-Aided diagnosis of COVID-19 from chest X-ray images using hybrid-features and random forest classifier, *Healthcare* 11 (2023) 837, <https://doi.org/10.3390/healthcare11060837>.
- [28] L. Hussain, S. Saeed, I.A. Awan, A. Idris, M.S.A. Nadeem, Q.-A. Chaudhary, Q.-A. Chaudhary, Q.-A. Chaudhary, Detecting brain tumor using machine learning techniques based on different features extracting strategies, *Curr. Med. Imaging* 14 (2019) 595–606, <https://doi.org/10.2174/1573405614666180718123533>.
- [29] S. Anjum, L. Hussain, M. Ali, A.A. Abbasi, T.Q. Duong, Automated multi-class brain tumor types detection by extracting RICA based features and employing machine learning techniques, *Math. Biosci. Eng.* 18 (2021) 2882–2908, <https://doi.org/10.3934/mbe.2021146>.
- [30] M.M. Eltahir, L. Hussain, A.A. Malibari, M.K. Nour, M. Obayya, H. Mohsen, A. Younis, M. Ahmed Hamza, A bayesian dynamic inference approach based on extracted gray level Co-occurrence (GLCM) features for the dynamical analysis of congestive heart failure, *Appl. Sci.* 12 (2022) 6350, <https://doi.org/10.3390/app12136350>.
- [31] L. Hussain, W. Aziz, A.S. Khan, A.Q. Abbasi, S.Z. Hassan, Classification of electroencephalography (EEG) alcoholic and control subjects using machine learning ensemble methods, *J. Multidiscip. Eng. Sci. Technol* 2 (2015) 126–131.
- [32] L. Hussain, A. Ali, S. Rathore, S. Saeed, A. Idris, M.U. Usman, M.A. Iftikhar, D.Y. Suh, Applying Bayesian network approach to determine the association between morphological features extracted from prostate cancer images, *IEEE Access* 7 (2019) 1586–1601, <https://doi.org/10.1109/ACCESS.2018.2886644>.
- [33] 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.combiomed.2015.03.004>.
- [34] 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>.
- [35] 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](https://doi.org/10.1016/S0140-6736(99)06093-6).
- [36] 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>.
- [37] W. de Wever, J. Coolen, J.A. Verschakelen, Imaging techniques in lung cancer, *Breath* 7 (2011) 338–346, <https://doi.org/10.1183/20734735.022110>.
- [38] L. Hussain, W. Aziz, A.A.A. Alshabani, 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>.
- [39] 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>.
- [40] H. Golnabi, A. Asadpour, Design and application of industrial machine vision systems, *Robot. Comput. Integrat. Manuf.* 23 (2007) 630–637, <https://doi.org/10.1016/j.rcim.2007.02.005>.
- [41] 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>.
- [42] 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>.
- [43] 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/IPHDW.2018.8388338>.
- [44] 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>.
- [45] 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>.
- [46] 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>.
- [47] 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>.
- [48] Qiujuan 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>.
- [49] Z. Nematollahi, 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>.
- [50] I. Tsamardinos, E. Greasidou, G. Bourboudakis, 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>.
- [51] 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.combiomed.2013.12.010>.
- [52] 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.
- [53] Automated Colon Cancer Detection Using Hybrid of Novel Geometric Features and Some Traditional Features, 2016, <https://doi.org/10.1016/j.combiomed.2015.03.004>.
- [54] 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/TrustComBigDataSE.2018.00057>.
- [55] 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>.
- [56] 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>.
- [57] 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., IEEE, n.d., 2005, pp. 1005–1008, <https://doi.org/10.1109/ICASSP.2005.1416476>.
- [58] 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>.
- [59] 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.
- [60] 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>.
- [61] 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](https://doi.org/10.1016/S0010-4485(02)00101-X).
- [62] 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/10.15886/irecos.v10i1.5052>.
- [63] Y. Eroglu, M. Yildirim, A. Cinlar, 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.combiomed.2021.104407>.

- [64] 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.combiomed.2015.03.004>.
- [65] 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>.
- [66] M. Madhubala, M. Seetha, Hybrid feature extraction and selection using bayesian classifier, in: Natl. Conf. Adv. Era Multi Discip. Syst. AEMDS, 2013, Technol. Educ. Res. Integr. Institutions, Kurukshetra, Haryana, India, 2013, pp. 449–453.
- [67] 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>.
- [68] 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>.
- [69] 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>.
- [70] 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>.
- [71] 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.2023.32235619>.
- [72] E. Selvi, 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>.
- [73] 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>.
- [74] 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>.
- [75] Wei Guo, Wei Ying, 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>.
- [76] 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>.
- [77] 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.combiomed.2009.10.005>.
- [78] 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 Random Complex Media* (2021) 1–26, <https://doi.org/10.1080/17455030.2021.1905912>.
- [79] E. Dandil, 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>.
- [80] J.R.F. da Silva Sousa, 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>.